NOTTINGHAM TRENT UNIVERSITY

In silico Identification of Novel Genetic Factors Associated with Longevity in *Drosophila*

A thesis by Bethany Hall submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of Doctor of Philosophy.

Department of Physics & Mathematics

Nottingham Trent University

September 2019

Acknowledgements

Firstly, I would like to thank my supervisor, Professor Nadia Chuzhanova. The confidence she has had in me since during my undergraduate degree, alongside the guidance and support she has provided consistently during my PhD, have helped me get to where I am today. The enthusiasm and passion she has for her own research and her strong successful persona have influenced me so positively in my own studies. Nadia is my academic role model. I also wish to thank my second supervisor, Dr Jonathan Crofts, and my third supervisor, Professor Yvonne Barnett, for their kind attitudes and important contributions towards my PhD studies. Both have helped me to expand my knowledge in different areas of my PhD study, in areas that I would have otherwise felt less confident. Staff and PhD students in the Mathematics Department of Nottingham Trent University, particularly those who I have shared an office with during my time here, have kept me sane and motivated even when times have got tough, so thank you. I would also like to thank the Vice Chancellor's bursary for providing me with the funding to undertake this postgraduate study at NTU.

Finally, I would like to thank my best friends and family for their love and encouragement, and their ability to cope with me when my PhD has been difficult and made me moody. I want to extend a special gratitude to my parents, grandparents, Sarah and Martin. Genuinely, without their reassurance, dependability and general amazingness as people, this PhD experience would have been so much more difficult.

Publications and Presentations

The material presented in Chapter 4 as well as some of the material presented in Chapter 5 has been published in *Aging* journal:
Hall BS, Barnett YA, Crofts JJ, Chuzhanova N. Identification of novel genes associated with longevity in Drosophila melanogaster - a computational approach. *Aging* (*Albany NY*). 2019 Dec 3;11(23):11244-11267. doi: 10.18632/aging.102527. Epub 2019 Dec 3.

In addition, the material in Chapter 4 was reported as various presentations, as follows:

The 67th BRSA (British Society for Research on Ageing) Annual Scientific Meeting
University of Exeter, UK, July 2017
Contributed talk: *In silico* Modelling of Longevity of *Drosophila melanogaster* – a Network Approach.

School of Science and Technology Annual Research Conference
Nottingham Trent University, UK, May 2018
Contributed talk: *In silico* Modelling of Longevity of *Drosophila melanogaster* – a Network Approach.

ISCB (International Society for Computational Biology) Conference
Chicago, USA, July 2018
Contributed talk: *In silico* identification of novel genetic factors associated with longevity in *Drosophila*.
Poster presentation: *In silico* identification of novel genetic factors associated with longevity in *Drosophila*.

RECOMB (Research in Computational Molecular Biology) Conference
Washington DC, USA, May 2019
Poster presentation: *In silico* identification of novel genetic factors associated with longevity in *Drosophila*.

RECOMB (Research in Computational Molecular Biology) Satellite on Computational Methods in Genetics
Washington DC, USA, May 2019
Poster presentation: *In silico* identification of novel genetic factors associated with longevity in *Drosophila*.

Abstract

To determine genetic factors causing variation in survival into old age, several genome-wide association studies (GWAS) have been carried out on panels of long-lived individuals. The findings from a number of these GWAS studies were somewhat inconclusive, owing to the small sample sizes investigated. It is for this reason that model organisms such as *Drosophila melanogaster* have become increasingly important in identifying genetic factors underlying longevity.

In this study we hypothesised that co-location of novel genes/genomic regions with genes, known to be associated with longevity, that share biological function with co-located genes, make them good candidates for novel genomic regions, linked to longevity. We further hypothesised that single nucleotide polymorphisms (SNPs) residing within these co-located regions may influence longevity either individually (when a SNP in one of these genes causes a particular phenotype) or collectively (when one or several SNPs in these regions occur in the same individual thus causing the phenotype). Summary statistics of datasets of SNPs generated by two GWAS (Burke et al., 2013; Ivanov et al., 2015) which include position of each SNP and a corresponding statistic (D or P- value) showing the strength of association with longevity were used in this study to guide the initial choice of genes/loci strongly associated with longevity.

First, a network approach was applied to predict novel genes/genomic regions/SNPs, playing a role in longevity, which integrated three-dimensional (3D) chromosome conformation data (Hi-C) and two GWAS datasets. Networks were created using genes/genomic regions, known to associate with longevity, as original nodes with additional nodes (regions) later added to these networks if they strongly interacted (i.e. came into close proximity as measured by the Hi-C data) with the original nodes. Various network measures were calculated, in order to identify important previously unknown regions. These previously unknown regions were further explored and longevity associated genes were found including *Rim* and *Tpi* with a 'long-lived' phenotype, and some newly found regions were observed to be common between both GWAS datasets. A human ortholog search of genes found in this analysis resulted in matches to human genes with functions related to lifespan. Subnetworks of these GWAS-based networks were sought for enrichment in GO terms and several genes

with no previous association with longevity but enriched in longevity-related terms were identified.

Second, SNPs residing in non-coding regions, e.g. within transcription factor binding sites (TFBSs) recognised by transcription factors (TF) and borders between Topologically Associated Domains (TADs) were analysed. Each TF typically recognises a collection of often dissimilar DNA motifs. Here we hypothesised that TFs may recognise a certain structure, e.g. non-B DNA structures, rather than sequence motifs. Structures such as slipped, cruciform, triplexes and tetraplexes, formed on direct, inverted and mirrored repeats and G-quartets were considered and SNPs residing within these structures were analysed. For the study of SNPs in TAD borders we hypothesised that SNPs residing in these border regions may cause a severe disruption to the way in which regulation usually occurs within these TADs. We found that a significant proportion (~2%) of non-coding SNPs, reported in the DGRP GWAS dataset, resided in TAD border regions on the *Drosophila* genome, when compared to a match control dataset ($P = 1.0376 \times 10^{-75}$).

Finally, potential target genes for non-coding SNPs were explored, taking a different approach to the assumption that these target genes are usually the nearest on the linear genome. Using intra-chromosomal Hi-C data with finer resolution we identified regions highly interacting with those non-coding SNPs. These interacting regions were further analysed, focusing on the genes they harboured and the functions of these genes. Many regions found to have long-range interactions in both GWAS datasets analysed were observed to harbour genes not directly associated with longevity but with phenotypes displaying longevity-related functions. Eleven genes were found in common between the two GWAS datasets, where several were matched to human orthologs.

In conclusion, the network and other bioinformatics approaches which have been used in this study have enabled the identification of factors associated with longevity that were previously unknown, or overlooked.

List of Figures

List of Tables

List of Abbreviations

[3C]	Chromosome conformation capture
[3D]	Three-dimensional
[4C]	Chromosome conformation capture on ChIP
[5C]	Chromosome conformation capture carbon copy
[A]	Adenine
[bp]	Base pair
[ChIP]	Chromosome immunoprecipitation
[C]	Cytosine
[DGRP]	*Drosophila* genome reference panel
[DNA]	Deoxyribonucleic acid
[EGS]	Extragonadal seminoma
[G]	Guanine
[GAIN]	Genetic association information network
[GO]	Gene ontology
[GSEA]	Gene-set enrichment analysis
[GWAS]	Genome-wide association study
[HDL]	High-density lipoprotein
[IIS]	Insulin/IGF-1 signalling
[ISC]	International Schizophrenia Consortium
[JNK]	Jun-N-terminal kinase
[Kb]	Kilobase
[LD]	Linkage disequilibrium
[MAGENTA]	Meta-analysis gene-set enrichment of variant associations
[MAGMA]	Multi-marker analysis of geno-mic annotation
[Mb]	Megabase
[mRNA]	Messenger ribonucleic acid
[SIDS]	Sudden infant death syndrome
[SMR]	Standardised mortality ratio

[SNP]	Single nucleotide polymorphism
[SPSS]	Statistical package for the social sciences
[T]	Thymine
[TAD]	Topologically associated domain
[TF]	Transcription Factor
[TFBS]	Transcription factor binding site
[TOR]	Target of rapamycin
[VLDP]	Very low-density lipoprotein

Contents

Chapter 1

INTRODUCTION

Given the increasing rate of survival into advanced (≥85 years) and exceptionally advanced (≥100 years) age in humans, achieving healthy ageing is becoming increasingly important. With healthy ageing being described as the process of developing and maintaining the functional ability that enables wellbeing through to older age, this is important to ensure that humans are able to continue what it is that they value, meeting basics needs, being mobile and contributing to society. Human twin studies suggest that 20-30% of variation in survival, besides maintaining a healthy life style, is determined by heritable genetic factors (Herskind et al., 1996). Studies have looked at the variation in the age at death between monozygotic and dizygotic twins. These twin studies have found that around 25% of the variation in human longevity can be attributed to genetic factors, with the genetic component being higher at older ages and more important in males than females (Herskind et al., 1996; Skytthe et al., 2003; Hjelmborg et al., 2007). The identification of these genetic factors causing such variation has therefore become of great interest in the research area of ageing.

Model organisms such as *Drosophila melanogaster* have become increasingly important for studying and understanding genetic factors affecting longevity; their short, simple reproduction cycle and large number of offspring make it an ideal organism for the study of human genetics (He and Jasper, 2014).

Fairly recent genome-wide association studies (GWAS) have enabled to identify DNA alterations in the *Drosophila* genome, resulting in phenotypic change. Dependant on the rate that these changes occur, we refer to them as either single nucleotide polymorphisms (SNPs) which occur in more than 1% of the population, or mutations, which occur in less than 1% of the population. Although it is acknowledged that a single mutation/SNP often does not cause phenotypic change by itself, the specific accumulation of these SNPs may result in such change. These SNPs are often scattered across the genome and located in different regions, where the biological relationship between these SNP harbouring regions is not well understood. Therefore it is interesting to explore whether these regions share a common feature or biological function.

SNPs are individual base pair changes on a chromosome and are not often observed to cause positive/negative effects to organisms. However, those SNPs that do induce such effects are important to note, no matter the severity of its effect. The impact of SNP effects (or their significance) have been obtained via GWAS, in which SNPs are identified and analysed across the genomes of individuals from both healthy and diseased populations, as well as long lived and non-long lived populations as studied here. These latter GWAS datasets can provide detail as to which SNPs are most significantly associated with longevity.

In the following sections of this chapter, the ageing process is introduced along with theories of ageing proposed to attempt to explain such a complex process. Longevity studies in both humans and *Drosophila* are also discussed, with findings from previous ageing studies. SNPs are introduced and genome-wide association studies (GWAS) and their main focus is explained. GWAS carried out in previous studies on both humans and *Drosophila* are also discussed, along with their findings, and specific bioinformatics techniques used for identifying important SNPs in GWAS data are described. Finally this chapter introduces the hypotheses, aims and objectives of this study.

1.1 AGEING

1.1.1 The Ageing Process

Ageing is associated with changes in several processes of an organism, ranging from social and environmental to biological and physiological. Such age-associated changes in humans include those considered to have very small or zero impact on the way in which a person lives their normal life, for example hair turning grey, or age-spots spearing on the skin. However, one of the most important reasons for the study of ageing is to target the changes which are found to reduce the quality of life a person can live. This includes changes which result in a decline in the functioning of senses and daily life activities, and very importantly those changes which result in increased susceptibility to disease, frailty and disability. Ageing is already known to be associated with the development of serious chronic diseases in humans (Fontana, 2009), which is one main reason for studies aiming to uncover theories and mechanisms that can be used to explain ageing.

Systematic research in ageing came into existence only 50-60 years ago, and in this relatively short time a great deal has been discovered and understood about human ageing. However, one of these discoveries was the actual complexity of ageing and the difficulty of controlling the factors by which it is influenced. For example, the rate at which an individual ages has been observed to largely depend on interactions with environmental factors (Geller and Zenick, 2005), for which every human is different. Such dependency on environment means that the rate of ageing for humans is highly individualised, causing problems in studies of longevity for which it is impossible to compare mean values of measured variables across populations.

Many different definitions of biological ageing have been proposed over the years. Some describe ageing as an increased risk of mortality or death (Medawar, 1952), which would be an appropriate definition for some species in which death and ageing are the same. For example, in the Mayfly, after the completion of adult development death occurs very soon after, complicating measurement of the rate at which Mayflies age. However, such definition is not useful for those who try to correlate biological events to ageing outcomes in individuals.

For example, hair turning grey in an elderly woman is a clear sign of ageing, but this change in the hair colour not significantly increase mortality risk of this individual.

Those who correlate biological events with the rate of ageing are likely to use functional-based definitions of ageing. Some suggest that ageing is a result of 'deteriorative changes with time during post maturational life that underlie an increasing vulnerability to challenges, thereby decreasing the ability of the organism to survive' (Masoro, 1995). 'Senescence is a term mainly used to describe age-related changes in an organism that adversely affect its vitality and functions but most importantly, increase the mortality rate as a function of time. Senility represents the end stage of senescence, when mortality risk is approaching 100%' (Finch, 1994). Both of these considerations advantageously specify a time period to look for ageing, which is once an organism has reached full growth. These definitions also identify processes associated with advanced age, which can be measured and tracked over time. The limitation to these definitions however, is that they both address only the ageing of an organism as a whole, as opposed to ageing at a lower level of organization, for example cellular function. There is also no consideration in either definition for events that occur during development which may have a direct impact on post maturational life. Finally, in these definitions it is uncertain when ageing starts, as it is possible that whilst some physiological functions begin, others will still be developing.

The discovery of the cause of cellular ageing, reflecting the random accumulation of damaged proteins that result from an organism's interaction with the environment (López-Otín et al., 2013), allows more conclusive definitions to be drawn about biological ageing. Acknowledging these points, we will consider that ageing is the result of both the passing of time and an individual's interaction with the environment, causing random changes in the structure and function of molecules, cells and therefore an organism as a whole. The probability of the death of an organism is positively correlated with the process of ageing.

1.1.2 Theories of Ageing

In an attempt to explain the process of ageing, many theories have been proposed, of which none have individually succeeded in explaining the process. Most theories have been found

to fall into two main categories, with the first category stating that ageing is a natural process and programmed into the body (programmed theories), and the second category stating that ageing is a result of accumulation of damage to the body over time, known as error theories (Jin, 2010). Although considered as two different categories, neither are necessarily mutually exclusive, and many agree that as well as ageing varying across different species, the build-up of damage suggested in the second category can be accelerated by programmed senescence suggested in the first category (Weinert and Timiras, 2003). As well as these two categories, evolutionary theories of ageing have also been proposed in which, according to these theories, ageing is a by-product of natural selection.

Evolutionary Theories of Ageing

The concept of the evolution of ageing was first introduced by Ronald Fisher in 1930 (Fisher, 1930). Since this initial introduction, innovative thinkers like Peter Medawar (Medawar, 1952), George Williams (Williams, 1957), Thomas Kirkwood (Kirkwood, 1977), and others, have established cogent evolutionary theories to help explain why ageing evolved (Johnson et al., 2019). Theories under this concept believe that extrinsic mortality, for example predation, disease and starvation, is a primary evolutionary determinant of the rate at which an organism will age. Key evolutionary theories include the 'mutation accumulation' theory, 'antagonistic pleiotropy' theory and 'disposable soma' theory, where all three theories predict than an increase in extrinsic mortality should select for the evolution of shorter lifespans and *vice versa*.

According to the mutation accumulation theory, from the evolutionary perspective, ageing is an inevitable result of the declining force of natural selection with age (Medawar, 1952). For example, a mutant gene that kills young children will be strongly selected against and therefore will not be passed to the next generation, whereas a lethal mutation which effects only people over the age of 70 will experience no selection because by that age, people with that mutation will have already passed it to their offspring. Over successive generations, late-acting deleterious mutations will accumulate, leading to an increase in mortality rates late in life (Bengtson and Settersten, 2016).

The antagonistic pleiotropy theory suggests that late-acting deleterious genes may even be favoured by selection and be actively accumulated in populations if they have any beneficial effects early in life (Williams, 1957). The main difference between both theories is that in the mutation accumulation theory, genes with detrimental effects at old age accumulate passively from one generation to the next whereas in the antagonistic pleiotropy theory, these genes are actively kept in the gene pool by selection (Le Bourg, 2001). It is important to note that both of these theories are not mutually exclusive, and both mechanisms may operate at the same time. These theories were also later formalised mathematically and further developed by Hamilton (1966).

The antagonistic pleiotropy theory was further studied in an attempt to specify in more details how one gene could have both deleterious and beneficial effects. In doing so, the disposable soma theory was proposed (Kirkwood, 1977), postulating a special class of gene mutations with antagonistic pleiotropic effects. This theory is more mechanical and energy-focused, emphasising that because resources are limited, most organisms will be better off investing their finite energy into mechanisms that increase reproduction instead of non-reproductive mechanisms. Both the antagonistic pleiotropy and disposable soma theories expect a trade-off between ageing and fecundity (Flatt and Partridge, 2018).

Programmed Theories of Ageing

Theories stating that ageing is programmed claim that an organism is designed to age, and that for each organism there is a specific biological timeline that an organism follows (Kirkwood and Melov, 2011). Such theory defends ageing as an essential and innate part of the biology of organisms, to prevent living forever. The main claim of this theory is that ageing is about evolution as opposed to biology, which supports claims of ageing being inherent in an organism and dismisses those of the environment or disease being important factors. Evidence supporting this theory includes the following question: if the body is able to repair and renew itself, on all biological levels, then there is no reason for the body wearing out. The lack of variation in overall lifespan within species, other than in cases in which factors such as nutrition or medical care are heavily influential, is another argument for the theory of programmed ageing.

Within this theory, there are more questions about the way in which an organism is designed to age. For example, there are those who support the Endocrine theory that believe it is hormones controlling the function of organs, which are the cause of ageing (Van Heemst, 2010). The endocrine system is vital in the secretion and controlling of hormones regulating many of the body's processes, including metabolism, growth and development. During ageing, the efficiency of such systems decreases, and it is changes such as this that are suggested by the hormone theory to be the cause of the effects of ageing. Evidence that supports this theory includes studies in which the pituitary gland, which controls much of the endocrine system, is removed from mice and all hormones identified in mice are supplemented. Results concluded that mice without the pituitary gland survived longer than those mice in the control group that did have their pituitary gland. This result leads to the conclusion that this gland must secrete another hormone, currently unknown, which has a negative impact on ageing (Brown-Borg, 2007). Other research, on a variety of organisms, has shown that mutations reducing levels of insulin-like growth factor 1 (IGF-1) have resulted in extended lifespan. However, this reduction in IGF-1 has had inconsistent effects when observed in age-related diseases in humans. Observations of a reduction in IGF-1 have led to reduced risk of age-related diseases in some, but lead to increased risk in others, specifically metabolic syndrome which includes cardiovascular disease (Aguirre et al., 2016).

Another belief under the theory of programmed ageing is that the immune system is programmed to decline over time, resulting in an organism's increased susceptibility to disease, also known as the immunological theory (Walford, 1964). The belief that the rate of ageing is controlled mainly by the immune system, reverses the suggestion that changes in the immune system in the elderly are a result of the ageing process, and instead suggests that symptoms of ageing such as chronic disease are in fact caused by these changes in the immune system. It is widely known that changes in the immune system that accompany old age are able to impact a person's longevity directly, and that the potential for such changes to cause damage to the body increases with age. However, although there are strong suggestions that dysfunction of the immune system related to old age may cause some of the known aspects of the ageing process (Franceschi and Campisi, 2014), the triggers of changes in immune system and the way in which they develop and progress remain largely unknown

(Barnett and Barnett, 1998; Montecino-Rodriguez et al., 2013; Castelo-Branco and Soveral, 2014; Fuentes et al., 2017).

Error Theories of Ageing

As well as beliefs in ageing being programmed, there are also theories that state that ageing is a result of damage to our body's systems caused by the environment, over time. In contrast to the programmed theory of ageing, error theories state that ageing is a result of a series of errors, as opposed to an event which is programmed. Error theories include the popular belief that ageing is due to wear and tear, that over time cells and body systems become progressively damaged and in the end our body is 'worn out', and therefore unable to function adequately (Jin, 2010). Such damage can be caused by a number of processes, for example damage of genes due to exposure to radiation, or the inability for proteins to work efficiently due to oxidative damage resulting in the cross-linking of proteins. Even just the essential functioning of our body, for example the metabolism of oxygen, can result in the damage of cells and tissues due to free radical production. This theory, makes logical sense as it fits very closely with the perceived sense in which we age, and on a cellular level functions are observed to decline with age. However, this theory also faces strong criticisms, especially by the knowledge that we have of our bodies ability to repair the damage (Mitteldorf, 2010). Our DNA contains DNA repair genes with a specific function to repair genetic damage. There is also the argument that organisms are observed to grow stronger in their growth phase, building strength and resilience with age, however in a wear and tear theory it would be logical to assume that an organism would start life at the peak of performance. Finally, the wide variation of lifespans between different species of animal is also questioned under this theory.

Cross-linking is another theory, which is a complicated process observed to happen slowly between proteins, DNA and other structural molecules in the body (Bjorksten and Tenhu, 1990). Evidence has suggested that a variety of post-translational modifications and oxidative stress may contribute to cross-linking, where crosslinking mechanisms have been proposed dependent on the cross-linked compounds observed (Wang et al., 2014). These mechanism proposals include identified compounds such as advanced glycation end products (Nagaraj et al., 1991) and c-glutamyl-e-lysine (Lorand et al., 1981). The cross-linking theory describes

bonding occurring between structural molecules in the body, causing chemical changes, over time leading to ageing. Once molecules are cross-linked, their ability to function properly deteriorates, and if enough cross-links form, they can accumulate in tissues causing further problems. The cross-linking of molecules results in the stiffening of the tissues in which these molecules reside, and many symptoms related to ageing are associated with the stiffening of tissues. For example, it is believed that heart attack and stroke risk is increased with the hardening of arteries (Kumosani et al., 2011).

Finally, and probably the most important theory in relation to the research in this thesis is the somatic mutation theory, in which it is stated that what happens to our genes once they are inherited is an important determinant in ageing. A mutation can be caused by a variety of factors, and results from incorrect copying of a gene during cell division. It is also important to note that mutations can also occur outside of genes. Often the body is able to correct or destroy any mutations, however when this is not the case, mutated cells can accumulate, copy themselves and cause age-related issues in the functioning of the body systems. Like all ageing theories discussed here, and those not covered, each only seems to explain one part of the question of ageing. Although for many theories there is a supporting evidence, it cannot be proved that any of these factors play the most important role in ageing.

1.2 LONGEVITY STUDIES IN HUMANS

The search for genetic determinants of human longevity has, and most likely will always be, challenged due to lack of studies producing datasets with large numbers of participants reaching extremely old age. Therefore, despite the evidence given for a genetic contribution to longevity, the identification of specific genes that associate with human longevity has been difficult (Christensen et al., 2006). The survival and health characteristics in families of long-lived subjects have been studied, and the evidence from such studies suggest that healthy ageing and longevity have a hereditary component. One of these studies involved conducting a sib pair study in very old subjects to map longevity loci, and then looked at families with at least two long-lived siblings (Schoenmaker et al., 2006). Statistical analysis included the calculation of standardised mortality ratios (SMRs), comparing the mortality of various generations with the general population; the mortality between different groups then being

compared using Cox regression analysis. From the results obtained it was concluded that familial clustering of extended survival was unlikely to be caused by ascertainment bias or environmental factors, and therefore it was concluded that the long-lived individuals were genetically enriched for extreme survival. Similar conclusions, to that of genetic enrichment for extreme survival, have also been drawn from other studies (Perls et al., 2002; Montesanto et al., 2011). A study by Kerber et al. (2012) found regions of interest in the vicinity of D3S3547 on chromosome 3p24.1; interestingly, this chromosome loci had also previously been reported in relation to longevity by Boyden and Kunkel (2010), on an almost identical region of the chromosome. This study by Kerber et al. (2012) also corroborated the linkage of exceptional longevity to 3p22-24, another loci also found by Boyden & Kunkel (2010), strengthening the case that genes found in these regions affect variation in longevity, and therefore play a role in the regulation of human lifespan. In this linkage study by Boyden & Kunkel (2010), several additional novel loci were identified as having significant association with longevity, e.g. on chromosomes 9q31-34, 12q24 and 4q22-25.

1.3 SINGLE NUCLEOTIDE POLYMORPHISMS

SNPs are individual base pair changes on a chromosome, and in the human genome are the most common type of genetic variation. An example of such base pair change is shown in Figure 1.1. The effect a SNP has on a protein in which it resides will depend on the type of SNP, of which there are two types: synonymous (silent) or non-synonymous. Synonymous SNPs are considered to have no effect on the proteins it is involved with, whereas non-synonymous SNPs result in an alteration to an amino-acid sequence. This alteration occurs either via missense polymorphisms, which alter an amino acid sequence or via nonsense polymorphisms that stop the function of proteins after inducing a premature stop codon. The regions of the genome on which SNPs reside, as well as the number of SNPs, also indicates the effect in which they may have (Shastry, 2009). For example, some SNPs may occur in non-coding regions such as an enhancer, therefore affecting the expression of specific genes. Regions on the genome in which SNPs are found in greater abundance would often be expected to experience larger mutational effect, for example more important phenotypic changes, and in this situation the probability of the mutations being deleterious would also increase.

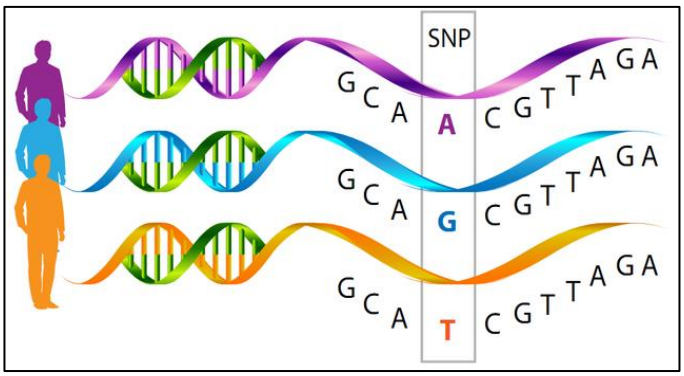


Figure 1.1 A visual representation of an individual base pair change (SNP) across chromosomes for three humans (https://www.whatisdna.net/wp-content/uploads/2016/11/SNP.png).

1.4 GENOME-WIDE ASSOCIATION STUDIES

A genome-wide association study (GWAS) is an approach that involves rapidly scanning many genomes with markers, with the intent to find common genetic variations associated with a particular phenotype; in many cases, this is a particular disease. Such association testing on an unbiased, genome-wide scale is possible due to advances in genotyping technology and statistical methods, as well as improved understanding of genomic variation. Often, GWAS focuses on associations between single-nucleotide polymorphisms (SNPs), which are changes of a single base in the DNA known to vary between individuals, and a particular disease or, in the case of this PhD investigation, longevity. These studies are designed to enable a comparison between the frequencies of an allelic variant in a case versus a control sample. For example, SNPs in the genome in those with a particular disease may be compared against those without the disease. This comparison enables the recognition of any genotypes that may occur more often or less frequently in individuals with the disease. As a very basic example in Figure 1.2, the genomes of a population of individuals found to have a specific disease have been scanned to find all SNPs that these genomes harbour. The same is then done for a population that does not have this disease, and the quantity of each SNP found in the population with the disease is then compared with the quantity of the same SNP found in the population without. For example, in Figure 1.2 the red dots in some individuals represent a specific SNP, in this example SNP ‘A’, residing in the genome of this individual. For SNP ‘A’, observations showed that both populations had the same number of individuals harbouring

this SNP, and therefore it would not be logical to assume that SNP ‘A’ had any association with this disease. However, for SNP ‘B’, it is clear to see by looking at the yellow dots representing this SNP in the genome of the individual, that SNP ‘B’ occurs more frequently in individuals in the population with disease than the population without. As SNP ‘B’ is found to be more common in the individuals with disease, the assumption that this SNP may have association with this specific disease can be made.

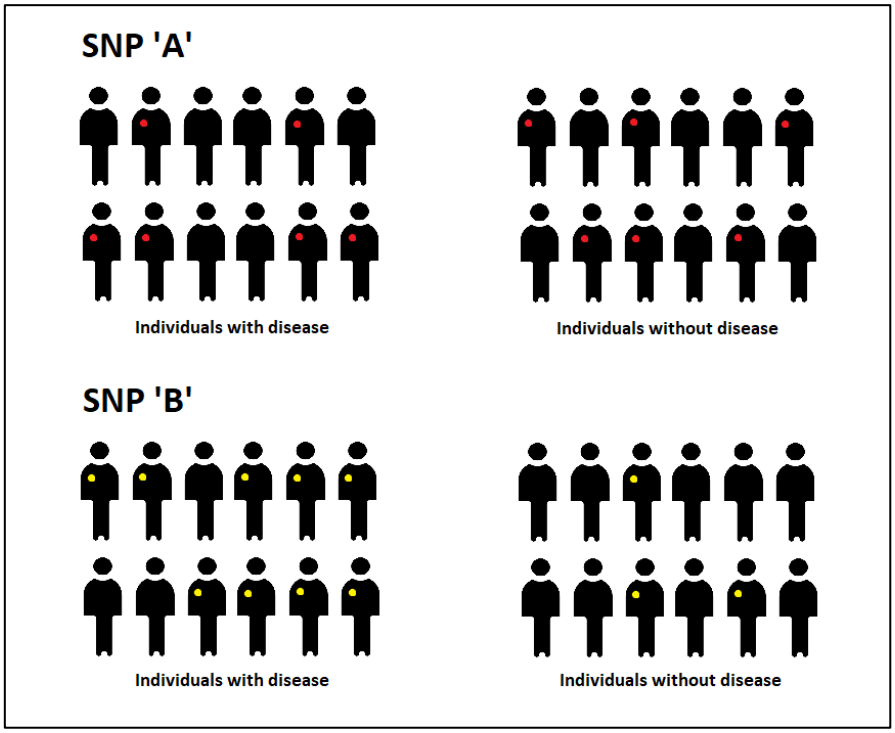


Figure 1.2 Four groups of individuals, in which two groups contain diseased individuals and two do not. Red dots in some individuals represent a specific SNP, in this example SNP ‘A’, residing in the genome of this individual. Yellow dots in some individuals represent a specific SNP, in this example SNP ‘B’, residing in the genome of this individual.

This example described is more simplistic than the process actually is, and in reality, there are more statistics involved and in fact once a well-defined phenotype has been selected for a study population and genotype data is collected, a series of single-locus statistical tests are carried out. These tests involve the individual examination of each SNP to look for association with the phenotype selected, and often in dichotomous cases (also referred to as the primary trait/phenotype) logistic regression is used for analysis. Logistic regression, shown in the equations below (equation 1.1), predicts the probability of each SNP having disease status

given a genotype class. Logistic regression also includes a measure of effect size, which provides adjusted odds ratios and allows for adjustment for clinical covariates.

Wanting to test whether there is a difference in the distribution of genotypes between case and control groups in the study, hypotheses can be written in terms of the conditional probabilities of genotype given case/control status, where the null hypothesis is parameterized by $p = \{p_0, p_1, p_2\}$. To describe this regression analysis, the following notation is introduced: GWAS data consists of healthy or diseased phenotypes ($y_i \in \{0, 1\}$) and the genotype at the typed locus ($x_i \in \{0, 1, 2\}$) with an effect size of β_i. The effect size is simply a quantifier for the size of difference between two groups, in this case, the difference between the healthy and diseased groups. In logistic regression, odds ratios are often used to calculate effect size, determining whether a particular exposure is a risk factor for a particular outcome, for example disease. In the calculation of logistic regression, the regression coefficient β_1 is the estimated increase in the log odds of the outcome per unit increase in the value of the exposure. The relationship between y and x is then modelled using the likelihood method:

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{(1-y_i)} \quad \text{where} \quad \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i. \tag{1.1}$$

This whole-genome analysis of genetic variants in humans has provided insight into the architecture of complex traits and along the way revealed dozens of longevity-associated loci.

1.5 GWAS OF HUMAN LONGEVITY

To determine genetic factors of longevity in humans, several genome-wide association studies have been carried out on panels of (exceptionally) long-lived individuals, with ages of participants ranging from 90+ in two studies, and between 95 years to 119 years in another (Puca at el., 2001; Newman et al., 2010; Sebastiani et al., 2012). Variation in many loci, e.g. near the *D4S1564* (Puca et al., 2001), *MINPP1* (Newman et al., 2010), *HLA-DQA1/DRB1* and *LPA* (Joshi et al., 2017) genes, have been identified as contributing to survival into old age, but only SNPs in TOMM40/*APOE* and *FOXO3* loci were found to robustly associate with longevity.

The two loci commonly associated with longevity in previous independent studies, *APOE* (Apolipoprotein E) and *FOXO3* (Forkhead Box O3), are also identified by GWAS (Lu et al., 2014; Broer et al., 2014; Zeng et al., 2016). The *APOE* gene combines with lipids (fats) in the body, forming lipoproteins, which play a role in packaging and transporting fats, including cholesterol, through the bloodstream. Maintaining cholesterol levels is essential to prevent cardiovascular diseases, for example strokes and heart attacks. The *FOXO3* gene is a transcriptional activator, which, in the absence of survival factors, triggers apoptosis. This includes neuronal cell death upon oxidative stress.

Newman et al. (2010) confirmed the association of rs4420638 SNP on chromosome 19q13.32, representing the TOMM40/APOE/APOC1 locus, with longevity. Both the *TOMM40* (Translocase of Outer Mitochondrial Membrane 40) gene and *APOC1* (Apolipoprotein C1) gene are associated with Alzheimer Disease. *APOC1* is also related to lipoproteins, playing a central role in high density lipoprotein (HDL) and very low density lipoprotein (VLDL) metabolism. The *TOMM40* gene is part of the mitochondrial outer membrane (TOM) complex, which is essential for the importing of protein precursors into mitochondria. Flachsbart et al. (2016) used a combined sample of 3208 long-lived individuals and 8919 younger controls of European origin, and performed a large-scale case control study, targeting known immune-associated loci. The first part of analysis by Flachsbart et al. (2016) performed a large-scale association study on 1458 German long-lived individuals (mean age 99 years) and 6368 controls (mean age 57.2 years). Findings in the German groups of this study further supported the association of the TOMM40/APOE region with longevity, finding significantly associated SNP rs2075650 in this region. The same study also reported a novel locus for longevity, RAD50/IL13 region on chromosome 5q31.1, harbouring rs2706372 SNP. RAD50 is known to be involved in the biological processes DNA repair and inflammation, making it a credible longevity candidate.

Deelen et al. (2014) performed GWAS meta-analysis of 7729 long-lived individuals of European descent ($\geq$85 years) and 16,121 younger controls ($<$65 years), observing genome-wide significant association of rs2149954 SNP (intergenic) with longevity at the novel locus on chromosome 5q33.3. The same SNP was also found in the Han Chinese and European population (Zeng et al., 2016). The *APOE* locus has also been associated with longevity in large

population-based linkage studies, alongside TOMM40/APOC1 loci (Beekman et al., 2013). Beekman et al. (2013) performed the largest genome-wide linkage scan reported so far, observing four regions showing linkage with longevity. A fixed-effect meta-analysis approach at the APOE/TOMM40/APOC1 gene locus identified a single SNP, rs4420638, showing a significant (P-value = 9.6×10^{-8}) association with longevity.

Despite the use of GWAS to investigate human longevity, a lot of studies were underpowered, due to the availability of small sample sizes only. For this reason, model organisms such as *Drosophila melanogaster* have become increasingly important for studying and understanding genetic factors affecting longevity. GWAS have proven as extremely useful in the discovery of genomic variants responsible for traits in species such as *Drosophila*. GWAS have been used for identification of susceptibility loci for phenotypes including aggression (Shorter et al., 2015), brain size (Zwarts at al., 2015) and longevity (Ivanov et al., 2015) in *Drosophila*.

1.6 LONGEVITY STUDIES IN DROSOPHILA

Drosophila melanogaster, known generally as the common fruit fly or vinegar fly, is an extensively studied model organism in the field of genetics and developmental biology. Technology has advanced in ways that allow for further unfolding of the biology of ageing using fly as a model organism (Paaby and Schmidt, 2009). The lifespan of *Drosophila melanogaster* is affected by several factors including differences in environmental conditions, diet and overcrowding. In a controlled environment (including temperature, humidity, diet and exposure to carbon dioxide) in the laboratory the average lifespan of *Drosophila* is found to be >50 days, where the lives of adult males were observed to be shorter than those of females (Linford et al., 2013). Mutations in specific genes have been found to increase the lifespan of *Drosophila*. For example, a mutation within the *mth* (Methuselah) G protein-coupled receptor gene, which leads to the partial loss-of-function, has been found to extend the average lifespan of *Drosophila* by 35% (Lin et al., 1998). Mutant versions of the *Indy* (I'm not dead yet) gene, which encodes an amino acid transporter, has been shown to double the average lifespan (Rogina et al., 2000). In addition, it was shown that single gene mutations in the target of rapamycin (*TOR*) and the insulin/IGF-1 signalling (IIS) pathways can slow down

the ageing process in model organisms (Fontana et al., 2010). The *Drosophila* GWAS have identified millions of naturally occurring SNPs that potentially influence longevity; however, none of these SNPs reached genome-wide significance level prompting the hypothesis of possible combined effect of sets of SNPs on longevity.

Commonly in statistics, P-values are used as an indication of the significance of a result, whereby if a P-value calculated is less than 0.05, it is said to be statistically significant. In hypothesis testing, there is always a small chance (usually around 5%) that a false significant result will be produced in a single test. As the number of tests run increases, the number of these false significant results increases dramatically, and this is referred to as the multiple testing problem. This can be corrected for by adjusting P-values, taking into account how many hypothesis tests are actually running. Due to the scale and size of statistical analysis in GWAS, significance levels such as P<0.05 are adapted, for which a P-value threshold of 5×10^{-8} has become a standard for GWAS. This threshold is calculated by considering the commonly used P-value of 0.05 at a 5% significance level in hypothesis testing, and taking into account the number of SNPs that are observed in studies, typically up to 10^{6}.

The identification of many ageing genes in *Drosophila* has been made possible because of the association made between stress and lifespan. The best determination of whether ageing is altered as a result of oxidative stress or damage is an alteration in lifespan. Observations in relation to this have been contradictory, which could mean that oxidative stress plays a very limited, if any, role in ageing or that the role of oxidative stress in ageing is dependent on environment. In relation to environmental factors, environments in which minimal stress results in oxidative damage are reported to play little role in ageing, whereas under chronic stress, oxidative stress plays a much greater role. Under chronic stress, enhanced antioxidant defences exert an 'anti-ageing' action, resulting in changes in lifespan (Salmon et al., 2010).

Increasing the expression of genes that promote antioxidant defences have demonstrated increased organismal longevity. Over-expression of Catalase (*Cat*), a gene found to reduce oxidative damage to biomolecules and protect cells from toxic effects of reactive oxygen species, is one example (Orr and Sohal, 1994). Another example is Superoxide dismutase (*SOD*), a gene found to detoxify superoxide radicals in mitochondria, in which a loss of this gene generates endogenous oxidative stress resulting in reduced activity of critical mitochondrial enzymes (Orr and Sohal, 1994). Over-expression of glucose-6-phosphate

dehydrogenase (*G6PD*) has been shown to increase lifespan. *G6PD* is an enzyme that participates in a metabolic pathway, its main function is to produce NADPH, which is an electron donor that defends against oxidizing agents and plays an important role in reductive biosynthetic reactions (Legan et al., 2008). Screening for genes in *Drosophila* that show differences in gene expression between normal and stress conditions has identified two new loci affecting lifespan, heat shock protein (*hsp*) genes *hsp26* and *hsp27* (Wang et al., 2004). Several loci that have already been shown to affect lifespan were also identified in these screenings, including *hsp70*, *Cu/ZnSOD* and *catalase*.

In *Drosophila*, the *14-3-3-epsilon* gene has been found to antagonize *dFOXO* function whereby in oxidative stress, the loss of this *in vivo* interaction results in the loss of *14-3-3-epsilon*, causing increased stress-induced apoptosis and growth repression and, in turn, extends lifespan (Nielsen et al., 2008). In the nervous system, the overexpression of *Eip71CD* has been observed to increase lifespan by up to 70% through increase of resistance to oxidative stress and delaying the onset of senescence-induced decline in activity levels (Ruan et al., 2002).

The relationship between ageing and DNA repair was studied by measuring lifespan of *Drosophila melanogaster* males in the absence of the *mei-41* excision repair and comparing it to transgenic flies with 1 or 2 extra copies of the *mei-41* wild-type gene. In the absence of repair, the lifespan of *Drosophila* was significantly reduced whereas with an extra copy of the gene coding for excision repair, the lifespan of *Drosophila* was significantly increased (Symphorien and Woodruff, 2003).

It has been demonstrated that continuous over-expression of the *dFOXO* (Forkhead box class O transcription factor) gene, a homolog of the *FOXO3* gene, in adult fat body reduces mortality rate throughout adulthood (Giannakou et al., 2007). Further, Hwangbo et al. (2004) showed that limited activation of *dFOXO* reduces expression of the *Drosophila* insulin-like peptide dilp-2, as well as represses endogenous insulin-dependent signalling in peripheral body fat. This finding, along with the previously mentioned observations of *dFOXO* over-expression in humans reducing mortality rate, suggests that the combining of autonomous and non-autonomous roles of insulin signalling can control ageing.

The silent information regulator 2 (*Sir2* or Sirtuin) family of proteins has been shown to affect various aspects of physiology, including that of stress response (Imai and Guarente, 2010;

Haigis and Sinclair, 2010). Such findings have led to the questioning of the effects that *Sir2* may have on lifespan in *Drosophila*. Evidence has suggested that *Sir2* in *Drosophila* can mediate life span extension through caloric restriction and research findings have implicated *Sir2* in a number of beneficial effects of caloric restriction that have been shown to extend lifespan (Frankel et al., 2011). Rogina and Helfand (2004) demonstrated that an increase in expression of the *Sir2* gene extends lifespan, and in cases of a decrease in *Sir2* the lifespan-extending effect of calorie reduction is blocked. Whitaker et al. (2013) showed that increased expression of *Sir2* extended lifespan in a dose-dependent manner; life span is consistently extended when expression is increased to moderate levels (approximately 2-5 fold increase over normal) and expression below this range or slightly above it inconsistently extends lifespan. It was also found that significantly higher levels of expression are detrimental to lifespan; for example, over-expression can induce JNK signalling which is generally a 'death' signalling pathway, controlling the cell response to harmful extracellular stimuli, including programming cell death (apoptosis).

First documented by McCay et al. (1935) over 80 years ago, caloric restriction has continued to stand out as the most effective dietary intervention to extend both average and maximum lifespan, as well as to delay the onset of age-related pathologies (Anderson and Weindruch, 2012).This effect of caloric restriction on longevity is conserved across a diverse range of species, where numerous factors have been associated with the beneficial factors. Caloric restriction has been observed to effect multiple signalling pathways that regulate growth, metabolism, oxidative stress response, damage repair, inflammation, autophagy and proteostatis, in turn modulating the ageing process (Lopez-Lluch and Navas, 2016). Focussing on caloric restriction in flies, with appropriate techniques it has been possible to extend the lifespan of *Drosophila* by reducing food intake. A table provided in supplementary data in Piper and Partridge (2007) summarises various dietary restriction experiments performed with flies (Table S1: Piper and Partridge, 2007). Dietary restriction experiments have obtained divergent results. Some studies involving food dilution or nutrient manipulation resulted in no lifespan extension by food reduction but the majority of observations resulted in lifespan extension.

The *GADD45* protein family plays an important role in stress signalling. A single *D-GADD45* ortholog, when over-expressed in the *Drosophila* nervous system, has been shown to significantly increase lifespan without a decrease in fecundity and locomotor activity (Plyusnina et al., 2011). This effect is more noticeable in males than females, with the median lifespan of males extending by 73-77%, and that of females by 22-46%, in line with other findings showing effects to be dependent on factors such as sex. Over-expression of this gene results in more efficient recognition and repair of DNA damage (Barreto et al., 2007) and is therefore assumed to be the cause for increased longevity in *Drosophila*.

It is thought that the insulin/insulin-like signalling (IIS) pathway regulates various physiological processes that include the regulation of lifespan as well as stress responses, growth and development. Examination of the IIS pathway in *Drosophila* has determined genes that, by reduction of insulin signalling, are able to extend lifespan (Giannakou and Partridge, 2007). Processes regulating lifespan extensions by dietary restriction include metabolism, nutrient sensing and determination of body size, all of which are regulated by the IIS pathway, leading to the interest in testing the role of this pathway in lifespan extensions in mice (Weindruch et al., 1986). The identification of the nematode insulin receptor homolog *daf-2* as an ageing gene also contributed to the interest in testing this pathway in worms (Kenyon et al., 1993). Insulin signalling can be reduced via independent disruption of the Insulin-like Receptor (*InR*), a gene which regulates body and organ size and is involved in the development of the embryonic nervous system.

The IIS pathway of *Drosophila* is made up of many components, these include the insulin/IGF receptor (*dInR*) and the insulin receptor substrate (*chico*); the independent disruption of either the insulin/IGF receptor or substrate reduces insulin signalling and, in turn, increases lifespan. This pathway also includes the *Drosophila* transcription factor *FOXO* (*dFOXO*) which, when phosphorylated, through activation of *dInR*, reduces insulin signalling due to over-expression and, in turn, extends lifespan. Nuclear localization of *dFOXO* is promoted by the *PTEN* gene which, when over-expressed, causes genes involved in longevity to be upregulated by *dFOXO* (Altintas et al., 2016). Figure 1.3 demonstrates how in the IIS pathway, transcription factors eventually regulate the expression of target genes, contributing to longevity. Insulin-like peptides (ILPs) bind to insulin/IGF-1 receptor (*dInR*) resulting in its phosphorylation.

Binding to the insulin receptor substrate, *chico*, is reduced due to inhibition of the insulin/IGF-1 receptor, decreasing the activity of *PI3K* and the levels of PIP_3 converted. This decrease leads to decreased activities of *dPDK1* and *dAkt1* as well as causing the activation of the transcription factor *dFOXO*, which regulates the expression of target genes which contribute to longevity.

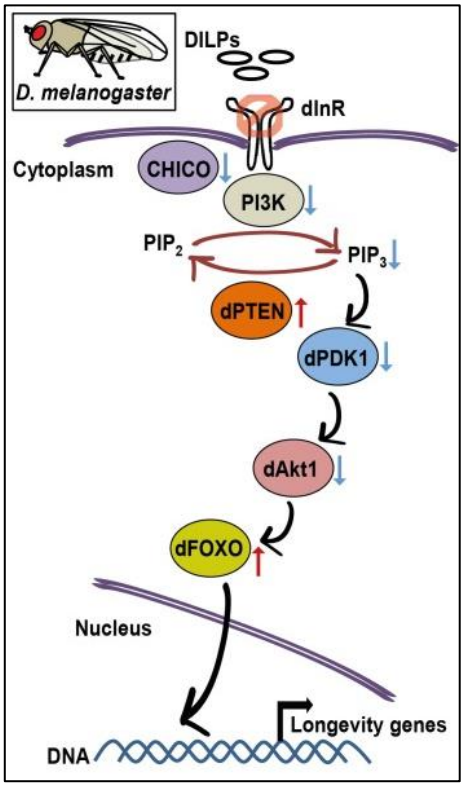


Figure 1.3 Conserved longevity-regulatory components of insulin/IGF-1 signalling pathway in *Drosophila* (image reproduced from Altintas et al. (2016)).

As well as genes involved in the IIS pathway, potential ageing genes in other pathways have been evaluated. The common property of the genes found is that they are members of pathways that appear to connect with the IIS pathway. The target of rapamycin (TOR) pathway, a regulator of body size, is one of these pathways connecting with IIS. Inhibition of TOR signalling by single gene modulation (modulation of a single gene from this pathway) has been found to increase lifespan in *Drosophila*; this includes expression of dominant-negative forms of *dTOR* or *dSK6* or over-expression of *dTsc1* or *dTsc2* (Kapahi et al., 2004).

A pathway activated in response to stress, which antagonises IIS is the Jun-N-terminal Kinase (JNK) pathway. JNK is known to phosphorylate a variety of transcription factors, enhancing

their transcriptional activation potential. This pathway causes nuclear localization of *dFOXO*, for which an increase in JNK signalling is dependent. Extension in lifespan due to this increase in JNK signalling has been demonstrated several times including by over-expression of both JNK kinase hemipterous (*hep*), and *hsp68* which is induced by JNK signalling (Wang et al., 2003). JNK signal transduction in *Drosophila* is summarised in Figure 1.4.

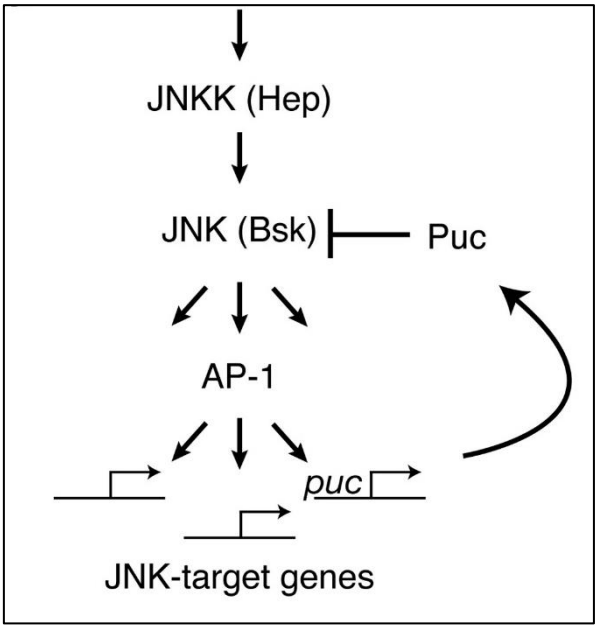


Figure 1.4 JNK signal transduction in *Drosophila*, in which JNKK is encoded by *hemipterous* (*hep*) and JNK is encoded by basket (*Bsk*). Phosphorylation activates the AP-1 transcription factor complex, where puckered (*puc*), a phosphatase of *Bsk*, is induced by AP-1 causing it to downregulate the pathway (Wang et al., 2003).

Disruption of the JNK phosphatase, puckered (*puc*), has also been demonstrated to extend lifespan; *puc* is activated when *Drosophila* Plenty of SH3s (*DPOSH*) is over-expressed, which also extends lifespan (Seong et al., 2001). A potential ageing gene evaluated based on its role in hypothesised mechanisms of ageing, is the elongation factor *EF-1alpha*, which is required for protein synthesis. Reduction of *EF-1alpha* is associated with senescence whilst over-expression can extend lifespan (Shepherd et al., 1989).

Research in *Drosophila* has also identified a critical role for Ras-Erk-ETS signalling in the ageing process (Slack et al., 2015). It was shown that inhibition of *Ras* downstream of IIS signalling causes an increase in *Drosophila* lifespan. The direct reduction of *Ras* or *Erk* activity was shown to lead to an increased lifespan. In the study by Slack et al. (2015), *aop* was identified

as playing central role in lifespan, as a result of reduced *IIS* or *Ras* attenuation. Treatment using trametinib, a highly specific inhibitor of the *Mek* kinase, was found to extend lifespan in *Drosophila* through the prevention of activation of *Erk* by *Ras*. A graphical abstract of these activities in the Ras-Erk-ETS signalling pathway are shown in Figure 1.5.

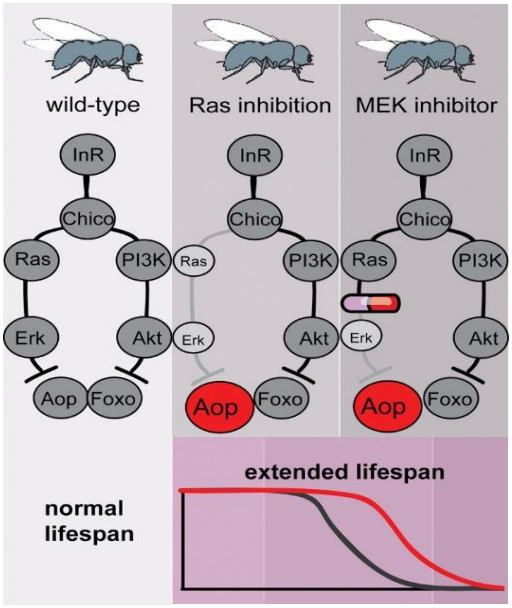


Figure 1.5 A visual representation of Ras and MEK inhibition in Ras-Erk-ETS signalling and the effect that they have on lifespan of *Drosophila* (extracted from a figure in Slack et al. (2015), https://creativecommons.org/licenses/by/4.0/).

Histone deacetylases, *Sir2* and *rpd3,* are genes through which lifespan extension has been demonstrated, both of which may operate through mechanisms related to IIS as well as dietary restriction. Increased levels of *Sir2* have been found to increase lifespan, whilst it is a reduction in the levels of *rpd3* that have been found to result in lifespan extension. A common mechanism between the processes of IIS and dietary restriction has been suggested due to the observation of reduced *rpd3* levels extending lifespan when diet is not restricted, and not having the same effect under dietary restriction (Rogina et al., 2002). RNA levels of *Sir2* are increased due to reduction of *rpd3* expression, directly extending lifespan (Rogina and Helfand, 2004). The activity between *Sir2* and *rpd3*, and their relationship with IIS, is further observed in findings including the lifespan extension in *C. elegans* by *sir-2.1*, an ortholog of *Sir2* (Partridge et al., 2005).

Candidates mediating lifespan have been identified through mutation screening for longevity genes. The disruption of both the G-protein coupled receptor *methuselah* (*mth*) and the Krebs cycle co-transporter *Indy* have been found to extend lifespan in *Drosophila* (Lin et al., 1998; Rogina et al., 2000). This effect on *Indy*, however, does not seem to be due to the activity of the gene itself but an artefact of genetic background and *Wolbachia* infection (Toivonen et al., 2007). Endogenous peptide ligands of *mth*, a G protein-coupled receptor, are encoded by the *stunted* (*sun*) gene, which when mutated have been shown to cause an increase in lifespan (Cvejic et al., 2004). The characterization of two endogenous peptide ligands of *Methuselah*, *Stunted A* and *B* have been reported. An increase in lifespan is observed where flies with mutations in the *mth* gene encode these ligands, as well as an observation of resistance to oxidative stress. The Stunted-Methuselah system is therefore concluded to have an involvement in the control of animal ageing.

1.7 GWAS IN DROSOPHILA

Single-gene association studies have been carried out, comparing the allele frequency between long-lived subjects and younger controls and allowing for genotype-specific relative mortality risks to be estimated. *APOE* and *FOXO3A* are the only genes consistently replicated across *Drosophila* populations in such studies (Gerdes et al., 2000; Bathum et al., 2006).

Burke et al. (2013) used next-generation DNA sequencing and compared estimated allele and haplotype frequencies in the oldest surviving *Drosophila* with those of randomly sampled *Drosophila*. Gene Ontology (GO) enrichment analysis was used to make sense of the findings, which provides a system for classifying genes based on their molecular functions, biological processes and cellular components. In this analysis, each gene may be described by multiple terms, and allows for the grouping of genes that may share common functions or processes. Findings in the Burke et al. (2013) study showed GO enrichment terms ‘defence response’ and ‘glutathione metabolic process’ being most common in their genes found under significant peaks, leading to association between bacterial defence and glutathione transferase genes with extreme longevity. This study observed that five out of eight regions with significant effects on longevity were in regions of suppressed recombination, which are much more likely

to harbour unconditionally deleterious alleles of large effect compared to regions of normal recombination. Such regions of suppressed recombination include telomeric and centromeric regions of the chromosome, suggesting that in future longevity research, these would be areas of interest to further investigate.

Ivanov et al. (2015) used lines from the *Drosophila melanogaster* Genetic Reference Panel (DGRP) to perform GWAS, observing considerable genetic variation in lifespan and broad-sense heritability. Polygenic score analysis was used to find the additive effects of common SNPs, causing a small proportion of the lifespan variation observed (~4.7%). Several of the longevity associated genes found by this study are involved in processes which are known to impact ageing, however the function of others is not known but provide opportunity for further, promising experimental examination. Several genes were identified in this study, including through gene-based analysis, in either gene regions or gene regions extended into $\pm$5 Kb of flanking sequences. These genes included *CG11523*, a gene found to have a GSK3β interaction domain that is known to be a crucial component of the TOR pathway in human cell lines and the *Neprilysin* gene that has been suggested to be essential for female fitness (Ivanov et al., 2015). Among the top-ranked 100 genes in this study (P<4.79 x 10^{-6}), *Chrb, slif, mipp2, dredd, RpS9* and *dm* were also found to contribute significantly to the enrichment of the TOR pathway with GO enrichment analysis highlighting genes involved in carbohydrate metabolism as important for lifespan. However, none of these SNPs reached genome-wide significance level prompting the hypothesis of possible combined effect of sets of SNPs on longevity.

Longevity associated genes in *Drosophila*, discussed in this chapter, have been summarised below in Table 1.1.

Table 1.1 Summary of longevity genes in *Drosophila* discussed in Chapter 1.

Longevity gene	Additional observations	Genetic Manipulation	Lifespan increase	Publication
mth (Methuselah)	Overexpression of Methuselah (*mth*) in the insulin-producing cells of the brain, extends lifespan and enhances stress resistance in flies.	Over expression	9-28% in males and 25-33% in females maximum lifespan	Gimenez et al., 2013
	Reducing signalling through Methuselah (*mth*), targeted to the insulin-producing cells of the brain, extends lifespan and enhances stress resistance in flies.	RNA interference	Maximum lifespan extended by 37% and 50% in males and females respectively	Gimenez et al., 2013
	Mutants displayed enhanced resistance to various forms of stress.	Mutation	Maximum lifespan approximately 35% higher	Lin et al., 1998
chico (chico)		Knockout	Lifespan extension up to 48% increase	Clancy et al., 2001
Indy (I'm Not Dead Yet)	Flies heterozygotic for a disruption have extended maximum lifespan of 45% and those homozygotes for the disruption show only a 10-20% increase in mean lifespan.	Knockout	Extended maximum lifespan of 45% and a 10-20% increase in mean lifespan	Rogina et al., 2000
	Reduction of Indy leads to a significant lifespan extension.	Mutations	Average female lifespan is 11% higher; Average male lifespan is 26% higher	Wang et al., 2008
Cat (Catalase)	Overexpression of catalase results in a slower rate of mortality acceleration and a delayed loss in physical performance.	Over expression	Lifespan is one-third higher	Orr et al., 1994
Sod (Superoxide dismutase)	Overexpression of catalase results in a slower rate of mortality acceleration and a delayed loss in physical performance.	Over expression	Lifespan is one-third higher	Orr et al., 1994
G6PD (Zwischenferment)		Over expression		Luckinbill et al., 1990
Hsp26 and Hsp27 (Heat shock protein 26 and 27)	Overexpression increased stress resistance.	Over expression	Average lifespan is 30% higher	Wang et al., 2004
mei-41 (meiotic 41)		Over expression		Symphorien & Woodruff, 2003
dFOXO (forkhead box, sub-group O)	Age-specific mortality analysis showed that overexpression of *dFOXO* in the fat-body of adult females increased median lowered the age-specific	Over expression	Median lifespan is 21-33% higher	Giannakou et al., 2007

	mortality compared to control flies at all ages. The effects of removal of *dFOXO* overexpression at different ages closely mirrored those of induction of expression.			
Sir2 (Sirtuin 2)	Moderate (3-fold) Sir2 overexpression in the fat body during adulthood only can promote longevity in both sexes.	Over expression	Average lifespan is 13% higher	Hoffmann et al., 2013
	Decreased expression of the Sir2 gene in all cells caused lethality during development. Suppression of the Sir2 in neurons (10-30% median lifespan reduction) and ubiquitous silencing of the Sir2-like genes shortened lifespans.	RNA interference	Median lifespan is 10-30% higher	Kusama at al., 2006
	A decrease in Sir2 blocks the life-extending effect of caloric reduction or rpd3 mutations.	Over expression	Lifespan is up to 57% higher	Rogina & Helfand, 2004
GADD45 (Growth arrest and DNA damage-inducible 45)	Overexpression in the nervous system leads to a significantly increase of *Drosophila* lifespan without a decrease in fecundity and locomotor activity.	Over expression	Maximum lifespan is up to 59% and 50% higher in males and females respectively. Median lifespan is up to 77% and 46% higher in males and females respectively.	Plyusnina et al., 2011
dInR (Insulin-like receptor)	Mutations result in dwarf females with extended lifespan of up to 85% and dwarf males with reduced late age-specific mortality	Mutation	Average female lifespan is up to 85% higher	Tatar et al., 2001
dTOR (Target of rapamycin)		Dominant negative mutation		Kapahi et al., 2004
dSk6		Dominant negative mutation		Kapahi et al., 2004
dTsc1 (dTsc1)		Over expression		Kapahi et al., 2004
dTsc2 (dTsc2)		Over expression		Kapahi et al., 2004
Hep (hemipterous)		Over expression		Wang et al., 2003
hsp68 (Heat shock protein 68)		Over expression		Wang et al., 2003
	Genetic manipulations that improve proliferative homeostasis extend lifespan.	Over expression	Average lifespan is 20% higher	Biteau et al., 2010
puc (puckered)	Heterozygous loss-of-function mutations extend lifespan and increase resistance to oxidative stress.	Mutation		Wang et al., 2003
EF-1alpha (eukaryotic translation	The decrease in protein synthesis that accompanies ageing is preceded by a decrease in elongation factor EF-1 alpha protein and mRNA.	Over expression		Shepherd et al., 1989

elongation factor 1 alpha 1)				
Rpd3 (Histone deacetylase 1)		Mutation	Mutations extend lifespan by 33% and 52% in males and females respectively.	Rogina et al., 2002
DPOSH (Plenty of SH3s)		Over expression	Average lifespan is 14% higher	Seong et al., 2001
Eip71CD (Methionine sulfoxide reductase A)	Overexpression in the nervous system extended lifespan by up to 70%, increased resistance to oxidative stress, and delayed the onset of senescence-induced decline in activity levels and reproductive capacity.	Over expression	Lifespan is up to 70% higher	Ruan et al., 2002
14-3-3-epsilon (14-3-3-epsilon)	Loss of 14-3-3e results in increased stress-induced apoptosis, growth repression and extended lifespan of flies, in a *dFOXO* dependent manner.	Mutation	Average male lifespan is up to 25% higher; average female lifespan is up to 49% higher	Nielsen et al., 2008
Cat (Catalase)	Overexpression of catalase and Sod1 result in a one-third lifespan extension, a slower rate of mortality acceleration, and a delayed loss in physical performance.	Over expression	Lifespan is one third higher	Orr and Sohal, 1994
Cu/ZnSOD (Copper- and zinc-containing superoxide dismutase)		Over expression	Average lifespan is up to 48% higher	Sun and Tower, 1999
sun (stunted)	Mutations increase lifespan and resistance to oxidative stress.	Mutation		Cvejic et al., 2004

1.8 BIOINFORMATICS TECHNIQUES USED IN ANALYSES OF GWAS DATA

GWAS have identified novel genetic variants for a wide variety of phenotypes, however these genetic variants often only account for a small percentage of the inherited component of phenotype. In recent years, longevity GWAS data have been meta-analysed, revolutionizing the field of human genetics by allowing the quantitative combination of data from multiple studies. This data combination improves the power to detect more associations to longevity and investigates the heterogeneity of these associations across diverse datasets and study populations.

Single SNP association analysis is the most commonly applied approach in GWAS, however, often SNPs identified have small effects from which limited biological insight can be inferred. More advanced approaches have therefore been used to interpret GWAS data, analysing the combined effect of a SNP set grouped per pathway or gene region. Gene and gene-set analysis are more powerful types of analyses, in comparison with single-SNP analysis (Wang et al., 2011). Whereas gene analysis tests the joint association of all SNPs in a gene that share the same phenotypes, gene-set analysis tests phenotypic association with genetic variants in a group of functionally related genes. An advantage of grouping genes as described is that this significantly reduces the amount of testing required, and potentially allows for the detection of any effects that weaker associations may have but would otherwise have been missed.

In gene-set analysis for all species, gene boundaries are to be set for which the criteria often differ between approaches. For example, it has been proposed by Wang et al. (2007) to use 500 Kb both upstream and downstream of the gene coding regions, to incorporate SNPs in non-coding regions, whereas the region proposed by Chen et al. (2010) was much smaller at 5 Kb. A gene region is often defined to include both genic region and boundary region, with linkage disequilibrium (LD) and gene regulation pattern taken into consideration. Approaches have also been proposed which involve the inclusion of SNPs in LD with the gene; such strategies proposed by Bush et al. (2009) and Hong et al. (2009) are aimed to cover SNPs playing regulatory roles in gene expression and/or linking to causal variants within the same LD block. However, the signal strength for a gene set using this approach may be reduced due

to the unavoidable inclusion of additional irrelevant SNPs or exclusion of regulatory regions located outside 5 Kb or 500 Kb flanking regions (e.g. enhancers).

An approach, based on multiple linear regression, has recently been developed incorporating LD between markers and detecting multi-marker effects, and is known as MAGMA (Multi-marker Analysis of GenoMic Annotation) (de Leeuw et al., 2015). This is a tool used for gene- and gene-set analysis of GWAS data, mapping the SNP matrix for a gene onto its principle components, eliminating those principle components with very small contribution. The linear regression model then uses the remaining principle components as predictors for phenotypes, and an F-test to compute the gene P-value. The results of a study by de Leeuw et al. (2015) showed MAGMA being not only quicker than other methods used, but also obtaining greater statistical power. When raw genotype data is not available for analysis, MAGMA also provides a SNP-wise model for so called summary statistics, for cases in which only SNP P-values are available. This analyses the individual SNPs in a gene, then combines the resulting SNP P-values into a gene test-statistic.

When individual genotype data is available, an approach suggested in Lips et al. (2012) could be used. In their study the authors used two data samples, ISC (International Schizophrenia Consortium) case control sample and GAIN (Genetic Association Information Network) schizophrenia dataset, in which all SNPs associations were analysed using additive models of allele counts. For association analyses of both datasets, Cochran-Mantel-Haenszel tests implemented in PLINK (Purcell et al., 2007), a tool set that allows rapid manipulation and analysis of large datasets. Analysis was carried out separately for the datasets, with GAIN being split into two samples. Using Stouffer's weighted Z-transport method (Whitlock, 2005), empirical P-values from the three datasets were combined and an overall P-value was obtained.

Often it is the results calculated to be most significant in mutation studies that are reported by researchers, however, in some cases SNPs may be assigned weights, dependent on the probability that they seem plausible in biological terms. For example, previously to assign weights to SNPs in GWAS, information on genome-wide linkage has been incorporated

(Roeder et al., 2006) as well as the knowledge of previous probabilities of disease association (Pe'er et al., 2006).

Wang et al. (2007) demonstrated that pathway-based approaches might complement the most significant SNPs/genes approach for interpreting GWAS data in complex studies. This study tested two alternative approaches to achieve single P-values for each gene. The first approach looked at all SNPs surrounding a gene and assigned the SNP with the most significant P-value to this gene. However, this approach increased bias in the study, as there were likely to be more-significant P-values around larger genes. The second approach computed a P-value from multiple SNPs, using a Simes method (Sarkar & Chang, 1997). For *N* SNPs ranked by their P-value, $p_{(1)}, \ldots, p_{(N)}$, the Simes P-value is calculated as min $\{p_{(i)}N/i\}$, where $1 \leq i \leq N$ (Simes, 1986). This approach calculates an overall P-value for the group of ranked SNPs, however the use of this over-conservative approach may lead to loss of power.

Genetic linkage analysis is a tool often used to detect the chromosomal location of genes which cause disease. Often linkage studies require the identification of genetic markers, usually SNPs, on a region of a chromosome. The length of this selected region is then reduced and narrowed until the gene of genetic variant of interest is identified. A whole-genome scan for genetic linkage was performed by Kerber et al. (2012) on individuals from the Utah Population Database, in which high levels of both familial longevity and individual longevity were exhibited.

For the enrichment of modest associations with a complex disease or trait, pre-specified gene sets can be evaluated, using Meta-Analysis Gene-set Enrichment of variant Associations (MAGENTA) (Segrè et al., 2010). This approach maps SNPs onto genes, with each gene then being assigned a gene association score, where this score is a function of its regional SNP association P-values. Next, cofounding effects on gene association scores are identified and corrected for, enabling the use of meta-analyses without requirement of genotype data. The final step involves a Gene Set Enrichment Analysis (GSEA)-like statistical test being applied to predefined biologically relevant gene sets, comparing these to randomly sampled gene sets from the genome, to determine whether any of the gene sets are enriched for highly ranked gene association scores.

These bioinformatics techniques described, although providing some powerful analyses of SNPs, do not explain their distribution across the genome or consider the physical 3D organization of the genome that governs co-location and co-regulation of seemingly distant regions within the 3D space of the cell nucleus. The knowledge of the physical structure of the genome generated by chromosome conformation capture techniques (so called Hi-C interaction datasets) and available for the *Drosophila* genome, combined with GWAS datasets via a network approach is proposed in this study aiming to shed a new light onto novel genetic factors associated with longevity.

1.9 HYPOTHESES, AIMS AND OBJECTIVES OF THIS STUDY

To predict novel genomic regions associated with longevity, we hypothesised that the 3D architecture of the genome governs the co-location of longevity-associated genes/genomic regions with novel unknown regions that may share biological functions of importance to the process of longevity.

To identify novel longevity-associated genes we further hypothesised that SNPs in genes, residing within co-located genomic regions and sharing the same biological function, may influence longevity either independently or have a cumulative effect on longevity i.e. alterations in one or several genes may be responsible for the same longevity-related phenotype.

To explore the role of non-coding SNPs we hypothesised that SNPs residing in topologically associated domain (TAD) border regions may cause disruption to TADs and a change in expression of the nearby gene(s) by forming looping interactions with regions in adjacent TADs and “hijacking” regulatory elements residing within these adjacent interacting regions.

To assess the occurrence of non-coding SNPs in transcription factor binding sites (TFBSs), we hypothesised that transcription factors may recognise a certain genomic structure, e.g. non-B DNA structures, rather than specific sequence motifs.

We further hypothesised that both non-coding SNPs and their potential target genes also reside within co-located loci.

The aims of this study were to develop a mathematical model based on network approaches that utilise the knowledge that we have of the 3D structure of the *Drosophila* genome and two GWAS datasets, along with known longevity-associated genes for predicting novel genes/genomic regions that may play a role in longevity. This approach considered the biological functions of genes observed, focussing on those that are longevity related.

To explore the occurrence of SNPs in real datasets for TADs and TFBSs, their occurrences were compared against matched control datasets. For TFBSs, their occurrence in non-B DNA structures such as slipped, cruciform, triplexes and tetraplexes, formed on direct, inverted and mirrored repeats and G-quartets were considered and over-representation of non-coding SNPs in these structures was explored.

To identify target genes for non-coding SNPs, Hi-C data with a higher resolution was used. This analysis sought for regions that had the strongest interaction frequencies with regions containing non-coding SNPs, where genes residing in these highly interacting regions were then further explored as target genes for these non-coding SNPs.

1.10 STRUCTURE OF DISSERTATION

Following from this introduction chapter, in Chapter 2 the data used in the analyses of this dissertation is described and the laboratory techniques used to obtain this data are discussed. In Chapter 3, networks are introduced and various network measures are discussed. In this chapter, the statistical tests and bioinformatics techniques used in the subsequent analyses performed in this study are also described, as well as the software used for biological interpretation. The methods and results of network analysis for coding SNPs are discussed in Chapter 4. The method and results for the analysis of non-coding SNPs in TAD borders and TFBS regions and for identification of target genes for non-coding SNPs are discussed in Chapter 5. The results from all analyses described in Chapters 4 and 5 are then summarised in Chapter 6, and their implications are discussed.

Chapter 2

DATA DESCRIPTION

This study focuses on the use of two *Drosophila* longevity GWAS datasets along with Hi-C data providing fine resolution chromatin structure insight of *Drosophila*. We discuss how the GWAS data was obtained and also look at different molecular techniques used over the last few decades which have allowed for interactions between loci in the genome to be quantified and consider their strengths and weaknesses. In this study, a number of datasets have been used to look at specific genes or regions on the genome for further exploration of their structures and how they interact. In this chapter we describe the datasets used to map positions of Transcription Factor Binding Sites (TFBS) and Topologically Associated Domains (TADs) to the *Drosophila* genome, and also explain how matched control datasets were created for comparative analysis.

2.1 GWAS DATASETS OF DROSOPHILA USED IN THIS STUDY

The published datasets generated by Burke et al. (2013) and Ivanov et al. (2015), containing respectively ~2.3 million and ~2 million SNPs, were used as references sources in this study. In both datasets, the data presented allowed for the calculation of, or provided information about, each recorded SNP's significance of association with longevity.

2.1.1 Dataset 1: Synthetic GWAS Dataset

The first dataset was obtained from a study by Burke et al. (2013) in which a "synthetic" population of *Drosophila* was derived from a small number of inbred founders; it will be referred to as the Synthetic GWAS dataset. These founders consisted of two independent sets of seven inbred *Drosophila* lines with another founding line added to both sets were crossed to initiate two synthetic recombinant populations, A and B. Populations A and B were then maintained as four independent large populations (A1/A2, B1/B2). Next-generation sequencing was used to identify allele frequencies in the 'young' control group, comprising 120 14-day-old females, and the last surviving ~2% of females from the same cohort (an 'old' group) (see Figure 2.1 for a timeline summary). The occurrence of SNPs in each of the eight 'old' samples and eight 'young' control samples was recorded, resulting in ~1.2M SNPs in the A populations and ~1.1M SNPs in the B populations [see Burke et al. (2013) for details]. The SNPs for both populations were combined and for this dataset, when duplicates of SNPs were observed across populations, each was recorded only once by combining the haplotype data for each and summing the allele frequencies for calculation.

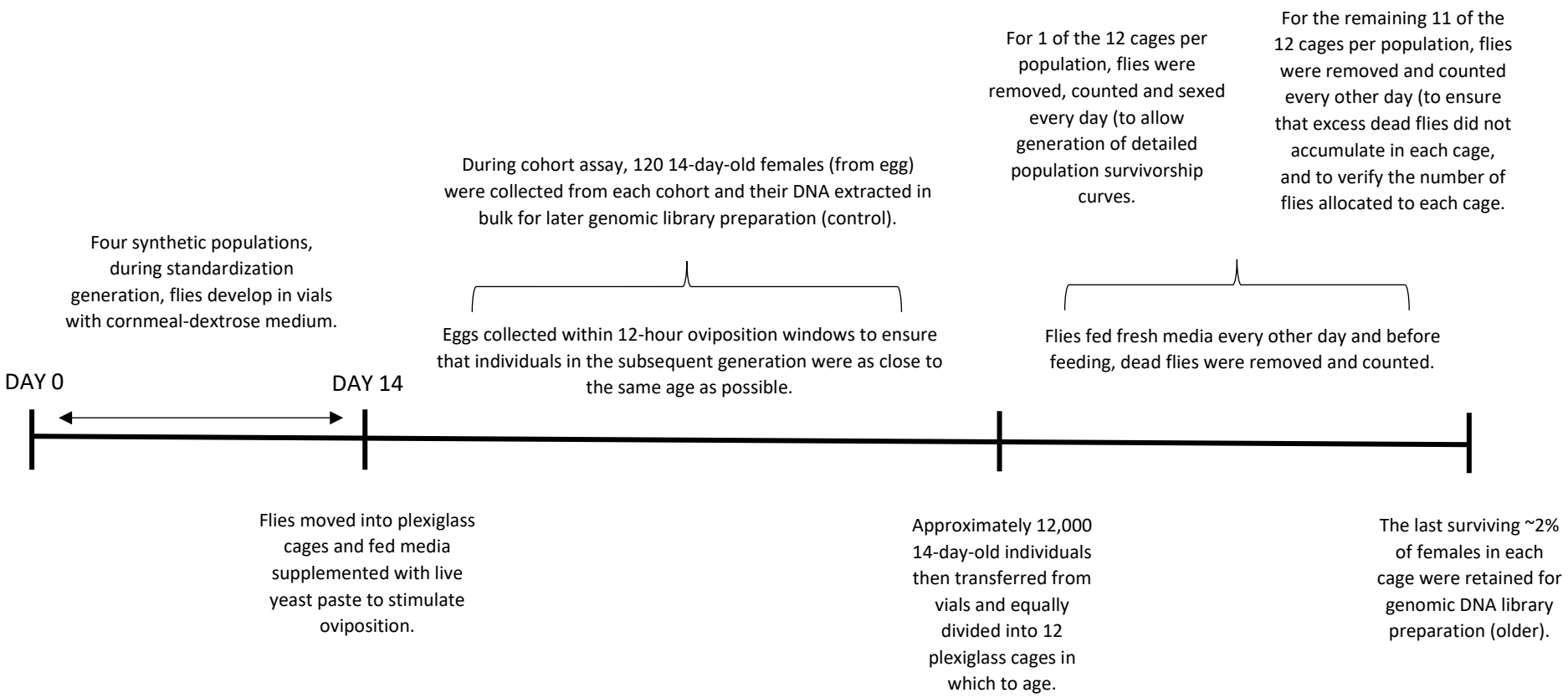


Figure 2.1 A timeline summarising the way in which DNA was extracted to form the control and older dataset in the Synthetic GWAS dataset (produced using information from Burke et al. 2013).

2.1.2 Dataset 2: *Drosophila* Genetic Reference Panel (DGRP) GWAS Dataset

The second dataset was obtained from a study by Ivanov et al. (2015) in which GWAS was performed on the *Drosophila* Genetic Reference Panel (DGRP), Freeze 2.0 (Mackay et al., 2012, Huang et al., 2014). This comprises 205 *D. melanogaster* lines derived from 20 generations of full-sib mating from inseminated wild-type females caught from Raleigh, North Carolina. Lifespan data was available for virgin females for 197 DGRP lines, with ~25 females per line. A total of 2,193,745 SNPs were recorded together with the corresponding P-values, quantifying association with lifespan. P-values were calculated using linear regression under an additive model with four first principal components and the presence/absence of *Wolbachia pipientis* infection included as a covariate [see Ivanov et al. (2015) for details]. Henceforth, this dataset will be referred to as the DGRP GWAS data.

2.2 BIOLOGICAL TECHNIQUES FOR UNRAVELLING 3D CHROMATIN STRUCTURE

Chromosome conformation capture (3C) techniques have been developed to study chromatin structure at a much finer resolution than by using microscopy. Chromosome conformation capture (3C) technique is a set of molecular biology methods used to quantify interactions between loci in the genome. The protocol determines DNA contact frequencies by quantifying the number of interactions between loci which is inversely proportional to the distance between loci within the 3D cell nucleus. These interactions include those between loci separated on the linear genome by up to thousands of nucleotides or that even occur on different chromosomes. Such interactions may result in biological functions, for example promoter-enhancer interaction may influence expression. In 2002, Dekker et al. (2002) were the first to report 3C assay, and since then it has become the most frequently used method to demonstrate interactions between two unique loci (Dekker et al., 2002).

2.2.1 Chromosome Conformation Capture (3C)

Evolving both qualitatively and quantitatively, 3C based methods have gradually improved as technology has advanced. There are now several 3C methods available, with the main difference between them being their scope. The method with the smallest scope is 3C, known

as a ‘one versus one’ strategy, which quantifies the interactions between two predefined loci on the genome, which can be either inter- or intra- chromosomal (Fullwood et al., 2009). The interaction frequency measured between two fragments correlates to their spatial closeness in the 3D structure of the genome.

The initial step in 3C and 3C-derived methods is to establish a representation of the 3D organization of the DNA (de Wit & de Laat, 2012). Chromatin in the DNA is fixed, often using formaldehyde as a fixative (Dekker et al., 2002) and restriction enzymes recognizing a certain sequence of base pairs (bp) are then used to cut the fixed chromatin. A simple example of this cutting is given in Figure 2.2, in which it is shown where DNA would be cut at a selected 2 bp and 3 bp sequence pattern recognised by restriction enzymes. The sticky ends of the cross-linked DNA fragments are re-ligated under diluted conditions to promote intramolecular ligations, for example between cross-linked fragments. This allows for the ligation of those DNA fragments that co-localise in 3D space but are not in close proximity on a linear genome. The 3D conformation of a locus is then established by measuring the number of ligation events between neighbouring or non-neighbouring sites. In 3C, this is done by PCR amplification of selected ligation junctions (de Wit & de Laat, 2012). The steps of this method, and all other 3C methods discussed in this chapter, are illustrated in Figure 2.3.

Resolution of this technique will depend on the size of restriction enzyme used. A large base cutter results in less frequent digestion of fragments of DNA, resulting in larger distances between cuts of DNA and therefore lower resolution of data. Although more complex, a small base cutter provides a higher resolution of analysis due to more frequent digestion of fragments of DNA, resulting in shorter distances between cuts.

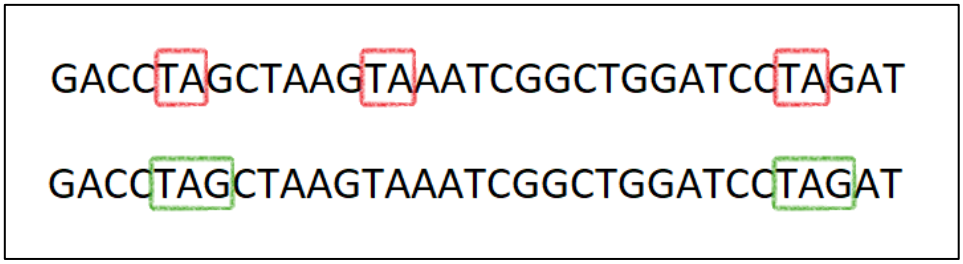


Figure 2.2 A hypothetical 2 bp cutter shown on the first string of DNA, in which the DNA is cut at a recognised pattern ‘TA’, resulting in shorter DNA fragment cuts and therefore higher 3C resolution than the second string using a 3 bp cutter, recognising the base pattern ‘TAG’.

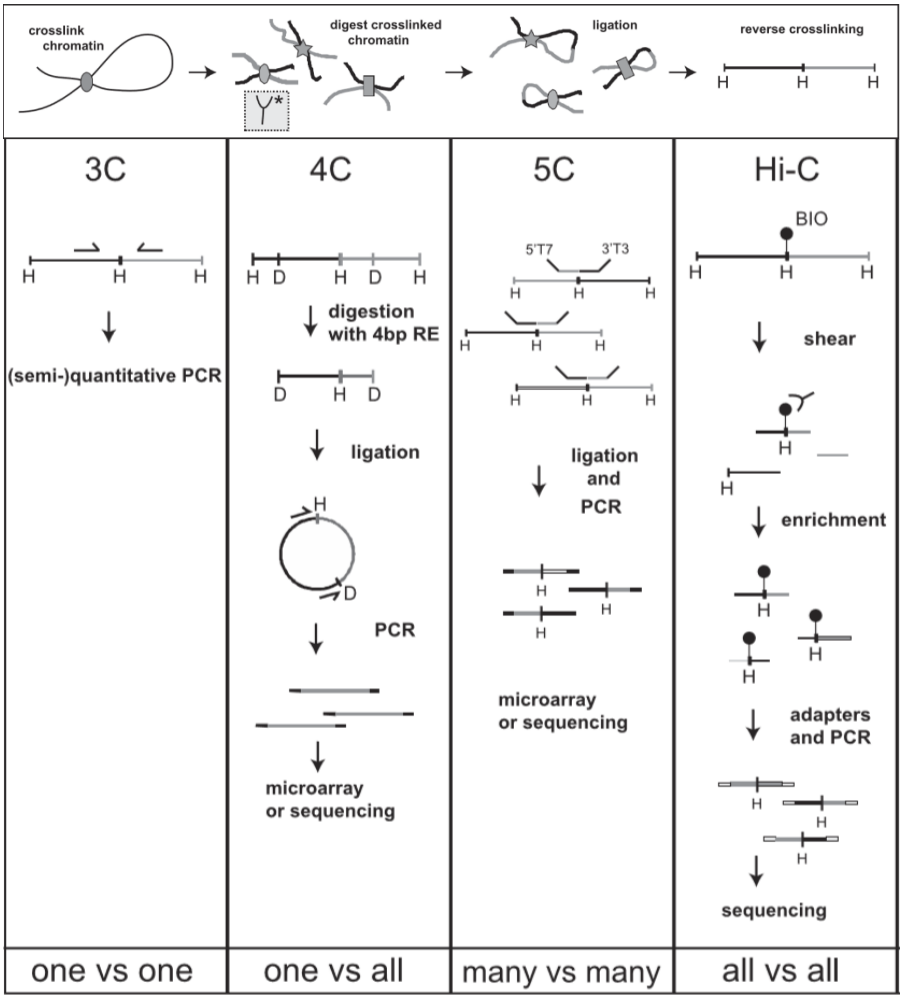


Figure 2.3 An overview of the 3C-derived methods described in this chapter, where the top panel shows the steps common to all methods, and the vertical panels below explain the following steps, which are specific to each different 3C method. 'H' represents a 6 bp cutter restriction enzyme and 'D' a 4 bp cutter restriction enzyme (de Wit & de Laat, 2012).

3C is most effectively used in cases when there is prior knowledge about the region/s of interest as this technique works most efficiently when just one region (or a small number of regions) are selected to explore. As well as 3C's inefficiency to create large libraries of interaction frequencies with ease, there are other limitations of this technique, which includes the inability to determine the proximity of individual haplotype chromosomes – lacking the ability to distinguish whether the long-range contact is made between the paternal or the maternal chromosome, or both (Barutcu et al., 2016). The accuracy of detection of interactions is also limited, those within a range of approximately 1 megabase (Mb) are attainable, however as interaction distance range increases, accuracy of technique decreases.

2.2.2 Chromosome Conformation Capture (4C)

The larger the distances become between separated sites, the more infrequent ligation products become to allow accurate quantifying by 3C. A new technique was required whereby if there was no prior knowledge about interacting region/s, for example, if we consider target genes for SNPs in non-coding regions, in which the position of a SNP is known but their target genes are not. The position of a SNP's target gene is not necessarily that lying directly next to each SNP on the genome, and so an approach was required in which the interactions between one region with all other regions are specified.

A solution for this was the combination of 3C technology with microarrays, used to analyse the contacts of a selected genomic site with all of the genomic fragments that are represented on the array. A number of approaches referred to as 'one versus all' were developed, all referred to as 4C methods. These methods include chromosome conformation capture-on-chip (ChIP), open-ended 3C, olfactory receptor 3C and circular 3C, in which interactions are captured between one locus and all other genomic loci. The most popularly of these methods is ChIP which, in brief, processes the ligated 3C template with a second round of DNA digestion and ligation to create small DNA circles (de Wit & de Laat, 2012). These DNA circles are then used to perform inverse PCR, allowing the known sequence to amplify the unknown sequence ligated to it. These sequences are then analysed using microarrays (de Wit & de Laat, 2012).

Results obtained from a 4C experiment should be carefully looked at to ensure that for all cases, high interaction frequencies correspond to existing long-range interactions. If this is not the case, it may be that frequent ligation events have caused poor crosslinking during experiment, resulting in a high false positive rate from frequent random ligation events (Barutcu et al., 2016).

2.2.3 Chromosome Conformation Capture Carbon Copy (5C)

Similar to 3D methods, prior knowledge of regions is required for 5C techniques, however with the difference, and advantage, being the size of the predefined region/s of interest. Unlike the 3C method, this 5C method is able to explore a group of neighbouring regions,

hence why this approach is known as a 'many versus many' strategy. This method processes the ligated 3C template again, with detections then made by ligating universal primers to all fragments. DNA is recombined back into double-helix form at ligation junctions, as it was found in the original 3C library, with the use of hundreds of primers. Finally, DNA fragments are amplified using PCR and deep sequencing is used to detect genomic location (Ferraiuolo et al., 2012).

Such technique design has useful biological implications, for example for identifying important processes in the genome occurring as the result of physical interactions between enhancers and promotors with genes, with a detection range from a few base pairs up to approximately 1Mb. However, one disadvantage of this technique due to a somewhat limited detection range is its inability to measure more remote interactions, with another disadvantage being that with a relatively low coverage, it is deemed unsuitable for conducting genome-wide complex interaction search, and therefore in order to derive useful results, prior knowledge is again required.

2.2.4 Hi-C

Finally, the first of the 3C methods to be truly genome-wide is Hi-C. Using high-throughput sequencing to find the nucleotide sequence of fragments, this was the first 'all versus all' strategy to be developed. In this method, the procedure for creating a 3C template is slightly adjusted, and before ligation, the restriction enzymes are filled in with biotin-labelled nucleotides (Lieberman-Aiden et al., 2009). This facilitates selective purification of ligation junctions that are then directly sequenced (Dekker et al., 2013). This 'all versus all' approach is ideal for the exploration of genome folding in which there is no prior knowledge of regions of interest in a study, especially in the case of my studies in which the aim is to discover new regions of the genome containing genes with potential to influence longevity.

Despite being the most advanced of the techniques previously discussed, Hi-C still has its limitations, with the first being due to the way in which the interaction frequencies are obtained. Such interactions are measured between millions of cells at one time, with data from these interactions resulting from a so-called 'snapshot' or a population average and

therefore from this data it is not possible to draw conclusions about 3D proximity within specific cells. Another aspect of this technique to consider is the heterogeneity of fragment digestion, which does not guarantee the specific whereabouts of restriction enzymes cutting at their target sequence. The inability to control the lengths of fragments being cut is something that should be taken into consideration when choosing the size of binned regions to explore interactions for, with smaller bin regions having a higher frequency resolution due to the way in which technique is performed and interactions are counted. All chromosome conformation capture techniques discussed are summarised in Figure 2.4, in which the coloured boxes represent different interacting fragments on the genome and the coloured arrows between these fragments indicate the interactions identified by different techniques.

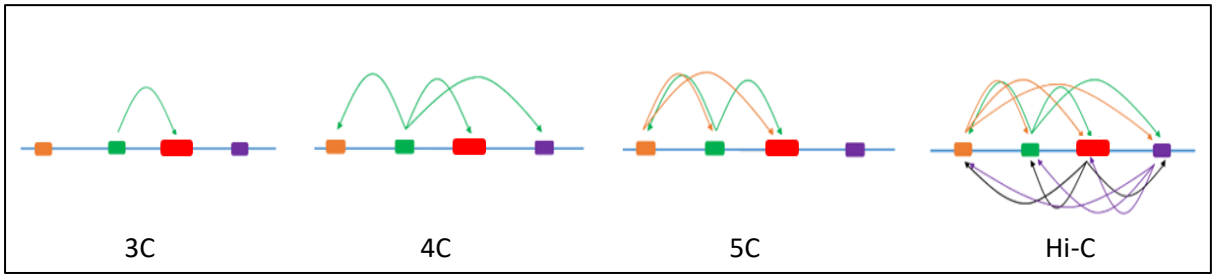


Figure 2.4 Schematic representation of fragment interactions measured by chromosome conformation capture techniques, 3C, 4C, 5C and Hi-C.

Interaction frequencies presented in Hi-C datasets take values greater than 1 due to the way in which these frequencies are measured, as shown in Figure 2.5. Chromosomes are cut using restriction enzymes recognizing a specific pattern, as previously mentioned in Figure 2.3 as 'H' and 'D', in this example 'AATAGC'. The interactions between these unevenly cut portions are then measured. These unevenly cut portions are then binned (grouped and separated) into 80,000 base regions, in this example. The frequency of the interactions previously measured by 3C techniques are then counted between these 80 Kb bins. In the example in Figure 2.5, there would be one interaction counted between bin 1 and itself, one interaction between bin 1 and bin 2, two interactions between bin 2 and itself and between bin 2 and 3 there would also be two interactions counted.

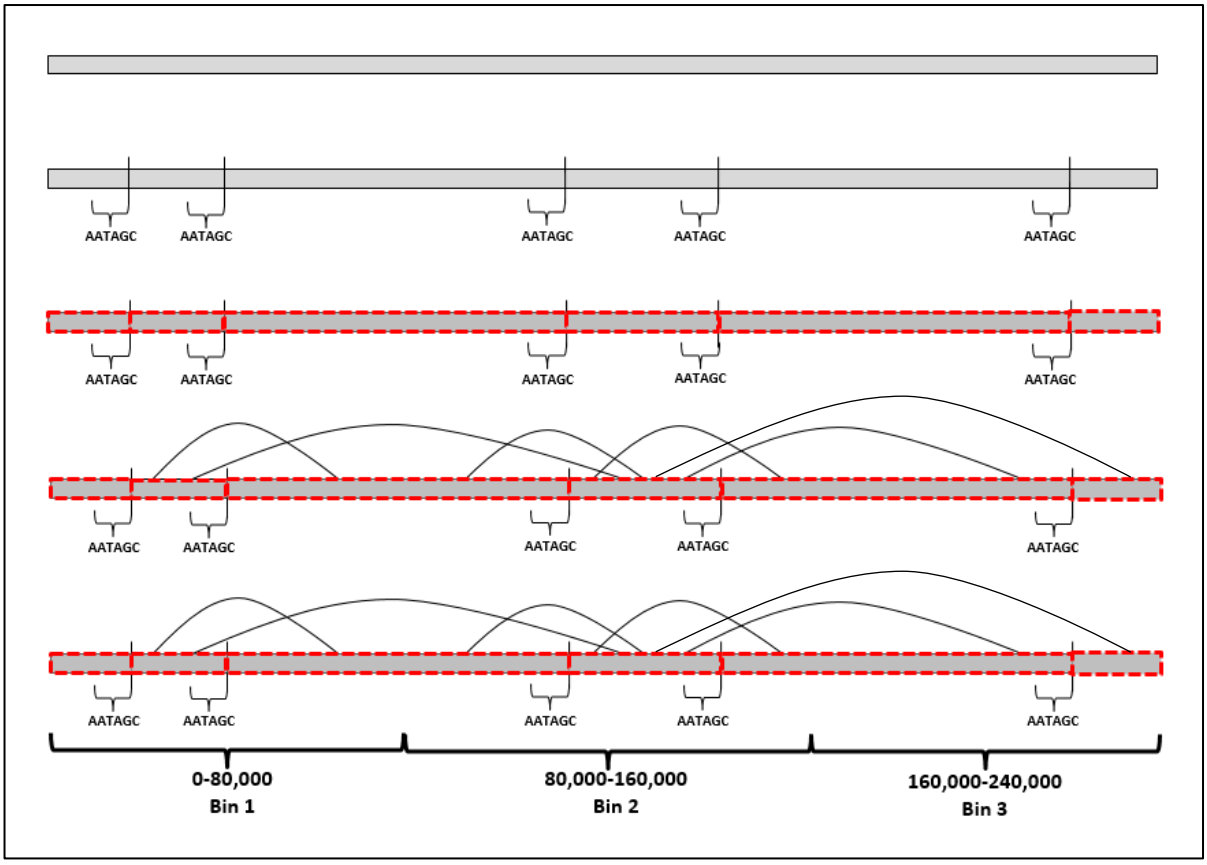


Figure 2.5 Counting interaction frequencies, a basic example of how a restriction enzyme would be used to cut a chromosome and the way in which interaction frequencies are measured between the cut portions.

2.3 HI-C INTERACTION DATASET FOR DROSOPHILA

To measure co-location of genomic regions, datasets of intra- and inter-chromosomal interactions at 10 Kb and 80 Kb resolution, obtained by Sexton et al. (2012) and normalised to avoid any biases introduced by the experimental procedure, were downloaded from GEO database (accession number GSM849422). A low-frequency Hi-C read processing was used to normalise this interaction data (Yaffe and Tanay, 2011). A dataset of chromosome positions and their corresponding bins (example for 80 Kb bins shown in Table 2.1), is used to interpret Hi-C data and to identify the exact positions for which interaction data is provided.

Table 2.1 An example of Hi-C bin position data and corresponding genomic regions.

Bin number	Chromosome	Start position	End position
1	2L	0	80000
2	2L	80000	160000
3	2L	160000	240000
4	2L	240000	320000
5	2L	320000	400000
6	2L	400000	480000
7	2L	480000	560000
8	2L	560000	640000
9	2L	640000	720000
10	2L	720000	800000

*Note that in this Hi-C dataset, bins 1-287 correspond to chromosome 2L, bins 288-551 to chromosome 2R, bins 552-858 to chromosome 3L, bins 859-1207 to chromosome 3R, bins 1208-1223 to chromosome 4 and bins 1224-1503 to chromosome X.

An example of a Hi-C interaction dataset shown in Table 2.2, consists of three columns. The first two columns describing 80 Kb regions, are represented by the corresponding bin numbers, with bin 1 containing positions 0-80,000, bin 2 containing positions 80,000-160,000 and so on as recorded in Table 2.1. The third column contains the observed frequency of interactions between these two bins stated.

Table 2.2 Format in which Hi-C interaction data is recorded.

bin1	bin2	normalised count
1	1	2457
1	2	984
1	3	204
1	4	152
1	5	92
1	6	83
1	7	78
1	8	56
1	9	44
1	10	36

2.4 TRANSCRIPTION FACTOR BINDING SITE/CIS-REGULATORY MODULES DATASET

A comprehensive database of experimentally verified *Drosophila* regulatory sequences, comprising of both *cis*-regulatory modules (CRMs) such as enhancers, silencers and proximal promotor sequences and TFBSs is provided by RedFly (http://redfly.ccr.buffalo.edu). The CRMs dataset, in which there are currently 23,990 CRMs associated with 1604 genes, provides genomic positions for all modules. The TFBS dataset provides the names of all recorded TFBSs in *Drosophila*. This database includes all experimentally verified fly regulatory elements, including recorded start and end genome positions of each TFBS, its corresponding binding TF and in some cases, a target gene. Sequences for each TFBS were also given by this database, with sizes ranging from 3 bp to over 1000 bp, and for many TFs, there were a number of corresponding TFBSs with sequences recorded. The total number of TFBS sequences recorded in this dataset at the time of analysis was 2209. These 2209 recorded TFBSs recognised by 192 transcription factors, acting on 248 target genes.

Each TFBS sequence was extended by 50 bp both up- and down-stream during the pre-processing of data, using TFBS positions and whole chromosome datasets. This extension of sequences enabled the taking into account of flanking regions and allowing for consideration of the true size of the TFBS not just the region to which a specific TF binds.

2.5 TOPOLIGICALLY ASSOCIATED DOMAINS DATASET

A database from Supplementary Data 1 in Ramírez et al. (2018) was used, in which TAD region positions (beginning and end) were recorded for the *Drosophila* genome. The authors achieved a high sequencing depth, with the use of *DpnII* as a restriction enzyme to identify these TADs. A total of 2846 TAD regions were recorded with a median TAD length being 26 Kb. This study by Ramírez et al. (2018) obtained Hi-C data for Kc167 cells from Li et al. (2015) and Cubeñas-Potts et al. (2017), producing corrected Hi-C contact matrices at restriction-fragment resolution. This TAD data was post-processed to compile TAD border regions for further exploration in this study.

The post-processing of TAD data in this study involved selecting the borders of each TAD by taking the base position at which a TAD ends and the next adjacent TAD to this begins and adding 100 bp to each side of this position. Therefore, a total of 2847 TAD borders were

selected, where all TAD border regions had a length of 200 bp. This selected length took into consideration the smallest and largest lengths of TAD regions recorded, ensuring that the TAD border region allowed for observation of a large enough region without any overlapping with other TAD border regions. Matched controls (see below) were also created for the “real” TAD border regions. A total of 100 controls were created for each TAD border region, each of length 200 bp selected from the same chromosome but not overlapping with any “real” TAD borders.

2.6 CREATION OF MATCHED CONTROL DATASETS

To assess the significance of findings, matched control datasets of TFBSs and TAD boundary regions were created for statistical analysis. For each sequence constituting TFBS or TAD boundary regions recorded in the “real” dataset, a control sequence was selected from the same chromosome from which the real data was derived and matching the length of the real sequence. These control sequences were also selected from regions on the genome which do not harbour a gene or region of interest (TFBSs or TAD boundary regions). This process was repeated 100 times creating 100 control datasets. The frequencies of searched features were then averaged (divided by 100) to obtain a matched control frequency to compare with the corresponding frequency of the searched feature in the real dataset. A simple example of how a dataset of controls is created is shown in Figure 2.6.

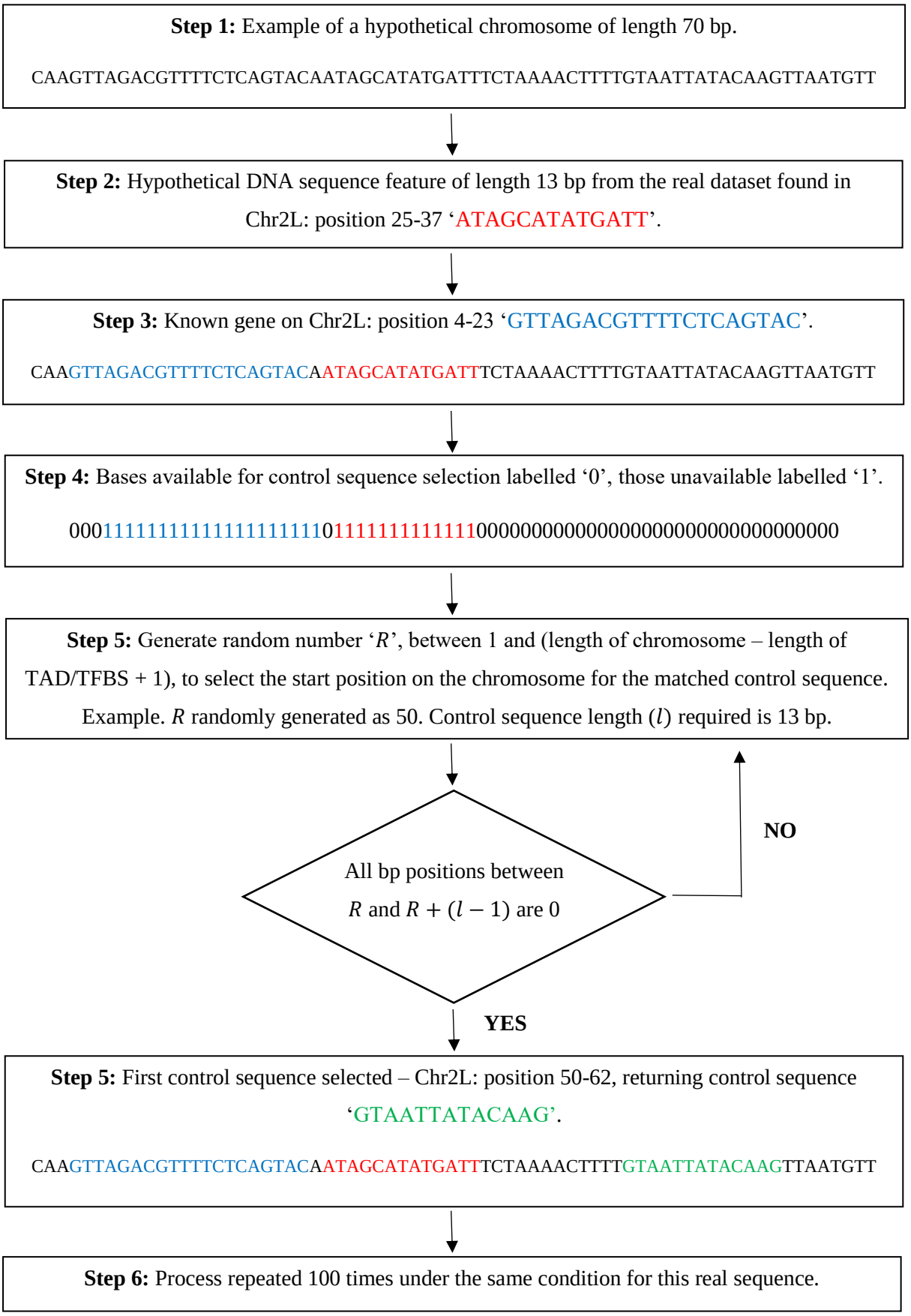


Figure 2.6 Flow chart summarising how matched control datasets are obtained for a given DNA sequence from the real dataset.

Chapter 3

NETWORKS AND BIOINFORMATICS TECHNIQUES USED IN THIS STUDY

In this chapter we introduce both the fundamentals of network theory and a specific statistical test that was carried out in the analyses of this study, which are used for identifying novel genomic regions and target genes (see section 1.1). To ease the understanding we accompany our explanations with examples. We also discuss additional statistical and bioinformatics techniques that were used in this study. The bioinformatics techniques used for the pre-processing of GWAS data and enabling further analysis together with an explanation how the genes suggested for exploration by our analysis results can be analysed to enable important biological interpretation are discussed in this chapter. The software used for biological interpretation is also described.

3.1 NETWORKS

3.1.1 Introduction to Networks

Mathematically speaking, a network, N, is a pair of disjoint sets (V, E) in which V is a non-empty finite set of elements referred to as nodes or vertices and E is a finite set of distinct pairs of elements of these vertices, referred to as edges (Newman, 2018). In networks, nodes are the points of interest, for example representing people, and edges exist between these nodes if there is a relationship or connection between what these nodes represent, therefore in this example this could be friendships. Figure 3.1 shows a simple example of a network, in which nodes are labelled by numbers from 1 to 6, and the lines joining these nodes are edges. In a more biological setting, the nodes of a network could represent genomic regions and the edges connecting these nodes would then represent some biological relationship between such regions. Similarly, edges can also be points of interest, e.g., in some graph-based genome assembly methods. Information such as this is most likely to be displayed in what is known as a simple network, in which a node cannot have a loop, which is an edge that connects a node to itself, and any two nodes cannot be connected by more than one edge. Networks can be labelled as either directed or undirected; directed networks are networks where all edges are directed from one node to another and undirected are networks in which all edges are bidirectional.

3.1.2 Adjacency Matrices

Networks can be represented in matrix form. The most commonly used representation is the *adjacency matrix*, also referred to as the connection matrix, which contains rows and columns labelled by graph nodes. Consider an example of a network shown in Figure 3.1, in which $N = (V, E)$ where $V = \{1,2,3,4,5,6\}$ and $E = \{(1,2), (2,3), (2,4), (2,6), (4,5), (4,6), (5,6)\}$, $(|V| = 6)$ where, for example, the first edge (1,2) represents a connection between node 1 and 2. The elements of the adjacency matrix depend upon whether the network's edges are weighted or not. In non-weighted cases, the matrix is binary, in which a '1' or '0' indicates the presence of a corresponding edge in a network or not. For cases in which edges are weighted, whereby each edge is assigned a weight according to the strength of the interaction between

two connecting nodes, the matrices are also weighted, for example if nodes represented cities then the edge weights could represent the distance between these cities.

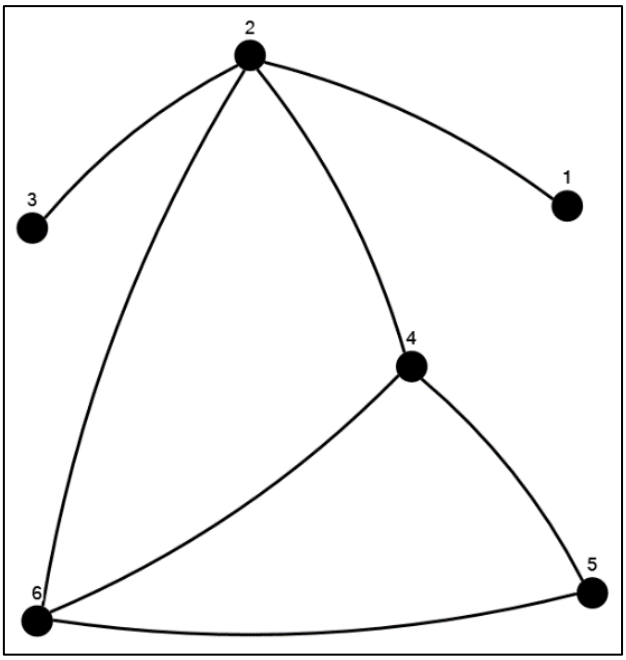


Figure 3.1 An example of a simple, undirected and non-weighted network.

The adjacency matrix of the network shown in Figure 3.1 is given below:

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \end{pmatrix}.$$

Looking at this matrix, it is clear that it is symmetric, with $A = A^T$, which is a property of matrices for undirected networks.

Directionality can also be incorporated into networks, this can happen when edges existing between i,j do not necessarily exist between j,i and *vice versa*. An example of a directed network is shown in fo.

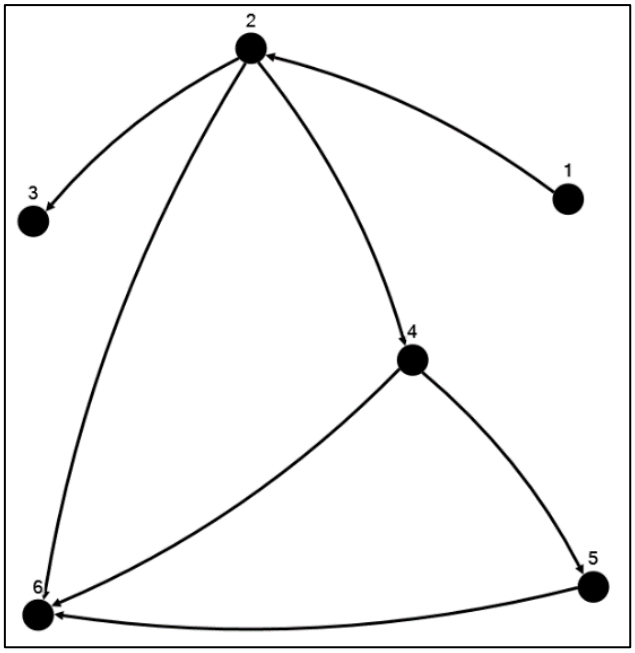


Figure 3.2 An example of a directed, non-weighted network.

The adjacency matrix of the network shown in Figure 3.2 is given below:

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Note that this matrix is not symmetric, which is common in a lot of cases for directed networks. However, if a network is weighted, the elements in the corresponding matrices take values from the real number line. If we were to take a simple example of a group of friends (each friend represented by a node) and how often they visit each other every month (represented by the interactions between each of the nodes), in a network for this example, the higher weights represent more frequent visits between two friends, and are shown on a network by the thickness of the edge between the two nodes, with a thicker edge representing a larger number of visits. Therefore we can see in the network example in Figure 3.3 that the friends who visit each other most often are friends 4 and 6.

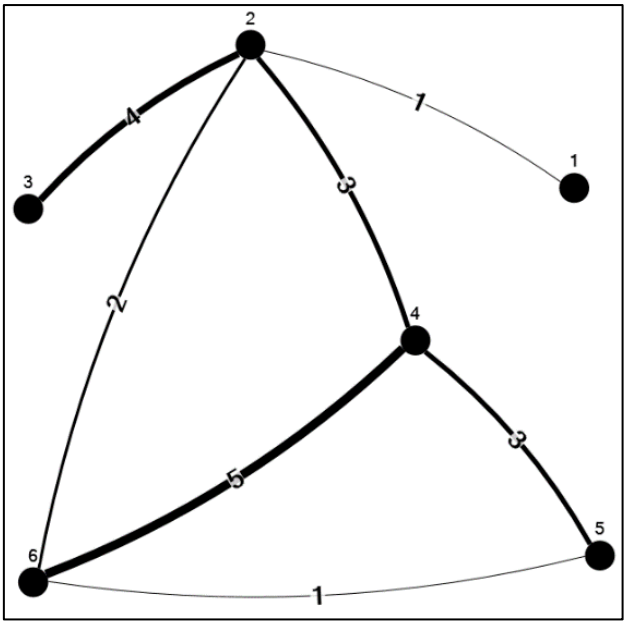


Figure 3.3 An example of weighted, undirected network in which the thickness of an edge indicates the weight.

The adjacency matrix of a network shown in Figure 3.3 is given below:

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 4 & 3 & 0 & 2 \\ 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 3 & 5 \\ 0 & 0 & 0 & 3 & 0 & 1 \\ 0 & 2 & 0 & 5 & 1 & 0 \end{pmatrix}.$$

3.1.3 Network Properties and Statistics

Degree distribution

In a network, the degree of a node is the number of connections it has to other nodes. Relating this back to adjacency matrices, the degree of node v_i can be calculated by summing the i^{th} row of the adjacency matrix. In the case of a directed network, an in- and out-degree for each node can be calculated, in which the in-degree sums the number of incoming edges and the out-degree sums the number of outgoing edges. From the adjacency matrix, the in-degree for each node can be computed by summing the entries of the corresponding column and the out-degree can be computed by summing the entries of the corresponding row.

The degree distribution shows the degrees of nodes across the whole network, and plots the probabilities of choosing a node with degree *k* as

$$P(k) = \frac{n_k}{n},$$

where *n* is the total number of nodes in the network and n_k is the number of nodes with degree *k*. Separate degree distributions, in- and out- degrees, are calculated for directed networks.

Path length

A path is defined as a finite sequence of edges connecting node v_i to node v_j through a chain of distinct nodes. It is possible for many paths to exist connecting these nodes together, especially when the network is big, however in unweighted networks the path containing the fewest number of edges is known as the shortest path, denoted by d_{ij}. In weighted networks d_{ij} is calculated by taking into consideration not the number of edges, but the weights attached to these edges. For example, if edges were weighted according to the distance between buildings, there may be a direct path between nodes v_i and v_j of let say 3km, but there may also be two shorter edges at 1km each connecting v_i and v_j, via a node v_z, in which case the shortest path would not have the smallest number of edges. In this case, the edge weights have been summed and the smallest was chosen to select the shortest pathway. However, there are some networks in which the most important edges connecting nodes will be represented by the highest weights. In cases such as these, often the reciprocals of these weights are calculated, and the weights of all edges between v_i and v_j are summed, with the shortest path still representing the desired path.

Characteristic path length

The most commonly computed path-length of a network is the 'characteristic path-length', which for any connected graph G, is defined as the average distance between pairs of nodes. In this computation, for any two nodes v and v' in $V(G)$, let $L(v, v')$ denote the shortest path length connecting v to v'. $L(v)$ then denotes the average of $L(v, v')$ across all nodes v' in

$V(G)$ where v' is not equal to v. The characteristic path length $L(G)$ of G is then defined as the average of $L(v)$ across all nodes v in $V(G)$.

Clustering Coefficient

The clustering coefficient is a measure of how a node's neighbours are connected with each other in a network or the degree to which these nodes tend to cluster together. There are two main clustering coefficients used in the studies of networks: Watts-Strogatz (Watts and Strogatz, 1998), which can be defined both locally and globally; and the transitivity index (Wasserman and Faust), which is a global measure of clustering. In this study, only Watts-Strogatz clustering was considered.

The Watts-Strogatz clustering coefficient of a node, with a degree ≥ 2, is the probability that any two randomly chosen neighbours of this node are linked together. A node's clustering coefficient can be calculated by counting the number of triangles formed including this node and dividing by the number of possible edges between its neighbours.

The formula for the clustering coefficient can be written as:

$$C(i) = \frac{t_i}{k_i(k_i - 1)/2} = \frac{2t_i}{k_i(k_i - 1)}$$

in which k_i is the degree of the i^{th} node and t_i is defined as the total number of triangles centred on node i.

A simple example can be used to show how to calculate the local clustering coefficient of reference node, *N,* the red node shown in Figure 3.4. The three steps shown in Figure 3.4 explain how to calculate local clustering coefficient, looking at the original connections of a network, specifically those with the reference node, finding all possible edges that would connect the neighbouring nodes of the reference node and observing which of these possible edges actually exist in the network. Using this example, for node *N*, a local clustering coefficient is calculated as $C(N) = \frac{5}{6} = 0.833$.

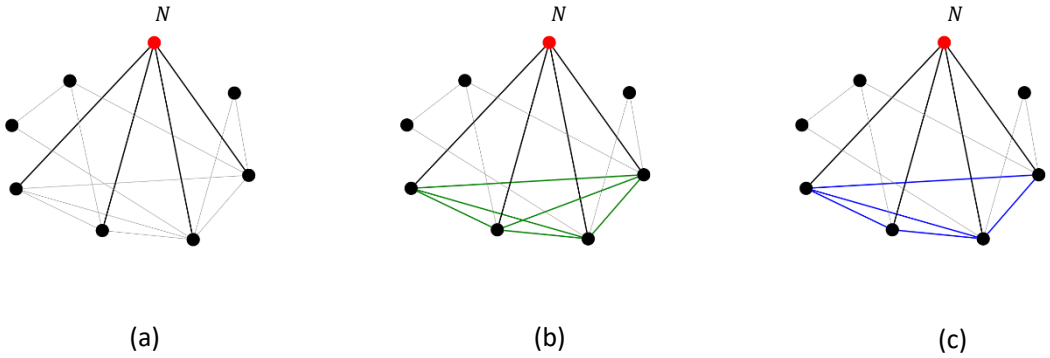


Figure 3.4 An example of how to calculate the local clustering coefficient of red reference node, N. The three steps shown calculate as follows: (a) the original network in which the edges between N and its neighbours are shown with bolder black lines. (b) All possible edges connecting the four neighbouring nodes of N, of which there are six, shown in green. (c) The five connections actually existing in the network, between the four neighbouring nodes of N, shown in blue.

The global clustering coefficient ($\bar{C}$) does not look at each node individually, but instead looks at how connected a network is relative to its number of nodes. The global clustering coefficient of a network is calculated by taking the average of local coefficients:

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C(i),$$

where n is the number of nodes in the network and $C(i)$ is a local coefficient. The calculation of global clustering coefficients enable comparison between different networks and assess the extent to which they cluster.

This global measure can be applied to both undirected and directed networks, however cannot be applied to weighted networks. Instead, a generalization of the global clustering coefficient to weighted networks was proposed by Opsahl and Panzarasa (2009). This generalization requires a triplet value to be defined, for which there are several methods proposed, including the arithmetic mean, geometric mean, and the maximum and minimum of the two tie weights that make up the triplet. The method chosen for defining this triplet value should be appropriately selected due to its impact on the outcome of the coefficient and should bear in mind the way in which the strength of the ties are incorporated into

weights. The value of closed triplets is then divided by the value of all triplets of the weighted network.

Modularity

Modularity measures the structure of a network by looking at the number of clusters formed by nodes, also referred to as modules. By looking at the way in which nodes group together, the strength of the division of a network can be calculated; therefore, a higher modularity score for a network suggests more dense connections between nodes that have been categorized in the same module, but fewer connections between nodes belonging to different modules. To extract the community structure of networks, Blondel et al. 2008 developed a heuristic method based on modularity optimization.

The problem of community detection requires the partition of a network into communities of densely connected nodes, where the nodes that belong to different communities are only sparsely connected. A number of algorithms have been proposed to find these network partitions required, and the quality of these selected partitions is commonly measured using the so-called modularity of the partition. For unweighted networks, the modularity of a partition is given as a scalar value between -1 and 1 where this value measures the density of links inside communities as opposed to links between communities. However, in the case of weighed networks, the modularity of a partition can be defined as

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j),$$

where A_{ij} represents the weight of the edge between nodes i and j, $k_i = \sum_j A_{ij}$ is the sum of the weights of the edges attached to node i, c_i is the community to which node i is assigned, the δ-function $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise, and $m = \frac{1}{2}\sum_{ij} A_{ij}$.

The community detection algorithm by Blondel et al. 2008 is divided into two phases. The first assigns a different community to each node of the network, and therefore in the first partition the number of communities is equal to the total number of nodes in the network. The

neighbours of each node i are then considered, j, and the gain of modularity that would occur if i was removed from its community and instead placed in the community of j is evaluated. Once evaluated, the node i is placed in the community for which this gain is maximum, but this is only done in cases for which the gain is positive. If there is no positive gain, i remains in its original community. This process is repeated sequentially for all nodes and this first phase is completed when a local maxima of the modularity is attained and can no longer be improved. This gain in modularity ΔQ can be achieved by moving an isolated node i into a different community C can easily be computed by

$$\Delta Q = \left[\frac{\Sigma_{\text{in}}+2k_{i,\text{in}}}{2m} - \left(\frac{\Sigma_{\text{tot}}+k_i}{2m}\right)^2\right] - \left[\frac{\Sigma_{\text{in}}}{2m} - \left(\frac{\Sigma_{\text{tot}}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right],$$

where Σ_{in} is the sum of the weights of the links inside C, Σ_{tot} is the sum of the weights of the links incident to nodes in C, k_i is the sum of the weights of the links incident to node i, $k_{i,\text{in}}$ is the sum of the weights of the links from i to nodes in C and m is the sum of the weights of all the links in the network (Blondel et al., 2008).

The second phase of the algorithm builds a new network in which the nodes are now the communities found through applying the first phase. This is done by calculating the weights of the links between the new nodes as the sum of the weight of the links between nodes in the corresponding two communities. Links between nodes of the same community lead to self-loops for this community in the new network (see Blondel et al. 2008 for more details). Once this second phase is completed, the first phase of the algorithm can then be reapplied to the resulting weighted network and iterated.

PageRank

The PageRank measure of a network node indicates the importance of the node, not only by considering its degree, but also the influence that neighbouring nodes may have on the node's importance. The basic idea being that a node having lots of neighbours isn't enough for a node to be considered important, but that it also depends on how important those neighbours are. The PageRank algorithm was first introduced by Brin and Page (Page et al., 1999), as a concept employed to rank webpages on the World Wide Web (WWW). Google

uses this algorithm, modelling the WWW as a network in which the webpages are nodes and hyperlinks existing between these webpages are edges, with each edge weighted according to the out-degree of each node. PageRank scores for each node are obtained by using a Markov chain on a network, with the calculated score being proportional to the amount of time spent by a surfer at that node. Other than for use in Web search algorithms, PageRank is well defined for any given network. As a simple example in Figure 3.5, we show the results of applying PageRank to a toy network with six nodes. We see that despite the red node containing more incoming links than the orange node, the orange node has a higher PageRank score as those incoming links that the orange node does have, have a larger influence than those that the red node has.

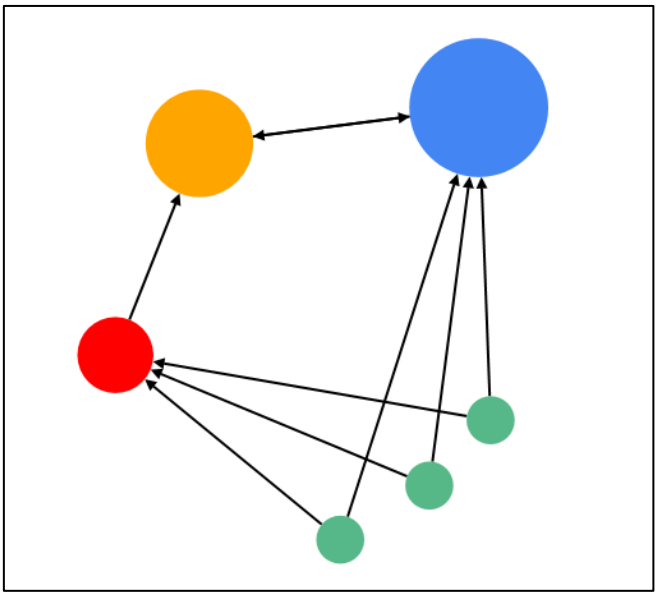


Figure 3.5 A network in which the size of each node (webpage) is roughly proportional to the probability that a surfer is at that webpage. The orange node has only two incoming links, but both are from nodes of a larger size than any of the three incoming links to the red node; therefore, the orange node has a higher PageRank than the red node.

Weighted edges of the network $N = (V, E)$ in PageRank lead to a weighted adjacency matrix of the form

$$H = D_{out}^{-1}A, \quad (\text{here } D_{out} = \text{diag}\,(k_1^{out}, k_2^{out}, \dots, k_n^{out})) \tag{3.1}$$

which is a row stochastic matrix referred to as the hyperlink matrix. The structure of the WWW causes issues to arise, including the existence of web pages with no out-going links. These are referred to as 'dangling nodes' and act as dead-ends from which the random surfer cannot escape. To avoid this issue, the above equation (3.1) was modified and the actual Google matrix is given by

$$G = \alpha S + \frac{1 - \alpha}{n}\mathbf{1}\mathbf{1}^T$$

where $S = H + \mathbf{a}\,(1/n\mathbf{1}^T)$, $a_i = 1$ if page i is 'dangling', and $\mathbf{1}_i = 1, \forall i$. Resulting in a matrix which is both stochastic and irreducible.

This modification, however, had not been addressed from a computational point, where this matrix representation for a network the size of the WWW would result in a completely dense matrix. To overcome this problem, G can be rewritten as a rank-one update to the sparse hyperlink matrix H as

$$G = \alpha(H + 1/n\mathbf{a}\mathbf{1}^T) + (1 - \alpha)1/n\mathbf{1}\mathbf{1}^T$$

$$= \alpha H + (\alpha\mathbf{a} + (1 - \alpha)\mathbf{1})1/n\mathbf{1}^T.$$

3.2 NETWORK APPROACH TO IDENTIFY NOVEL CANDIDATE REGIONS ASSOCIATED WITH LONGEVITY

3.2.1 Creation of Original Networks

To identify novel candidate regions associated with longevity, we hypothesised that the 3D architecture of the genome governs the co-location of longevity-associated genes/genome regions with novel unknown regions that share biological functions of importance to the process of longevity. Networks were created using Hi-C data at a resolution of 80 Kb and SNP information from the Synthetic and DGRP GWAS dataset. The Hi-C data provided interaction information between regions 80,000 bp in length (see section 2.3), and it was these regions which were represented as nodes on the Synthetic and DGRP GWAS-based networks. For

both the Synthetic and DGRP GWAS dataset, each SNP recorded has a calculated D or P-value, indicating its potential significance of association with longevity (see section 3.3.1). Regions were selected as nodes on each of the original networks if they harboured SNPs that satisfied selected predefined thresholds of D and P-values (see section 4.2). Next, intra- and inter-chromosomal interaction frequency distributions were analysed and interactions with frequencies exceeding a threshold corresponding to 1% of the strongest interactions were selected as strong interactions. Edges, between the nodes in the original networks, were then added if there were strong interactions between these nodes. The original network was therefore produced only using regions known to harbour SNPs with high association to longevity.

3.2.2 Extension of Original Networks (Extended Networks)

The Hi-C data at a resolution of 80 Kb was then further utilised to add nearest neighbouring nodes, i.e. residing in close proximity within the cell nucleus, to the original networks. The same thresholds for intra- and inter- chromosomal interactions, as above, were used. However, this time original networks were extended by adding new nodes/regions, which were those connected to original nodes with frequencies exceeding identified thresholds. These new regions do not necessarily contain any SNPs identified by GWAS studies, and it is these additional nodes that are further explored as our novel nodes. These extended networks are therefore now produced using original regions known to contain SNPs with longevity association, as well as novel regions selected due to their strong physical interactions with these original regions.

3.3 ADDITIONAL BIOINFORMATICS TECHNIQUES USED IN THIS STUDY

3.3.1 Identifying Significant SNPs in Synthetic GWAS Dataset

To identify SNPs with divergent haplotype frequencies in the control and old groups in the Synthetic GWAS dataset, Euclidean distances between the control and old groups were calculated for haplotype data for populations A1/A2 and B1/B2 combined. All duplicates were removed.

The Euclidean distance for a given SNP was calculated as suggested in (Burke et al. 2013):

$$D = 100.\sqrt{\frac{\sum_{j=1}^{n}\left(h_{O,j} - h_{Y,j}\right)^2}{n}},$$

where $h_{O,j}$ is the haplotype frequency of the j^{th} founder in the old samples, $h_{Y,j}$ is the haplotype frequency of the j^{th} founder in the young control sample, n is the number of haplotypes found at that position. SNP positions with the largest calculated D values were those showing the largest differences between haplotype frequencies in the control and old groups, and it was therefore these SNPs that were indicated as most likely to have association with longevity. Using a stringent selection process, as suggested by Burke et al. (2013), SNPs with a genome-wide significance corresponding to $D \geq 7.9$ (genome-wide alpha of 5% significance), were used in the following study.

3.3.2 Lift-over of Gene Positions from BDGP Release 6/dm6 to BDGP Release 5/dm3

To identify *Drosophila* genes, residing within regions of interest, a list of genes and their genomic coordinates according to the BDGP Release 6/dm6 assembly (dos Santos et al., 2014) was downloaded from the FlyBase database (http://flybase.org/). To align the Hi-C data and GWAS SNP positions, all gene positions were lifted over to BDGP Release 5/dm3. This was done using a LiftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver) that converts genomic coordinates between assemblies.

3.3.3 Gene Ontology Enrichment Analysis

To further analyse specific groups of genes, the FlyBase database was also used to find phenotypes for genes of interest. The FlyMine software (http://www.flymine.org/) was used to analyse the enrichment of the set of identified genes in Gene Ontology (GO) terms for cellular component, biological process and molecular function. Genes used in GO analysis were those genes residing in influential nodes found when network measures were calculated for each network. Genes found in both original and novel regions were taken into consideration during analysis, with those genes in novel regions of most interest, as these

were new regions with no previously published association with longevity. Each gene was also compared with a list of longevity genes recorded in the GenAge database (http://genomics.senescence.info/genes/models.html), to determine if any novel genes, not present in the original datasets but found by network analysis, were previously known as associated with longevity.

3.4 STATISTICAL APPROACHES USED

3.4.1 Test for Difference in Proportions

The Chi-squared test is used for testing for difference between proportions of an event considered as a "success" are the same across different populations/groups. Hypotheses for this Chi-squared test are stated as H_0: $p_{real} = p_{control} = p$ (there is no difference in proportions) and H_1: $p_{real} \neq p_{control}$, H_1: $p_{real} > p_{control}$ or H_1: $p_{real} < p_{control}$ (there is a difference in proportions).

Observations of the proportions for analysis are expressed in a 2x2 contingency table (Table 3.1), in which the observed frequencies recorded are then tested against expected frequencies calculated from generated controls.

Table 3.1 Contingency table showing observations for case and control samples.

	Success	Failure	Total
Real Case 1	a	b	$a + b$
Control Case 2	c	d	$c + d$
Total	$a + c$	$b + d$	n

Fisher's Exact Test, which is a modification of the chi-squared test, can be used for 2x2 contingency tables and is most commonly used in cases of small observation counts. In this test, a P-value is given which corresponds to the probability of obtaining the observed proportion of frequencies under the assumption that the proportions between these two samples are equal.

The formula used to calculate this P-value is given as:

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}.$$

A calculated P-value greater than significance level of 0.05 means that we are unable to reject the null hypothesis, concluding that there is no significant difference between two proportions. If a P-value less than or equal to 0.05 was calculated, this would allow the null hypothesis to be rejected in favour of the alternative hypothesis, concluding that there is a significant difference between the two proportions observed.

3.5 IMPLEMENTATION AND SOFTWARE USED

MATLAB Programming Language

MATLAB was used for implementation of all software tools for analyses. For calculating PageRank and other network measures of the networks produced, the Brain Connectivity Toolbox was downloaded from brain-connectivity-toolbox.net and used.

Gephi

To visualize and explore our data, an open-source software, Gephi, was downloaded from https://gephi.org/. Using this software, data was able to be visualised as networks, and tools in Gephi also allowed calculation of network measures, for example modularity and clustering coefficient.

IBM SPSS Statistics

IBM SPSS Statistics v24 software package was used for statistical analyses of data, whether simple or complex. SPSS offers several programs for exploring and analysing data, including a Statistics Program providing basic statistical functions, including frequencies, cross tabulation and Chi-square statistics.

Chapter 4

NOVEL LONGEVITY-ASSOCIATED CANDIDATE REGIONS IDENTIFIED VIA NETWORK APPROACH

In this chapter we describe novel genomic regions identified by network approaches, described in Chapter 3. Various network measures were calculated, identifying important previously unknown regions, with some regions observed to be common between both GWAS-based networks. Using literature search and other bioinformatics resources, these newly discovered genes/regions were investigated with the aim to find association with longevity. Subnetworks of these networks were also explored, and Gene Ontology enrichment analysis was performed with the aim of identifying genes/regions, enriched in longevity-related terms, with no previously known association with longevity.

4.1 EXPLORATION OF HI-C DATA

4.1.1 Heat Maps

To visually explore the strength of inter- and intra-chromosomal interactions, Hi-C data was visualised using heat maps, as shown in Figure 4.1. Here, both the X and Y axis represent bin numbers from the Hi-C data. Each pixel represents interaction frequency between an 80 Kb region with another 80 Kb region. More dense areas of redness are regions in which there are high interaction frequencies between bins. From observation, these denser areas occur intra-chromosomally, meaning that more interactions are observed between bins closer together on a linear model of the genome within the same chromosome, which is what would be expected. The areas of the heat maps in which the points plotted are lighter shades indicate low interaction frequencies between these bins, and as expected these occur more commonly inter-chromosomally.

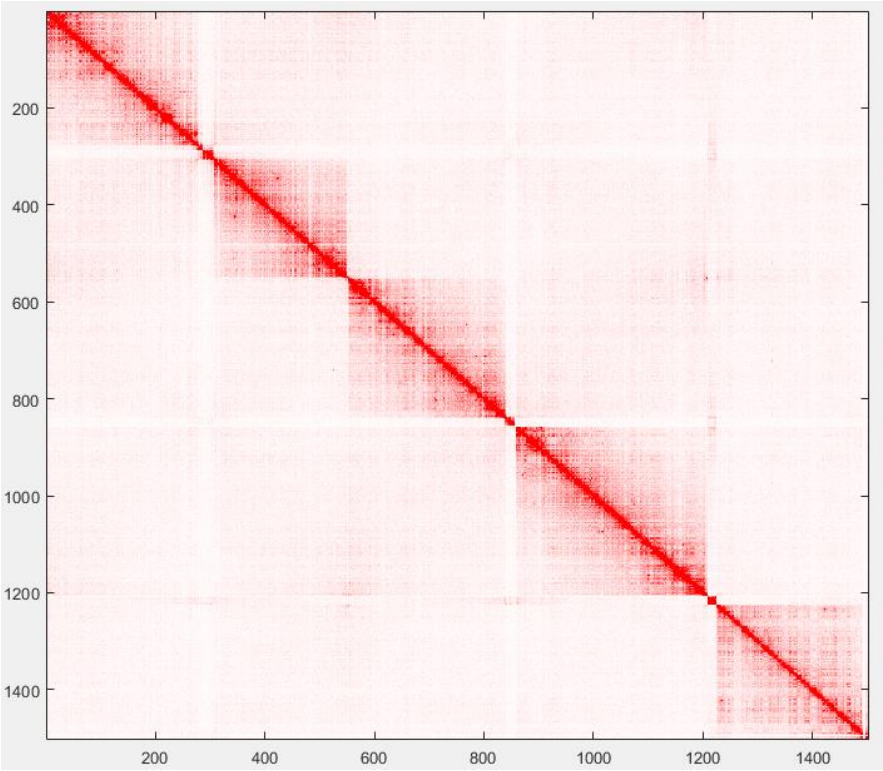


Figure 4.1 A heat map produced using normalised Hi-C data at a resolution of 80 Kb for the *Drosophila* genome.

4.1.2 Hi-C Interaction Frequency Distributions

Interaction frequency distributions, plotted as histograms in Figure 4.2, show right-skewed distributions for all intra-chromosomal and inter-chromosomal graphs. This means that for all chromosomes in the *Drosophila* genome, intra- and inter-chromosomal interactions of smaller interaction frequencies are much more frequent than higher interaction frequencies. Distributions for chromosome 2, 3 and X all presented similar patterns in the plotted data and therefore only one histogram, which shows the data for chromosome 2, was shown in Figure 4.2(a). Note that interaction frequencies are inversely proportional to the distance between regions within the cell nucleus.

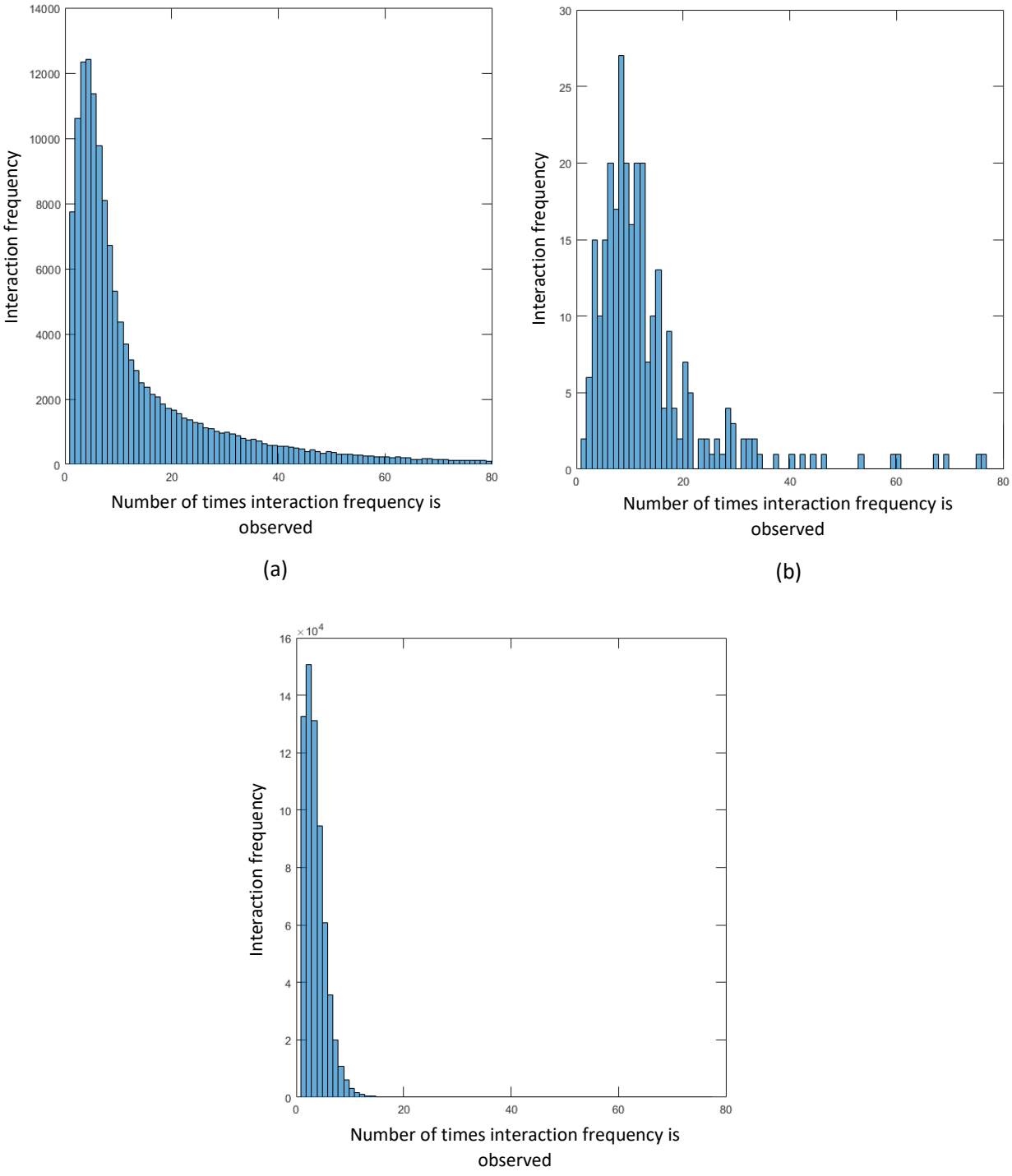


Figure 4.2 Intra-chromosomal interaction frequency distribution histograms for (a) chromosome 2, (b) chromosome 4 and (c) inter-chromosomal interaction frequency distribution histogram for all chromosomes, in which the x-axes represents interaction frequencies and the y-axes represent the number of times this interaction frequency is observed between 80 Kb regions in this chromosome.

4.2 CHOICE OF INTERACTION FREQUENCY THRESHOLDS AND GENOME-WIDE SIGNIFICANCE LEVELS

To assess the strength of the interactions between intra- and inter-chromosomal genomic regions, distributions of interacting frequencies described in section 4.1 were analysed (Figure 4.2). Values corresponding to the 1% of the strongest intra-chromosomal interactions were calculated individually for each chromosome. This 1% of strongest interactions corresponds to interactions frequencies ≥ 247 for chromosome 2, ≥ 215 for chromosome 3, ≥ 1308 for chromosome 4 and ≥ 342 for chromosome X. The threshold for interaction frequency for inter-chromosomal interactions, corresponding to 1% strongest interactions, was found to be ≥ 10; the highest frequency was 111. Interactions with frequencies exceeding threshold are referred to as "strong" interactions. The maximum number of intra-chromosomal interactions for each chromosome varied being 3098 for chromosome 2, 3708, for chromosome 3, 1732 for chromosome 4 and 1976 for chromosome X.

Genome-wide significance level, required for finding association between ~10^6 SNPs, is usually set to $P < 5\times10^{-8}$. This value corresponds to 0.05 level of significance after Bonferroni correction for multiple testing. In our case, each SNP is binned into an 80 Kb region and there are 1503 distinct 80 Kb regions recorded in the *Drosophila* Hi-C data. Taking this into account, the required significance level was corrected as $0.05/1503=3.33\times10^{-5}$ for the selection of SNPs in the DGRP GWAS dataset. The threshold used for SNPs in the Synthetic GWAS dataset was $D \geq 7.9$, which corresponds to genome-wide alpha of < 0.05 [see Burke et al. (2013) for details].

4.3 PROPERTIES OF GWAS-BASED NETWORKS

4.3.1 Network of Interactions Originated from the Synthetic GWAS Dataset

Each 80 Kb region harbouring at least one SNP with $D \geq 7.9$ was represented by a node; these nodes will be referred to from now on as original nodes. Edges between nodes were added if the frequency of interaction lay in the highest 1% of frequency interaction data found in the Hi-C dataset, either intra- or inter- chromosomally, calculated individually for each chromosome. Hi-C data for all genomic interactions was then further utilised to add nearest neighbouring nodes to the current network, using the same thresholds for intra- and inter-

chromosomal interactions as used previously (see Figure 4.3). These nodes will be referred to from now on as novel nodes not necessarily genotyped by GWAS. The network for the Synthetic GWAS dataset, including both original and novel nodes, from now on will be referred to as the extended Synthetic GWAS-based network.

Clustering coefficient and PageRank network measures were calculated for the Synthetic GWAS-based network shown in Figure 4.3. This network consisted of 1099 nodes and 4489 edges and comprised a single component. These 1099 nodes harboured ~75% (69,951) of SNPs recorded in the Synthetic GWAS dataset, of which 2409 resided in a coding region. The node with the highest degree was bin 547 corresponding to region Chr2R: 20800000-20880000, with a degree of 150. The clustering coefficient scores for nodes in this network ranged from 0 to 1, with an average score of 0.256. In total, 89 nodes had a score of 1. PageRank scores for nodes in the Synthetic GWAS network ranged from 0.000139 to 0.009349.

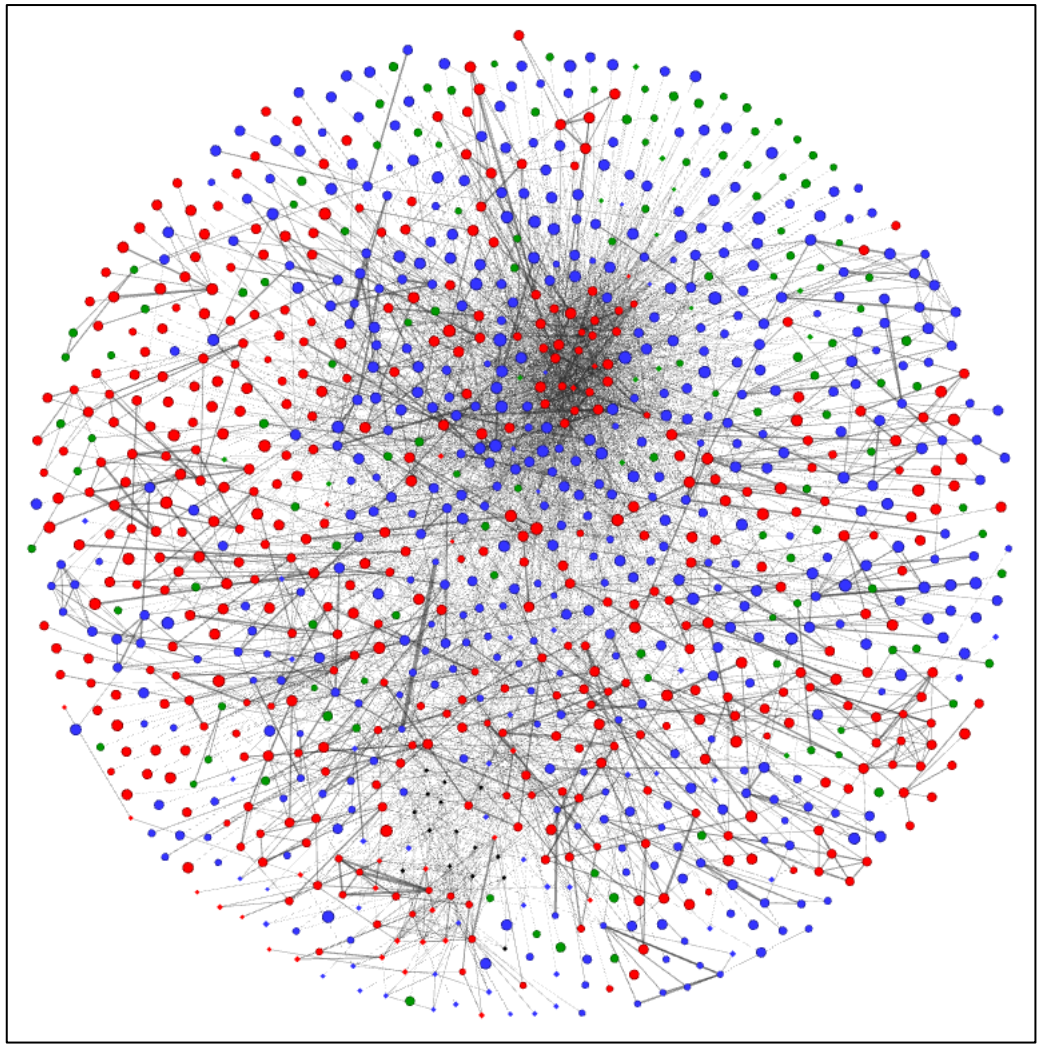


Figure 4.3 Extended network of interactions between genomic regions harbouring significant SNPs identified in the Synthetic GWAS dataset. Nodes colour coded, with red nodes corresponding to chromosome 2, blue nodes corresponding to chromosome 3, black nodes corresponding to chromosome 4 and green nodes corresponding to chromosome X.

4.3.2 Network of Interactions Originated from the DGRP GWAS Dataset

Each 80 Kb region harbouring at least one SNP, satisfying $P \leq 3.33 \times 10^{-5}$ was represented by a node. As in the Synthetic GWAS-based network (see section 4.3.1), edges were added to this network, if they met interaction frequency cut offs, and nearest neighbouring nodes were also included (see Figure 4.4). The network for the DGRP GWAS dataset, including both original and novel nodes, from now on will be referred to as the extended DGRP GWAS-based network.

Clustering coefficient and PageRank network measures were calculated for the extended DGRP GWAS-based network shown in Figure 4.4. This network consisted of 671 nodes and 1137 edges, where the number of connected components in this network was six. The 671 nodes harboured ~50% (1,093,533) of SNPs recorded in the DGRP GWAS dataset where 114 of these SNPs resided in coding regions. The node with the highest degree in this network was bin 1183 corresponding to region Chr3R: 25920000-26000000, with a degree of 68. The clustering coefficient scores for nodes in this network ranged from 0 to 1, with an average score of 0.157. In total, 38 nodes had a score of 1. PageRank scores for nodes in the extended DGRP GWAS-based network ranged from 0.000232 to 0.020202.

Nodes in both extended GWAS-based networks with clustering coefficients equal to 1, with 1 being the highest possible score, were selected as clustering coefficient influential nodes and therefore all genes residing in these selected regions were grouped together for further analysis. As mentioned previously, the PageRank score of a node indicates its importance in a network relative to the other nodes scores, and therefore those nodes in both extended GWAS-based networks with the highest PageRank scores were considered most important. Once PageRank scores had been calculated for all nodes in each of the networks, they were then ranked in descending order and the top 10% of nodes with the highest scores were selected as PageRank influential nodes and all genes residing in these regions were grouped together for further analysis.

4.3.3 Novel Nodes with the Highest Degrees in Extended GWAS-Based Networks

In both extended GWAS networks, novel nodes that had the highest degrees were selected for further analysis along with the nodes in the clusters centred around these novel regions.

All genes residing within these interacting nodes in each cluster were grouped together for GO enrichment analysis, with the aim to find any common biological function between these genes, to infer that this novel region may play a role in the way in which these genes with common biological functions work together or that the function of genes found in these novel regions may be influenced by the common biological processes of genes with which it interacts.

Novel region corresponding to bin 928 in the Synthetic GWAS-based network (Table 4.1), was found to interact with regions harbouring genes that enriched in the GO terms 'apoptotic process' and 'nervous system development', shown in Table 4.2. Fourteen genes, residing in seven regions interacting with bin 928, were enriched in the GO term 'apoptotic process'. Sixteen genes were enriched in the GO term 'nervous system development', most of which were different genes than in the previous enrichment groups; these genes resided in nine regions interacting with bin 928. All 23 genes found to reside in bin 928 were explored, where *trbd* and *CG8412* had the phenotype 'short-lived'. The loss of the *trbd* gene, a negative regulator of the *Drosophila* immune-deficiency pathway, has previously been observed to reduce lifespan (Fernando et al., 2014). A number of genes in this novel region, including *dmt*, *hyd*, *CG16908* and *CG9471*, were found to have phenotypes 'increased mortality' and 'lethal'. The *MED6* gene was found to have a phenotype of 'cell lethal' and is known to be required for elevated expression of a distinct set of developmentally regulated genes. This gene is essential for viability and/or proliferation of most cells and mutants of this gene have previously been observed to fail to pupate, dying in the third larval instar with severe proliferation defects in imaginal discs and other larval mitotic cells (Gim et al., 2001). Finally, bin 928 also contained the *FoxP* gene, a protein that encodes a transcription factor expressed in the nervous system. This gene has recently been shown to be important for regulating several neurodevelopmental processes and behaviours that are related to human disease or vertebrate disease model phenotypes (Castells-Nobau et al., 2019). Bin 928 was further explored for any enhancers that reside in this genomic region, where there were several found. One enhancer had a known target gene *alphaTub85E* residing in bin 928, this gene is known to affect the pattern of proprioceptive chordotonal organs (ChO) cell elongation (Hassan et al., 2018). ChOs are a group of specialised sensory organs that innervate the joints of an insect body, and therefore involved with the nervous system which has previously been

discussed in this thesis in relation to association with longevity. All other target genes for this enhancer were unspecified. It can be speculated that these enhancers target one or more of these genes discussed that share a common longevity-associated phenotype or biological process, influencing their expression.

Exploration of the genes listed in Table 4.2 showed that some of these genes have previously been found to associate with ageing or have phenotypes which could link to association with longevity. All genes with this association were found to have a negative effect on longevity, with genes *sidpn*, *hook* and *CG12935* having a 'short-lived' phenotype. Loss-of-function mutation in the *hook* gene has been found to reduce maximum lifespan by up to 30% (Simonsen et al., 2007). Mutant flies lacking mitochondrial *Top3alpha* have also been found to show decreased maximum lifespan by up to 25%, in which a premature ageing phenotype was demonstrated and mobility defects were observed (Tsai et al., 2016). Several genes in Table 4.13 were also found to have an 'increased mortality' phenotype, e.g. *RpL30*, *Eps-15*, *Nipped-B* and *RPA2*.

The novel region, bin 1220, also in the Synthetic GWAS-based network, was found to interact with regions harbouring genes enriched in the GO term 'DNA repair'. Interestingly, this novel region is positioned on chromosome 4 of the *Drosophila* genome, a chromosome seen as an anomaly because of its small size compared to other chromosomes in the genome, and its chromatin structure. Due to its size, this chromosome is often ignored, however it is known to harbour at least 16 genes where many of them are thought to have male-related functions, and these genes also include the well-known *eyeless* gene (Carvalho, 2002). A search for enhancers in this novel region on chromosome 4 found enhancers targeting lncRNA *sphinx* and the transcription factor *toy*, as well as enhancers with unknown target genes. Therefore it can be speculated that these enhancers, with unknown target genes, could target genes co-located in the 3D organization, meaning those regions residing within the same cluster centred around bin 1220.

Table 4.1 A summary table of novel nodes with the highest degree in the extended Synthetic GWAS-based network with interacting regions enriched in longevity-related GO terms.

Novel node	Degree	Interacting nodes within the extended Synthetic GWAS-based network	Total number of genes in interacting regions
1220	20	244, 295, 262, 245, 923, 255, 270, 271, 272, 273, 275, 276, 277, 302, 305, 334, 359, 799, 848, 920	188
928	15	11, 545, 233, 536, 531, 456, 409, 370, 234, 238, 265, 343, 360, 361, 366	290

Table 4.2 Genes enriched in longevity-related GO terms interacting with novel regions with the highest degree in the extended Synthetic GWAS-based network.

Novel node	GO term enrichment for genes in interacting regions	P-value	Genes enriched in GO term	Number of gene harbouring regions/total number of interacting regions
928	Apoptotic process	2.27E-04	*E2f2, lola, egr, Ret*, Vps25, TER94, ptc, eEF5*(*CG3186*), *snama, ninaA, yki, sigmar, l(2)tid, Mcm10*	7 bins/16 bins
928	Nervous system development	4.37E-04	*CG10339, amos, CG10431,* ***Sidpn*****, **RpL30**, **hook**, Dap160, enok, lola, dgo, egr,* ***CG12935***, *Ret, Pka-R2,* ***Eps-15***, *Galphao*	9 bins/16 bins
1220	DNA repair	0.0294	***Top3alpha***, *PCNA2*(*CG10262*), *Nipped-B, CG9272,* ***RPA2***	4 bins/21 bins

* Genes residing within original nodes, i.e. harbouring SNPs with D>7.9 are underlined.

** Genes previously found to have association with longevity as recorded in FlyBase or GenAge resources are shown in bold.

Novel region, bin 28, in the DGRP GWAS-based network (Table 4.3), was found to interact with regions harbouring genes that enriched in the GO term 'Immune System Process' shown in Table 4.4. Exploration of the enriched genes showed some of these genes have previously been found to associate with ageing or have phenotypes which could link to association with

longevity. Flies heterozygous for the *Stat92E* mutation have been found to have a maximum lifespan up to 30% shorter than those of wild-type control flies (Larson et al., 2012). Lifespan of *Drosophila* was found to be increased through post developmental RNA interference of *GlyP*, causing an increase in mean lifespan by up to 17.1% (Bai et al., 2013). Enhancers found residing in novel regions included those with target genes *CG34172*, *ush* and the transcriptional-repressor protein *aop*; the latter has been strongly associated with longevity previously and is found to play a crucial role in lifespan extension caused by reduced IIS or Ras attenuation (Slack et al., 2015). The *aop* gene was a gene enriched in the 'immune system process' GO term, however *CG3417* and *ush* were not. Both *aop* and *CG34172* also reside in the novel bin 28, whereas the *ush* gene resides in bin 6 which is a region not interacting with bin 28 in the extended DGRP GWAS-based network. Enhancers in this region were also found with unspecified target genes, which can be speculated to target other co-located genes residing within the same cluster centred around bin 28.

Novel region, bin 2, in the DGRP GWAS-based network, was found to interact with regions harbouring genes that enriched in the GO term 'cellular response to stress' shown in Table 4.4. The genes residing in this novel region, bin 2, were explored using a phenotype search. Despite no genes in this novel region being found to be enriched in the GO term 'cellular response to stress', bin 2 harboured 21 genes in total, where several genes had the 'lethal' and 'increased mortality' phenotypes, including genes *net*, *Sam-S*, *ND-15*, *CG4822* and *Gs1*. This region also harbours the *Zir* gene, with the phenotype 'immune response defective', and also previously found to play an important role in cellular immune response through the activation of the Rho-family GTPases *Rac2* and *Cdc42* (Sampson et al., 2012). The *Nhe1* gene with phenotypes 'lethal' and 'short-lived' was also found to reside in this region. Bin 2 was then further explored for any enhancers that reside in this genomic region, in which three enhancers were found for which all target genes were unspecified. One can speculate that these enhancers could target other co-located genes residing within the cluster, i.e. in close proximity within the cell nucleus. Among these genes which have been found to have a positive effect on lifespan is *Cat*, where an overexpression of this gene has been found to result in an increase in lifespan by up to a third (Orr et al., 1994). Several genes enriched in the GO term 'cellular response to stress' were also found to have phenotypes associated with ageing, including *Clbn* and *Atg16* with a 'short-lived' phenotype, and the *BI-1* gene which as

well as having a 'short-lived' phenotype, also had 'long-lived' phenotype. Genes *kay* and *HipHop* also had phenotypes for increased mortality, and the *aop* gene, previously discussed, was also found in this analysis. These genes with ageing phenotypes residing in regions in the same cluster as novel region bin 2 can be speculated to influence genes residing in novel region bin 2 that have similar biological functions, to have an effect on ageing.

Table 4.3 A summary table of novel nodes with the highest degree in the extended DGRP GWAS-based network with interacting regions enriched in longevity-related GO terms.

Novel node	Degree	Interacting nodes within the extended Synthetic GWAS-based network	Total number of genes in interacting regions
28	7	29, 30, 1063, 1124, 1152, 1179, 27	114
2	14	1178, 699, 1120, 736, 1131, 1132, 787, 1152, 670, 655, 660, 576, 1183, 1179	255

Table 4.4 Genes enriched in longevity-related GO terms interacting with novel regions with the highest degree in the extended DGRP GWAS-based network.

Novel node	GO term enrichment for genes in interacting regions	P-value	Genes enriched in GO term	Number of gene harbouring regions/total number of interacting regions
28	Immune system process	0.021515	*Vps16B, Cad99C,* ***aop****, DPCoAC*(*CG4241*)*,* ***Stat92E****, Mtl,* ***GlyP***	4 bins/8 bins
2	Cellular response to stress	0.006104	*CG11498,* ***Clbn****, CG13473, CG14130, Sld5, mu2,* ***Atg16****,* ***kay****, CG3448, Rad9, Mtl, Grx1*(*CG6852*)*,* ***Cat****,* ***HipHop****,* ***Bl-1****, Wdr24*(*CG7609*)*, Drice*	13 bins/15 bins

* Genes residing within original nodes, i.e. harbouring SNPs with D>7.9 are underlined.

** Genes previously found to have association with longevity as recorded in FlyBase or GenAge resources are shown in bold.

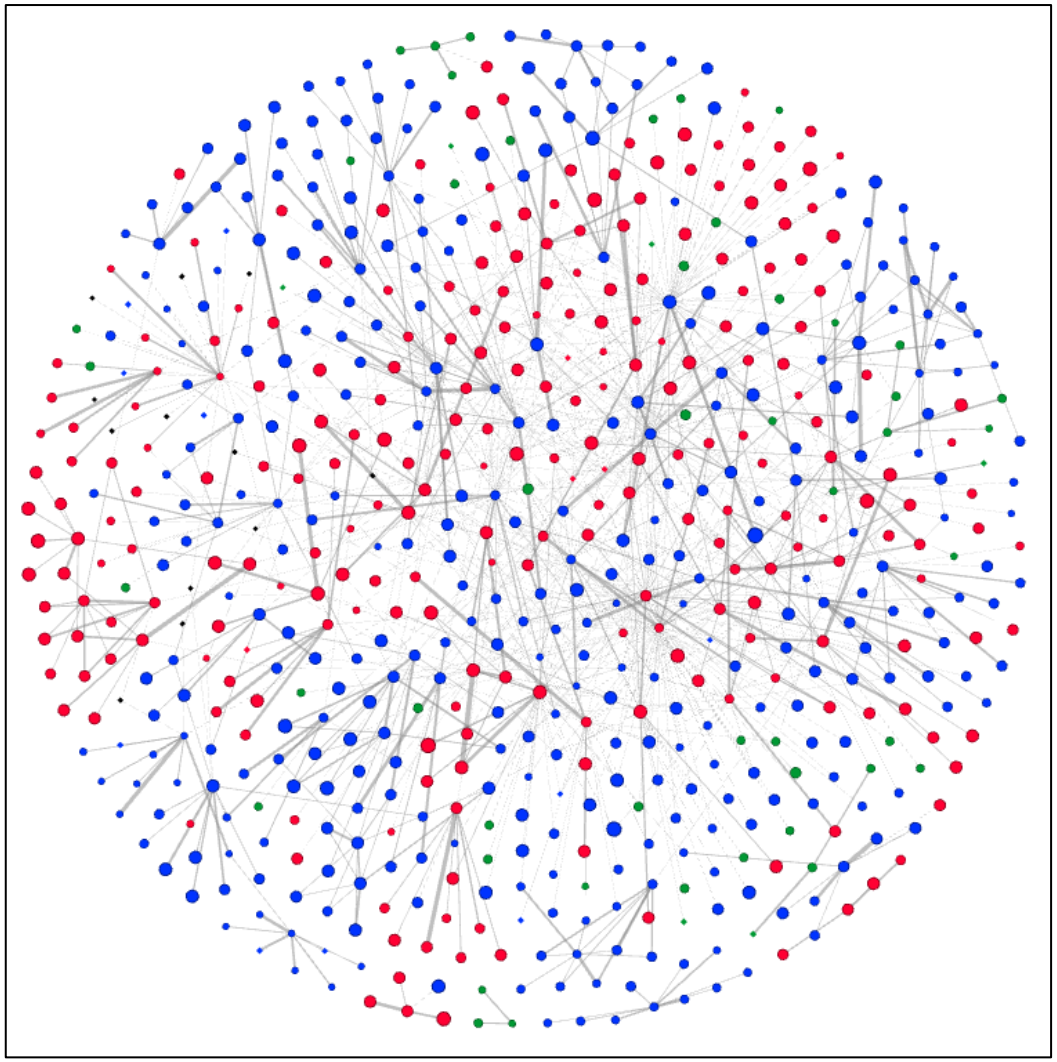


Figure 4.4 Extended network of interactions between genomic regions harbouring significant SNPs identified in the DGRP GWAS dataset. Nodes colour coded, with red nodes corresponding to chromosome 2, blue nodes corresponding to chromosome 3, black nodes corresponding to chromosome 4 and green nodes corresponding to chromosome X.

4.4 COMMON REGIONS/GENES IDENTIFIED BY EXTENDED SYNTHETIC AND DGRP GWAS-BASED NETWORKS

Several regions selected as original regions, and therefore containing significant SNPs, for each extended GWAS-based network were found to be in common. A total of 14 common nodes were shared between the GWAS-based networks, covering 1.12 Mb of the *Drosophila* genome. All 14 regions were observed to harbour genes, with a total of 168 genes residing between these regions. However, only five of these genes were found in the FlyBase database to have a phenotype of 'long-lived', including genes *Rim2*, *GlyP*, *aop*, *HDAC1* and *Tpi*. The majority of these 14 regions were found to harbour SNPs, where the highest mutated genes were the same in both GWAS datasets, including genes *nmo*, *sima*, *axo*, *CG9967*, *eys*, *chinmo* and *dpr3* (full list in Appendix Table S4.1). Of all 168 genes, 43 genes were found to harbour 91 significant SNPs (D≥7.9) in the Synthetic GWAS dataset, where genes *CG4168*, *axo* and *aop* harboured the highest number of significant SNPS (≥ 8). Only 10 of these 168 genes harboured significant SNPs in the DGRP GWAS dataset, in which a total of 19 SNPs were found.

Additional regions extending these GWAS-based networks increased the number of regions in common from 14 to 527, covering 42.16 Mb of the *Drosophila* genome. In this case, 15 common regions were found to harbour no genes, and the regions that did harbour genes contained from 1 gene in many of the regions to 43 genes in bin 518. A total of 7413 genes resided in these 527 common regions, where almost 30% of the genes did not harbour any SNPs in the Synthetic GWAS dataset and approximately 3% of the genes did not harbour any SNPs in the DGRP GWAS dataset. Again, in both GWAS datasets, a number of genes found in common extended regions were highly mutated including *kirre*, *Ptp61F*, *CG45186*, *Sema-1a, Con* and *Ptp99A*. Only a small percentage of genes found in these common regions were found to harbour significant SNPs from each GWAS dataset. In the extended Synthetic GWAS-based network, a total of 433 genes harboured 717 significant SNPs in total, where *CG42732* contained the highest number of significant SNPs, 40. In the extended DGRP GWAS-based network, a total of 48 genes harboured 57 significant SNPs in total, where previously mentioned *sima* contained the highest number of significant SNPs. The only genes found common to both extended GWAS-based networks that also contained significant SNPs from both GWAS datasets were *CG42389*, *Fur1*, *sima* and *Lpt*.

4.5 MODULARITY MEASURES OF EXTENDED SYNTHETIC AND DGRP GWAS-BASED NETWORKS

The modularity measures of the extended networks in Figure 4.3 and Figure 4.4 were calculated, for which modular structures were produced and are shown in Figure 4.5 and Figure 4.6 respectively. These network modules were detected using Gephi software, which uses the Louvain modularity method (Blondel et al., 2008) as a community detection algorithm. This algorithm had a tunable 'resolution', that enabled the changing of focus towards the number of detected of communities, where a smaller resolution meant more communities and a greater resolution meant less communities [for more information on resolution factors, see Lambiotte et al. (2008)]. This resolution parameter was explored in our study, and a resolution of 0.1 was selected, at which point the quantity and sizes of modules in each network appeared to remain at similar values as the resolution was further decreased. These networks in Figures 4.5 and 4.6 present a better visualisation of the community structure in each network and, from visual inspection alone, patterns emerge in the clustering of nodes. For example, the majority of regions found to cluster as a group all appear to be of the same colour node, and therefore are regions residing on the same chromosome. The extended Synthetic GWAS-based network comprised of 81 communities, with the smallest consisting of 3 nodes and the largest of 72 nodes. The extended DGRP GWAS-based network comprised of 61 communities, where the smallest also consisted of 3 nodes and the largest of 42 nodes. For both networks, overlapping of communities was observed in most cases, especially between the larger communities harbouring many nodes. Communities were also observed to stand alone, however these tended to be those with sizes of less than 5 nodes.

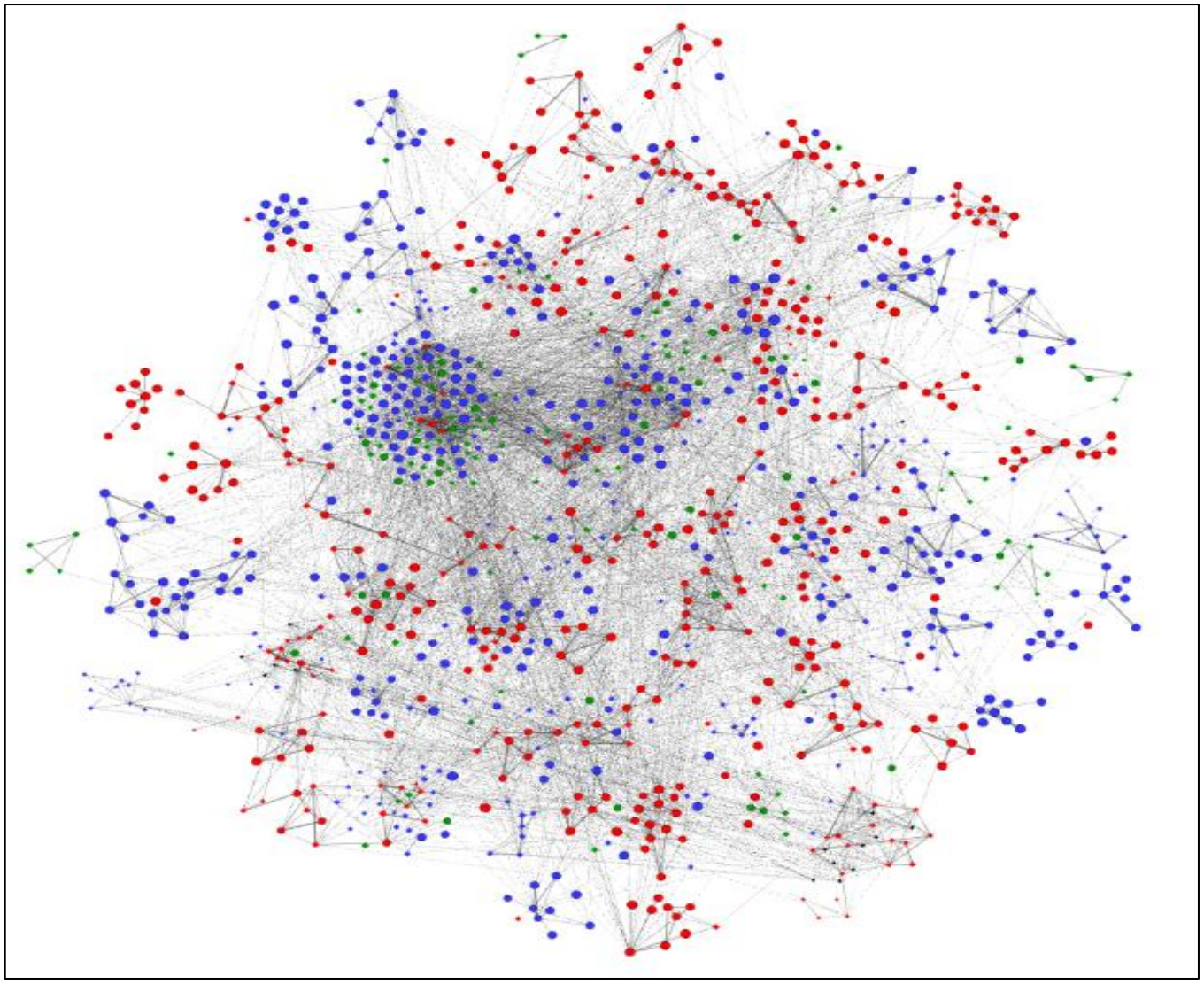


Figure 4.5 Modular structure of the extended Synthetic GWAS-based network.

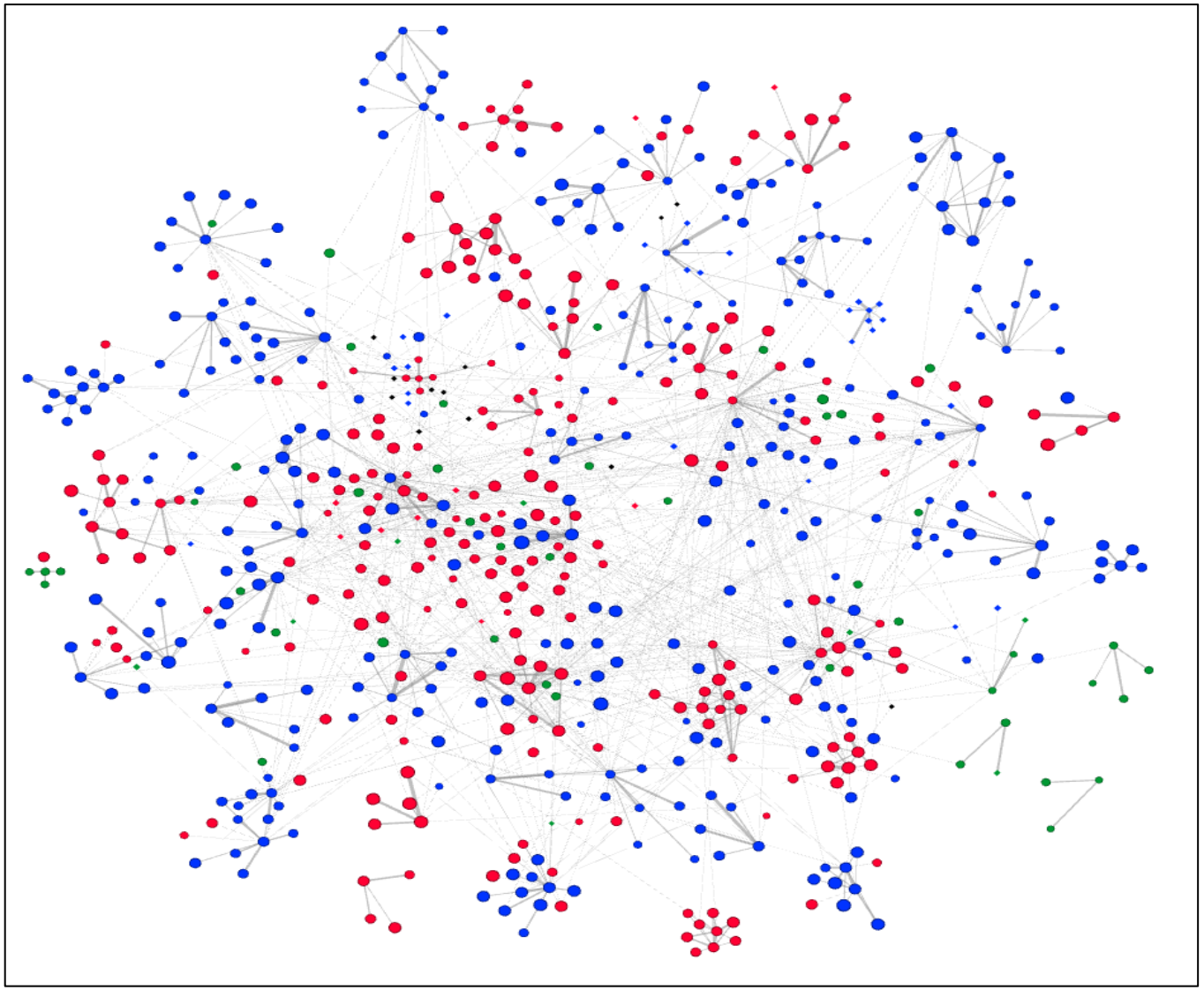


Figure 4.6 Modular structure of the extended DGRP GWAS-based network.

The largest of the communities in the extended Synthetic GWAS-based dataset (Figure 4.5), modularity class 21, consisted of 218 nodes in which two of the significant nodes common between our datasets were found using PageRank analysis [bins 534 (Chr2R: 19760000-19840000) and 535 (Chr2R: 19840000-19920000)]. Out of the 12 common bins found based on PageRank measure, nine of these bins were found to reside in the larger communities in the extended Synthetic GWAS-based network (modularity classes 0, 21, 23 and 41). Bin 32 found in common between both networks based on clustering coefficient measure also resided in modularity class 0. The largest of the communities in the extended DGRP GWAS-based network (Figure 4.6), modularity class 38, consisted of 72 nodes, in which four of the significant nodes common between our datasets identified by PageRank analysis were found: bins 1180 (Chr3R: 25680000-25760000), 1182 (Chr3R: 25840000-25920000), 1183 (Chr3R: 25920000-26000000) and 1184 (Chr3R: 26000000-26080000). The same is true for the extended DGRP GWAS-based network, of the 12 common bins found based on PageRank measure, nine of these bins were found to reside in the larger communities of this network (modularity classes 0, 14, 20 and 38). Bin 32 (Chr2L: 2480000-2560000) found in common between both datasets based on clustering coefficient measure also resided in modularity class 0.

4.6 GENE ONTOLOGY ENRICHMENT ANALYSIS

Tables 4.5-4.8 below show the results of Gene Ontology enrichment analyses, in which GO terms are stated along with their P-values generated and the number of genes from each group found to share this GO term. Note that P-values are not corrected for multiple testing.

4.6.1 GO Term Enrichment Analysis for Genes Residing in Nodes Defined Using Clustering Coefficient Measure

Genes, found in nodes characterised by high clustering coefficient and enriched in longevity-related GO terms, most often resided in novel nodes of the extended Synthetic GWAS-based network (Table 4.5). For all GO enrichment analysis in this study, P-values are not corrected for multiple testing. None of the genes found enriched in these GO terms in original regions

had been previously reported to be associated with longevity, only two genes found in novel regions, genes *esg* and *Sema-5c*, were found to have previous reports on associations with longevity. GO term analysis for genes residing in regions found using clustering coefficient measure included terms 'respiratory system development' and 'defence response', biological processes which could potentially have an effect on longevity in *Drosophila*.

Table 4.5 Gene Ontology (GO) enrichment analysis of genes found in regions with high clustering coefficient in the extended Synthetic GWAS-based network.

GO Term	Total number of genes enriched in GO term	P-value	Genes found within novel regions
Respiratory system development	13	0.049031	***esg****, Rac2, aPKC, ct, Tor, Vha68-2, cold, sano*
Defence response	20	0.045648	*Mtk, SPE, PPO1, sphinx1, sphinx2, akirin, Atg18a, Hat1, Rac2, DptB, Dronc, ECSIT, GILT2, Glt, Mst57Da, Rm62, aPKC, polyph*
System process	40	0.005027	*nan, Or22a, Or22b, Or49b, CG2121, Fas3, Obp47a, Obp83a, Obp83b, Or67a, Or67c, CG32698, HEATR2, Obp18a, PrBP, SKIP, brv1, ct, dpr10, dpr13, dpr6, f, sbb*
Multicellular organismal process	158	0.003284	*aPKC, Dronc, raw, Rac2, sna, vg, D, Tor,* ***esg****, cv-c, Fas3, sbb, C15, jeb, lilli, nan, nclb, MED24, dpn, elF4E3, f, pip, CSN7,* ***Sema-5c****, bic, jagn, pcs, stwl, tra2, CG17575, Glt, Mst57Da, Mst57Db, Mst57Dc, Or22a, Or22b, Or49b, Ote, PPO1, ProtA, Ptp4E, Ptth, Rbp9, Rdl, RecQ4, antr, Atg18a, CG2121, CG9932, HEATR2, Hr3, Lcp1, Lcp2, Lcp3, Mes2, Obp47a, Obp83a, Obp83b, Or67a, Or67c, PrBP, Vha68-2, ckn, dgrn, dor, kat-60L1, nuf, nwk, rols, sano, sdk, BG642167, Blos1, BoYb, CG10131, CG10257, CG11131, CG12404, CG15283, CG30486, CG31661, CG31926, CG32698, CG34129, CG34130, CG3740, CR45727, Cpr76Ba, Cpr76Bb, Cpr76Bc, Drep1, ERR, Fim, Gld2, Ipk2, Lcp4, LpR2, NPFR, Nmt, Obp18a, SKIP, Sec61beta, Sfp96F, Slh, Sp1, a6, akirin, brv1, cold, disp, dpr10, dpr13, dpr6, ect, haf, hfw, hll, lace, lectin-21Ca, mir-1, ms(2)35Ci, ms(3)76Ba, mthl7, mtt, nht, nrm, oaf, pnut, polyph, scramb1, soti, sphinx1, sphinx2, tut, vnc*

* Genes previously found to have association with longevity are shown in bold.

Genes, found in nodes defined using clustering coefficient measure, in the extended DGRP GWAS-based network are enriched in different GO terms (Table 4.6) to those of the extended

Synthetic GWAS-based network, with many of the genes found residing in novel bins. Here, the longevity associated gene *esg* was also enriched in GO terms in the clustering coefficient measure for the extended DGRP GWAS-based network. This analysis found enrichment in the 'detoxification' and 'cellular chemical homeostasis' GO terms, which again are biological processes that could potentially play a role in longevity.

Table 4.6 Gene Ontology (GO) enrichment analysis of genes found by using clustering coefficient measure in the extended DGRP GWAS-based network.

GO Term	Total number of genes enriched in GO term	P-value	Genes found within novel regions
RNA modification	9	$3.036714e^{-4}$	*Nop60B, l(2)35Bd, CG3808, CG6745, THG, snoRNA:Me28S-G1083a, snoRNA:Me28S-G1083b, snoRNA:Me28S-G1083c, snoRNA:Me28S-G1083d*
Detoxification	6	0.003181	*CG5948, GstO1, GstO2, GstO3, Txl, se*
Cellular chemical homeostatis	8	0.009417	*CG11619, CG18135, CG3942, CG6125, Dop1R2, MCO3, MICU1, foi*
Negative regulation of gene expression	21	0.022001	*Spn-E, Su(H),* ***esg****, CG15262, Elba2, NC2beta, drm, sob, stc, Atx2, PCID2, Rh6, SmydA-2, Usp47, hdc, insv, lin-52, mir-2c, rhea, svp, vig*
Negative regulation of cellular biosynthetic process	18	0.023316	***esg****, Su(H), Elba2, NC2beta, drm, insv, sob, spn-E, stc, Atx2, CG15262, GABA-B-R1, Rh6, Usp47, lin-52, mir-2c, rhea, svp*

* Genes previously found to have association with longevity are shown in bold.

4.6.2 GO Term Enrichment Analysis for Genes Residing in Nodes Identified Using PageRank Measure

Genes, found in nodes identified using PageRank measure, enriched in GO terms were mostly located in original bins in the extended Synthetic GWAS-based network (Table 4.7). Some of these original bins contained genes already known as associated with longevity, including *ATPCL* and *chm* (Peleg et al., 2016) and *Nf1* (Tong et al., 2007). These genes, along with genes

found in novel regions, were found to be enriched in GO terms for processes involved in regulation processes of the organism. The *Atg7* gene, along with genes found in original regions and one gene found in a novel region, was also enriched in the GO term 'regulation of apoptotic process', an important function in the ageing process.

Table 4.7 Gene Ontology (GO) enrichment analysis of genes found by using PageRank measure in the extended Synthetic GWAS-based network.

GO Term	Total number of genes enriched in GO terms	P-value	Genes found within novel regions	Genes with longevity relation found within original regions
Regulation of biosynthetic process	141	0.032552	*CG3328, dimm, Anp, CG33786, pcs*	***chm*, dpp, Nf1***
Regulation of glucose metabolic process	27	0.011477	*Eno*	***ATPCL, chm***
Regulation of membrane potential	15	0.034068	*CG7912*	
Regulation of apoptotic process	26	0.036777	*app*	***Atg7***

* Genes previously found to have association with longevity are shown in bold.

Genes, found in nodes identified using PageRank measure, in the extended DGRP GWAS-based network were enriched in different GO terms (Table 4.8) to those of the extended Synthetic GWAS-based network, including the GO term 'ageing', but many of the genes enriched came from original bins. Here, more longevity associated genes were found, with the well-known longevity genes *Indy* (Rogina and Helfand, 2013) and *chico* (Clancy et al., 2001) occurring among enrichments for GO terms 'ageing', 'developmental process' and 'immune response'. A longevity gene, *EcR* (Tatar et al., 2003), was also found to be enriched in some of these terms along with other novel regions.

Table 4.8 Gene Ontology (GO) enrichment analysis of genes found by using PageRank measure in the extended DGRP GWAS-based network.

GO Term	Total number of genes enriched in GO terms	P-value	Genes found within novel regions	Genes with longevity relation found within original regions
Ageing	18	0.007864	***EcR****	***chico, GlyP, Nf1, Thor, mle***
Developmental process	192	0.027609	***EcR**, Chi, drm, Axn, spn-A, CSN1b, Sox14, HSPC300, Sry-delta, sob, Unc-89, sima, Alas, CG15515, CecA1, CecB, EbpIII, Phm, RpS28a, Sry-alpha, Tbce, janA, janB, mr*	***Pten, chico, Nf1, Thor, mle, GlyP, Indy***
Immune response	32	$5.083e^{-4}$	*Anp, CecA1, CecA2, CecB, CecC, Sr-CIV, sima*	***Thor, GlyP, Pten, chico***
Regulation of gene expression	95	0.011513	***EcR**, sob, orb, Ars2, drm, Chi, sima, Sry-delta, SmydA-5, Sox14, ZIPIC, thoc5*	***mle, Thor, bsk***

* Genes previously found to have association with longevity are shown in bold.

4.7 COMPARISON OF NETWORKS FOR BOTH GWAS STUDIES

4.7.1 Common Regions Identified between extended DGRP and Synthetic GWAS-based Networks using Network Measures

Nodes identified using network measures were found in common between the extended DGRP and Synthetic GWAS-based network. These common regions included original bins in each network, as well as novel bins. Among the significant bins found using clustering coefficient measures, bins 32 (Chr2L: 2480000-2560000), 192 (Chr2L: 15280000-15360000), 195 (Chr2L: 15520000-15600000), 702 (Chr3L: 12000000-12080000), 1129 (Chr3R: 21600000-21680000) and 1134 (Chr3R: 22000000-22080000) were found in common between the extended Synthetic GWAS-based network and extended DGRP GWAS-based network. All six of these bins were novel in both networks. All genes in each region were analysed one-by-one, using FlyBase database tools to find their genetic phenotypes. All genes were also compared to the list of longevity genes in the GenAge database. In these six bins, two genes already associated with longevity, *esg* in bin 192 (Chr2L: 15280000-15360000) and *Sema-5c*

in bin 702 (Chr3L: 12000000-12080000), as well as genes *oaf* (Chr2L:2492955-2498846) and *Ald* (chr3R:22079791-22087313) were found to have the phenotype 'increased mortality'. All genes found in these six common bins were grouped together and analysed for GO term enrichment, however the results were not significant for any terms that may be related to longevity.

Twelve bins identified using PageRank Measure were found in common between the two extended GWAS-based networks: 22 (Chr2L: 1680000-1760000), 30 (Chr2L: 2320000-2400000), 534 (Chr2R: 19760000-19840000), 535 (Chr2R: 19840000-19920000), 660 (Chr3L: 8640000-8720000), 989 (Chr3R: 10400000-10480000), 1097 (Chr3R: 19040000-19120000), 1131 (Chr3R: 21760000-21840000), 1180 (Chr3R: 25680000-25760000), 1182 (Chr3R: 25840000-25920000), 1183 (Chr3R: 25920000-26000000) and 1184 (Chr3R: 26000000-26080000). However, only one of these bins was a novel bin in both datasets, bin 1184. Bin 1184 contained 31 genes, for which phenotypic data was not readily available, however the *CG9747* gene with a phenotype of 'increased mortality' was present. Bin 1184 also contained five genes known to play a role in immune response, four of these being antimicrobial peptides known as cecropins: *CecA1*, *CecA2*, *CecB* and *CecC*, as well as the *Anp* gene. Antimicrobial peptides are known to be important defence molecules of the innate immune system, and in *Drosophila* cecropins are synthesised as a response to infections (Kylsten et al., 1990). The induced expression of antimicrobial peptides *Drosocin* and *CecropinA1* have previously been found to significantly prolong lifespan of adult flies (Loch et al., 2017).

Bin 535 was common to both networks but only novel in the extended DGRP GWAS-based network. This bin was found to contain seven genes: *Chi*, *Alas*, *mr*, *Gadd34*, *Sox14*, *Phm* and *Adk2*, with a phenotype of 'increased mortality'. The phenotypes of genes found in the remaining ten common bins were also searched for, for which numerous ageing related phenotypes were observed. These included *frtz*, *Atxn7*, *CG5339*, *CG4434* and *Zip99C* genes found to have phenotypes of 'short lived' and *Rim2* and *Tpi* found to have phenotypes of 'long lived'. Sixteen genes analysed were also found to have a phenotype of 'increased mortality': *cpb*, *Eno*, *VGlut*, *DCP1*, *Lpt*, *SERCA*, *Galphas*, *h*, *stumps*, *put*, *l(3)L1231*, *Cdc16*, *E(spl)mdelta-HLH*, *dgt1*, *spn-A* and *sima*. All genes found in the twelve bins common between both networks were grouped together and analysed for enrichment in GO terms. Among the

significant terms found in this analysis were ‘phagocytosis’, ‘processes in the tracheal system’ and ‘regulation of homeostatic processes’.

4.7.2 Human Ortholog Search

Drosophila are commonly used as a model organism for reasons previously discussed in this thesis, with one main reason for which they are so useful being due to the ability to apply findings in this organism to humans. So far, all findings in this study have been related only to genes in *Drosophila*, and so the next step was to see if we are able to map these findings onto humans, by searching for human orthologs of the genes found to be significant in the studies reported above.

For all genes found in the common regions observed in both extended GWAS-based networks, each *Drosophila* gene was analysed for the mapping of any human orthologs onto this gene using an Integrative Ortholog Prediction Tool available at https://www.flyrnai.org/cgi-bin/DRSC_orthologs.pl. All human orthologs found to match with *Drosophila* genes were further analysed by looking at the function of this ortholog and exploring any association that this could have with longevity. Numerous *Drosophila* genes were found to have a human ortholog, of which several were found to have biological functions, which could suggest a link to longevity/ageing.

The *CG5886* gene in *Drosophila*, an uncharacterized protein, was found to have the human ortholog *TXLNA*, which is a gene related to the Innate Immune System pathway. One major function of the innate immune system is to control inflammation and maintain the cytokine balance; therefore, defects in this system may diminish the ability to combat infection. At the site of infection, cellular components of the innate immune system, for example, neutrophils and macrophages are found and therefore any alterations in innate immunity role of such cellular components in inflammatory response could therefore have an impact of human health, consequently longevity (Solana et al., 2006). The *esg* gene in *Drosophila*, enriched in ‘respiratory system development’, was found to have the human ortholog *SNAI2*, a protein that has anti-apoptotic activity. Apoptosis is a process that goes on continuously throughout life, eliminating unneeded, damaged and senescent cells from the body and is therefore essential in optimising cell functions. Dysregulation of apoptosis, therefore, could lead to a

decline in immune function, therefore effecting chances of longevity (Joaquin & Gollapudi, 2001).

The *CG15262 Drosophila* gene, another uncharacterized protein, was found to have the human ortholog *CNOT2,* a gene that encodes a subunit of the multi-component CCR4-NOT complex. This complex regulates mRNA synthesis and degradation which is an essential determinant in the regulation of gene expression. Gene expression control is achieved at the level of the mRNA clearance, as well as mRNA stability and accessibility to other molecules. The assembly and function of specific mRNA granules that harbour the mRNA decay machinery can be modulated to promote stress resistance to adverse conditions (Borbolis & Syntichaki, 2015), and therefore over time affect the ageing process and lifespan of an organism. The *Ald* gene, involved in glucose homeostasis and related to phenotypes 'increased mortality' and 'lethal', was found to have the human ortholog *ALDOA*, a gene related to the metabolic pathway. Biological processes that can be associated to longevity, such as altered metabolism, are considered not just as a consequence of old age, but also a potential driving force of longevity (Häsler et al., 2017). The *ALDOA* gene is also related to the Sudden Infant Death Syndrome (SIDS) pathway, otherwise known as cot death or crib death, which is the sudden unexplained death of a child less than one year of age.

Another gene found to have a human ortholog which relates to a pathway associated with lifespan regulation is the *Slh* gene in *Drosophila*, a gene involved with response to toxic substance and phenotypes 'lethal' and 'neuroanatomy defective'. The *Slh* gene has human ortholog *SCFD1*, which is related to pathways including subsequent modification of proteins, and for all organisms from yeast to human, lifespan is known to be regulated by protein modification. For example, the gene encoding sirtuin, modifies proteins as a protein deacetylase and is a well-known marker of life span regulation (Lee et al., 2018).

Several of the *Drosophila* genes searched were found to have human orthologs which are associated with diseases that can potentially have an effect on the lifespan of a human. These included the *Drosophila* gene *CG17770*, which is associated with the biological process calcium-mediated signalling, having the human ortholog *CALML6*. Among the related pathways for this ortholog is the Apelin signalling pathway. Apelin is the endogenous ligand of APJ, the orphan G protein-coupled receptor. The apelin–APJ signal transduction pathway is widely expressed in the cardiovascular system and is crucial in cardiovascular homeostasis,

such pathway is often related to heart diseases with high morbidity in the elderly, including heart failure and atrial fibrillation (Zhou at al., 2017). The human ortholog found for the *Drosophila* genes *CG9743* and *CG9747* was *SCD*. These *Drosophila* genes are both involved in the biological process oxidation-reduction process and have phenotypes 'increased mortality' and 'lethal'. The human gene *SCD* is associated with many diseases including Reye Syndrome and Fatty Liver Disease. In Reye Syndrome abnormal accumulations of fat begin to develop in the liver and other organs of the body, as well as a dramatic increase of fluid pressure in the brain. Unless diagnosed and treated successfully, death is common, often within a few days, and even a few hours. The *CG1983 Drosophila* gene, an uncharacterized protein, was found to have the human ortholog *PLPBP*, which is known to have a tumour suppressor effect on hepatocellular carcinoma, which is the most common type of primary liver cancer. Liver cancer is rarely detectable early, at which point it is most treatable, making this cancer very difficult to cure and leading to fairly low survival rates in humans.

The *Drosophila* gene *kek3*, for which the biological processes are unknown, was found to have the human ortholog *LRRC4C*, which is associated with Extragonadal Seminoma. Primary extragonadal seminoma (EGS) is a rare tumour found in young adults, which often presents with bulky primary tumours and metastatic disease where primary cancer cells break away and travel through the blood or lymph system, forming new tumours in other parts of the body. The *LRRC4C* gene is related to the Cell Adhesion Molecules pathway; cell adhesion molecules are glycoproteins expressed on cell surfaces and play critical roles in biological processes including the immune response and inflammation. The *Drosophila* gene *Sema-5c* was found to have the human ortholog gene *SEMA5B*. This *Sema-5c* gene is associated with biological processes including axon guidance and central complex development, its phenotypes include 'long lived' and this gene has previously been observed in a screening for longevity genes in *Drosophila* (Seong et al., 2001). Its ortholog *SEMA5B* encodes a member of the semaphorin protein family. This protein family is known to regulate axon growth during development of the nervous system, in which the axon carries all data humans use to sense environment and carry out behaviours. The nervous system plays an important role in processing complex information from the environment, which could have a major influence on an animal's ageing and longevity (Alcedo et al., 2013), meaning that any factor effecting the way in which a nervous system functions may have an effect on longevity.

4.7.3 SNP Counts in Significant Regions/Genes

For the novel bins common between these extended GWAS-based networks when applying network measures, the total number of SNPs recorded in both the Synthetic GWAS and DGRP GWAS datasets were counted for each novel region; counts displayed in Table 4.9. In both SNP datasets, the number of SNPs found to reside in each of the bins is fairly high, with bin 1184 harbouring the highest number of SNPs as compared to any of the total 1503 80 Kb bins recorded.

Table 4.9 Number of SNPs recorded in novel bins identified by using various network measures and common for both networks.

Bin number	Bin position	Number of SNPs recorded in Synthetic GWAS dataset	Number of SNPs recorded in DGRP GWAS dataset
32	Chr2L: 2480000-2560000	92	2158
192	Chr2L: 15280000-15360000	76	2225
195	Chr2L: 15520000-15600000	61	1592
702	Chr3L: 12000000-12080000	106	2706
1129	Chr3R: 21600000-21680000	75	1774
1134	Chr3R: 22000000-22080000	91	1783
1184	Chr3R: 26000000-26080000	123	3102

For all common bins in both extended GWAS-based networks that have high clustering coefficient or PageRank scores, all genes residing in these bins were found and the number of SNPs residing in each of the genes was counted. The 20% of genes with the highest number of SNPs are listed separately for common regions found by each network measure in Tables 4.10 and 4.11 (see Appendix Tables S4.2 and S4.3 for full lists). These lists contained genes that have been previously found to associate with longevity, including *Sema-5c*, *Ald*, *oaf*, *pnt* and *Nf1*.

Table 4.10 Genes in common regions identified by clustering coefficient measure between extended Synthetic and DGRP GWAS-based networks, containing the highest number of SNPs.

Gene	Gene position	Bin	Number of SNPs recorded in Synthetic GWAS dataset	Number of SNPs recorded in DGRP dataset	Percentage of total number SNPs in Synthetic GWAS dataset	Percentage of total number of SNPs in DGRP dataset
rols	Chr3L:12001325-12057959	702	76	1884	0.0811	0.0860
CR44320	Chr3R:22020972-22055990	1134	34	711	0.0363	0.0324
CR46061	Chr3R:22061047-22078654	1134	25	466	0.0267	0.0212
kek3	Chr2L:15552723-15583978	195	22	591	0.0235	0.0269
Sema-5c	Chr3L:12060611-12074930	702	18	503	0.0192	0.0229
Ald	Chr3R:22079791-22087313	1134	9	158	0.0096	0.0072
oaf	Chr2L:2492955-2498846	32	4	114	0.0042	0.0052
CG6793	Chr3L:12035487-12037317	702	4	75	0.0042	0.0034
CG12290	Chr3R:22055970-22060422	1134	4	71	0.0042	0.0032
Slh	Chr2L:2488197-2492671	32	4	69	0.0042	0.0031
CG3515	Chr2L:2551145-2552825	32	4	50	0.0042	0.0022
CG15263	Chr2L:15285000-15286047	192	4	50	0.0042	0.0022

Table 4.11 Genes in common regions identified by PageRank measure between extended Synthetic and DGRP GWAS-based networks, containing the highest number of SNPs.

Gene	Gene Position	Bin	Number of SNPs recorded in Synthetic GWAS dataset	Number of SNPs recorded in DGRP dataset	Percentage of total number SNPs in Synthetic GWAS dataset	Percentage of total number of SNPs in DGRP dataset
sima	Chr3R:25884033-25947520	1182	81	1865	0.086532	0.085145
pnt	Chr3R:19115953-19171889	1097	57	1316	0.060893	0.060081
CG2970	Chr2R:19836724-19839982	534	38	1107	0.040595	0.038036
stumps	Chr3R:10402939-10433801	989	28	803	0.029912	0.03666
AdoR	Chr3R:25960578-25975328	1183	24	536	0.025639	0.018417
CG17646	Chr2L:1732524-1750612	22	23	501	0.024571	0.022873
CG4467	Chr3R:19053449-19074625	1097	22	609	0.023503	0.020925
CG31038	Chr3R:25702781-25728953	1180	22	442	0.023503	0.015187
VGlut	Chr2L:2391660-2410662	30	21	514	0.022434	0.023466
kcc	Chr2R:19795177-19812589	534	19	418	0.020298	0.014362
CG10904	Chr2R:19856464-19857596	535	17	568	0.018161	0.025932
orb	Chr3R:19086332-19106578	1097	17	245	0.018161	0.011185
Axn	Chr3R:25848564-25861033	1182	16	403	0.017093	0.013847
Adk2	Chr2R:19911730-19913400	535	15	390	0.016024	0.0134
DNA-ligI	Chr2R:19778035-19780700	534	14	381	0.014956	0.017394
Rim2	Chr2L:1716244-1724388	22	14	247	0.014956	0.011277
CG7886	Chr3R:10452398-10468285	989	13	367	0.013888	0.01261
Nap1	Chr2R:19792698-19794817	534	12	311	0.01282	0.010686
EbpIII	Chr2R:19913695-19915041	535	12	279	0.01282	0.012738
CG11498	Chr3R:25978744-25986195	1183	11	279	0.011751	0.012738
Mgat2	Chr3R:25841235-25847943	1182	11	255	0.011751	0.011642
Nf1	Chr3R:21808735-21821459	1131	11	200	0.011751	0.009131

Cad88C	Chr3R:10456200-10467498	989	10	295	0.010683	0.013468
CG17658	Chr2R:19782586-19784520	534	10	277	0.010683	0.009518
CG3121	Chr2R:19904944-19907419	535	10	260	0.010683	0.01187
CG42261	Chr3R:21795802-21808458	1131	10	241	0.010683	0.011003
CG9743	Chr3R:26022082-26028720	1184	10	218	0.010683	0.009953
Alas	Chr2R:19852295-19854280	535	9	210	0.009615	0.009587
Gadd34	Chr2R:19863122-19864778	535	9	186	0.009615	0.008492
CG31036	Chr3R:25733829-25738614	1180	8	144	0.008546	0.006574
mr	Chr2R:19859954-19862931	535	8	136	0.008546	0.006209
trp	Chr3R:25740243-25746011	1180	7	163	0.007478	0.007442
CG14853	Chr3R:10435364-10444709	989	6	273	0.00641	0.00938
CG9747	Chr3R:26011145-26016974	1184	6	164	0.00641	0.007487
CR44806	Chr2R:19795997-19796398	534	6	137	0.00641	0.006255
Zip99C	Chr3R:25689423-25694617	1180	6	131	0.00641	0.005981
CG15528	Chr3R:25874357-25877834	1182	6	120	0.00641	0.004123
Eno	Chr2L:1724768-1729636	22	6	115	0.00641	0.00525
frtz	Chr2L:1711302-1715851	22	6	114	0.00641	0.003917
CR46082	Chr3R:25960730-25963303	1183	6	90	0.00641	0.004109
CG9733	Chr3R:26069700-26072425	1184	5	145	0.005341	0.00662
CG31028	Chr3R:25992353-25996929	1183	5	139	0.005341	0.006346

4.8 EXPLORING SUBNETWORKS

For each extended GWAS network, subnetworks were found using modularity scores calculated as described in Chapter 4.5. These subnetworks produced using the Louvain modularity measure (see Figures 4.5 and 4.6) were further explored with the aim to identify novel genes/regions that co-located with genes, known to be associated with a specific phenotype, and enriched in the same GO terms as known genes. First, we compiled a list of genes residing in the regions of these individual subnetworks, and then performed Gene Ontology Enrichment analysis on these genes. The idea of this approach was that if genes, residing in original or novel regions, in the same subnetworks were found to be enriched in the same GO term, this would act as a proof of concept that there may be one SNP residing in one gene that has an effect by itself on the biological function shared or that more than one SNP in these functionally-related genes may have a cumulative effect on the shared biological function and therefore on longevity.

A division into dominant and other nodes in the next part of this analysis is done simply to distinguish between different subnetworks; despite this subdivision all regions corresponding to nodes within subnetworks occur in close proximity within the cell nucleus.

4.8.1 Subnetworks of Extended Synthetic GWAS-Based Network

A modularity score for the extended Synthetic GWAS-based network was 0.868 and produced 81 communities (modules). Figure 4.7 shows a modularity size distribution graph, where for each module in the network, the number of nodes within the module is recorded. Several subnetworks that appeared to have one dominant node were further explored for each dataset, where all genes found in both the dominant node and all neighbouring nodes in the same module were grouped together for each subnetwork and entered for GO enrichment analysis. The extended Synthetic GWAS-based network was found to have 7 subnetworks in which a dominant node was apparent; Table 4.12 summarises the nodes found in each of these subnetworks, along with the number of genes found in these nodes which were then used for GO term enrichment analysis.

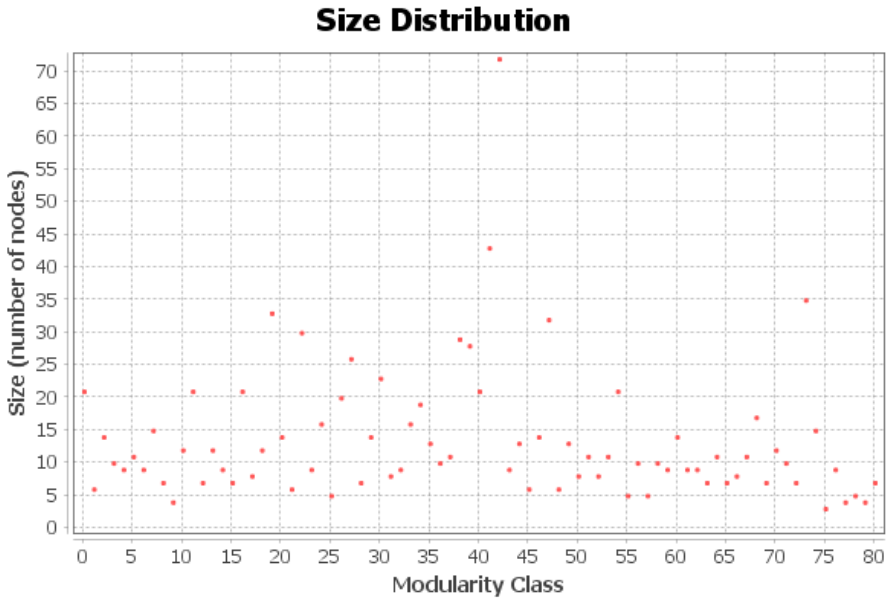


Figure 4.7 The distribution of sizes of modularity classes for the extended Synthetic GWAS-based network.

Table 4.12 A summary table of subnetworks in the extended Synthetic GWAS-based network chosen for further analysis.

Modularity class	Main node	All other nodes in module	Total nodes in module (including main node)	Total genes in module
4	93	94, 92, 89, 90, 91, 95, 1319, 96	9	92
5	128	125, 129, 130, 131, 711, 929, 124, 126, 127, 740	11	229
11	210	578, 203, 205, 209, 211, 206, 207, 916, 204, 208, 982, 1431, 930, 1436, 198, 199, 200, 201, 202, 1049	21	249
23	334	335, 332, 337, 330, 336, 331, 333, 749	9	141
29	397	401, 400, 393, 395, 394, 396, 399, 398, 402, 403, 961, 1044, 1338	14	222
60	822	818, 821, 476, 1218, 820, 826, 824, 823, 827, 313, 385, 819, 825	14	174
67	1082	1081, 1086, 502, 1079, 1080, 1078, 1083, 1084, 1085, 1386	11	150

Once GO enrichment analysis was performed on groups of genes found in these subnetworks, those with most significant relation to longevity were selected, and genes found to be enriched in selected longevity related GO terms are displayed in Table 4.13. The number of nodes in which these genes were found was also counted, to show that genes found to be enriched in the same GO term reside in different nodes (regions). Novel genes residing in close proximity to, and sharing the same longevity related biological processes with known longevity genes can be considered as good longevity candidate genes.

Table 4.13 Genes residing within subnetworks of the extended Synthetic GWAS-based network and enriched in longevity-related GO terms.

Modular class	GO term	P-value	Genes enriched in GO term	Number of gene harbouring nodes/total number of nodes
4	cellular response to stimulus	0.004503	*Rab39, Tom40, santa-maria,* ***Mnn1****, sem1, Pvf2,* ***Gr28b****, Pvf3, Ziz, RapGAP1, Wnt4, wg, Wnt6, Wnt10**, ninaC, CG5160, CG5181, mir-305*	6/9
4	localization	0.007119	*Rab39, Tom40, Sem1, Pvf2, CG13793, CG13794, CG13795, CG13796, CG31904, CG31907, CG33296, Pvf3, Ndae1, Wnt4, ninaC, Ntl, ATPsynGL, Nuf2*	5/9
4	cell communication	0.023993	*Rab39, santa-maria,* ***Mnn1****, Pvf2, Gr28b, Pvf3, Ziz, RapGAP, Wnt4, wg, Wnt6, Wnt10, ninaC, CG5160, mir-305*	7/9
5	macromolecule modification	0.003413	***Atg1****, Ptp69D, Cnot4, RluA-1, CG32847, CG33303, CG34183, CG42366, Fkbp59, CG4839, Ror, CG4968, Sps2, gny, STUB1, Sp27A, LManI, Bug22, Cdk1, Cand1, Usp14,* ***CYLD****, Utx,* ***Pten****,* ***bsk****, Dref, RluA-2, LMannII, FBXO11*	9/11
5	cellular catabolic process	0.020971	***Atg1****, lft, CG32847, CG4592, CG4594, CG4598, yip2, Prosalpha6, RpS27A, CG5367, Usp14, Utx,* ***Pten****, CG5676,* ***bsk****,* ***chico****, CG5731, CG8526, FBXO11*	9/11
11	DNA repair	0.021953	*CG17329,* ***ku80****, CG31807, CG33552, EndoGI, CG5316*	5/21
11	developmental process	0.010492	*cact, Cas, chif, cni, crp, dac,* ***foxo****, fzy, glue, goe, grp, heix, her, mdy, mir-9b, mir-9c, sing, squ, twe, wek, yellow-b, BicC, BuGZ, CG17328, CG32572, CG4793, CG5953,* ***Ca-alpha1D****, Cyp303a1, Cyt-c-d, EndoGI, GMF, Idgf1, Idgf2, Idgf3, Mhc, Npc2b, Syx5, TwdlX, TwdlY, TwdlZ,*	14/21

			*TwdIalpha, **VhaSFD**, beat-Ia, beat-Ib, beat-Ic*	
23	apoptotic process	0.033954	***azot**, tor, **cathD**, Cul1, fwe, mir-263b*	4/9
23	positive regulation of gene expression	0.028920	*CG12769, Rpt1, Kdm4A, udd, **Nup50**, nito, CG6244, Lpin, lig*	6/9
29	negative regulation of transcription, DNA-templated	0.024472	*CG10038, spt4, wuc, Iz, seq, Kdm4B, **sug**, **Psc**, **Su(z)2**, Iswi*	7/14
60	gene expression	$5.603132e^{-4}$	*CG10474, Rpb8, sa, CG11906, mip40, Pc, croc, barc, CRIF, Hr78, wbl, rib, Tsr1, eg, CycH, CG7414, Nopp140, **mub**, RpLPO, Cdk12, TfAP-2 , **rho-7***	9/14

* Genes previously found to have association with longevity as recorded in FlyBase or GenAge resources are shown in bold.

** Genes residing within original nodes, i.e. harbouring SNPs with D>7.9 are underlined.

GO analysis found an enrichment in the term 'DNA repair' for modular class 11 (P-value = 0.021953), for which six genes that share this term were found to reside in five different regions of the subnetwork. The ability for a cell or tissue, and therefore organism, to function efficiently relies on the maintenance of stability of its unit components. Such stability is dependent on the differentiated state of the system, whereby the higher the differentiated state, the greater amount of stability is required. The process of DNA repair is one way in which stability can be achieved, therefore making DNA repair essential in order to avoid problems effecting organism function. For example, it has been shown that unrepaired DNA damage that arises in stem cells over time leads to stem cell exhaustion which has been proposed to be a principle mechanism of ageing (Nijnik A et al., 2007). The gene *ku80*, enriched in 'DNA repair', is identified as a candidate gene in longevity, found to have a positive effect on the lifespan of *Drosophila* via DNA repair (Shaposhnikov et al., 2015). A phenotype search for this gene also states 'short-lived'. The *ENDOGI* gene is involved in positive regulation of the Notch signalling pathway, and an 'increased mortality' phenotype. Notch signalling is important for cell-cell communication, and therefore plays a role in important processes such as neuronal function and development (Gaiano and Fishell, 2002). Notch signalling is also dysregulated in many cancers (Bolós et al., 2007) and faulty signalling is implicated in many diseases (Harper et al., 2003). The other four genes *CG17329*, *CG31807*, *CG33552* and *CG5316* enriched in the 'DNA repair' GO term are genes that are currently uncharacterized. In relation to DNA repair, longevity associated gene *FOXO3a*, discussed in Chapter 1 of this thesis, has also been found to stimulate the DNA Repair Pathway through

the *Gadd45* protein which has also been associated with ageing (Tran et al., 2002). It has also been revealed that most premature ageing syndromes are due to mutations found in genes that encode proteins involved in DNA repair (Karanjawala and Lieber, 2004).

Six genes in modular class 23 were enriched in the 'apoptotic process' GO term (P-value = 0.033954). One major role of apoptosis is to remove cells in the organism, which may harm the way in which an organism functions as a whole. The apoptotic process is almost an alternative to the previously discussed GO term 'DNA repair', whereby when DNA is damaged, the checkpoint protein *p53* is activated and the decision is made as to whether replication should be stopped and the DNA should be repaired, or made to die by apoptosis (Warner, 1999). The decline in immune response with ageing has been well established, with the decline in cellular immune response, which is mediated by T lymphocytes, being observed (Miller, 1996). Apoptosis has been found to regulate the size of the lymphocyte population in an organism (Mountz et al., 1996), and impaired functions as a result of age associated immune decline are known to concern mainly T lymphocyte, for which the relationship between autoimmune diseases increasing with age, and age-related autoreactivity has been also studied in relation to T cell reactivity and autoantibodies production (Makinodan and Kay, 1980; Candore et al., 1997). Studies have also found that in mammals, at least in part, apoptosis plays an important role in the process of ageing and tumorigenesis and that age-enhanced apoptosis may work as a protective mechanism against age-associated tumorigenesis (Higami and Shimokawa, 2000). The six genes in this enrichment group include *CathD*, a gene with phenotypes including those associated with apoptosis such as 'increased cell death' as well as longevity associated phenotype 'short-lived'. The *Drosophila* gene *azot* has the phenotype 'long-lived' and has previously been observed to result in a longer lifespan (Proshkina et al., 2015). Another gene enriched in this GO term is *Cul1*, this gene has phenotypes of 'increased mortality' and 'neuroanatomy defective'. The *Fwe* gene encodes a transmembrane protein that mediates win/lose decisions in cell competition and neuronal culling curing development and ageing, again this gene has longevity related phenotypes 'increased mortality' and 'lethal'. Genes residing in modularity class 23, sharing the same biological function process 'apoptotic process' and coming into close proximity with longevity genes *CathD* and *azot* can be speculated to have an effect on longevity in the same way as these known longevity genes.

Some genes enriched in the GO terms displayed in Table 4.13 have been shown previously to have association with longevity or display phenotypes associated with ageing in which most cases this association is with increased lifespan. This includes *chico*, a gene encoding an insulin receptor substrate that functions in an insulin/insulin-like growth factor (IGF) signalling pathway and is found to increase lifespan by up to 48% (Clancy et al., 2001). The *Pten* gene, when activity is increased, has been shown to delay the process of proteostatis and therefore result in a decrease in the loss of muscle strength during muscle ageing. This increase in *Pten* activity has been found to increase maximum lifespan by up to 7.7% in comparison with matched controls (Demontis and Perrimon, 2010). In the same modularity class, the *Atg1* gene was found. Neuronal-specific upregulation of this *Atg1* gene during adult-onset has been found to result in increased median lifespan by up to 25% (Ulgherai et al., 2014). These three genes residing in the modular class 5 subnetwork, *chico*, *Pten* and *Atg1* harbour 80, 61 and 506 SNPs from the DGRP GWAS dataset, however none of these residing SNPs had a significant P-value. We can speculate that the SNPs residing in these longevity-associated genes, could work collectively or with the SNPs that are found to reside in the genes not yet associated with longevity but enriched in the 'cellular catabolic process' GO term. As a result of this, the SNPs in these three longevity genes can influence genes found in the same subnetworks, sharing the same biological functions to also having ageing effects. The *dFOXO* gene, a transcription factor also involved in the regulation of the insulin signalling pathway, is a commonly known longevity gene in *Drosophila* (Giannakou et al., 2004; Giannakou et al., 2007; Hwangbo et al., 2004). In all cases it was found that overexpression of this gene results in an increase in maximum lifespan. Genes *VhaSFD* and *sug* were both identified in the same study (Landis et al., 2003), when overexpressed, to result in an increase in mean life span by 5-10%. The *mub* gene has previously been found to have association with longevity when mutated; a study showed that the insertion of a p-element in the gene results in longer lived *Drosophila*. These longevity genes *foxo*, *Mnn1*, *VhaSFD* and *mub* were all found to harbour a number of SNPs from the DGRP GWAS dataset. The highest SNP enriched gene of these was *mub*, which harboured 1649 SNPs where the lowest P-value was 1.04×10^{-5}. *Foxo* also harboured a large number of SNPs, 715, with a lowest P-value of 4.88×10^{-5}. Genes *Mnn1* and *VhaSFD* harboured 118 and 92 SNPs respectively, however their minimum P-values were not significant. Novel genes residing in close proximity to, and sharing the same longevity

related biological processes with known longevity genes *mub* and *foxo* can be considered as good longevity candidate genes.

Several genes in Table 4.13 were found to display the phenotype 'increased mortality' in a search using the FlyBase database. These included the *CYLD* gene, a cancer consensus gene responsible for tightly limiting the immune response duration (Zerofsky et al., 2005). Generated *Drosophila CYLD* (*dCYLD*) mutant has been proven to be essential for JNK-dependent oxidative stress resistance and normal lifespan and has also been indicated to play a critical role in modulating TNF-JNK-mediated cell death (Xue et al., 2007). The *Mnn1* gene also displays this phenotype, for which this gene has been suggested to have a role in the regulation of stress response in *Drosophila* (Papaconstantinou et al., 2005). The association between stress and lifespan has often been made, and previous studies have observed differences in gene expression when comparing normal and stress conditions which has resulted in the identification of ageing genes in *Drosophila*. The genes found to reside in the same subnetworks as these genes previously shown to play roles in biological processes associated with longevity were found to harbour a number of SNPs themselves. We can speculate that the SNPs residing in these genes and SNPs residing in genes *CYLD* and *Mnn1* have a cumulative effect, which results in these genes being involved with the same biological process of 'regulation of immune system process'

There are several genes in Table 4.13 which have shown previously to be associated with decreased lifespan in *Drosophila*. Two genes were found to have a 'short-lived' phenotype when searched for in the FlyBase database, these being *CathD* and *ku80*. This study observed an increase in lifespan by up to 21.4% (Magwire et al., 2010). The *rho-7* gene has also been found to decrease lifespan where a study showed that knockout flies were found to have severe neurological defects as well as much reduced lifespan (McQuibban et al., 2006). Another gene found to cause a decrease in lifespan of *Drosophila*, this time in cases of RNA interference, was the *bsk* gene. Such interference in intestinal stem cells results in short life due to impaired intestinal homeostasis and tissue regeneration and has been found to reduce mean lifespan by 16.4% and 10.2% in males and females, respectively (Biteau et al., 2010). These genes known to have association with longevity can be speculated to interact with novel genes found in the same subnetworks sharing the same GO terms. Novel genes residing

in close proximity to, and sharing the same longevity related biological processes with these known longevity genes can be considered as good longevity candidate genes.

For all subnetworks shown in Table 4.12, all nodes selected as the main node were original node regions in the extended Synthetic GWAS-based network and the majority of other nodes in the subnetworks were novel (additional) regions of the extended Synthetic GWAS-based network. Therefore, many of the genes shown in Table 4.13, enriched in the specific GO terms, reside in novel regions of the extended Synthetic GWAS-based network.

4.8.2 Subnetworks of Extended DGRP GWAS-Based Network

For the extended DGRP GWAS-based network, a modularity score was 0.917 and 61 communities (modules) were observed. Figure 4.8 shows a modularity size distribution graph, where for each module in the network, the number of nodes within each subnetwork is recorded. Subnetworks that appeared to have one dominant node (node with a high degree compared to all other nodes in the same subnetwork) were then further explored, where genes found in both the dominant node and all neighbouring nodes in the same module were grouped together for each subnetwork and entered for GO enrichment analysis. The extended DGRP GWAS-based network was found to have 9 subnetworks in which a dominant node was apparent; Table 4.14 summarises the nodes found in each of these subnetworks, along with the number of genes found in these nodes which were then used for GO term analysis.

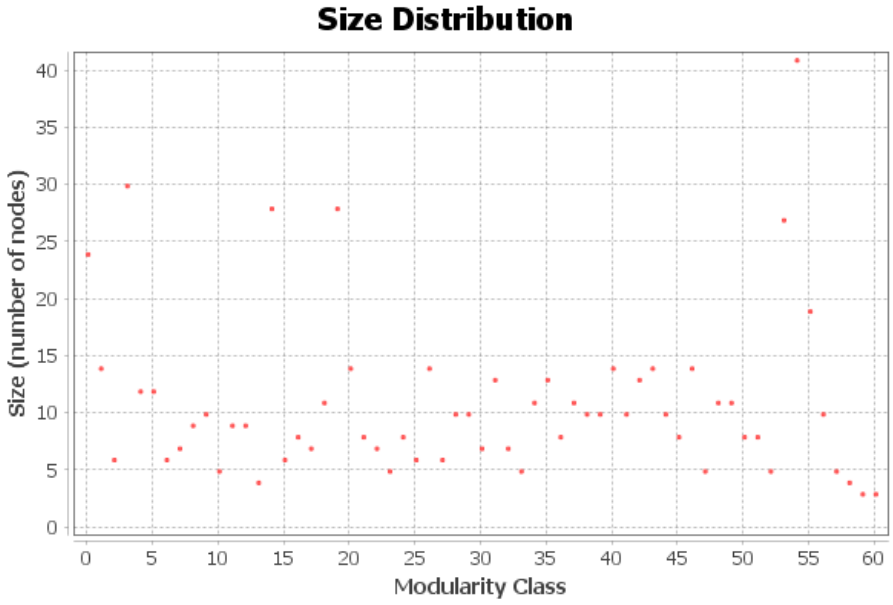


Figure 4.8 The distribution of sizes of modularity classes for the extended DGRP GWAS-based network.

Table 4.14 A summary table of subnetworks in the extended DGRP GWAS-based network chosen for further analysis.

Modularity class	Main node	All other nodes in module	Total nodes in module (including main node)	Total number of genes in module
4	56	53, 55, 57, 58, 61, 54, 59, 60, 1078, 1079, 1338	12	125
18	529	530, 523, 527, 531, 560, 524, 525, 526, 528, 1253	11	155
20	576	521, 575, 104, 453, 573, 578, 105, 478, 572, 574, 577, 579, 580, 581, 1227	16	256
26	660	329, 412, 499, 658, 661, 321, 445, 659, 662, 663, 1324, 1367, 1418	14	223
28	670	672, 330, 669, 673, 344, 366, 413, 671, 1419	10	177
34	778	776, 785, 318, 777, 779, 780, 781, 782, 783, 784	11	103
40	989	1210, 103, 458, 426, 983, 985, 986, 987, 988, 990, 991, 992, 1336	14	156
44	1091	448, 1095, 1089, 1094, 361, 429, 1090, 1092, 1093	10	156
49	1152	1151, 1147, 1153, 1154, 1155, 63, 457, 1148, 1149, 1150	11	142

Once GO analysis was performed on groups of genes residing within these subnetworks, those with most significant relation to longevity were selected, and the genes found to be enriched in selected longevity related GO terms are displayed in Table 4.15.

Table 4.15 Genes residing within subnetworks of the DGRP GWAS-based network and enriched in longevity-related GO terms.

Modularity class	GO term	P-value	Genes enriched in GO term	Number of gene harbouring nodes/total number of nodes
4	development growth	0.030823	*Elp3, ine, bdl,* ***ft****, CASK, tsl*	4/12
18	nervous system process	0.033071	*Or59c, bw, Gr59c, Gr59a, Gr59b, Gr59d, Gr59e, Gr59f, Or59b, Or59a, tko*	4/11
20	organelle assembly	0.016230	*Oseg2, Pp2A-29B, Rcd4, sls, Oseg4, CG42787, hts, Cnb, RpL11, Ar16, mtsh, RpL23A*	9/14
20	immune system process	0.035555	*CG10764, asrij, HBS1, sls, Rap1, ac, ecd, cnk, Ostgamma, Bgb, Bro,* ***Btk29A****, par-1*	9/14
26	regulation of immune system process	0.019532	***Traf6****,* ***PGRP-SA****, CG1572, Cyt-b5, GNBP3, GstO2,* ***Sod2****, Spn42Dd*	8/14
34	response to stimulus	0.031506	*geko, skl, AstC-R2, Adf1, Dic4, Trap1, geminin, Bap170, Debcl,* ***Chmp1****, GNBP2, not, CG4306, rpr,* ***grim****, hid, CG6893, GNBP1*	8/11
40	open tracheal system development	0.001555	*stumps, Cad88C, cv-c, grh, btsz, thr, put, scb*	5/14

* Genes previously found to have association with longevity as recorded in FlyBase or GenAge resources are shown in bold.

** Genes residing within original nodes, i.e. harbouring SNPs with $P \leq 3.33 \times 10^{-5}$ are underlined.

GO analysis found an enrichment in the terms 'immune system process' for modular class 20 (P-value = 0.035555) and 'regulation of immune system process' for modular class 26 (P-value = 0.019532), for which eight genes found to reside in eight different regions of the subnetwork shared this term. Immune senescence is the deterioration of immune function with age, in which the body's resistance to infection is highly reduced. As well as resistance to infection, immune senescence may also reduce resistance to cancer in humans and reduce chronic activation of the immune system as a result of infection of cancer. Immune senescence has

also been found to induce changes in immune response in humans that are very similar to those changes in elderly individuals (Tarazona et al., 2002). Alterations in the innate immune system were found to contribute to age-associated morbidity and mortality, which allowed for determination of the relative roles in which these immune pathways play (Huang et al., 2005). In response to ageing, most physiological functions are altered, for example the decline in cellular and humoral immunity. The most sensitive immune cells to ageing appeared to be T cells, and the most critical component of immunological ageing is known to be changes in the T lymphocyte compartment, concluded by studies on ageing humans (Goronzy et al., 2015), documenting significant changes in the functional and phenotypic profiles of T cells. Further analysis of literature has also suggested that the inability of the innate immune system working efficiently is a contributing factor to the development of many diseases observed in the elderly (Gomez et al., 2008).

Some genes enriched in the GO terms displayed in Table 4.15 have been found previously to have association with longevity, with many of them being associated with a decrease in life span. It has been found that *Drosophila*, heterozygous for the tumour suppressor gene *ft,* had a shorter lifespan, where it was proposed that this mortality effect was associated with the interaction between this *ft* tumour suppressor and signal transduction pathways mediated by the Hippo pathway (Kopyl et al., 2014). Phenotype searches for genes in this table found a number of genes to express the phenotypes 'increased mortality' and 'lethal' including genes *grim*, *Btk29A, Rap1, cnk, ecd, Ostgamma* and *tko.* The *sls* gene had a 'stress response defective' phenotype and *Chmp1* was found to express the phenotype 'short-lived'. An increase in the proapoptotic protein *grim* has been shown to significantly reduce lifespan in female *Drosophila* by up to 34% in median lifespan and 25% in maximum lifespan (Bauer et al., 2005). *Btk29A*, as well as *Traf6*, are genes found in Table 4.15 that are d*FOXO* targets in the JNK (Jun-N-terminal Kinase) signalling pathway. This signalling pathway is stress-activated and involved in developmental and metabolic regulation, immune responses and lifespan extension (Biteau et al., 2011; Karpac et al., 2009).

The *Sod2* gene has been observed, in separate studies, to have both a positive and negative effect on lifespan in *Drosophila*. When over expressed, the gene was found to result in a 20% increase in both mean and maximum lifespan (Curtis et al., 2007), and RNA interference-mediated silencing of *Sod2* caused an increase in oxidative stress which lead to early-onset

mortality in young adults (Kirby et al., 2002). The *PGRP-SA* gene has also been observed as one of few genes to show age-related changes in expression without being affected by diet, allowing this gene to be considered a candidate marker of ageing (Doroszuk et al., 2012). Other genes found within this subnetwork and sharing the same biological function as known longevity-associated genes may influence longevity in the same way as the known gene discussed.

4.8.3 Further Analysis of Subnetworks

Subnetworks found using GO enrichment analysis, containing genes which shared common biological processes with potential association to longevity, were then further explored. A search for longevity significant SNPs, highlighted previously in this study, was carried out within each subnetwork and they were further investigated.

Subnetworks with a dominant region not containing genes enriched in longevity related GO term

Subnetworks were observed in this analysis, whereby a dominant region was present, however this region itself did not contain any genes found to be enriched in the GO term observed in this subnetwork. The term 'dominant' is used here to indicate its relationship with other regions of the network. One has to remember that all interacting regions reside within close proximity to each other within the cell nucleus, e.g. all genes enriched in this GO term resided in the regions coming into close proximity with this dominant region. This was the case for a subnetwork in the extended Synthetic GWAS-based network (modularity class 11), which was found to contain six genes enriched in the GO term 'DNA repair'. In this example, the dominant region (bin 210) did not contain genes enriched in this GO term. All genes enriched in this GO term contained SNPs, but none with a calculated significant D value. This dominant region (bin 210) was then explored for presence of any enhancers which may reside in this region. Only enhancers found to target the *Mhc* gene, a gene found in bin 210, that harbours several longevity significant SNPs ($D \geq 7.9$). Using GO enrichment analysis, the *Mhc* gene did not show any enrichment, however a study has previously identified mutations

in the *Mhc* gene that lead to hypercontraction and subsequent degeneration of flight muscles in *Drosophila*, ultimately causing remodelling of the muscle cytoskeleton. This kind of alteration in muscle contraction is known to lead to a large array of diseases (Montana and Littleton, 2006). This enhancer can be speculated to also influence the expression not just of this gene but also other genes in close proximity.

A second example was in the DGRP GWAS-based network, in a subnetwork (modularity class 20) in which bin 576 was the dominant region. Thirteen genes in this subnetwork were enriched in the GO term 'immune system process', but none of these genes resided within the dominant node. However, the dominant node contained fifteen genes, and further exploration of these genes found several to have longevity associated phenotypes including *CG5687* and *Mfap1* with 'increased mortality' and 'lethal', and *Dbx* with the phenotype 'neuroanatomy defective'. The gene *mv* was also found to reside in the dominant node, which has the phenotypes 'immune response defective', 'stress response defective' and 'increased mortality' and in a previous study mutants in this gene are shown to enhance susceptibility to infections, cause a defect in the cellular immune response as well as affect autophagy due to growth of auto phagosomes beyond their normal size (Rahman et al., 2012). The 13 genes enriched in this immune system process GO term resided in nine regions of the subnetwork. These regions were all found to contain many SNPs, with some of these SNPs found to reside in all genes enriched in the GO term, however none with significant P values. This dominant region was further explored for any enhancers, which may be residing in this area, however none were found.

A final example of this case in the DGRP GWAS-based network was in a subnetwork (modularity class 26) found to contain eight genes enriched in the GO term 'regulation of immune system process'. In this subnetwork the dominant region (bin 660) again did not contain genes enriched in this GO term but came into close proximity to eight regions which did. All eight regions were found to contain SNPs, some of which resided in all the GO term enriched genes, however none of them with a longevity significance. We speculate that SNPs coming into close proximity work together, resulting in the genes in which these SNPs reside to be involved with the same biological process of 'regulation of immune system process'. Next, this dominant region was further explored for enhancers, in which many were found. This included enhancers for which the target genes were known, the *h* gene and enhancers

which had unspecified target genes. These enhancers found to have unspecified target genes can be speculated to influence the expressions of the genes sharing this longevity related biological process, found within this subnetwork. A search on genes residing in bin 660 found eight genes other than *h*, where a phenotype search on these genes found gene *h* to have phenotypes 'increased mortality during development' and 'partially lethal', and the *SrpRbeta* and *Cp18* genes to have a 'lethal' phenotype. This bin also harboured three uncharacterized genes *CR44526*, *CG6511* and *CG43965*. Another gene residing in this bin *Pex7*, a peroxisome biogenesis gene, is responsible for matrix enzyme import and receptor recycling in *Drosophila* (Fujiki et al., 2014). Mutations in many peroxin genes have been observed to lead to various forms of peroxisome biogenesis disorder (PBD), also known as Zellweger syndrome (ZS) in humans (Wanders and Waterham, 2006).

Subnetworks with dominant region containing genes enriched in GO term, where these genes do not contain significant SNPs

The second group comprises subnetworks in which the dominant region was found to contain genes enriched in the GO term. However, all SNPs that resided in these genes did not show significant association with longevity. For a subnetwork (modularity class 23) found in the extended Synthetic GWAS-based network, the dominant region (bin 334) and three regions coming into close proximity contained six genes enriched in the GO term 'apoptotic process'. Despite this dominant region containing significant SNPs, none of these SNPs residing in the genes enriched in this GO term were. All genes enriched in the 'apoptotic process' GO term were found to harbour SNPs, however none with a longevity significance ($D \geq 7.9$). This dominant region was then further explored for presence of any enhancers that may reside in this bin, and two enhancers were found. One of these enhancers was found to target the *Or43b* gene, a chemoreceptor that mediates response to volatile chemicals, which resides in bin 334. The target gene for the other enhancer was unspecified, and so can be speculated to influence the expression of any of the genes found in this subnetwork to be enriched in this longevity related biological process, potentially more than just one gene.

A similar observation was made for a subnetwork (modularity class 4) in the extended DGRP GWAS-based network, which contained six genes enriched in the GO term 'developmental

growth'. Bin 56 was the dominant region in this subnetwork, and this region harboured significant SNPs, however none of them resided in any of the genes enriched in this GO term. Three regions coming into close proximity to this dominant region were also found to harbour genes enriched in the 'developmental growth' GO term, where all genes enriched harboured non-significant SNPs. Non-significant SNPs residing in all genes enriched in this GO term are therefore speculated to work together, as these six genes come into close proximity, playing a role in developmental growth. This dominant region, bin 56, was then explored further for presence of any enhancers, in which four were found, all with an unspecified target gene. These enhancers could potentially target genes residing in other regions of the subnetwork, which are associated with the 'development growth' GO term.

Subnetworks with genes harbouring significant SNPs

A subnetwork found in the extended Synthetic GWAS-based network (modularity class 11), harbouring 46 genes enriched in 'developmental process' was found to contain significant SNPs (D>7.9) in three of the genes residing in the dominant region of this subnetwork: *yellow-b*, *BuGZ* and *glu* (bin 210). This region was found to be in close proximity to 13 regions, also containing genes enriched in the 'developmental process' GO term. Four of these regions were also found to harbour genes containing significant SNPs, including *mdy*, *Cas*, *VG5953*, *grp* and *Ca-alpha1D* with the remaining nine regions found to contain genes harbouring SNPs, but none with significant P-values. Therefore from this observation we conclude that genes *BuGZ*, *glue*, *mdy*, *Cas*, *VG5953*, *grp* and *Ca-alpha* in this subnetwork, harbouring significant SNPs, are influencing the biological function 'development process' of genes which reside in the nine regions coming into close proximity, containing no significant SNPs.

A subnetwork (modularity class 18), found in the extended DGRP GWAS-based network with its dominant region bin 529, contained 11 genes enriched in the GO term 'nervous system process'. All 11 genes harboured SNPs, with two of these genes, *Orc59c* and *bw*, residing in the dominant region and containing SNPs with a low P-value ($P = 8.11e^{-05}$ and $P = 2.26e^{-05}$). This dominant region was found to be in close proximity to three novel regions, also containing genes enriched in this GO term, in which SNPs also resided but had less significant P-values. One can speculate that the significant SNPs residing in genes *Orc59c* and

bw may work independently to influence the biological processes 'nervous system process' whereas other genes in close proximity which contain SNPs with no longevity significance may influence this longevity related biological process collectively.

4.9 EXPLORATION OF NOVEL REGIONS IN ONE EXTENDED GWAS-BASED NETWORK HARBOURING SIGNIFICANT SNPS IN THE OTHER GWAS DATASET

Both GWAS networks were compared to find regions, which were selected as original nodes in one GWAS network but added as novel nodes in the other. Note that original nodes are 80 Kb regions harbouring at least one significant SNP (either with $D \geq 7.9$ or$P \leq 3.33 \times 10^{-5}$). This comparison found 43 regions selected as novel regions in the extended Synthetic GWAS-based network that harbour significant SNPs in the DGRP GWAS dataset (see Appendix Table S4.4), and 85 novel regions in the extended DGRP GWAS-based network that harbour significant SNPs in the Synthetic GWAS dataset (see Appendix Table S4.5).

For those novel regions in the extended Synthetic GWAS-based network, the significant SNPs in the DGRP GWAS dataset residing in these regions were found. The genes that harboured these SNPs were searched for in all results from previous network analysis for both GWAS datasets. One of these genes, *bw*, was previously found to be enriched in the GO term 'nervous system process' in the subnetwork analysis for the extended DGRP GWAS-based network (see Table 4.15). Two genes, *Wnt4* and *sdk*, were also found to reside within a subnetwork of the extended Synthetic GWAS-based network enriched in 'cellular response to stimuli', 'localization' and 'cell communication' GO terms (see Table 4.13) and in GO enrichment analysis of genes found in regions with high clustering coefficient in the extended Synthetic GWAS-based network (see Table 4.5), where this gene was enriched in 'multicellular organismal process'. These genes harbouring significant DGRP SNPs, residing in novel regions of the Synthetic GWAS network, were then compared with those findings of Ivanov et al. (2015) in which 14 genes were found listed in their table of the top 50 important SNPs and their genes/nearby genes [Supplementary Table 3 in Ivanov et al. (2015)]. These genes included *CG10019*, *Blimp-1*, *CG10361*, *bmm*, *CG14073*, *CG32204*, *ATPsynD*, *CG4467*, *CG31510*, *Mlc1*, *Doa*, *CG7601*, *sdk* and *bves*. Genes found to harbour significant SNPs in the DGRP GWAS dataset that resided in novel regions of the extended Synthetic GWAS-based network were

also found listed in the top 30 genes found by gene-based analysis [Supplementary Table 4 in Ivanov et al. (2015)], this included *CG4972*, *CG10361* and *ATPsynD* as well as the *CG33700* gene listed in the top 30 genes found by gene-based analysis when genes $\pm$5 Kb of flanking regions were considered [Supplementary Table 5 in Ivanov et al. (2015)].

All genes found in novel regions of the extended Synthetic GWAS-based network harbouring significant SNPs from the DGRP GWAS dataset were also searched for the 'long-lived' phenotypes. The *ATPsynD* gene was found to have the 'long-lived' phenotype and is a gene known to be essential for normal development and interaction of this gene with TOR signalling has been shown to modulate protein homeostatis and lifespan in *Drosophila* (Sun et al., 2014).

For those novel regions in the extended DGRP GWAS-based network, the significant SNPs in the Synthetic GWAS dataset residing in these regions were also found. The genes that harboured these SNPs were searched for in all results from previous network analysis for both GWAS datasets. Several genes had been found in previous subnetwork analysis of the Synthetic GWAS-based network (see Table 4.13) including the *Wnt10* gene which was enriched in GO terms 'cellular response to stimulus' and 'cell communication' and the *CG5953* gene was enriched in 'developmental process' GO term. Also in the analysis for novel nodes with the highest degrees, both *Galphao* and *lola* genes were found enriched in GO terms (see Table 4.2) including 'apoptotic process' and 'nervous system development'. Finally, subnetwork analysis that focused on those subnetworks containing a dominant region (modularity class 11) in Table 4.13 also observed the *grp* gene to be enriched in the GO term 'developmental process'. The genes that harboured these significant SNPs from the Synthetic GWAS dataset were also searched for in all results from previous analysis for the DGRP GWAS-based network. In this case, four genes were found: *aop*, *vig*, *rhea* and *GstO2*. The *aop* gene was found in analysis of novel nodes with the highest degrees in the extended DGRP GWAS-based network, where it was enriched in the GO term 'immune system process' (see Table 4.4). The genes *vig* and *rhea* were found to be enriched the GO term 'negative regulation of gene expression' (see Table 4.6). Finally, the *GstO2* gene appeared in results of the analyses of the extended DGRP GWAS-based network as enriched in the GO term 'detoxification' (see Table 4.6) and 'regulation of immune system process' (see Table 4.15).

All genes found in novel regions of the extended DGRP GWAS-based network harbouring significant SNPs from the Synthetic GWAS dataset were also searched for the 'long-lived' phenotypes. Several genes were found to have the 'long-lived' phenotype including *aop*, *CG8677*, *hebe*, *magu* and *CG30427*. The *aop* gene and its relation to longevity has been previously discussed in this chapter (see Section 4.8.3). The *CG8677* gene has also been identified as a gene that is crucial for extension of longevity in *Drosophila* (Seong et al., 2001), and the *CG30427* gene known to be involved with the biological process of determination of adult lifespan was identified by Paik et al. (2012) who studied factors associated with controlling *Drosophila* ageing. Genes *hebe* and *magu* are also involved with the biological process of determination of adult lifespan, where both genes when over-expressed in adults have been observed to increase life span and modulate late-age female fecundity (Li and Tower, 2009).

Chapter 4 Conclusions

To identify novel regions associated with longevity, networks were created using genes/genomic regions that are quantified to associate with longevity as original nodes (regions) of the network, with additional nodes (regions) later added to these networks if they strongly interacted (co-localise) with original nodes. Various network measures were calculated, identifying important previously unknown regions. All of the important regions and genes they harbour were further explored using Gene Ontology enrichment analysis. A number of these genes were found to be enriched in biological processes with longevity relation. These processes included: 'ageing', 'immune response' and 'defence response', where genes enriched in these GO terms were found to reside in both novel and original regions of the networks, and included genes already known to have association with longevity, such as *Indy* and *chico*.

Regions were found in common between both extended GWAS-based networks when their clustering coefficient and PageRank scores were considered, where these regions included both original and novel nodes of the networks. Further exploration of these common regions found genes with longevity related phenotypes. These included the genes *Rim* and *Tpi* with 'long-lived' phenotypes, and *frtz*, *Atxn7*, *CG5339*, *CG4434* and *Zip99C* with 'short-lived'

phenotypes, as well as numerous genes found to have a phenotype of 'increased mortality'. It was also found that these genes are enriched in 'phagocytosis', 'processes in the tracheal system' and 'regulation of homeostatic processes'. Many of the newly found genes in this analysis were observed to harbour SNPs that did not reach the predefined genome-wide significance level, and therefore speculation was made that the SNPs residing within genes enriched in the same GO term may influence longevity collectively as opposed to a SNP causing a phenotype by itself.

A human ortholog search taken on genes found to reside in common regions showed several matches to human genes with functions related to the lifespan. The *Drosophila* gene *CG5886* was found to have the human ortholog, gene *TXLNA*, a gene related to the Innate Immune System pathway. The *Drosophila* gene *esg* had the human ortholog, human gene *SNAI2*, which is a protein with anti-apoptotic activity. The *Drosophila* gene *Ald* was found to have a human ortholog that related to the metabolic pathway, human gene *ALDOA*. The *Drosophila* gene *Slh* was found to have human ortholog *SCFD1* which is a gene that plays a role in protein modification. *Drosophila* genes *CG17770*, *CG9743*/*CG9747*, *CG1983*, and *Kek3* had human orthologs *CALML6*, *SCD*, *PLPBP* and *LRRC4C*, respectively, which are all genes known to be involved with diseases known to affect the lifespan of humans. Finally, the *Drosophila* gene *Sema-5c* was found to match the human gene *SEMA5B*, a gene known to be involved in the development of the nervous system.

Further analysis of subnetworks of the previously produced networks was also undertaken. The results showed that genes with no previous association with longevity, were found to be enriched in longevity-related GO terms. These subnetworks analysed harboured genes that enriched in GO terms including 'DNA repair', 'apoptotic process', 'developmental process', 'nervous system process' and 'immune system process'. Some of these enrichments included genes that have previously been found to associate with longevity, including *ku80*, *foxo*, *VhaSFD*, *CathD, ft, grim* and *Chmp1*. Genes in these subnetworks were observed to harbour both significant and non-significant SNPs, leading to speculation that SNPs residing in genes involved in the same biological processes may influence longevity either independently or cumulatively. Enhancers residing in regions of these subnetworks were also explored.

Chapter 5

SNPS IN NON-CODING REGIONS

In this chapter we explore SNPs occurring in non-coding regions of the *Drosophila* genome where a total of 26,499 non-coding SNPs were recorded in the Synthetic GWAS dataset and 653,030 non-coding SNPs in the DGRP GWAS dataset. These non-coding SNPs were first searched in regions including topologically associated domain (TAD) borders and transcription factor binding sites (TFBSs). Alterations of TAD borders and their structure are known to cause disruption to the way in which regulations occurs within these TAD regions on the genome, therefore in this study we hypothesised that SNPs residing in these border regions may cause disruption to the regulation that usually occurs within these TADs via looping interactions. Alterations to TFBSs may also cause disruption, by affecting the binding ability of transcription factors (TFs) which play a crucial role in controlling important processes in the genome. Whilst it is believed that TFs recognise a specific sequence pattern to which it binds, this may not be the case and it has been suggested that TFs, instead, recognise certain genomic structures. In this study we hypothesised that TFs may recognise a certain structure, e.g. non-B DNA structures, rather than sequence motifs.

First, we introduce topologically associated domains (TADs) and calculate the proportion of SNPs occurring within TADs' borders. Disruption to these borders as a result of residing SNPs may lead to interactions between TADs that would not usually occur because of the borders by which they are separated. To assess the significance of findings, a matched control dataset of non-coding SNPs was generated as described in section 2.6 and Chi-square test for proportion (Fisher's Exact Test) was used. In addition, non-coding SNPs found to reside in binding architectural proteins were also quantified and discussed.

We then explore the occurrence of SNPs in TFBSs, the sites to which TFs bind, by first considering DNA sequence patterns through the use of consensus sequence logos, then turning focus onto non-B DNA structures. These non-B DNA structures form when base pair sequence repeats occur in the genome. In this study we considered four types of repeats - direct, inverted, mirror repeats and G-quartets – described in section 5.3. SNPs in repeats

were quantified for all TFBSs recorded in a TFBS dataset, and then compared to a matched control dataset produced. This comparison to a matched control dataset allowed statistical analysis to determine if there was any significant difference in the number of SNPs in sequence repeats found in the binding sites recorded in the TFBS dataset.

Finally, we explored potential target genes for non-coding SNPs, taking a different approach to the usual assumption that these target genes are those that are nearest to each SNP on the linear genome. This is explored using intra-chromosomal Hi-C data with finer resolution and considering those non-coding SNPs which have highest interactions with regions not in the immediate vicinity of SNPs. These interacting regions are then selected for further analysis, focusing on the genes they harbour and the functions of these genes.

5.1 TOPOLOGICALLY ASSOCIATED DOMAINS (TADs)

Topologically associating domains (TADs) are genomic regions, in which DNA sequences residing within each TAD physically interact with each other more frequently than with DNA sequences in other TADs. TADs range from thousands to millions of DNA bases in length, varying between ~880 Kb and 1Mb in mammals and up to 60 Kb in *Drosophila*, and have been proven to play important roles in genome organization and gene regulation. TAD borders, when disrupted, have also been shown to lead to disease through disruption of gene regulation through changes in 3D organization of genomes (Lupiáñez et al., 2016).

TADs can be clearly visualised using heat maps (Figure 5.1), in which areas of dense red colour represent high frequency of interactions between regions on the genome, and areas of less dense red colour represent low frequency of interactions between regions on the genome. It is clear that these denser areas are found between regions that lie close to each other along the genome, with those regions further apart sharing fewer interactions. It is these areas of low frequencies that create the borders between different TADs. Figure 5.1 shows an example of a weak border, in which interactions are present between regions in different TADs, but not at high frequencies. A much stronger border is shown further along the genome, whereby there are fewer or zero interactions between regions across the TADs that this border separates. TADs can be further split into subTADs, in which triangular patterns of interactions formed are less obvious, but recognisable.

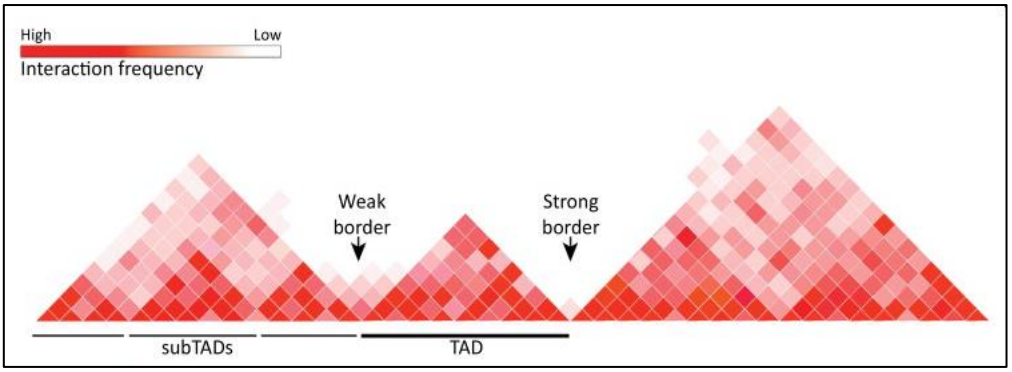


Figure 5.1 A diagram showing TADs, regions of the genome characterised by high frequency of local interactions, and their separation by borders whereby regions of the genome with low frequency of interaction (Gómez-Díaz and Corces, 2014).

5.1.1 Counting the Number of SNPs in TAD Border Regions

A total of 2846 TAD regions were recorded in Supplementary Data 1 (Ramírez et al., 2018), and after post-processing (see Section 2.5) a total of 2847 TAD border region positions were recorded. A matched control dataset was produced (described in Section 2.6) to create the exact same number of regions for comparative analysis. Using SNP position data, non-coding SNPs residing in each TAD border region in the real dataset were counted and non-coding SNPs residing in the matched control dataset for each TAD border region were counted. The total number of non-coding SNPs found to reside within all TAD borders was totalled separately for the real and matched control TAD dataset. This number of SNPs, for each dataset, was then subtracted from the total number of non-coding SNPs found across the whole genome. This created Table 4.16, on which Chi-Squared test was able to be carried out, to test for any differences in proportions between the number of non-coding SNPs observed in the TAD borders in the real dataset and the matched control dataset. Table 5.1 shows counts for non-coding SNPs found only in the DGRP GWAS dataset. Similar calculations were performed for the Synthetic GWAS dataset, but the totalled values were so low that the results were not considered for further analysis.

Table 5.1 Number of non-coding SNPs residing in TAD border regions and outside TAD borders in real and control datasets.

	Real dataset	Control dataset
Non-coding SNPs residing in TAD borders	11982	9321.24
Non-coding SNPs not residing in TAD borders	641048	643708.76

Fisher's Exact Test shows that the proportion of non-coding SNPs residing within TAD borders in the real dataset is significantly ($P = 1.0376 \times 10^{-75}$) higher than in the control dataset. This higher than expected count of SNPs in TAD border regions can be speculated to cause disruption in regulation pattern of gene/genes by forming looping interactions with distal enhancer, and as discussed previously such disruption can affect processes such as the expression of nearby genes, as well as enable more long-range interactions to occur across the genome, allowing regions that wouldn't usually come into contact to interact.

Alternatively, these observations could result in speculation that reduced interaction frequency at TAD borders may predict neutrality of SNPs accumulating where they are less likely to be injurious to the carrier. This speculation would contradict our hypothesis previously made but, of course, could be plausible. Table 1.1 (Section 1.8) shows examples of genes which have an effect on longevity when overexpressed; the overexpression of these genes could occur due to mutations in TAD regions that reside near to said genes.

5.2 SNPS IN TRANSCRIPTION FACTOR BINDING SITES

5.2.1 Transcription Factor Binding Sites

Transcription factors (TFs) are proteins in the genome that are crucial for controlling many important processes. A TF recognises and binds to a specific site on the DNA known as a transcription factor binding site (TFBS). SNPs could disrupt the interactions between these TF proteins and the DNA (TFBSs) and any defects in these interactions can contribute to the occurrence of various diseases, or in the case for our study, longevity. For this reason, it is important to look at the SNPs that occur within TFBSs.

It is known that TFs interact with DNA in two primary ways: sequence-specific interactions with the genomic bases and interactions with the back bone of the DNA structure known as a non-sequence-specific interaction. The role of DNA shape as a determinant of protein-DNA binding specificity has previously been shown to be important (Rohs et al., 2009; Rohs et al., 2010; Parker & Tullius, 2011). Many sequence-specific interactions occur via what is known as the major groove of DNA, as illustrated in Figure 5.2 for B-DNA, major grooves occur when the backbones of DNA are far apart. Such interactions are commonly linked to direct hydrogen bonding between specific DNA bases and DNA-binding domain amino acids (Garvie and Wolberger, 2001). Opposite the major groove, also illustrated in Figure 5.2, is the minor groove, showing a shorter distance measured between the two backbone strands of DNA. Both grooves run continuously along the entire DNA molecule and are due to the antiparallel arrangement of the backbone strands, however these grooves are not just consequence of the way in which the backbones align but are an actual structural feature.

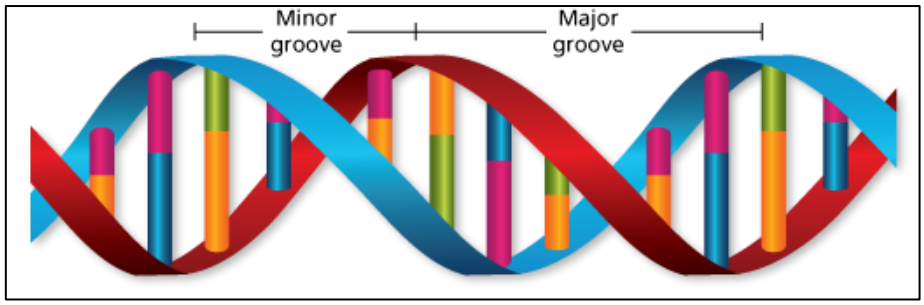


Figure 5.2 The major and minor groove labelled in a segment of DNA (https://www.scalarlight.com/articles/?p=454).

The role of numerous DNA shapes in binding sites on the genome recognised by TFs has been demonstrated (Slattery et al., 2011; Dror et al., 2014; Gordan et al., 2013). Structures such as the minor groove in human DNA influencing the way in which interactions between TFs and TFBSs occur along the genome, include a type of interaction, which is dependent on the shape of the minor groove in DNA, also conferring partial sequence specificity. Such DNA structure-based-binding motif occurs in cases whereby the binding sequence of a TF is flanked by a stretch of either A or T bases and is very common in different TF families (Jolma et al., 2013). Interactions such as these may be important in regulatory elements, in relation to the formation of consecutive TFBSs. This recognition of DNA is based on the shape of DNA, therefore the base preferences of TFs in these regions flanked by A and T bases may also be affected by any shape changes induced in cases where multiple TFs bind in close proximity to each other (Jolma et al., 2013). Findings of interactions occurring between TFs and DNA, as a result of the recognition of DNA structures including the major and minor groove, has sparked interest in further exploration of the other structures that form in TFBSs along the DNA, and under what circumstances these occur. DNA shapes can also explain why sequences flanking TFBSs are important to consider when exploring binding specificity of TFs (Gordan et al., 2013).

5.2.2 Creation of Consensus Sequence Logos

Consensus sequence logos were produced with the aim to observe whether experimentally validated TFBSs have common consensus sequences. The extended sequences of experimentally validated TFBSs (see section 2.4) were aligned using Clustal Omega, a multiple sequence alignment program (https://www.ebi.ac.uk/Tools/msa/clustalo/) to allow for analysis of consensus sequences using consensus sequence logos. Such logos are a graphical representation, in this case of a DNA multiple sequence alignment, in which each position of a sequence (x-axis) is represented by a stack on the graph, formed using symbols of the nucleotides (A, T, C and G). The overall height of each stack indicates the sequence conservation at that position, and the height of each symbol in this stack indicates the frequency in which this nucleic acid occurs on this position. This analysis allowed for observation of any similarities between the positions of nucleic bases in sequences for the same TF.

WebLogo (Crooks et al., 2004; Schneider and Stephens, 1990) was used to generate sequence logos for each TF, an example of the logos produced for the binding sites of the TF *Adb-A* is shown in Figure 5.3. The addition of 50 bp to each side of the original sequences increased sequence lengths significantly; therefore, the sequence alignments for this TFBS were generated as five separate alignments, resulting in five different consensus sequence logos being produced.

5.2.3 Analysis of Consensus Sequence Logos

For all 2209 TFBS sequences recognised by 192 transcription factors, the consensus sequence logos produced for each TF did not show any highly conserved regions in which specific nucleotide bases were enriched at the same positions on sequences. An example of one consensus sequence logo produced for the TFBS sequences recorded for the TF *Adb-A* is shown in Figure 5.3. In this example there are no interesting observations, as in most positions of the alignment three or more different bases were observed. Low conservation of the sequences is also apparent in several areas of the logos where there is no stack for an x-axis value, or where stacks are short. It is also only in these shorter stacks where, in some cases,

just one nucleic base is recorded for this alignment position. Across all consensus sequence logos for all TFBSs analysed, no consensus sequences were found. These observations were the motivation for the exploration of structure recognition in TFBSs, and the reason for the approach taken for analyses described in this chapter.

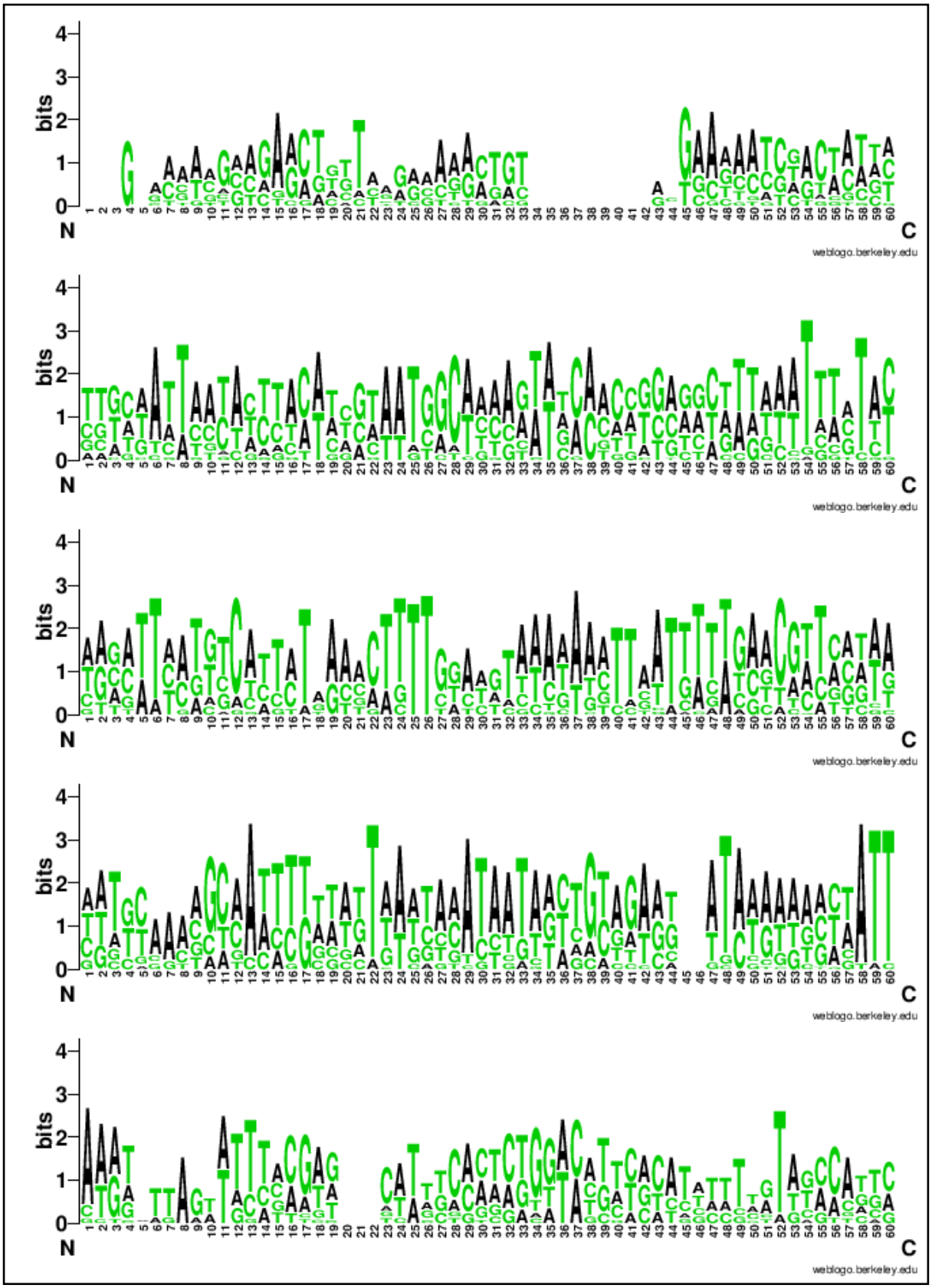


Figure 5.3 Consensus sequence logos for the TFBSs recorded for the TF *Adb-A* extended by ±50 bp.

5.3 THE ROLE OF DNA SHAPE IN TF-TFBS BINDING SPECIFICITY

Normal cell functions in many biological systems are highly dependent on the principle genetic molecule, DNA. As has been discussed previously in this thesis, the role of DNA is constantly challenged by mutations in the genome, ranging from mutations caused by environmental factors to those caused by chemical reactions. The majority of DNA in the genome is in the B conformation, in which sequences adopt the orthodox right-handed B form described as the Watson-Crick (W-C) Model of DNA (Watson and Crick, 1993). However, previous studies have demonstrated that non-B DNA conformations, of which there are at least ten, have been found to adopt at specific naturally occurring repeat sequences and are also known to be mutagenic (Mirkin, 2007; Wells et al., 2005).

Non-B DNA conformations have been found to form at specific sequence motifs; these structures include slipped structures (direct repeats), cruciform (inverted repeats), triplexes (mirrored repeats) and tetraplexes (G-quartets). These conformations, *in vivo*, are believed to form in higher energy states during metabolic processes in DNA such as transcription, replication and repair (Wang and Vasquez, 2014). All non-B DNA conformations are found to contain contorted bond angles or nucleotides which are unpaired, as is shown in Figures 5.4–5.11 below. The formation of such non-B DNA conformations requires that direct, inverted and mirrored repeats are greater or equal to 5 bp in length, and each repeated sequence is no more than 20 bp apart from each other (Ball et al., 2005). The formation of tetraplex structures require four runs of guanine bases, in which each is the same length of either 2, 3 or 4 bp with each being separated by 1-7 bp (Rouleau et al., 2014).

Direct Repeats – Slipped structure

Slipped structures are found to rely in part or exclusively on non-W-C interactions. Such structures form when complementary stands, containing direct repeats, pair together in a slipped fashion. An example of a direct repeat is shown in Figure 5.4.

Figure 5.4 Example of a direct repeat in DNA strands.

An example of this structure is shown in Figure 5.5, where the region containing the direct repeat can be seen to unwind and its complementary strand then pairs with the second direct repeat further down the sequence. Thus far, slipped structures have only been detected in DNA containing short nucleotide repeats.

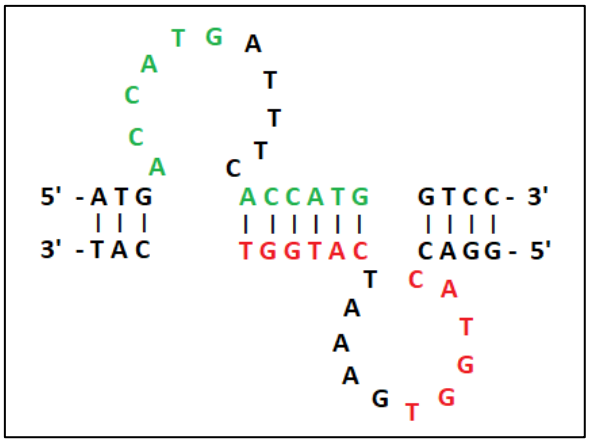


Figure 5.5 Slipped structure, corresponding to direct repeats.

Inverted Repeats – Cruciform structure

Cruciform structures form in regions where inverted repeats occur, such repeats are defined as reverse complement to each other on a single stranded DNA. An example of an inverted repeat is shown in Figure 5.6.

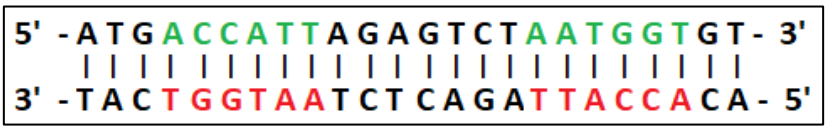


Figure 5.6 Example of an inverted repeat in DNA strands.

This structure is a non-B DNA structure that contains base pairs in the W-C conformation and, as shown in Figure 5.7, contains strands folding at the centre of symmetry at the inverted repeat where the organization of the strand then forms an intramolecular B-helix capped by a single-stranded hoop which can range from a few bp to several Kb in length (Bacolla and Wells, 2004).

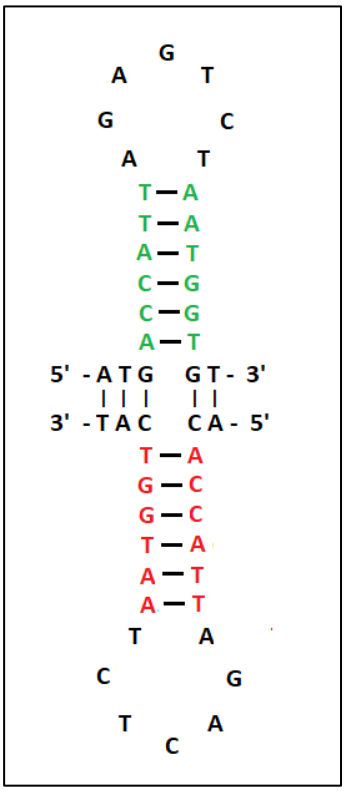


Figure 5.7 Secondary structure of a cruciform, corresponding to inverted repeats.

Symmetric Repeats – Triplex structure

Triplex DNA structures form in regions in which mirror symmetric repeats occur, of which an example is shown in Figure 5.8.

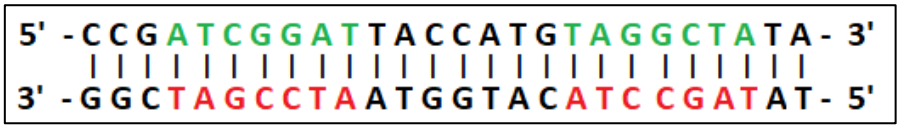
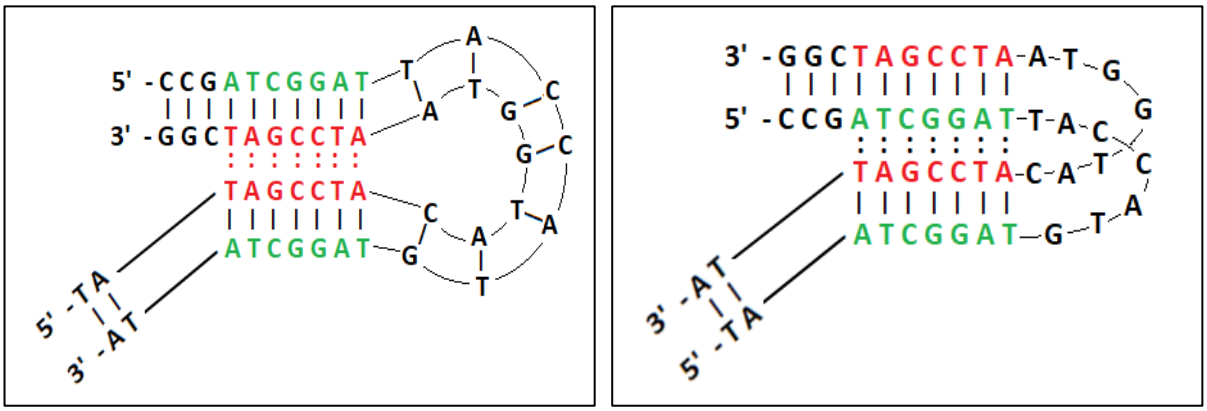


Figure 5.8 Example of a mirrored repeat in DNA strands.

A triplex structure is formed when the major groove of the DNA double helix is occupied by pyrimidine or purine bases on a single-stranded region, in turn forming a three-stranded helix. Triplex DNA is able to be classified dependant on the positioning and structure of this third helix strand, and whether it forms either Hoogsteen or reverse-Hoogsteen hydrogen bonds with the purine-rich strand of the duplex DNA. Such DNA is known as an R·R-Y triplex (Figure 5.9a) in cases where this third strand is purine-rich and anti-parallel to the complementary strand. In cases of a pyrimidine-rich third stand parallel to the complementary strand, this is known as a Y·R-Y triplex (Figure 5.9b). The conditions under which both triplexes form are also different, R·R-Y triplexes form under conditions of physiological pH whereas Y·R-Y triplexes are found to form under conditions of acidic pH.



(a) (b)

Figure 5.9 Secondary structure of a triplex corresponding to mirrored repeats. (a) R·R-Y triplex and (b) Y·R-Y triplex.

G-Quartets Repeats – Tetraplex structure

A tetraplex structure, also known as a guanine tetrad, is formed in regions of the genome where there are four repeats of sequences containing only the guanine base (known as G-quartets); an example is shown in Figure 5.10 in which there are four runs of guanine bases with a length of 3 bp.

5' - G G G T T A G G G C G C G G G T A G G G - 3'

Figure 5.10 Example of G-quartets in a DNA strand.

This structure is a square coplanar array of four guanines, which is formed of guanine tetrads stacked on top of each other, as shown in Figure 5.11. The number of guanine bases per repeat is important in terms of stack height, for which repeats of three or more guanine bases are favoured.

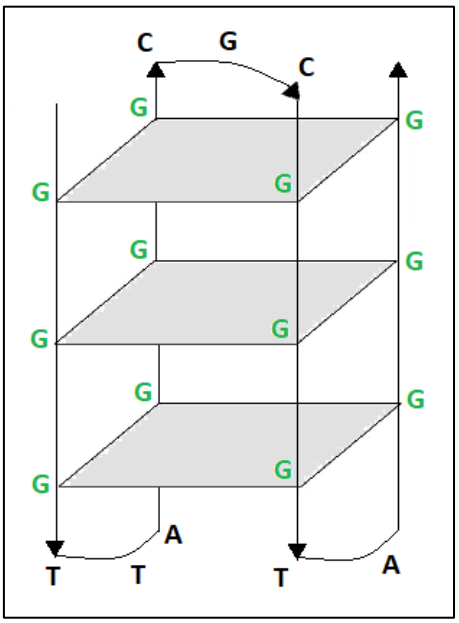


Figure 5.11 Secondary structure of a tetraplex, corresponding to G-quartets.

Code was written and implemented in Matlab for the identification of these repeats discussed in section 5.3.

5.3.1 Non-B DNA Conformation of Transcription Factor Binding Sites

The TFBS dataset previously used for the creation of consensus sequences was then sought for the occurrence of any non-B DNA conformations through a search for sequence repeats described in section 5.3. This search was done for all 2209 TFBSs recorded, for which direct repeats, inverted repeats and mirrored repeats of length 5 bp and 6 bp and separated by ≤ 20 bp were quantified individually, and G-quartets of length 2, 3 and 4 bp and separated by ≤ 7 bp were recorded.

For each TFBS sequence, the presence or absence of repeats was recorded. A summary of occurrence of repeats in TFBSs for a selected number of TFs is given in Table 5.2. A total of 192 TFs were recorded, not including those unspecified. This table included TFs for which different numbers of TFBSs were recorded, those of most interest were the ones with the highest number recorded. Therefore, of these 192 TFs, those with ≥ 10 binding sites recorded were selected and used as our real dataset. This resulted in a real TFBS dataset in which 49 TFs and their 1276 binding sites were recorded. For each of these binding sites, the number of SNPs residing within TFBSs were counted. An average SNP count for all sequences recorded under each of these 49 TFs was calculated, and the TFs with the binding sites containing the highest average number of SNPs are shown in Table 5.3 (see Appendix Table S5.1 for full list).

Table 5.2 A fragment of results summarising the total number of repeats found across TFBS sequences recorded for a given number of TFs, and the total number of sequences recorded for each of the TFs.

TF name	Number of direct repeats (5 bp)	Number of inverted repeats (5 bp)	Number of mirrored repeats (5 bp)	Number of G-quartets (2 bp)	Total number of TFBS recorded sequences
Abd-A	38	38	40	0	43
Antp	16	13	12	1	16
aop	7	8	6	1	9
ap	10	13	13	0	14
bcd	50	42	43	9	55
br	4	4	4	3	4
brk	19	12	18	4	21
cad	11	13	10	0	13

Table 5.3 TFs with TFBS sequences harbouring the highest average number of SNPs.

TF name	Total number of SNPs counted across all TFBS sequences recorded	Average number of SNPs per sequence
ap	69	4.928571429
vvl	54	3.857142857
sd	67	3.045454545
sna	37	2.916666667
ey	43	2.866666667
Dref	80	2.857142857
kni	111	2.846153846
fkh	30	2.727272727
pnr	38	2.714285714
bcd	143	2.6
pho	26	2.6
grh	31	2.583333333
Br-Z3	43	2.529411765
srp	61	2.44
usp	26	2.363636364

Overrepresentation of non-B DNA forming repeats in TFBSs

To assess the significance of findings, a matched control dataset was produced for the real TFBS dataset. Before such sequences were selected, all positions on which genes or TFBSs were already known to reside were masked from the genome, meaning that the control sequences could not overlap these positions. 100 matching control datasets were created as described in Section 2.6. Repeat counts were taken for each of the control sequences, creating 100 tables similar to that of Table 5.2. All counts for each repeat were then averaged across the 100 controls, creating one table used as the matched control dataset to carry out statistical analysis and compare with our original TFBS table.

The Chi-Square test for proportion was used, in which P-values were calculated to find if there was any significant differences in proportion of each sequence repeat between those sequences from the original TFBS dataset and those from the matched control dataset produced. An example is shown in Table 5.4 for direct repeats of base pair length five in sequences recorded for the TF *Br-Z2*, for which a total of 22 sequences were recorded in the original TFBS dataset, in which 20 of these sequences contained direct repeats, and two sequences did not. In the matched control dataset, an average of 11.82 sequences were found to contain this direct repeat of base pair length five, with an average of 10.18 sequences found not to contain these repeats.

Table 5.4 An example of frequency data used for Chi-Squared analysis, for direct repeats with a bp-length of five in the TF Br-Z2.

	Real dataset	Control dataset
Contains direct repeat 5 bp	20	11.82
Does not contain direct repeat 5 bp	2	10.18

Chi-Squared tests were calculated using SPSS software, selecting chi-squared testing under cross tabulation in the Descriptive Statistics tab. P-values (one-sided) were calculated for each sequence repeat (direct, inverted and mirrored) for all TFBS sequences recorded under each TF name, for lengths 5 bp and 6 bp. Results that showed to be significant are summarised in Table 5.5.

Table 5.5 TFs for which significant enrichment in sequence repeats in TFBSs were identified.

TF name	TFBSs enriched in	P-value (corrected for multiple testing*)	Number of TFBSs used for analysis	DGRP GWAS dataset SNP count for TF	Synthetic GWAS dataset SNP count for TF	Lowest P-value recorded	Highest D value recorded
ap	Direct (6 bp)	0.036	14	241	9	0.005105	4.295596
	Mirrored (6 bp)	0.044	14				
Br-Z2	Direct (5 bp)	0.014	22	877	24	0.005082	3.746626
	Mirrored (5 bp)	0.002	22				
	Direct (6 bp)	0.032	22				
	Mirrored (6 bp)	0.014	22				
EcR	Mirrored (6 bp)	0.008	19	743	37	0.000285	5.821549
fkh	Direct (5 bp)	0.022	11	45	2	0.01481	4.959279
hb	Direct (5 bp)	0.044	95	47	3	0.01698	3.813478
Mad	Direct (6 bp)	0.046	64	387	13	0.026691	5.908299
tin	Inverted (6 bp)	0.018	24	42	1	0.1088	3.547556
vnd	Inverted (5 bp)	0.026	13	106	6	0.03056	2.807962
	Mirrored (5 bp)	0.06	13				
zen	Direct (6 bp)	0.044	23	7	2	0.146	4.541991

*For correction for multiple testing Bonferroni correction was used.

In Table 5.5, all TFs were found to have a significantly higher frequency of the stated repeats in their binding sequences as compared to the control dataset. This observation confirms our hypothesis that the non-B DNA structures formed by the sequence repeats in these TFBSs recorded, may be what is recognised by these TFs as opposed to them binding due to sequence recognition. SNPs were also counted for each of these TFs recorded, where several TFs such as *ap*, *Br-Z2*, *EcR*, *Mad* and *vnd* were found to be highly mutated in the DGRP GWAS dataset, however no SNPs were found to be significant. Although no SNPs were found to be significant, the number of SNPs residing in these TFs can be speculated to cause a mutational effect through accumulation of these SNPs, which could affect the binding affinity of these

TFs. The TF *EcR* was previously found and discussed in results in section 4.5.2, under the enrichment of the GO term 'ageing'. No other TFs in Table 5.5 have previously been found to show any association in relation to longevity.

5.3.2 Overrepresentation of non-coding SNPs in non-B DNA forming TFBSs

For the TFBSs found to contain repeats with significant differences to their matched control data, SNPs were counted for each of the sequences to which they bind. SNPs were also counted in each of the 100 controls, and then divided for an average to be compared with. The SNP counts were then grouped and averaged for each TF, for which the results for these SNP counts are shown in Table 5.6. Results for SNP counts in TFBSs for TFs in both the real data and control data were found to be similar, with only the *ap* TF showing any real difference between these counts.

Table 5.6 Average number of SNPs for the TFBS sequences in the real dataset and the control dataset, under each of the TFs with significant P-values.

TF name	**Average DGRP GWAS dataset SNP count for real sequences recorded under TF**	**Average DGRP GWAS dataset SNP count for control sequences recorded under TF**
ap	4.928571429	1.326428571
Br-Z2	2.045454545	1.798636364
EcR	2.071428571	1.544285714
fkh	2.727272727	1.4
hb	2.136842105	1.554210526
Mad	1.5	1.5728125
tin	1.666666667	1.75125
vnd	0.615384615	1.287692308
zen	2.347826087	1.885652174

5.4 SNPS IN ARCHITECTURAL PROTEINS

Architectural proteins, also referred to as insulator proteins, appear to play a vital role in the three-dimensional organisation of the genome (Cubeñas-Potts & Corces, 2015). Architectural proteins have the ability to facilitate the formation of long-range contacts between DNA sequences, which makes sense as TAD borders have been observed to be enriched in binding sites for architectural proteins, as well as actively transcribed genes (Hou et al., 2012). Eleven different DNA binding architectural proteins, for which positions and SNP counts are shown in Table 5.7, have been known in *Drosophila*, with each recognizing a unique DNA motif. All eleven architectural proteins were observed to harbour SNPs in the DGRP GWAS dataset, and although their P-values were not significant, could potentially cause disruption to not only their own function of facilitating interactions, but also the functions of the TAD border in which their binding site resides in. Such disruption could cause physical interactions to occur more frequently between genomic regions in different TADs, changing 3D organization of the genome and as a result affect gene regulation which could influence longevity.

Table 5.7 Number of SNPs residing in architectural proteins in *Drosophila*.

Architectural Protein	Reference	Chromosome position	DGRP GWAS dataset SNP count	Synthetic GWAS dataset SNP count	Minimum SNP P-value
CCCTC-binding factor (*CTCF*)	Bushey et al., (2009)	chr3L:7346678-7349813	20	0	0.05816
Suppressor of Hairy-wing (*Su(Hw)*)	Gurudatta et al., (2013)	chr3R:10130177-10134311	117	5	0.000597
Boundary Element Associated Factor 23 (*BEAF-32*)	Gurudatta et al., (2013), Yang et al. (2012)	chr2R:10657974-10660135	39	3	0.1993
DNA Replication Related Element binding Factor (*DREF*)	Gurudatta et al., (2013)	chr2L:9964147-9967229	61	2	0.1072
Transcription Factor IIIIC (*TFIIIC*)	Van Bortle K et al., (2014)	Gene has not been mapped to the genome sequence			
putzig (*pzg*)	Eggert et al., (2004)	chr3L:21279199-21283410	22	0	0.09363
Early Boundary Activity DNA-binding Factor (*Elba*) (see below)	Aoki et al., (2012)				
Elba2		chr2L:2577687-2579179	25	0	0.0411
Elba 3		chr2L:4687511-4688956	32	1	0.01376
Pita	Maksimenko et al., (2015)	chr2R:19436490-19442235	87	0	0.001291
Zinc Finger Interacting with CP190 (*ZIPIC*)	Maksimenko et al., (2015)	chr3R:25860962-25862592	16	1	0.04711
Insulating binding factor 1 (*Ibf1*)	Cuartero et al., (2014)	chr3R:5084686-5086122	10	0	0.005227
Insulating binding factor 2 (*Ibf2*)	Cuartero et al., (2014)	chr3R:5083449-5084223	6	0	0.5573

5.5 TARGET GENES FOR NON-CODING SNPS

Longevity-associated nucleotide changes in non-coding regions of the genome, both in *Drosophila* and humans is an area of research yet to be explored in depth. However, previous GWAS studies identified that disease-associated nucleotide changes are often found in non-coding regions of the human genome (Freedman et al., 2011; Blattler et al., 2014), in many cases corresponding to promoters or enhancers (Yao et al., 2014). Many studies have been carried out, comprising a range of analytical and experimental steps to find the way in which a non-coding SNP can be associated with an increased risk of a specific disease. Some of these approaches focus on the identification of target genes of non-coding SNPs, which has been found in many cases to be a difficult task. It is often assumed that a SNP found to reside within a gene may influence the expression/function of that particular gene, and similarly a non-coding SNP may target the gene residing closest to the promotor or enhancer on which the SNP resides. In many cases, if this corresponding gene carries out biological functions which are related to the disease being studied, this gene is automatically of great interest to the researcher. However, due to the observation of looping interactions between different promoters (Sahlen et al., 2015) and the postulation that promoters can have enhancer activity which may influence the expression of other genes (Andersson et al., 2015), it is not possible to say for sure that SNPs are only involved in regulation of the nearest genes.

To identify potential target genes Hi-C data with finer resolution, of 10,000 bp regions, was utilised by assigning each non-coding SNP, separately for the Synthetic GWAS and DGRP GWAS dataset, to a 10 Kb bin. This then allowed for the identification of the regions along the genome with which the region harbouring this non-coding SNP interacts most strongly. This was done to enable observation of regions harbouring non-coding SNPs that have strong long-range interactions with regions that may not necessarily be in close proximity, to prove the hypothesis that SNPs in non-coding regions may in fact target genes which are not nearest in distance. For each 10 Kb bin harbouring non-coding SNPs, the top three interacting bins for each was selected and the distance between them was calculated. For example, if bin 25 corresponding to regions Chr2L:240,000-250,000, was found to harbour SNPs and bins with the most interactions were bins 24, 27 and 28, the three distances calculated would be 10,000 bp, 20,000 bp and 30,000 bp. For all regions harbouring non-coding SNPs, the distances

calculated between them and their first, second and third strongest interacting regions were grouped separately and the number of times each distance was recorded was quantified. Histograms of distances are shown in Figures 5.12 and 5.13. These histograms proved what was to be expected in most cases, i.e. the regions being most highly interactive are either those 10 Kb regions interacting with themselves (distance equal to 0) or interacting with adjacent regions.

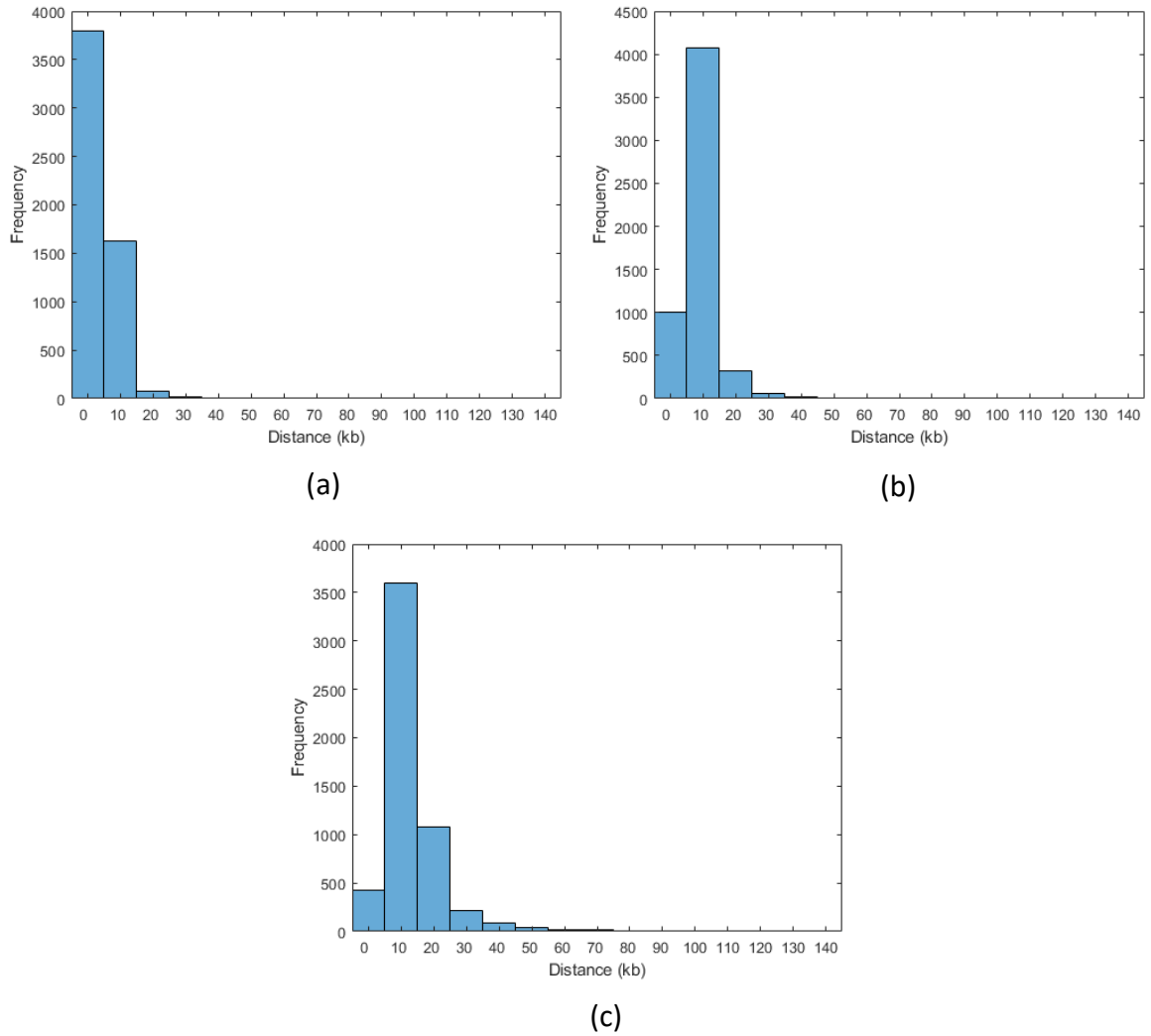


Figure 5.12 Histograms of interactions depending on the distances between regions containing non-coding SNPs in the Synthetic GWAS dataset and the (a) first highest interacting region, (b) second highest interacting region and (c) third highest interacting region with regions containing non-coding SNPs in the Synthetic GWAS dataset.

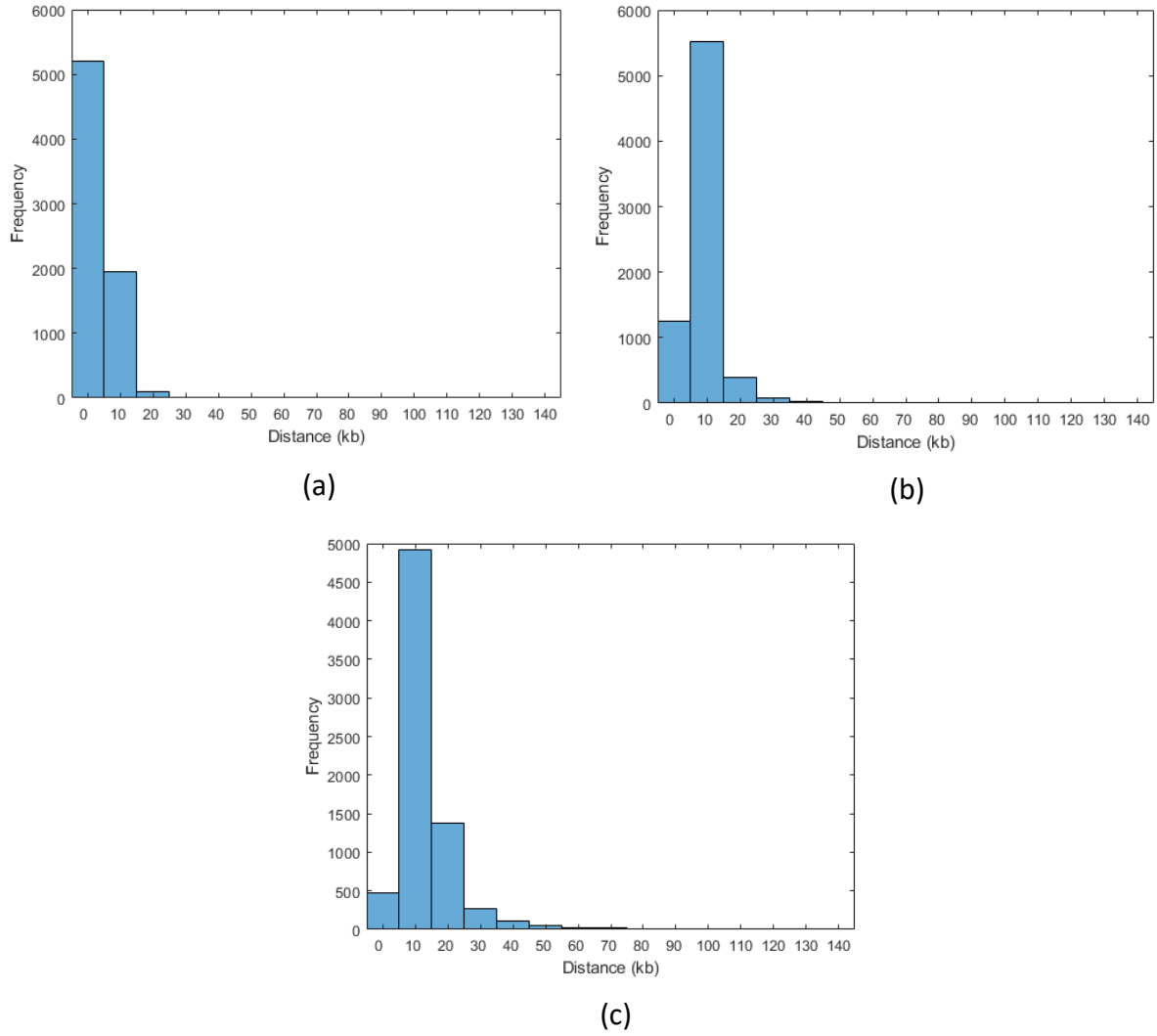


Figure 5.13 Histograms of interactions depending on the distances between regions containing non-coding SNPs in the DGRP GWAS dataset and the (a) first highest interacting region, (b) second highest interacting region and (c) third highest interacting region with regions containing non-coding SNPs in the DGRP GWAS dataset.

For all 10 Kb bins in which non-coding SNPs resided, in both the Synthetic GWAS dataset and DGRP dataset, the bin with which it had the strongest recorded interaction frequency was selected. For many of these bins, as shown in Figures 5.12 and 5.13, the strongest interacting region was itself or adjacent. However, there were regions containing non-coding SNPs for which the strongest interacting regions was as distant as 50,000 bp in the Synthetic GWAS dataset and 100,000 bp in the DGRP GWAS dataset. The top 30 regions containing non-coding SNPs with their most strongly interacting regions being long-range were selected for both datasets. For each of these regions, the 10 Kb bins containing non-coding SNPs are recorded

in Tables 5.8 and 5.9, along with their SNP counts, strongest interacting region and distance between interacting regions. Table 5.8 also includes the highest SNP D values and Table 5.9 includes the highest and lowest SNP-values recorded in each SNP containing region. The D values and P-values of non-coding SNPs residing in the strongly interacting regions were also recorded, the highest and lowest values are shown in the same tables, next to each of their corresponding regions.

For all non-coding SNPs residing in regions selected for the Synthetic GWAS dataset, many SNPs did not have calculated D values that reached the genome-wide level of significance of $\geq$7.9 used previously. However, one region, bin 2006, contained non-coding SNPs with a highest recorded D value of 12.009, for which the strongest interacting region was 30,000 base pairs upstream. Non-coding SNPs residing in regions selected for the DGRP GWAS dataset were observed to have what would be considered a low P-value, however after correcting the commonly used significance level value of 0.05 by taking into account the number of 10 Kb regions recorded in this Hi-C data, 0.05/11839 = 4.22×10^{-6}, no SNPs were found to reach this level of significance.

Genes residing in all regions strongly interacting with bins containing non-coding SNPs, recorded in Tables 5.8 and 5.9, were selected for further analysis. A total of 73 genes were found in the 30 regions selected for the Synthetic GWAS dataset, and 59 genes were found in the 30 regions selected for the DGRP GWAS dataset. No genes were found to reside in several of these bins across the findings for both the Synthetic and DGRP GWAS dataset. These bins included 1165, 4366, 7816 and 7887, which correspond to regions 2L: 11,680,000-11,690,000, 3L: 380,000-390,000, 3R: 10,760,000-10,770,000 and 11,470,000-11,480,000, respectively.

Table 5.8 Summary of regions containing non-coding SNPs from the Synthetic GWAS dataset and the distance to regions with which they have the strongest interaction.

Non-coding SNP region			**Interacting region**			
Bin number	**Number of SNPs residing**	**Highest recorded SNP D value**	**Strongest interacting bin**	**Distance (Kb)**	**Number of SNPs residing**	**Highest recorded SNP D value**
2367	3	2.727	2362	50	2	0
3982	7	5.421	3987	50	16	6.984
5470	9	5.765	5465	50	13	7.528
2459	4	5.170	2463	40	6	4.538
4370	12	7.189	4366	40	10	6.499
4560	4	3.814	4566	40	10	3.580
5656	9	6.284	5660	40	16	6.943
11643	2	5.029	11647	40	6	4.721
1605	7	7.657	1608	30	9	8.794
2006	3	12.009	2003	30	7	11.435
2461	1	4.592	2464	30	6	3.839
3146	13	3.468	3149	30	13	2.270
3447	1	2.160	3444	30	10	5.985
4837	8	6.213	4840	30	15	5.733
5702	1	3.582	5705	30	8	6.758
6981	6	4.577	6978	30	4	4.689
7819	4	3.189	7816	30	6	4.767
7883	6	5.029	7886	30	11	6.846
7884	4	5.989	7887	30	12	7.530
8958	11	4.396	8995	30	11	5.912
155	14	7.658	153	20	6	6.028
229	7	7.550	227	20	10	6.394
324	3	5.987	326	20	15	6.438
556	11	6.606	554	20	10	5.888
592	1	2.066	590	20	12	4.761
945	6	4.168	943	20	10	5.766
994	5	5.459	996	20	3	3.321
1167	10	6.455	1165	20	13	7.420
1173	7	6.222	1171	20	10	4.701
1174	11	5.694	1176	20	9	5.360

Table 5.9 Summary of regions containing non-coding SNPs from the DGRP GWAS dataset and the regions with which they have the strongest interaction.

Non-coding SNP region			**Interacting region**			
Bin number	**Number of SNPs residing**	**Lowest recorded SNP P-value**	**Strongest interacting bin**	**Distance (Kb)**	**Number of SNPs residing**	**Lowest recorded SNP D value**
2388	8	0.3858	2378	100	168	0.008851
3644	26	0.1303	3637	70	190	0.002221
2458	34	0.06431	2464	60	171	0.05381
2367	53	0.05865	2362	50	74	0.03473
2957	16	0.06397	2962	50	228	0.009989
3982	256	0.007077	3987	50	205	0.01519
5470	280	0.003503	5465	50	420	0.0122
2457	28	0.08089	2453	40	81	0.0515
2459	73	0.5731	2463	40	120	0.02168
4370	297	0.008632	4366	40	192	0.01473
4560	88	0.02264	4566	40	280	0.009818
5656	222	0.002052	5660	40	307	0.01401
11643	51	0.05711	11647	40	138	0.005184
1605	95	0.0236	1608	30	230	0.001396
2006	64	0.01037	2003	30	153	0.000732
2124	14	0.0354	2121	30	22	0.02479
2283	17	0.1617	2286	30	62	0.0776
2372	68	0.01846	2375	30	128	0.06902
2456	44	0.01432	2453	30	81	0.0515
2461	32	0.04341	2464	30	171	0.05381
3146	377	0.006937	3149	30	295	0.00241
3447	44	0.13	3444	30	285	0.008249
4837	181	0.007141	4840	30	359	0.02119
5702	27	0.01703	5705	30	322	0.006609
6981	87	0.00301	6978	30	83	0.000805
7639	11	0.09939	7636	30	126	0.003609
7819	238	0.001871	7816	30	203	0.003442
7883	108	0.007967	7886	30	234	0.001484
7884	120	0.006706	7887	30	255	0.002598
8958	222	0.001062	8955	30	197	0.01518

Further analysis of the potential target genes found in the long-range interacting regions was done using phenotype data, in which each gene was analysed individually. Lists of genes found to have phenotypes which could be associated with longevity, for both the Synthetic and DGRP GWAS datasets, including ‘increased mortality’, ‘lethal’ and ‘immune response defective’ are shown in Tables 5.10 and 5.11. For the interacting bins containing more than

one gene with longevity related phenotypes, we can speculate that a single enhancer that harbours non-coding SNPs may target several genes found in the same region, influencing their expressions and phenotypes.

Table 5.10 Summary table of phenotypes of genes found in regions most strongly interacting with regions containing non-coding SNPs found in Synthetic GWAS dataset.

Bin harbouring SNP(s)	**Possible target gene**	**Longevity related phenotypes**
4560	*CG45186*	lethal; increased mortality during development; increased mortality
	CG32298	partially lethal - majority die; flightless
5656	*SNCF*	lethal - all die during P-stage
	CG14107	partially lethal - majority die; some die during pupal stage; lethal - all die during P-stage
1605	***Ca-alpha1D****	increased mortality during development; lethal - all die before end of P-stage
2461	*jing*	locomotor behaviour defective; cell death defective
3146	*AttC*	partially lethal; some die during pupal stage; neuroanatomy defective
4837	*CG4597*	some die during pupal stage; partially lethal - majority die
	CG4611	lethal - all die during P-stage
5702	*Hml*	immune response defective
7883	*CG43335*	partially lethal - majority die; some die during pupal stage; partially lethal
556	*GluRIIA*	locomotor behaviour defective; neurophysiology defective; neuroanatomy defective; lethal
	GluRIIB	neuroanatomy defective; neurophysiology defective
945	*numb*	decreased cell number; some die during embryonic stage; increased mortality; increased cell number; lethal - all die before end of prepupal stage; flight defective; tumorigenic
994	*bib*	lethal - all die before end of pupal stage
1174	***crol***	locomotor behaviour defective; increased occurrence of cell division; increased mortality; cell death defective

* Genes previously found to have association with longevity as recorded in FlyBase or GenAge resources are shown in bold.

Table 5.11 Summary table of phenotypes of genes found in regions most strongly interacting with regions containing non-coding SNPs found in DRGP GWAS dataset.

Bin harbouring SNP(s)	**Possible target gene**	**Phenotypes**
2461	*jing*	locomotor behaviour defective; cell death defective
2957	*en*	lethal - all die during embryonic stage; size defective; planar polarity defective; increased cell death; some die during pupal stage; partially lethal - majority die
2457	*Pld*	developmental rate defective; partially lethal - majority die; some die during embryonic stage; neurophysiology defective; lethal - all die before end of embryonic stage
2459	*jing*	locomotor behaviour defective; cell death defective
4370	*trh*	neuroanatomy defective; partially lethal - majority die; lethal; some die during embryonic stage; lethal - all die before end of embryonic stage
4560	*CG45186*	lethal; increased mortality during development; increased mortality
	CG32298	some die during pupal stage; partially lethal - majority die; flightless
5656	*SNCF*	lethal - all die during P-stage
	CG14107	partially lethal - majority die; some die during pupal stage; lethal - all die during P-stage
1605	***Ca-alpha1D****	increased mortality during development; lethal - all die before end of P-stage
2283	*RpL38*	increased mortality; increased mortality during development; developmental rate defective
2372	*laccase2*	lethal; partially lethal; lethal - all die during embryonic stage;
2456	*Pld*	developmental rate defective; partially lethal - majority die; some die during embryonic stage; neurophysiology defective; lethal - all die before end of embryonic stage
2461	*jing*	locomotor behaviour defective; cell death defective
3146	*AttC*	partially lethal; some die during pupal stage; neuroanatomy defective
4837	*CG4597*	some die during pupal stage; partially lethal - majority die
	CG4611	lethal - all die during P-stage
5702	*Hml*	immune response defective
7639	*timeout*	increased mortality; lethal - all die before end of P-stage; some die during P-stage
7883	*CG43335*	partially lethal - majority die; some die during pupal stage; partially lethal
8958	*CG33970*	lethal; sleep defective; flightless

* Genes previously found to have association with longevity as recorded in FlyBase or GenAge resources are shown in bold.

5.6 SIMILARITIES BETWEEN REGIONS OBSERVED IN TARGET GENE ANALYSIS FOR SYNTHETIC AND DGRP GWAS DATASETS AND HUMAN ORTHOLOG SEARCH

A number of genes with longevity-associated phenotypes were found in common between both GWAS datasets: *CG45186, CG32298, SNCF, CG14107, Ca-alpha1D, jing, AttC, CG4597, CG4611, Hml* and *CG43335.* Searches for human orthologs were carried out on these genes, for which there were matches for five of the genes. The *Drosophila* gene *CG45186* matched human ortholog *SVIL*, *CG4611* matched *PTCD1*, *jing* matched *AEBP2*, *Ca-alpha1D* was found to match the two human genes *CACNA1D* and *CACNA1S* and finally *Hml* was found to match with four human genes: *SSPO*, *VWF*, *OTOG* and *MUC5B*. These human ortholog genes were then further explored to look for any known association that these genes may have with longevity, however this search only found that the *SSPO* gene is involved in the modulation of neuronal aggregation and has been suggested to be involved in developmental events during the formation of the central nervous system (Meiniel et al., 2003). Although there were no other human genes in this ortholog search found to previously have any biological association with longevity, we can speculate that the findings of their *Drosophila* orthologs having longevity-associated phenotypes in this analysis may indicate that their human orthologs may play a role in human longevity that is thus far unknown. This observation of SNPs in non-coding regions targeting genes that are not within closest proximity on the *Drosophila* genome is one that can be speculated to apply to non-coding SNPs on the human genome and their target genes.

A literature search was also taken on these *Drosophila* genes in Tables 5.10 and 5.11 for any previous research in which they may have been associated with longevity. A study by Doroszuk et al. (2012), previously mentioned in Chapter 4, found that the antimicrobial peptide encoding gene *AttC* shows age-related changes in expression and therefore considered it to be a candidate marker of ageing.

5.7 COMPARISON OF TARGET GENES SELECTED FOR NON-CODING SNPS IN DGRP GWAS STUDY WITH THOSE OBTAINED USING FINER RESOLUTION HI-C DATA

To incorporate SNPs with the highest calculated P-values residing in non-coding regions, Ivanov et al. (2015) performed a gene-based analysis with gene positions extended by 5 kb

upstream of the 5' and downstream of 3' ends in the DGRP GWAS study. Therefore, the 50 top SNPs with the lowest P-values identified by Ivanov et al. (2015) were assigned either to a gene (coding SNPs) if they reside within the extended gene sequence; otherwise non-coding SNPs were assigned to two nearby genes located upstream and downstream which this SNP was assumed to target. The non-coding SNPs in this list were further explored in this study by assigning each of these SNPs to a 10 Kb region and utilising the finer 10 Kb resolution Hi-C data once again, to identify the strongest interacting gene-harbouring regions interacting with regions that contain these non-coding SNPs. Table 5.12 shows these selected non-coding SNPs from the top 50 SNPs, their positions and the genes with which they were associated according to Ivanov et al. (2015). The two regions strongly interacting with the non-coding SNP containing region are displayed in the last two columns, where for many the strongest interacting or second strongest interacting regions coincide with the SNP-harbouring region. However, for a few cases, neither of the strongest interacting bins were the same as itself. This was the case for SNPs 2L: 1835028, 2L: 2279849 and 3L: 17762728. The strongest interacting regions with bin 182 were bins 181 and 184, and so genes residing within these regions were sought. Bin 181 harboured genes *CG31933* and *CG31664* and bin 184 harboured genes *wry* and *CG31663*. The *wry* gene was also selected for this SNP by Ivanov et al. (2015), however its association with Notch signalling was not mentioned or discussed. The *wry* gene has previously been identified as a notch ligand and notch signalling has been shown to be critically important for the maintenance of normal heart function in the adult fly (Kim et al., 2010). This signalling pathway has been previously discussed in Chapter 4 of this thesis when discussing other gene findings. A search for the other three genes residing in these strongest interacting regions found that no phenotypic data was available and biological processes in which these three genes are involved are unknown. For the SNP 2L: 2279849, both genes *CG17242* and *CG4271* assigned by Ivanov et al. (2015) were genes found in the strongest interacting regions selected in this analysis, however a number of other genes were also found in bins 225 and 227. These included genes *CG34049*, *CG4270*, *CR43754*, *CG31681*, *CG42658* and *CG17237*, however again a search for these genes resulted in unknown biological processes and lack of phenotypic data available. The third SNP, 3L: 17762728, was found in a region which strongly interacted with bins 6097 and 6096 which harboured six genes in total, of which only one of these genes was selected by Ivanov et al. (2015) as a target gene for this non-coding SNP, gene *Adgf-A*. Two other genes, *Adgf-A2* and *Adgf-B*, also resided

in these interacting regions, both of which are known to be involved in growth factor activity. Again other genes were found in this region for which biological processes were unknown, including genes *CG32182* and *CG32181*, but there was one gene found in bin 6096 for which more information was available. This was the *Ccn* gene, which is involved with negative regulation of cell death and has phenotypes including 'increased mortality' and 'lethal'. Lack of knowledge about many of the genes found in these strongly interacting regions gives room for speculation that the non-coding SNPs with which they strongly interact may be influencing the expression of these genes and their biological processes. The interaction between SNP 3L: 17762728 and gene *Ccn*, with longevity related features, may suggest that it is in fact this gene that could be of interest instead of the *CG42815* gene selected by Ivanov et al. (2015).

Table 5.12 Non-coding SNPs in the top 50 SNPs recorded in the DGRP GWAS dataset.

Non-coding SNP position	Chr	SNP bin start position (bin number)	P-value	5'	3'	1st strongest interacting region	2nd strongest interacting region
1632386	2L	1630000 (164)	$5.90\text{x}10^{-8}$	*chinmo*	*RFeSP*	164	165
1632388	2L	1630000 (164)	$3.74\text{x}10^{-7}$	*chinmo*	*RFeSP*	164	165
1835028	2L	1830000 (182)	$1.11\text{x}10^{-5}$	*c-cup*	*wry*	181	184
2279849	2L	2270000 (226)	$2.21\text{x}10^{-6}$	*CG17242*	*CG4271*	225	227
3480710	2L	3480000 (347)	$6.77\text{x}10^{-7}$	*CG15414*	*Thor*	347	348
4308343	2R	4300000 (2654)	$8.41\text{x}10^{-6}$	*CSN7*	*CG43296*	2653	2654
4308355	2R	4300000 (2654)	$7.86\text{x}10^{-6}$	*CSN7*	*CG43296*	2653	2654
17762728	3L	17760000 (6098)	$1.13\text{x}10^{-5}$	*Adgf-A*	*CG42815*	6097	6096
18140585	3L	18140000 (6135)	$6.51\text{x}10^{-6}$	*CG7330*	*gk*	6135	6134
5319539	3L	5310000 (4859)	$1.12\text{x}10^{-5}$	*shep*	*lama*	4858	4859
8650506	3L	8650000 (5188)	$6.13\text{x}10^{-6}$	*h*	*Pex7*	5187	5188
15950064	3R	15950000 (8334)	$1.17\text{x}10^{-5}$	*Gr92a*	*CG5023*	8333	8334
23482833	3R	23480000 (9085)	$9.26\text{x}10^{-6}$	*Mlc1*	*tau*	9085	9086
25189263	3R	25180000 (9255)	$1.05\text{x}10^{-5}$	*Cnx99A*	*Ptp99A*	9255	9256

A search for human orthologs for these genes previously discussed resulted in matches for the *Drosophila* genes *wry*, *CG17242*, *Adgf-A* and *Ccn* which had human orthologs *DNER*, *PRSS36*, *ADA2* and *CTGF,* respectively. Like its *Drosophila* ortholog, the human gene *DNER* is related to the notch signalling pathway in humans, where mutations in notch signalling pathway members have been shown to cause developmental phenotypes that affect the liver, skeleton, heart and vasculature (Penton et al., 2012) which can all have effects on lifespan and ageing. The human gene *ADA2* is related to the innate immune system pathway and associated with human diseases including Vasculitis, Autoinflammation, Immunodeficiency and Hematologic Defects Syndrome in which patients are known to suffer recurrent strokes resulting in neurologic dysfunction (Zhou et al., 2014). Finally, the human gene *CTGF*, also known as *CCN2*, has important roles in biological processes including skeletal development, tissue wound repair and cell proliferation, this gene is also critically involved in fibrotic disease and several forms of cancers

Chapter 5 conclusions

In this chapter we have shown that a significant proportion of non-coding SNPs, recorded in the DGRP GWAS dataset, were residing in TAD border regions on the *Drosophila* genome when compared to a match control dataset. Architectural proteins binding around these TAD border regions were also found to contain a reasonable number SNPs recorded in this GWAS datasets, with the Suppressor of Hairy-wing (*Su(Hw)*) and the DNA Replication Related Element binding Factor (*DREF*) being the most mutated. SNPs in experimentally identified TFBSs of 49 TFs were analysed. To find whether specific regions of the TFBSs were affected by SNPs, we created consensus sequence logos for TFBS sequences recorded. Although in some instances the logos showed clear positions in which there was a dominant nucleotide base present, overall no consistency was observed which could lead to the assumption that all binding sequences for each TF recorded had an obvious base pair pattern which could be recognised by each TF. Next, we looked whether there are specific structural features such as non-B DNA conformation. This analysis resulted in TFBSs for a number of TFs: *ap, Br-Z2, EcR, fkh, hb, Mad, tin, vnd* and *zen*, being enriched in non-B DNA conformation. SNP counts for each of the sequences recorded under each of these TFBSs, in both the real and control

dataset, were quantified, where only one TF *ap* was found to show any real difference between these counts.

In the analysis for target genes, many regions which were found to have strong long-range interactions with regions found to contain non-coding SNPs in the Synthetic and DGRP GWAS datasets were observed to harbour genes exhibiting longevity-related phenotypes such as 'increased mortality' and 'lethal'. Interacting regions were found to harbour genes not directly associated with longevity but they had longevity related phenotypes. Commonly observed phenotypes included 'lethal' and 'increased mortality', with 'immune response defective' and 'cell death defective' also found to be present in the genes in these interacting regions. These interacting regions were also then compared between the two GWAS datasets, in which there were a number of regions and therefore genes found in common. A total of 11 genes were found in common between these GWAS datasets, also having phenotypes that we could assume to have longevity association. These were genes: *CG45186, CG32298, SNCF, CG14107, Ca-alpha1D, jing, AttC, CG4597, CG4611, Hml* and *CG43335.* Further exploration of these genes found the *AttC* gene to have previously been considered as a candidate marker of ageing. Five of these *Drosophila* genes: *CG45186, CG4611, jing, Ca-alpha1D* and *Hml*, were also found to have strong human ortholog matches. These human orthologs were found to be *SVIL*, *PTCD1*, *AEBP2*, *CACNA1D*, *CACNA1S*, *SSPO*, *VWF*, *OTOG* and *MUC5B*. A search for any association these human genes had to longevity was carried out, with the only findings being the suggestion of the *SSPO* gene being involved in developmental events during the formation of the central nervous system.

CHAPTER 6

CONCLUSIONS

In this study we applied a network approach to predict novel genes/genomic regions/SNPs, playing a role in longevity by integrating three-dimensional chromosome conformation data and two GWAS datasets. We demonstrated that co-location of novel genes/genomic regions with genes, known to be associated with longevity, and their enrichment in the same biological function or pathway as known genes, make them good candidates for novel genomic regions, linked to longevity. We further demonstrated that SNPs residing within these regions may influence longevity either individually (when a SNP in one of these genes could cause a phenotype) or collectively (when one or several SNPs in these regions occur in the same patient to cause the phenotype). This study also analysed SNPs in non-coding regions, looking at TFBSs first, and found that TFs may recognise a certain structure, e.g. non-B DNA structures, rather than sequence motifs. Structures such as slipped, cruciform, triplexes and tetraplexes, formed on direct, inverted and mirrored repeats and G-quartets were considered. TADs were also explored, where we hypothesised that SNPs residing in these border regions may cause disruption to the way in which regulation usually occurs within these TADs via looping interactions. Such disruption may lead to interactions between TADs that would not usually occur because of the borders by which they are separated. Finally this study looked at potential target genes for non-coding SNPs, taking a different approach to the assumption that these target genes are usually the nearest on the linear genome.

In this chapter our results are summarised, their implications are discussed and ideas for future work are suggested.

6.1 NOVEL LONGEVITY-ASSOCIATED CANDIDATE REGIONS IDENTIFIED VIA NETWORK ANALYSIS

The ability to successfully identify new regions/genes on the *Drosophila* genome which we can say for sure have a direct effect on longevity is a challenge, however with the qualitative data available and hypotheses that make sense both mathematically and biologically, an attempt can be made. The first hypothesis we made in this thesis was that the 3D architecture of the *Drosophila* genome dictates the co-location of specific genes/genomic regions, both known to be associated with longevity and novel unknown regions that may be potentially important in longevity. Networks were created using genes/genomic regions, quantified to associate with longevity, as original nodes with additional nodes (regions) later added to these networks if they strongly interacted (co-localise) with original nodes. Various network measures were calculated, identifying important previously unknown regions of the genome. All regions found to be of interest using these network measures were further explored, and their residing genes were used in GO enrichment analysis, where they were found to enrich in biological processes with longevity relation.

Regions of each extended GWAS network were selected as important by network measures. For regions in the extended Synthetic GWAS-based network, GO enrichment found genes residing in these regions to be enriched in processes related to longevity such as 'respiratory system development', 'defence response', and 'regulation of apoptotic process'. Among these genes were previously longevity associated genes *Sema-5c* and *esg* found to reside in novel regions. For regions in the extended DGRP GWAS-based network, GO enrichment found genes residing in these regions to be enriched in processes such as 'ageing', 'immune response', 'detoxification' and 'defence response'. Again, among these genes enriched, longevity genes *Indy* and *chico* were also observed but this time in original regions of the network, the longevity gene *EcR* found in a novel region was also enriched.

Regions selected by network measures were found in common between both extended GWAS networks datasets, in which genes with longevity related phenotypes were found to reside. This included the genes *Rim* and *Tpi* with 'long-lived' phenotypes, and *frtz*, *Atxn7*, *CG5339*, *CG4434* and *Zip99C* with 'short-lived' phenotypes. A human ortholog search taken on genes found to reside in common regions showed several matches to human genes with

functions related to the lifespan of an organism. Genes were found to match human orthologs, where a number of these orthologs were found to relate to biological processes and functions known to relate to ageing/disease or effect lifespan of humans. These human genes included *TXLNA, SNAI2, ALDOA, SCFD1, CALML6*, *SCD*, *PLPBP*, *LRRC4C* and SEMA5B. Common regions between extended GWAS networks were also observed to harbour a fairly large number of SNPs.

Subnetworks in these extended GWAS networks were then further explored, where enrichment in Gene Ontology terms identified genes/regions with no previous association with longevity. This analysis was based on the hypothesis stating that SNPs residing within co-located genomic regions influence longevity either independently or have a cumulative effect. Subnetworks in the extended Synthetic GWAS-based network harboured genes that enriched under biological processes in GO analysis including 'DNA repair', 'apoptotic process' and 'developmental process'. Some of the genes had previously been found to associate with longevity, including *ku80*, *foxo*, *VhaSFD* and *CathD*. Other genes were found to have longevity related phenotypes including 'increased mortality', 'lethal' and the *azot* gene had a 'long-lived' phenotype. SNPs in genes of these subnetworks were counted, where genes such as *mub* and *foxo* harboured significant SNPs and were therefore speculated to independently influence genes in the same subnetwork in relation to longevity. Many genes in these subnetworks were observed to harbour a number of SNPs but none of which were significant, leading to speculation that in these cases longevity may be influenced through a cumulative effect of several SNPs.

Subnetworks in the extended DGRP GWAS-based network harboured genes that enriched in biological processes in GO analysis including 'developmental growth', 'nervous system process', 'immune system process' and 'regulation of immune system process'. Genes known to have longevity association, including genes *ft, grim* and *Chmp1* were enriched in these GO terms. These genes were found to be associated with a shortened lifespan in *Drosophila* as well as other genes including *Sod2*, *PGRP-SA*, *Btk29A* and *Trag6* involved in pathways and processes associated with ageing. These genes were observed to harbour SNPs, leading to speculation that they could influence the function of genes, coming into close proximity and sharing the same GO terms, in the same way as the known longevity associated genes.

Subnetworks found to contain genes which shared common biological processes associated with longevity were then further explored. Regions of each of these subnetworks selected were explored for the occurrence of significant SNPs which were highlighted earlier in this study. This further subnetwork exploration also looked at enhancers that reside in regions of these subnetworks.

6.2 SNPS IN NON-CODING REGIONS

In the genome, there are many different structures and features which have been proved to play important roles in the organization of the genome and gene regulation. The disruption of these structures or features have been hypothesised and proven to have a detrimental effect on the role they are supposed to play, and as a result lead to disease. This analysis first focussed on TAD boundary regions and non-coding SNPs within these regions. There was found to be a significantly higher proportion of non-coding SNPs, recorded in the DGRP GWAS dataset, residing in TAD border regions in the real dataset when compared to a matched control dataset. Architectural binding proteins known to bind around these TAD border regions were also found to contain a reasonable number of non-coding SNPs recorded in this GWAS dataset, with the Suppressor of Hairy-wing (*Su(Hw)*) and the DNA Replication Related Element binding Factor (*DREF*) being the most mutated.

Transcription factors are functional elements in the genome that are crucial for controlling many important processes, along with the sites of the DNA to which they bind, known as transcription factor binding sites. TFs are known to interact with DNA in two primary ways, through non-sequence-specific interactions or sequence-specific interactions and the role of DNA shape as a determinant of protein-DNA binding specificity has been shown to be important. Non-B DNA conformations have been found to form at specific sequence motifs and are believed to form in higher energy states during metabolic processes in DNA such as transcription, replication and repair. This chapter further explored TFBSs, by first considering base pair sequence patterns in DNA by producing consensus sequence logos, and then focussing on non-B DNA structures and their potential role in TF binding.

Using a dataset containing recorded binding site positions for TFs in *Drosophila*, consensus sequence logos were created for each TF. In some positions of the logos, dominant nucleotide bases were apparent, however this observation was not consistent enough to conclude that all binding sequences for each TF recorded had an obvious base pair pattern which could be recognised by this TF. The creation of a matched control TFBS sequence dataset enabled analysis, using chi-squared, between the frequencies of sequence repeat counts in TFBS sequences from the real dataset with those from the control dataset.

Statistical analysis found a number of TFBSs for TFs: *ap, Br-Z2, EcR, fkh, hb, Mad, tin, vnd* and *zen*, to show significant enrichment in the proportion of recorded sequences containing repeats between those in the real dataset when compared with those recorded in the control dataset. SNP counts for each of the TFBS sequences recorded under each of these TFs, in both the real and control dataset, were then quantified, where only TFBSs for the TF *ap* was found to show any real difference between these counts.

GWAS-identified disease-associated nucleotide differences are often found to reside in non-coding regions of the genome, where many have looked to find the way in which a non-coding SNP can be associated with an increased risk for a specific disease. Approaches to do so often focus on the identification of the target genes for these non-coding SNPs, where is has often been assumed that the target gene of such a SNP corresponds to the gene to which it resides closest in physical distance, but this is not possible to confirm. The strongest interacting regions with regions containing non-coding SNPs in the Synthetic and DGRP GWAS datasets, using higher resolution Hi-C data at 10 Kb, were found. Regions containing non-coding SNPs which have strongest interacting regions at the longest distances were of most interest, and were then selected for further analysis, focusing on the genes they harbour and the functions of these genes.

Interacting regions found to harbour genes were further explored, where it was found that numerous genes, although not found to be directly associated with longevity, had phenotypes displaying longevity related functions. Commonly observed phenotypes included ‘lethal’ and ‘increased mortality’, with ‘immune response defective’ and ‘cell death defective’ also found expressed in the genes in these selected regions. A total of 11 genes were found in common between the extended GWAS networks, also with phenotypes that we could assume to have longevity association, these were: *CG45186, CG32298, SNCF, CG14107, Ca-alpha1D, jing, AttC,*

CG4597, CG4611, Hml and *CG43335.* Further exploration of these genes found *AttC* to have previously been considered as a candidate marker of ageing and five of these *Drosophila* genes: *CG45186, CG4611, jing, Ca-alpha1D* and *Hml*, were also found to have strong human ortholog matches. These human orthologs were found to be *SVIL*, *PTCD1*, *AEBP2*, *CACNA1D*, *CACNA1S*, *SSPO*, *VWF*, *OTOG* and *MUC5B*. A search for any relation these human genes had to longevity was carried out, with the only findings being the suggestion of the *SSPO* gene being involved in developmental events during the formation of the central nervous system.

6.3 FUTURE WORK

All network analysis in this thesis utilised *Drosophila* longevity GWAS data from the Synthetic and DGRP GWAS datasets described. For future work, it would be of interest to use datasets that could be considered 'designed long-lived *Drosophila*' datasets to repeat the analyses performed in this thesis. If and when available, such datasets would include GWAS for *Drosophila* in environmentally controlled states, including specifically, calorie-restricted *Drosophila* and *Drosophila* in temperature-controlled environments. These conditions have commonly been shown to effect lifespan in *Drosophila*, and so the ability to analyse this data in a similar way to GWAS previously analysed in this thesis and comparison with the results reported in this thesis, could provide new information on genes and genomic regions of *Drosophila* in relation to longevity. Also in relation to networks approaches used in this thesis, future work could also be done to test the reliability of networks produced. This could be done by considering the percolation of networks, in which a percolation model is generated to assess the robustness of a network. In the percolation theory, the failure of a node/edge of network is modelled by removal of a subset of nodes with their adjacent edges and in short, this theory can help to understand the macroscopic failure behaviour of networks in relation to the microscopic states of the network components (Li et al., 2015). This assessment could be carried out for all networks produced in order to understand the behaviour of the networks and its ability to maintain its modularity.

Since the time that this study was started, results of human longevity GWAS have been published (Pilling et al., 2017), in which 25 genetic loci associated with longevity were identified through analysis of 389,166 UK biobank participants. The analysis carried out in this

PhD study of *Drosophila* has produced findings which have enabled the identification of a set of human homologs/orthologs/paralogs as potential longevity gene candidates. Therefore, direction for future work would be to utilise the recently published human longevity GWAS, comparing their findings with those found throughout this study. The human longevity GWAS would also be useful for justifying the hypothesis of the co-location of homologs/orthologs/paralogs in 3D space within the cell nucleus. The availability of human longevity GWAS would also allow the application of the modelling used in this PhD study of *Drosophila*, repeating the techniques developed and analysis used. Results in this PhD study using *Drosophila* longevity GWAS and the results from repetition of techniques and analysis using human longevity GWAS could then be compared for any similarities, where these similarities could strengthen the conclusions of the findings in this PhD study. For modelling using human longevity GWAS, identification of a suitable cell line would be required, for which chromosome conformation capture data is available, and analysis of available eQTL would be important.

BIBLIOGRAPHY

Aguirre, G.A., De Ita, J.R., de La Garza, R.G. and Castilla-Cortazar, I., 2016. Insulin-like growth factor-1 deficiency and metabolic syndrome. *Journal of Translational Medicine*, 14 (1), 3.

Alcedo, J., Flatt, T. and Pasyukova, E.G., 2013. The role of the nervous system in aging and longevity. *Frontiers in Genetics*, 4, 124.

Anderson, R.M. and Weindruch, R., 2012. The caloric restriction paradigm: implications for healthy human aging. *American Journal of Human Biology*, 24 (2), 101-106.

Altintas, O., Park, S. and Lee, S.J.V., 2016. The role of insulin/IGF-1 signalling in the longevity of model invertebrates, C. elegans and D. melanogaster. *BMB Reports*, 49 (2), 81.

Andersson, R., Sandelin, A. and Danko, C.G., 2015. A unified architecture of transcriptional regulatory elements. *Trends in Genetics*, 31 (8), 426-433.

Aoki, T., Sarkeshik, A., Yates, J. and Schedl, P., 2012. Elba, a novel developmentally regulated chromatin boundary factor is a hetero-tripartite DNA binding complex. *Elife*, 1, e00171.

Bacolla, A. and Wells, R.D., 2004. Non-B DNA conformations, genomic rearrangements, and human disease. *Journal of Biological Chemistry*, 279 (46), 47411-47414.

Bai, H., Kang, P., Hernandez, A.M. and Tatar, M., 2013. Activin signaling targeted by insulin/dFOXO regulates aging and muscle proteostasis in Drosophila. *PLoS Genetics*, 9 (11), 1003941.

Ball, E.V., Stenson, P.D., Abeysinghe, S.S., Krawczak, M., Cooper, D.N. and Chuzhanova, N.A., 2005. Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. *Human Mutation*, 26 (3), 205-213.

Barnett, Y.A. and Barnett, C.R., 1998. DNA damage and mutation: contributors to the age-related alterations in T cell-mediated immune responses?. *Mechanisms of Ageing and Development*, 102 (2-3), 165-176.

Barreto, G., Schäfer, A., Marhold, J., Stach, D., Swaminathan, S.K., Handa, V., Döderlein, G., Maltry, N., Wu, W., Lyko, F. and Niehrs, C., 2007. Gadd45a promotes epigenetic gene activation by repair-mediated DNA demethylation. *Nature*, 445 (7128), 671.

Barutcu, A.R., Fritz, A.J., Zaidi, S.K., van Wijnen, A.J., Lian, J.B., Stein, J.L., Nickerson, J.A., Imbalzano, A.N. and Stein, G.S., 2016. C-ing the genome: a compendium of chromosome conformation capture methods to study higher-order chromatin organization. *Journal of Cellular Physiology*, 231 (1), 31-35.

Bathum, L., Christiansen, L., Jeune, B., Vaupel, J., McGue, M. and Christensen, K., 2006. Apolipoprotein e genotypes: relationship to cognitive functioning, cognitive decline, and survival in nonagenarians. *Journal of the American Geriatrics Society*, 54 (4), 654-658.

Bauer, J.H., Poon, P.C., Glatt-Deeley, H., Abrams, J.M. and Helfand, S.L., 2005. Neuronal expression of p53 dominant-negative proteins in adult Drosophila melanogaster extends life span. *Current Biology*, 15 (22), 2063-2068.

Beekman, M., Blanché, H., Perola, M., Hervonen, A., Bezrukov, V., Sikora, E., Flachsbart, F., Christiansen, L., De Craen, A.J., Kirkwood, T.B. and Rea, I.M., 2013. Genome-wide linkage analysis for human longevity: Genetics of Healthy Aging Study. *Aging Cell*, 12 (2), 184-193.

Bengtson, V.L. and Settersten Jr, R. eds., 2016. *Handbook of Theories of Aging*. Springer Publishing Company.

Biteau, B., Karpac, J., Supoyo, S., DeGennaro, M., Lehmann, R. and Jasper, H., 2010. Lifespan extension by preserving proliferative homeostasis in Drosophila. *PLoS Genetics*, 6 (10), 1001159.

Biteau, B., Karpac, J., Hwangbo, D. and Jasper, H., 2011. Regulation of Drosophila lifespan by JNK signaling. *Experimental Gerontology*, 46 (5), 349-354.

Bjorksten, J. and Tenhu, H., 1990. The crosslinking theory of aging — Added evidence. *Experimental Gerontology*, 25 (2), 91-95.

Blattler, A., Yao, L., Witt, H., Guo, Y., Nicolet, C.M., Berman, B.P. and Farnham, P.J., 2014. Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes. *Genome Biology*, 15 (9), 469.

Blondel, V.D., Guillaume, J.L., Lambiotte, R. and Lefebvre, E., 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008 (10), 10008.

Bolós, V., Grego-Bessa, J. and de la Pompa, J.L., 2007. Notch signaling in development and cancer. *Endocrine Reviews*, 28 (3), 339-363.

Borbolis, F. and Syntichaki, P., 2015. Cytoplasmic mRNA turnover and ageing. *Mechanisms of Ageing and Development*, 152, 32-42.

Boyden, S.E. and Kunkel, L.M., 2010. High-density genomewide linkage analysis of exceptional human longevity identifies multiple novel loci. *PLoS One*, 5 (8), e12432.

Brin, S. and Page, L., 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30 (1-7), 107-117.

Broer, L., Buchman, A.S., Deelen, J., Evans, D.S., Faul, J.D., Lunetta, K.L., Sebastiani, P., Smith, J.A., Smith, A.V., Tanaka, T. and Yu, L., 2014. GWAS of longevity in CHARGE consortium confirms APOE and FOXO3 candidacy. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, 70 (1), 110-118.

Brown-Borg, H.M., 2007. Hormonal regulation of longevity in mammals. *Ageing Research Reviews*, 6 (1), 28-45.

Burke, M.K., King, E.G., Shahrestani, P., Rose, M.R. and Long, A.D., 2013. Genome-wide association study of extreme longevity in Drosophila melanogaster. *Genome Biology and Evolution*, 6 (1), 1-11.

Bush, W.S., Chen, G., Torstenson, E.S. and Ritchie, M.D., 2009. LD-spline: mapping SNPs on genotyping platforms to genomic regions using patterns of linkage disequilibrium. *BioData Mining*, 2 (1), 7.

Bushey, A.M., Ramos, E. and Corces, V.G., 2009. Three subclasses of a Drosophila insulator show distinct and cell type-specific genomic distributions. *Genes & Development*, 23 (11), 1338-1350.

Candore, G., Di Lorenzo, G., Mansueto, P., Melluso, M., Fradà, G., Vecchi, M.L., Pellitteri, M.E., Drago, A., Di Salvo, A. and Caruso, C., 1997. Prevalence of organ-specific and non organ-specific autoantibodies in healthy centenarians. *Mechanisms of Ageing and Development*, 94 (1-3), 183-190.

Carvalho, A.B., 2002. Origin and evolution of the Drosophila Y chromosome. *Current Opinion In Genetics & Development*, 12 (6), 664-668.

Castelo-Branco, C. and Soveral, I., 2014. The immune system and aging: a review. *Gynecological Endocrinology*, 30 (1), 16-22.

Castells-Nobau, A., Eidhof, I., Fenckova, M., Brenman-Suttner, D.B., Scheffer-de Gooyert, J.M., Christine, S., Schellevis, R.L., Van der Laan, K., Quentin, C., Van Ninhuijs, L. and Hofmann, F., 2019. Conserved regulation of neurodevelopmental processes and behavior by FoxP in Drosophila. *PloS One*, 14 (2), p0211652.

Chen, X., Wang, L., Hu, B., Guo, M., Barnard, J. and Zhu, X., 2010. Pathway-based analysis for genome-wide association studies using supervised principal components. *Genetic Epidemiology*, 34 (7), 716-724.

Christensen, K., Johnson, T.E. and Vaupel, J.W., 2006. The quest for genetic determinants of human longevity: challenges and insights. *Nature Reviews Genetics*, 7 (6), 436.

Clancy, D.J., Gems, D., Harshman, L.G., Oldham, S., Stocker, H., Hafen, E., Leevers, S.J. and Partridge, L., 2001. Extension of life-span by loss of CHICO, a Drosophila insulin receptor substrate protein. *Science*, 292 (5514), 104-106.

Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E., 2004. WebLogo: a sequence logo generator. *Genome Research*, 14 (6), 1188-1190.

Cuartero, S., Fresán, U., Reina, O., Planet, E. and Espinàs, M.L., 2014. Ibf1 and Ibf2 are novel CP190-interacting proteins required for insulator function. *The EMBO Journal*, 33 (6), 637-647.

Cubeñas-Potts, C. and Corces, V.G., 2015. Architectural proteins, transcription, and the three-dimensional organization of the genome. *FEBS letters*, 589 (20), 2923-2930.

Cubeñas-Potts, C., Rowley, M.J., Lyu, X., Li, G., Lei, E.P. and Corces, V.G., 2016. Different enhancer classes in Drosophila bind distinct architectural proteins and mediate unique chromatin interactions and 3D architecture. *Nucleic Acids Research*, 45 (4), 1714-1730.

Curtis, C., Landis, G.N., Folk, D., Wehr, N.B., Hoe, N., Waskar, M., Abdueva, D., Skvortsov, D., Ford, D., Luu, A. and Badrinath, A., 2007. Transcriptional profiling of MnSOD-mediated lifespan extension in Drosophila reveals a species-general network of aging and metabolic genes. *Genome Biology*, 8 (12), R262.

Cvejic, S., Zhu, Z., Felice, S.J., Berman, Y. and Huang, X.Y., 2004. The endogenous ligand Stunted of the GPCR Methuselah extends lifespan in Drosophila. *Nature Cell Biology*, 6 (6), 540.

de Leeuw, C.A., Mooij, J.M., Heskes, T. and Posthuma, D., 2015. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Computational Biology*, 11 (4), 1004219.

de Wit, E. and De Laat, W., 2012. A decade of 3C technologies: insights into nuclear organization. *Genes & Development*, 26 (1), 11-24.

Deelen, J., Beekman, M., Uh, H.W., Broer, L., Ayers, K.L., Tan, Q., Kamatani, Y., Bennet, A.M., Tamm, R., Trompet, S. and Guðbjartsson, D.F., 2014. Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. *Human Molecular Genetics*, 23 (16), 4420-4432.

Dekker, J., Rippe, K., Dekker, M. and Kleckner, N., 2002. Capturing chromosome conformation. *Science*, 295 (5558), 1306-1311.

Dekker, J., Marti-Renom, M.A. and Mirny, L.A., 2013. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*, 14 (6), 390.

Demontis, F. and Perrimon, N., 2010. FOXO/4E-BP signaling in Drosophila muscles regulates organism-wide proteostasis during aging. *Cell*, *143* (5), 813-825.

Doroszuk, A., Jonker, M.J., Pul, N., Breit, T.M. and Zwaan, B.J., 2012. Transcriptome analysis of a long-lived natural Drosophila variant: a prominent role of stress-and reproduction-genes in lifespan extension. *BMC Genomics*, 13 (1), 167.

dos Santos, G., Schroeder, A.J., Goodman, J.L., Strelets, V.B., Crosby, M.A., Thurmond, J., Emmert, D.B., Gelbart, W.M. and FlyBase Consortium, 2014. FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Research*, 43 (D1), 690-697.

Dror, I., Zhou, T., Mandel-Gutfreund, Y. and Rohs, R., 2013. Covariation between homeodomain transcription factors and the shape of their DNA binding sites. *Nucleic Acids Research*, 42 (1), 430-441.

Eggert, H., Gortchakov, A. and Saumweber, H., 2004. Identification of the Drosophila interband-specific protein Z4 as a DNA-binding zinc-finger protein determining chromosomal structure. *Journal of Cell Science*, 117 (18), 4253-4264.

Fernando, M.D.A., Kounatidis, I. and Ligoxygakis, P., 2014. Loss of Trabid, a new negative regulator of the Drosophila immune-deficiency pathway at the level of TAK1, reduces life span. *PLoS Genetics*, 10 (2), 1004117.

Ferraiuolo, M.A., Sanyal, A., Naumova, N., Dekker, J. and Dostie, J., 2012. From cells to chromatin: capturing snapshots of genome organization with 5C technology. *Methods*, 58 (3), 255-267.

Finch, C.E., 1994. *Longevity, senescence, and the genome*. University of Chicago Press.

Flachsbart, F., Ellinghaus, D., Gentschew, L., Heinsen, F.A., Caliebe, A., Christiansen, L., Nygaard, M., Christensen, K., Blanché, H., Deleuze, J.F. and Derbois, C., 2016. Immunochip analysis identifies association of the RAD 50/IL 13 region with human longevity. *Aging Cell*, 15 (3), 585-588.

Flatt, T. and Partridge, L., 2018. Horizons in the evolution of aging. *BMC Biology*, 16 (1), 1-13.

Fontana, L., 2009. Modulating human aging and age-associated diseases. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1790 (10), 1133-1138.

Fontana, L., Partridge, L. and Longo, V.D., 2010. Extending healthy life span—from yeast to humans. *Science*, 328 (5976), 321-326.

Franceschi, C. and Campisi, J., 2014. Chronic inflammation (inflammaging) and its potential contribution to age-associated diseases. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, 69 (Suppl_1), S4-S9.

Frankel, S., Ziafazeli, T. and Rogina, B., 2011. dSir2 and longevity in Drosophila. *Experimental Gerontology*, 46 (5), 391-396.

Freedman, M.L., Monteiro, A.N., Gayther, S.A., Coetzee, G.A., Risch, A., Plass, C., Casey, G., De Biasi, M., Carlson, C., Duggan, D. and James, M., 2011. Principles for the post-GWAS functional characterization of cancer risk loci. *Nature Genetics*, 43 (6), 513.

Fuentes, E., Fuentes, M., Alarcon, M. and Palomo, I., 2017. Immune system dysfunction in the elderly. *Anais da Academia Brasileira de Ciências*, 89 (1), 285-299.

Fujiki, Y., Okumoto, K., Mukai, S., Honsho, M. and Tamura, S., 2014. Peroxisome biogenesis in mammalian cells. *Frontiers In Physiology*, 5, 307.

Fullwood, M.J., Wei, C.L., Liu, E.T. and Ruan, Y., 2009. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Research*, 19 (4), 521-532.

Gaiano, N. and Fishell, G., 2002. The role of notch in promoting glial and neural stem cell fates. *Annual Review of Neuroscience*, 25 (1), 471-490.

Garvie, C.W. and Wolberger, C., 2001. Recognition of specific DNA sequences. *Molecular Cell*, 8 (5), 937-946.

Geller, A.M. and Zenick, H., 2005. Aging and the environment: a research framework. *Environmental Health Perspectives*, 113 (9), 1257-1262.

Gerasimova, T.I., Lei, E.P., Bushey, A.M. and Corces, V.G., 2007. Coordinated control of dCTCF and gypsy chromatin insulators in Drosophila. *Molecular Cell*, 28 (5), 761-772.

Gerdes, L.U., Jeune, B., Ranberg, K.A., Nybo, H. and Vaupel, J.W., 2000. Estimation of apolipoprotein E genotype-specific relative mortality risks from the distribution of genotypes in centenarians and middle-aged men: Apolipoprotein E gene is a “frailty gene,” not a “longevity gene”. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 19 (3), 202-210.

Giannakou, M.E., Goss, M., Jünger, M.A., Hafen, E., Leevers, S.J. and Partridge, L., 2004. Long-lived Drosophila with overexpressed dFOXO in adult fat body. *Science*, 305 (5682), 361-361.

Giannakou, M.E. and Partridge, L., 2007. Role of insulin-like signalling in Drosophila lifespan. *Trends in Biochemical Sciences*, 32 (4), 180-188.

Giannakou, M.E., Goss, M., Jacobson, J., Vinti, G., Leevers, S.J. and Partridge, L., 2007. Dynamics of the action of dFOXO on adult mortality in Drosophila. *Aging Cell*, 6 (4), 429-438.

Gim, B.S., Park, J.M., Yoon, J.H., Kang, C. and Kim, Y.J., 2001. Drosophila Med6 is required for elevated expression of a large but distinct set of developmentally regulated genes. *Molecular And Cellular Biology*, 21 (15), 5242-5255.

Gimenez, L.E., Ghildyal, P., Fischer, K.E., Hu, H., Ja, W.W., Eaton, B.A., Wu, Y., Austad, S.N. and Ranjan, R., 2013. Modulation of methuselah expression targeted to D rosophila insulin-producing cells extends life and enhances oxidative stress resistance. *Aging Cell*, 12 (1), 121-129.

Gomez, C.R., Nomellini, V., Faunce, D.E. and Kovacs, E.J., 2008. Innate immunity and aging. *Experimental Gerontology*, 43 (8), 718-728.

Gómez-Díaz, E. and Corces, V.G., 2014. Architectural proteins: regulators of 3D genome organization in cell fate. *Trends in Cell Biology*, 24 (11), 703-711.

Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R. and Bulyk, M.L., 2013. Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Reports*, 3 (4), 1093-1104.

Goronzy, J.J., Fang, F., Cavanagh, M.M., Qi, Q. and Weyand, C.M., 2015. Naive T cell maintenance and function in human aging. *The Journal of Immunology*, 194 (9), 4073-4080.

Gurudatta, B., Yang, J., Van Bortle, K., Donlin-Asp, P. and Corces, V., 2013. Dynamic changes in the genomic localization of DNA replication-related element binding factor during the cell cycle. *Cell Cycle*, 12 (10), 1605-1615.

Haigis, M.C. and Sinclair, D.A., 2010. Mammalian sirtuins: biological insights and disease relevance. *Annual Review of Pathological Mechanical Disease*, 5, 253-295.

Hamilton, W.D., 1966. The moulding of senescence by natural selection. *Journal of Theoretical Biology*, 12 (1), 12-45.

Harper, J.A., Yuan, J.S., Tan, J.B., Visan, I. and Guidos, C.J., 2003. Notch signaling in development and disease. *Clinical Genetics*, 64 (6), 461-472.

Häsler, R., Venkatesh, G., Tan, Q., Flachsbart, F., Sinha, A., Rosenstiel, P., Lieb, W., Schreiber, S., Christensen, K., Christiansen, L. and Nebel, A., 2017. Genetic interplay between human longevity and metabolic pathways—a large-scale eQTL study. *Aging Cell*, 16 (4), 716-725.

Hassan, A., Timerman, Y., Hamdan, R., Sela, N., Avetisyan, A., Halachmi, N. and Salzberg, A., 2018. An RNAi screen identifies new genes required for normal morphogenesis of larval chordotonal organs. *G3: Genes, Genomes, Genetics*, 8 (6), 1871-1884.

He, Y. and Jasper, H., 2014. Studying aging in Drosophila. *Methods*, 68 (1), 129-133.

Herskind, A.M., McGue, M., Holm, N.V., Sörensen, T.I., Harvald, B. and Vaupel, J.W., 1996. The heritability of human longevity: a population-based study of 2872 Danish twin pairs born 1870–1900. *Human Genetics*, 97 (3), 319-323.

Higami, Y. and Shimokawa, I., 2000. Apoptosis in the aging process. *Cell and Tissue Research*, 301 (1), 125-132.

Hjelmborg, J., 2007. Genetic Influence on Human Lifespan and Longevity. *Twin Research and Human Genetics*, 10 (Supplement), 34.

Hoffmann, J., Romey, R., Fink, C., Yong, L. and Roeder, T., 2013. Overexpression of Sir2 in the adult fat body is sufficient to extend lifespan of male and female Drosophila. *Aging (Albany NY)*, 5 (4), 315.

Hong, M.G., Pawitan, Y., Magnusson, P.K. and Prince, J.A., 2009. Strategies and issues in the detection of pathway enrichment in genome-wide association studies. *Human Genetics*, 126 (2), 289-301.

Hou, C., Li, L., Qin, Z.S. and Corces, V.G., 2012. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Molecular Cell*, 48 (3), 471-484.

Huang, H., Patel, D.D. and Manton, K.G., 2005. The immune system in aging: roles of cytokines, T cells and NK cells. *Frontiers in Bioscience*, 10 (4), 192-215.

Huang, W., Massouras, A., Inoue, Y., Peiffer, J., Ràmia, M., Tarone, A.M., Turlapati, L., Zichner, T., Zhu, D., Lyman, R.F. and Magwire, M.M., 2014. Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. *Genome Research*, 24 (7), 1193-1208.

Hwangbo, D.S., Gersham, B., Tu, M.P., Palmer, M. and Tatar, M., 2004. Drosophila dFOXO controls lifespan and regulates insulin signalling in brain and fat body. *Nature*, 429 (6991), 562.

Imai, S.I. and Guarente, L., 2010. Ten years of NAD-dependent SIR2 family deacetylases: implications for metabolic diseases. *Trends in Pharmacological Sciences*, 31 (5), 212-220.

Ivanov, D.K., Escott-Price, V., Ziehm, M., Magwire, M.M., Mackay, T.F., Partridge, L. and Thornton, J.M., 2015. Longevity GWAS using the Drosophila genetic reference panel. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, 70 (12), 1470-1478.

Jin, K., 2010. Modern biological theories of aging. *Aging and Disease*, 1 (2), 72.

Joaquin, A.M. and Gollapudi, S., 2001. Functional decline in aging and disease: a role for apoptosis. *Journal of the American Geriatrics Society*, 49 (9), 1234-1240.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G. and Palin, K., 2013. DNA-binding specificities of human transcription factors. *Cell*, 152 (1-2), 327-339.

Joshi, P.K., Pirastu, N., Kentistou, K.A., Fischer, K., Hofer, E., Schraut, K.E., Clark, D.W., Nutile, T., Barnes, C.L., Timmers, P.R. and Shen, X., 2017. Genome-wide meta-analysis associates HLA-DQA1/DRB1 and LPA and lifestyle factors with human longevity. *Nature Communications*, 8 (1), 910.

Kapahi, P., Zid, B.M., Harper, T., Koslover, D., Sapin, V. and Benzer, S., 2004. Regulation of lifespan in Drosophila by modulation of genes in the TOR signaling pathway. *Current Biology*, 14 (10), 885-890.

Karanjawala, Z.E. and Lieber, M.R., 2004. DNA damage and aging. *Mechanisms of Ageing and Development*, 125 (6), 405-416.

Karpac, J., Hull-Thompson, J., Falleur, M. and Jasper, H., 2009. JNK signaling in insulin-producing cells is required for adaptive responses to stress in Drosophila. *Aging Cell*, 8 (3), 288-295.

Kenyon, C., Chang, J., Gensch, E., Rudner, A. and Tabtiang, R., 1993. A C. elegans mutant that lives twice as long as wild type. *Nature*, 366 (6454), 461.

Kerber, R.A., O'Brien, E., Boucher, K.M., Smith, K.R. and Cawthon, R.M., 2012. A genome-wide study replicates linkage of 3p22-24 to extreme longevity in humans and identifies possible additional loci. *PloS One*, 7 (4), e34746.

Kim, I.M., Wolf, M.J. and Rockman, H.A., 2010. Gene deletion screen for cardiomyopathy in adult Drosophila identifies a new notch ligand. *Circulation Research*, 106 (7), 1233.

Kirby, K., Hu, J., Hilliker, A.J. and Phillips, J.P., 2002. RNA interference-mediated silencing of Sod2 in Drosophila leads to early adult-onset mortality and elevated endogenous oxidative stress. *Proceedings of the National Academy of Sciences*, 99 (25), 16162-16167.

Kirkwood, T.B., 1977. Evolution of Ageing. *Nature*, 270 (5635), 301-304.

Kirkwood, T.B. and Melov, S., 2011. On the programmed/non-programmed nature of ageing within the life history. *Current Biology*, 21 (18), R701-R707.

Kopyl, S.A., Omelyanchuk, L.V., Shaposhnikov, M.V. and Moskalev, A.A., 2014. Role of tumor suppressor genes in aging and longevity mechanisms in Drosophila melanogaster. *Russian Journal of Genetics: Applied Research*, 4 (1), 8-14.

Kumosani, T.A., Alama, M.N. and Iyer, A., 2011. Cardiovascular diseases in Saudi Arabia. *Prime Research Medicine*, 1, 1-6.

Kusama, S., Ueda, R., Suda, T., Nishihara, S. and Matsuura, E.T., 2006. Involvement of Drosophila Sir2-like genes in the regulation of life span. *Genes & Genetic Systems*, 81 (5), 341-348.

Kylsten, P., Samakovlis, C. and Hultmark, D., 1990. The cecropin locus in Drosophila; a compact gene cluster involved in the response to infection. *The EMBO Journal*, 9 (1), 217-224.

Lambiotte, R., Delvenne, J.C. and Barahona, M., 2008. Laplacian dynamics and multiscale modular structure in networks. *arXiv preprint arXiv:0812.1770*.

Landis, G.N., Bhole, D. and Tower, J., 2003. A search for doxycycline-dependent mutations that increase Drosophila melanogaster life span identifies the VhaSFD, Sugar baby, filamin, fwd and Cctl genes. *Genome Biology*, 4 (2), R8.

Larson, K., Yan, S.J., Tsurumi, A., Liu, J., Zhou, J., Gaur, K., Guo, D., Eickbush, T.H. and Li, W.X., 2012. Heterochromatin formation promotes longevity and represses ribosomal RNA synthesis. *PLoS Genetics*, 8 (1), 1002473.

Le Bourg, É., 2001. A mini-review of the evolutionary theories of aging. Is it the time to accept them?. *Demographic Research*, 4, 1-28.

Lee, M.C., Park, J.C., Yoon, D.S., Han, J., Kang, S., Kamizono, S., Om, A.S., Shin, K.H., Hagiwara, A. and Lee, J.S., 2018. Aging extension and modifications of lipid metabolism in the

monogonont rotifer Brachionus koreanus under chronic caloric restriction. *Scientific Reports*, 8 (1), 1741.

Legan, S.K., Rebrin, I., Mockett, R.J., Radyuk, S.N., Klichko, V.I., Sohal, R.S. and Orr, W.C., 2008. Overexpression of glucose-6-phosphate dehydrogenase extends the life span of Drosophila melanogaster. *Journal of Biological Chemistry*, 283 (47), 32492-32499.

Li, Y. and Tower, J., 2009. Adult-specific over-expression of the Drosophila genes magu and hebe increases life span and modulates late-age female fecundity. *Molecular Genetics and Genomics*, 281 (2), 147-162.

Li, D., Zhang, Q., Zio, E., Havlin, S. and Kang, R., 2015. Network reliability analysis based on percolation theory. *Reliability Engineering & System Safety*, 142, 556-562.

Li, L., Lyu, X., Hou, C., Takenaka, N., Nguyen, H.Q., Ong, C.T., Cubeñas-Potts, C., Hu, M., Lei, E.P., Bosco, G. and Qin, Z.S., 2015. Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing. *Molecular Cell*, 58 (2), 216-231.

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O. and Sandstrom, R., 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326 (5950), 289-293.

Lin, Y.J., Seroude, L. and Benzer, S., 1998. Extended life-span and stress resistance in the Drosophila mutant methuselah. *Science*, 282 (5390), 943-946.

Linford, N.J., Bilgir, C., Ro, J. and Pletcher, S.D., 2013. Measurement of lifespan in Drosophila melanogaster. *Journal of Visualized Experiments: JoVE*, (71).

Lips, E.S., Cornelisse, L.N., Toonen, R.F., Min, J.L., Hultman, C.M., Holmans, P.A., O'donovan, M.C., Purcell, S.M., Smit, A.B., Verhage, M. and Sullivan, P.F., 2012. Functional gene group analysis identifies synaptic gene groups as risk factor for schizophrenia. *Molecular Psychiatry*, 17 (10), 996.

Loch, G., Zinke, I., Mori, T., Carrera, P., Schroer, J., Takeyama, H. and Hoch, M., 2017. Antimicrobial peptides extend lifespan in Drosophila. *PloS One*, 12 (5), p.e0176689.

López-Lluch, G. and Navas, P., 2016. Calorie restriction as an intervention in ageing. *The Journal of Physiology*, 594 (8), 2043-2060.

López-Otín, C., Blasco, M.A., Partridge, L., Serrano, M. and Kroemer, G., 2013. The hallmarks of aging. *Cell*, 153 (6), 1194-1217.

Lorand, L., Hsu, L.K., Siefring, G.E. and Rafferty, N.S., 1981. Lens transglutaminase and cataract formation. *Proceedings of the National Academy of Sciences*, 78 (3), 1356-1360.

Lu, F., Guan, H., Gong, B., Liu, X., Zhu, R., Wang, Y., Qian, J., Zhou, T., Lan, X., Wang, P. and Lin, Y., 2014. Genetic variants in PVRL2-TOMM40-APOE region are associated with human longevity in a Han Chinese population. *PLoS One*, 9 (6), e99580.

Luckinbill, L.S., Riha, V., Rhine, S. and Grudzien, T.A., 1990. The role of glucose-6-phosphate dehydrogenase in the evolution of longevity in Drosophila melanogaster. *Heredity*, 65 (1), 29.

Mackay, T.F., Richards, S., Stone, E.A., Barbadilla, A., Ayroles, J.F., Zhu, D., Casillas, S., Han, Y., Magwire, M.M., Cridland, J.M. and Richardson, M.F., 2012. The Drosophila melanogaster genetic reference panel. *Nature*, 482 (7384), 173.

Magwire, M.M., Yamamoto, A., Carbone, M.A., Roshina, N.V., Symonenko, A.V., Pasyukova, E.G., Morozova, T.V. and Mackay, T.F., 2010. Quantitative and molecular genetic analyses of mutations increasing Drosophila life span. *PLoS Genetics*, 6 (7), 1001037.

Makinodan, T. and Kay, M.M., 1980. Age influence on the immune system. In *Advances in Immunology* (Vol. 29, 287-330). Academic Press.

Maksimenko, O., Bartkuhn, M., Stakhov, V., Herold, M., Zolotarev, N., Jox, T., Buxa, M.K., Kirsch, R., Bonchuk, A., Fedotova, A. and Kyrchanova, O., 2015. Two new insulator proteins, Pita and ZIPIC, target CP190 to chromatin. *Genome Research*, 25 (1), 89-99.

Masoro, E.J., 1995. Aging: current concepts. *Handbook of physiology*, 11, 3-21.

McCay, C.M., Crowell, M.F. and Maynard, L.A., 1935. The effect of retarded growth upon the length of life span and upon the ultimate body size: one figure. *The Journal of Nutrition*, 10 (1), 63-79.

McQuibban, G.A., Lee, J.R., Zheng, L., Juusola, M. and Freeman, M., 2006. Normal mitochondrial dynamics requires rhomboid-7 and affects Drosophila lifespan and neuronal function. *Current Biology*, 16 (10), 982-989.

Medawar, P.B., 1952. *An unsolved problem of biology* (24). College.

Meiniel, A., Meiniel, R., Gonçalves-Mendes, N., Creveaux, I., Didier, R. and Dastugue, B., 2003. The thrombospondin type 1 repeat (TSR) and neuronal differentiation: roles of SCO-spondin oligopeptides on neuronal cell types and cell lines. *International Review of Cytology*, *230*, 2-41.

Miller, R.A., 1996. The aging immune system: primer and prospectus. *Science*, 273 (5271), 70-74.

Mirkin, S.M., 2007. Expandable DNA repeats and human disease. *Nature*, 447 (7147), 932.

Mitteldorf, J., 2010. Aging is not a process of wear and tear. *Rejuvenation Research*, 13 (2-3), 322-326.

Montana, E.S. and Littleton, J.T., 2006. Expression profiling of a hypercontraction-induced myopathy in Drosophila suggests a compensatory cytoskeletal remodeling response. *Journal of Biological Chemistry*, 281 (12), 8100-8109.

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

Montesanto, A., Latorre, V., Giordano, M., Martino, C., Domma, F. and Passarino, G., 2011. The genetic component of human longevity: analysis of the survival advantage of parents and siblings of Italian nonagenarians. *European Journal of Human Genetics*, 19 (8), 882.

Mountz, J.D., Zhou, T., Su, X., Wu, J. and Cheng, J., 1996. The role of programmed cell death as an emerging new concept for the pathogenesis of autoimmune diseases. *Clinical Immunology and Immunopathology*, 80 (3), S2-S14.

Nagaraj, R.H., Sell, D.R., Prabhakaram, M., Ortwerth, B.J. and Monnier, V.M., 1991. High correlation between pentosidine protein crosslinks and pigmentation implicates ascorbate oxidation in human lens senescence and cataractogenesis. *Proceedings of the National Academy of Sciences*, 88 (22), 10257-10261.

Newman, A.B., Walter, S., Lunetta, K.L., Garcia, M.E., Slagboom, P.E., Christensen, K., Arnold, A.M., Aspelund, T., Aulchenko, Y.S., Benjamin, E.J. and Christiansen, L., 2010. A meta-analysis of four genome-wide association studies of survival to age 90 years or older: the Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, 65 (5), 478-487.

Newman, M. Networks. Oxford University Press. 2018.

Nielsen, M.D., Luo, X., Biteau, B., Syverson, K. and Jasper, H., 2008. 14-3-3ε antagonizes FoxO to control growth, apoptosis and longevity in Drosophila. *Aging Cell*, 7 (5), 688-699.

Nijnik, A., Woodbine, L., Marchetti, C., Dawson, S., Lambe, T., Liu, C., Rodrigues, N.P., Crockford, T.L., Cabuy, E., Vindigni, A. and Enver, T., 2007. DNA repair is limiting for haematopoietic stem cells during ageing. *Nature*, 447 (7145), 686.

Opsahl, T. and Panzarasa, P., 2009. Clustering in weighted networks. *Social Networks*, 31 (2), 155-163.

Orr, W.C. and Sohal, R.S., 1994. Extension of life-span by overexpression of superoxide dismutase and catalase in Drosophila melanogaster. *Science*, 263 (5150), 1128-1130.

Paaby, A.B. and Schmidt, P.S., 2009. Dissecting the genetics of longevity in Drosophila melanogaster. *Fly*, 3 (1), 29-38.

Page, L., Brin, S., Motwani, R. and Winograd, T., 1999. *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab.

Papaconstantinou, M., Wu, Y., Pretorius, H.N., Singh, N., Gianfelice, G., Tanguay, R.M., Campos, A.R. and Bédard, P.A., 2005. Menin is a regulator of the stress response in Drosophila melanogaster. *Molecular and Cellular Biology*, 25 (22), 9960-9972.

Parker, S.C. and Tullius, T.D., 2011. DNA shape, genetic codes, and evolution. *Current Opinion in Structural Biology*, 21 (3), 342-347.

Partridge, L., Piper, M.D. and Mair, W., 2005. Dietary restriction in Drosophila. *Mechanisms of Ageing and Development*, 126 (9), 938-950.

Pe'er, I., de Bakker, P.I., Maller, J., Yelensky, R., Altshuler, D. and Daly, M.J., 2006. Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature Genetics*, 38 (6), 663.

Peleg, S., Feller, C., Forne, I., Schiller, E., Sévin, D.C., Schauer, T., Regnard, C., Straub, T., Prestel, M., Klima, C. and Nogueira, M.S., 2016. Life span extension by targeting a link between metabolism and histone acetylation in Drosophila. *EMBO Reports*, 17 (3), 455-469.

Perls, T.T., Wilmoth, J., Levenson, R., Drinkwater, M., Cohen, M., Bogan, H., Joyce, E., Brewster, S., Kunkel, L. and Puca, A., 2002. Life-long sustained mortality advantage of siblings of centenarians. *Proceedings of the National Academy of Sciences*, 99 (12), 8442-8447.

Penton, A.L., Leonard, L.D. and Spinner, N.B., 2012, June. Notch signaling in human development and disease. *Seminars in Cell & Developmental Biology*, 23 (4), 450-457. Academic Press.

Pilling, L.C., Kuo, C.L., Sicinski, K., Tamosauskaite, J., Kuchel, G.A., Harries, L.W., Herd, P., Wallace, R., Ferrucci, L. and Melzer, D., 2017. Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. *Aging (Albany NY)*, 9 (12), 2504.

Piper, M.D. and Partridge, L., 2007. Dietary restriction in Drosophila: delayed aging or experimental artefact?. *PLoS Genetics*, 3 (4).

Plyusnina, E.N., Shaposhnikov, M.V. and Moskalev, A.A., 2011. Increase of Drosophila melanogaster lifespan due to D-GADD45 overexpression in the nervous system. *Biogerontology*, 12 (3), 211-226.

Proshkina, E.N., Shaposhnikov, M.V., Sadritdinova, A.F., Kudryavtseva, A.V. and Moskalev, A.A., 2015. Basic mechanisms of longevity: A case study of Drosophila pro-longevity genes. *Ageing Research Reviews*, 24, 218-231.

Puca, A.A., Daly, M.J., Brewster, S.J., Matise, T.C., Barrett, J., Shea-Drinkwater, M., Kang, S., Joyce, E., Nicoli, J., Benson, E. and Kunkel, L.M., 2001. A genome-wide scan for linkage to human exceptional longevity identifies a locus on chromosome 4. *Proceedings of the National Academy of Sciences*, 98 (18), 10505-10508.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J. and Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81 (3), 559-575.

Rahman, M., Haberman, A., Tracy, C., Ray, S. and Krämer, H., 2012. Drosophila mauve mutants reveal a role of LYST homologs late in the maturation of phagosomes and autophagosomes. *Traffic*, 13 (12), 1680-1692.

Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K.C., Grüning, B.A., Villaveces, J., Habermann, B., Akhtar, A. and Manke, T., 2018. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature Communications*, 9 (1), 189.

Roeder, K., Bacanu, S.A., Wasserman, L. and Devlin, B., 2006. Using linkage genome scans to improve power of association in genome scans. *The American Journal of Human Genetics*, 78 (2), 243-252.

Rogina, B., Reenan, R.A., Nilsen, S.P. and Helfand, S.L., 2000. Extended life-span conferred by cotransporter gene mutations in Drosophila. *Science*, 290 (5499), 2137-2140.

Rogina, B., Helfand, S.L. and Frankel, S., 2002. Longevity regulation by Drosophila Rpd3 deacetylase and caloric restriction. *Science*, 298 (5599), 1745.

Rogina, B. and Helfand, S.L., 2004. Sir2 mediates longevity in the fly through a pathway related to calorie restriction. *Proceedings of the National Academy of Sciences*, 101 (45), 15998-16003.

Rogina, B. and Helfand, S.L., 2013. Indy mutations and Drosophila longevity. *Frontiers in Genetics*, 4, 47.

Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S. and Honig, B., 2009. The role of DNA shape in protein–DNA recognition. *Nature*, 461 (7268), 1248.

Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B. and Mann, R.S., 2010. Origins of specificity in protein-DNA recognition. *Annual Review of Biochemistry*, 79, 233-269.

Rouleau, S.G., Beaudoin, J.D., Bisaillon, M. and Perreault, J.P., 2014. Small antisense oligonucleotides against G-quadruplexes: specific mRNA translational switches. *Nucleic Acids Research*, 43 (1), 595-606.

Ruan, H., Tang, X.D., Chen, M.L., Joiner, M.A., Sun, G., Brot, N., Weissbach, H., Heinemann, S.H., Iverson, L., Wu, C.F. and Hoshi, T., 2002. High-quality life extension by the enzyme peptide methionine sulfoxide reductase. *Proceedings of the National Academy of Sciences*, 99 (5), 2748-2753.

Sahlén, P., Abdullayev, I., Ramsköld, D., Matskova, L., Rilakovic, N., Lötstedt, B., Albert, T.J., Lundeberg, J. and Sandberg, R., 2015. Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. *Genome Biology*, 16 (1), 156.

Salmon, A.B., Richardson, A. and Pérez, V.I., 2010. Update on the oxidative stress theory of aging: does oxidative stress play a role in aging or healthy aging?. *Free Radical Biology and Medicine*, 48 (5), 642-655.

Sampson, C.J., Valanne, S., Fauvarque, M.O., Hultmark, D., Rämet, M. and Williams, M.J., 2012. The RhoGEF Zizimin-related acts in the Drosophila cellular immune response via the Rho GTPases Rac2 and Cdc42. *Developmental & Comparative Immunology*, 38 (1), 160-168.

Sarkar, S.K. and Chang, C.K., 1997. The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association*, 92 (440), 1601-1608.

Schneider, T.D. and Stephens, R.M., 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18 (20), 6097-6100.

Schoenmaker, M., de Craen, A.J., de Meijer, P.H., Beekman, M., Blauw, G.J., Slagboom, P.E. and Westendorp, R.G., 2006. Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. *European Journal of Human Genetics*, 14 (1), 79.

Sebastiani, P., Solovieff, N., DeWan, A.T., Walsh, K.M., Puca, A., Hartley, S.W., Melista, E., Andersen, S., Dworkis, D.A., Wilk, J.B. and Myers, R.H., 2012. Genetic signatures of exceptional longevity in humans. *PloS One*, 7 (1), e29848.

Segrè, A.V., Groop, L., Mootha, V.K., Daly, M.J., Altshuler, D., Diagram Consortium and Magic Investigators, 2010. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genetics*, 6 (8), e1001058.

Seong, K.H., Matsuo, T., Fuyama, Y. and Aigaki, T., 2001. Neural-specific overexpression of Drosophila Plenty of SH3s (DPOSH) extends the longevity of adult flies. *Biogerontology*, 2 (4), 271-281.

Seong, K.H., Ogashiwa, T., Matsuo, T., Fuyama, Y. and Aigaki, T., 2001. Application of the gene search system to screen for longevity genes in Drosophila. *Biogerontology*, 2 (3), 209-217.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A. and Cavalli, G., 2012. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, 148 (3), 458-472.

Shaposhnikov, M., Proshkina, E., Shilova, L., Zhavoronkov, A. and Moskalev, A., 2015. Lifespan and stress resistance in Drosophila with overexpressed DNA repair genes. *Scientific Reports*, 5, 15299.

Shastry, B.S., 2009. SNPs: Impact on gene function and phenotype. In *Single Nucleotide Polymorphisms.* Humana Press, Totowa, NJ, 3-22.

Shepherd, J.C., Walldorf, U., Hug, P. and Gehring, W.J., 1989. Fruit flies with additional expression of the elongation factor EF-1 alpha live longer. *Proceedings of the National Academy of Sciences*, 86 (19), 7520-7521.

Shorter, J., Couch, C., Huang, W., Carbone, M.A., Peiffer, J., Anholt, R.R. and Mackay, T.F., 2015. Genetic architecture of natural variation in Drosophila melanogaster aggressive behavior. *Proceedings of the National Academy of Sciences*, 112 (27), 3555-3563.

Simes, R.J., 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73 (3), 751-754.

Simonsen, A., Cumming, R.C., Lindmo, K., Galaviz, V., Cheng, S., Rusten, T.E. and Finley, K.D., 2007. Genetic modifiers of the Drosophila blue cheese gene link defects in lysosomal transport with decreased life span and altered ubiquitinated-protein profiles. *Genetics*, 176 (2), 1283-1297.

Skytthe, A., Pedersen, N.L., Kaprio, J., Stazi, M.A., vB Hjelmborg, J., Iachine, I., Vaupel, J.W. and Christensen, K., 2003. Longevity studies in GenomEUtwin. *Twin Research and Human Genetics*, *6* (5), 448-454.

Slack, C., Alic, N., Foley, A., Cabecinha, M., Hoddinott, M.P. and Partridge, L., 2015. The Ras-Erk-ETS-signaling pathway is a drug target for longevity. *Cell*, 162 (1), 72-83.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H.J. and Mann, R.S., 2011. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell*, 147 (6), 1270-1282.

Solana, R., Pawelec, G. and Tarazona, R., 2006. Aging and innate immunity. *Immunity*, 24 (5), 491-494.

Sun, J. and Tower, J., 1999. FLP recombinase-mediated induction of Cu/Zn-superoxide dismutase transgene expression can extend the life span of adult Drosophila melanogaster flies. *Molecular and Cellular Biology*, 19 (1), 216-228.

Sun, X., Wheeler, C.T., Yolitz, J., Laslo, M., Alberico, T., Sun, Y., Song, Q. and Zou, S., 2014. A mitochondrial ATP synthase subunit interacts with TOR signaling to modulate protein homeostasis and lifespan in Drosophila. *Cell Reports*, 8 (6), 1781-1792.

Symphorien, S. and Woodruff, R.C., 2003. Effect of DNA repair on aging of transgenic Drosophila melanogaster: I. mei-41 locus. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 58 (9), 782-787.

Tarazona, R., Solana, R., Ouyang, Q. and Pawelec, G., 2002. Basic biology and clinical impact of immunosenescence. *Experimental Gerontology*, 37 (2-3), 183-189.

Tatar, M., Kopelman, A., Epstein, D., Tu, M.P., Yin, C.M. and Garofalo, R.S., 2001. A mutant Drosophila insulin receptor homolog that extends life-span and impairs neuroendocrine function. *Science*, *292* (5514), 107-110.

Tatar, M., Bartke, A. and Antebi, A., 2003. The endocrine regulation of aging by insulin-like signals. *Science*, 299 (5611), 1346-1351.

Toivonen, J.M., Walker, G.A., Martinez-Diaz, P., Bjedov, I., Driege, Y., Jacobs, H.T., Gems, D. and Partridge, L., 2007. No influence of Indy on lifespan in Drosophila after correction for genetic and cytoplasmic background effects. *PLoS Genetics*, 3 (6), e95.

Tran, H., Brunet, A., Grenier, J.M., Datta, S.R., Fornace, A.J., DiStefano, P.S., Chiang, L.W. and Greenberg, M.E., 2002. DNA repair pathway stimulated by the forkhead transcription factor FOXO3a through the Gadd45 protein. *Science*, 296 (5567), 530-534.

Tsai, H.Z., Lin, R.K. and Hsieh, T.S., 2016. Drosophila mitochondrial topoisomerase III alpha affects the aging process via maintenance of mitochondrial function and genome integrity. *Journal of Biomedical Science*, 23 (1), 38.

Ulgherait, M., Rana, A., Rera, M., Graniel, J. and Walker, D.W., 2014. AMPK modulates tissue and organismal aging in a non-cell-autonomous manner. *Cell Reports*, 8 (6), 1767-1780.

Van Bortle, K., Nichols, M.H., Li, L., Ong, C.T., Takenaka, N., Qin, Z.S. and Corces, V.G., 2014. Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biology*, 15 (5), R82.

Van Heemst, D., 2010. Insulin, IGF-1 and longevity. *Aging and Disease*, 1 (2), 147.

Walford, R.L., 1964. The immunologic theory of aging. *The Gerontologist*, 4 (4), 195-197.

Wang, B., Goode, J., Best, J., Meltzer, J., Schilman, P.E., Chen, J., Garza, D., Thomas, J.B. and Montminy, M., 2008. The insulin-regulated CREB coactivator TORC promotes stress resistance in Drosophila. *Cell Metabolism*, 7 (5), 434-444.

Wang, G. and Vasquez, K.M., 2014. Impact of alternative DNA structures on DNA damage, DNA repair, and genetic instability. *DNA repair*, 19, 143-151.

Wang, H.D., Kazemi-Esfarjani, P. and Benzer, S., 2004. Multiple-stress analysis for isolation of Drosophila longevity genes. *Proceedings of the National Academy of Sciences*, 101 (34), 12610-12615.

Wang, Z., Lyons, B., Truscott, R.J. and Schey, K.L., 2014. Human protein aging: modification and crosslinking through dehydroalanine and dehydrobutyrine intermediates. *Aging Cell*, 13 (2), 226-234.

Wang, K., Li, M. and Bucan, M., 2007. Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics*, 81 (6), 1278-1283.

Wang, L., Jia, P., Wolfinger, R.D., Chen, X. and Zhao, Z., 2011. Gene set analysis of genome-wide association studies: methodological issues and perspectives. *Genomics*, 98 (1), 1-8.

Wang, M.C., Bohmann, D. and Jasper, H., 2003. JNK signaling confers tolerance to oxidative stress and extends lifespan in Drosophila. *Developmental Cell*, 5 (5), 811-816.

Warner, H.R., 1999. Apoptosis: a two-edged sword in aging. *Annals of the New York Academy of Sciences*, 887 (1), 1-11.

Wanders, R.J. and Waterham, H.R., 2006. Biochemistry of mammalian peroxisomes revisited. *Annu. Rev. Biochem.*, 75, 295-332.

Wasserman, S. and Faust, K., 1994. *Social Network Analysis: Methods and Applications* (Vol. 8). Cambridge University Press.

Watson, J.D. and Crick, F.H.C., 1993. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *JAMA*, 269 (15), 1966-1967.

Watts, D.J. and Strogatz, S.H., 1998. Collective dynamics of 'small world' networks. *Nature*, 393 (6684), 440.

Weindruch, R., Walford, R.L., Fligiel, S. and Guthrie, D., 1986. The retardation of aging in mice by dietary restriction: longevity, cancer, immunity and lifetime energy intake. *The Journal of Nutrition*, 116 (4), 641-654.

Weinert, B.T. and Timiras, P.S., 2003. Invited review: Theories of aging. *Journal of Applied Physiology*, 95 (4), 1706-1716.

Wells, R.D., Dere, R., Hebert, M.L., Napierala, M. and Son, L.S., 2005. Advances in mechanisms of genetic instability related to hereditary neurological diseases. *Nucleic Acids Research*, 33 (12), 3785-3798.

Whitaker, R., Faulkner, S., Miyokawa, R., Burhenn, L., Henriksen, M., Wood, J.G. and Helfand, S.L., 2013. Increased expression of Drosophila Sir 2 extends life span in a dose-dependent manner. *Aging (Albany NY)*, 5 (9), 682.

Whitlock, M.C., 2005. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology*, 18 (5), 1368-1373.

Williams, G.C., 1957. Pleiotropy, natural selection, and the evolution of senescence. *Evolution*, 398-411.

Xue, L., Igaki, T., Kuranaga, E., Kanda, H., Miura, M. and Xu, T., 2007. Tumor suppressor CYLD regulates JNK-induced cell death in Drosophila. *Developmental Cell*, 13 (3), 446-454.

Yaffe, E. and Tanay, A., 2011. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics*, 43 (11), 1059.

Yang, J., Ramos, E. and Corces, V.G., 2012. The BEAF-32 insulator coordinates genome organization and function during the evolution of Drosophila species. *Genome Research*, 22 (11), 2199-2207.

Yao, L., Tak, Y.G., Berman, B.P. and Farnham, P.J., 2014. Functional annotation of colon cancer risk SNPs. *Nature Communications*, 5, 5114.

Zeng, Y., Nie, C., Min, J., Liu, X., Li, M., Chen, H., Xu, H., Wang, M., Ni, T., Li, Y. and Yan, H., 2016. Novel loci and pathways significantly associated with longevity. *Scientific Reports*, 6, 21243.

Zerofsky, M., Harel, E., Silverman, N. and Tatar, M., 2005. Aging of the innate immune response in Drosophila melanogaster. *Aging Cell*, 4 (2), 103-108.

Zhou, Q., Yang, D., Ombrello, A.K., Zavialov, A.V., Toro, C., Zavialov, A.V., Stone, D.L., Chae, J.J., Rosenzweig, S.D., Bishop, K. and Barron, K.S., 2014. Early-onset stroke and vasculopathy associated with mutations in ADA2. *New England Journal of Medicine*, 370 (10), 911-920.

Zhou, Y., Wang, Y., Qiao, S. and Yin, L., 2017. Effects of Apelin on Cardiovascular Aging. *Frontiers in Physiology*, 8, 1035.

Zwarts, L., Broeck, L.V., Cappuyns, E., Ayroles, J.F., Magwire, M.M., Vulsteke, V., Clements, J., Mackay, T.F. and Callaerts, P., 2015. The genetic basis of natural variation in mushroom body size in Drosophila melanogaster. *Nature Communications*, 6, 10115.

APPENDIX

Table S4.1 All genes found to reside in common original regions identified between the extended Synthetic and DGRP GWAS-based networks. Genes with 'long-lived' phenotype are shown in bold.

Gene name	Bin	SNP count Burke	SNP count Ivanov	Significant SNP count Burke	Significant SNP count Ivanov
AdoR	1183	24	536	0	0
AIF	27	0	33	0	0
alrm	1124	1	30	0	0
aop	27	17	399	8	0
Arr2	660	3	60	1	0
Atxn7	30	2	83	1	0
axo	609	67	1747	10	0
Axud1	30	4	75	0	0
Cad88C	989	10	295	0	1
c-cup	23	0	21	0	0
CG10635	619	2	14	0	0
CG11498	1183	11	279	0	0
CG11723	27	3	34	0	0
CG12674	27	2	35	0	0
CG13562	534	3	48	1	0
CG13716	609	1	12	0	0
CG14853	989	6	273	0	0
CG15382	27	1	14	0	0
CG15390	30	0	13	0	0
CG15529	1183	4	67	0	0
CG16995	29	3	42	1	0
CG17234	29	1	3	0	0
CG17237	29	0	19	0	0
CG17239	29	2	21	0	0
CG17242	29	2	33	0	0
CG17646	22	23	501	0	0
CG17648	22	0	19	0	0
CG17650	22	2	15	0	0
CG17652	22	2	31	0	0
CG17658	534	1	26	1	0
CG17660	22	4	70	0	0
CG17712	22	3	36	0	0
CG2812	534	4	54	1	0
CG2970	534	1	56	1	0
CG31028	1183	5	139	0	0
CG31029	1183	2	89	0	0

CG31030	1183	5	79	0	0
CG31437	1124	1	19	0	0
CG31664	23	2	74	0	0
CG31681	29	2	32	0	0
CG31815	208	0	44	0	0
CG31933	23	2	54	0	0
CG31937	22	1	40	0	0
CG31949	29	0	26	0	0
CG32022	660	0	13	0	0
CG32246	609	2	33	1	0
CG34049	29	1	66	0	0
CG3557	30	1	13	0	0
CG3597	30	1	24	0	0
CG3609	30	1	54	0	0
CG3735	534	2	61	1	0
CG4168	190/191	21	784	12	0
CG42540	609	33	855	0	0
CG4259	27	1	24	0	0
CG42658	29	0	29	0	0
CG4267	29	4	83	0	0
CG4270	29	3	27	0	0
CG4271	29	0	8	0	0
CG43230	191	1	27	1	0
CG43750	30	3	61	0	0
CG43880	651	0	10	0	0
CG43965	660	0	3	0	0
CG44094	989	0	5	0	0
CG45072	1183	3	54	0	0
CG45073	1183	0	39	0	0
CG4631	208	4	97	0	0
CG4882	534	2	23	1	0
CG5339	534	0	15	0	0
CG5597	534	1	33	0	0
CG6511	660	0	91	0	0
CG7886	989	13	367	0	1
CG7987	989	2	65	0	0
CG8038	651	1	14	0	0
CG8042	651	3	57	0	0
CG8209	651	1	19	0	0
CG9870	30	3	71	0	0
CG9967	29/30	45	1225	0	1
chinmo	22	44	1142	0	2
Cp18	660	0	18	0	0
cpb	22	4	33	0	0

CR42859	29	2	44	0	0
CR43357	191	0	32	0	0
CR43682	191	2	57	0	0
CR43753	29	6	213	0	0
CR43754	29	0	44	0	0
CR43853	190	0	21	0	0
CR43854	190	0	11	0	0
CR44055	27	1	32	0	0
CR44073	23	1	7	0	0
CR44151	208	0	3	0	0
CR44196	191	1	7	0	0
CR44515	619	0	20	0	0
CR44516	619	0	7	0	0
CR44526	660	1	23	0	0
CR44706	190	0	16	0	0
CR44770	190	2	13	1	0
CR44771	190	0	15	0	0
CR44787	29	0	13	0	0
CR44788	29	0	10	0	0
CR44806	534	0	12	0	0
CR44808	27	0	4	0	0
CR44976	22	2	16	0	0
CR44982	30	0	7	0	0
CR45438	609	0	15	0	0
CR45743	619	0	12	0	0
CR45926	534	1	26	0	0
CR46082	1183	6	90	0	0
CR46112	1183	1	18	0	0
CR46153	989	0	1	0	0
Cul3	191	2	90	1	0
dao	191	5	107	2	0
DCP1	534	0	11	0	0
DIP-delta	619	24	526	3	0
DNA-ligI	534	4	40	1	0
dpr3	27	42	1095	10	0
Eno	22	6	115	0	0
Eogt	30	1	29	0	0
exex	651	3	115	0	0
eys	29/30	43	1190	0	1
frtz	22	6	114	0	0
fzr2	534	1	20	1	0
Galphas	534	2	89	1	0
GlyP	27	4	99	1	0
Gr22a	23	4	61	0	0

Gr22b	23	4	32	1	0
Gr22c	23	1	44	1	0
Gr22d	23	1	42	0	0
Gr22e	23	0	21	0	0
Gr22f	22	3	34	0	0
h	660	3	59	0	0
HDAC1	609	4	76	0	1
His4r	989	0	11	0	0
HP4	651	1	16	1	0
Ir64a	619	8	193	0	1
kcc	534	6	195	2	0
ken	534	4	31	4	0
l(2)35Cc	191	1	21	1	0
l(3)L1231	989	1	159	0	0
Lpt	534	4	113	1	1
Membrin	619	0	16	0	0
mir-2280	23	0	1	0	0
Mlc2	1183	3	55	0	0
mRpL48	22	2	33	0	0
Muc96D	1124	1	13	0	0
Nap1	534	1	36	0	0
nmo	651	74	2128	2	0
Orcokinin	534	3	43	1	0
Pex7	660	5	136	1	0
PHDP	534	2	19	0	0
Pol32	191	3	25	0	0
Ppi1	1183	4	91	0	0
put	989	4	91	0	0
Rab5	30	2	117	0	0
Rim2	22	14	247	1	0
RNaseX25	651	1	29	0	0
robl22E	29	1	20	0	0
Rrp40	22	2	24	0	0
Send1	29	0	18	0	0
Ser12	29	1	8	0	0
SERCA	534	6	153	2	0
Sfp35C	191	0	7	0	0
sima	1183	81	1865	1	3
Src64B	609	38	917	3	0
Srp9	651	3	25	0	0
SrpRbeta	660	0	41	0	0
stumps	989	28	803	1	0
syd	651	10	224	0	0
Taldo	534	4	33	1	0

TBCD	27	4	77	1	0
Tengl1	29	1	22	0	0
tho2	27	4	77	0	0
TM4SF	534	0	11	0	0
Tpi	1183	1	27	0	0
tRNA:L:35C	190	0	0	0	0
UK114	191	0	22	0	0
Upf3	534	0	41	0	0
VGlut	30	21	514	2	0
wry	23	30	650	0	0
yuri	191	6	138	2	0
ZnT35C	191	19	476	5	0

Table S4.2 SNP counts for genes residing in common novel regions between both GWAS dataset networks identified using a clustering coefficient network measure.

Gene	Bin	Number of SNPs recorded in Synthetic GWAS dataset	Number of SNPs recorded in DGRP dataset	Percentage of total number SNPs in Synthetic GWAS dataset	Percentage of total number of SNPs in DGRP dataset
rols	702	76	1884	0.08119051	0.08601263
CR44320	1134	34	711	0.03632207	0.03246018
CR46061	1134	25	466	0.0267074	0.02127489
kek3	195	22	591	0.02350252	0.02698167
Sema-5c	702	18	503	0.01922933	0.02296409
Ald	1134	9	158	0.00961467	0.00721337
oaf	32	4	114	0.00427318	0.00520459
CG6793	702	4	75	0.00427318	0.00342407
CG12290	1134	4	71	0.00427318	0.00324145
Slh	32	4	69	0.00427318	0.00315014
CG3515	32	4	50	0.00427318	0.00228271
CG15263	192	4	50	0.00427318	0.00228271
CR46069	32	4	28	0.00427318	0.00127832
CR43847	1129	3	49	0.00320489	0.00223706
CR43764	192	3	46	0.00320489	0.0021001
CG5886	1129	3	33	0.00320489	0.00150659
esg	192	3	32	0.00320489	0.00146094
CG3528	32	3	24	0.00320489	0.0010957
CG31380	1129	2	50	0.00213659	0.00228271
alpha4GT2	1129	2	45	0.00213659	0.00205444

CG15256	195	2	31	0.00213659	0.00141528
CG15262	192	2	29	0.00213659	0.00132397
Acam	1129	2	24	0.00213659	0.0010957
CG17770	1129	2	24	0.00213659	0.0010957
CG6073	1134	2	24	0.00213659	0.0010957
CG4956	1129	2	22	0.00213659	0.00100439
CG31086	1134	2	15	0.00213659	0.00068481
CR45652	1129	2	11	0.00213659	0.0005022
CR43632	1129	2	10	0.00213659	0.00045654
CR43761	195	1	39	0.0010683	0.00178052
CR45654	1129	1	26	0.0010683	0.00118701
CG15260	192	1	25	0.0010683	0.00114136
CG4960	1129	1	25	0.0010683	0.00114136
BG642167	1129	1	20	0.0010683	0.00091309
CG17195	1129	1	19	0.0010683	0.00086743
CG34130	1134	1	19	0.0010683	0.00086743
CG5024	1129	1	18	0.0010683	0.00082178
CG43993	702	1	16	0.0010683	0.00073047
CG14354	1129	1	16	0.0010683	0.00073047
CR43631	1129	1	15	0.0010683	0.00068481
nht	192	1	14	0.0010683	0.00063916
CR45655	1129	1	12	0.0010683	0.00054785
CR46062	1134	1	12	0.0010683	0.00054785
CR43994	702	1	11	0.0010683	0.0005022
Mst57Dc	1129	1	8	0.0010683	0.00036523
CR44607	32	1	5	0.0010683	0.00022827
CR44949	1129	0	38	0	0.00173486
CG17196	1129	0	26	0	0.00118701
Skadu	195	0	20	0	0.00091309
CG43638	702	0	18	0	0.00082178
ms(2)35Ci	192	0	17	0	0.00077612
CR44113	32	0	16	0	0.00073047
CR45349	192	0	16	0	0.00073047
CG34129	1134	0	15	0	0.00068481
CG14545	1129	0	14	0	0.00063916
CR43634	1129	0	14	0	0.00063916
CR44112	32	0	12	0	0.00054785
CR45656	1129	0	11	0	0.0005022
Mst57Da	1129	0	10	0	0.00045654
CG31093	1129	0	9	0	0.00041089
CR45653	1129	0	8	0	0.00036523
CR45728	195	0	4	0	0.00018262

CR45657	1129	0	4	0	0.00018262
Sfp96F	1129	0	4	0	0.00018262
Mst57Db	1129	0	3	0	0.00013696
CR43633	1129	0	2	0	9.1309E-05

Table S4.3 SNP counts for genes residing in common novel regions between both GWAS dataset networks identified using a PageRank network measure.

Gene	Bin	Number of SNPs recorded in Synthetic GWAS dataset	Number of SNPs recorded in DGRP dataset	Percentage of total number SNPs in Synthetic GWAS dataset	Percentage of total number of SNPs in DGRP dataset
sima	1182	81	1865	0.08653199	0.085145199
pnt	1097	57	1316	0.060892882	0.060081009
CG2970	534	38	1107	0.040595255	0.038036322
stumps	989	28	803	0.029912293	0.036660372
AdoR	1183	24	536	0.025639108	0.018416864
CG17646	22	23	501	0.024570812	0.022872785
CG4467	1097	22	609	0.023502516	0.020925131
CG31038	1180	22	442	0.023502516	0.015187041
VGlut	30	21	514	0.02243422	0.023466291
kcc	534	19	418	0.020297627	0.014362405
CG10904	535	17	568	0.018161035	0.025931621
orb	1097	17	245	0.018161035	0.011185294
Axn	1182	16	403	0.017092739	0.013847008
Adk2	535	15	390	0.016024443	0.01340033
DNA-ligI	534	14	381	0.014956146	0.017394274
Rim2	22	14	247	0.014956146	0.011276603
CG7886	989	13	367	0.01388785	0.012610055
Nap1	534	12	311	0.012819554	0.010685905
EbpIII	535	12	279	0.012819554	0.012737539
CG11498	1183	11	279	0.011751258	0.012737539
Mgat2	1182	11	255	0.011751258	0.011641837
Nf1	1131	11	200	0.011751258	0.009130852
Cad88C	989	10	295	0.010682962	0.013468007
CG17658	534	10	277	0.010682962	0.009517671
CG3121	535	10	260	0.010682962	0.011870108
CG42261	1131	10	241	0.010682962	0.011002677
CG9743	1184	10	218	0.010682962	0.009952629
Alas	535	9	210	0.009614666	0.009587395

Gadd34	535	9	186	0.009614666	0.008491693
CG31036	1180	8	144	0.008546369	0.006574214
mr	535	8	136	0.008546369	0.00620898
trp	1180	7	163	0.007478073	0.007441645
CG14853	989	6	273	0.006409777	0.009380231
CG9747	1184	6	164	0.006409777	0.007487299
CR44806	534	6	137	0.006409777	0.006254634
Zip99C	1180	6	131	0.006409777	0.005980708
CG15528	1182	6	120	0.006409777	0.004123179
Eno	22	6	115	0.006409777	0.00525024
frtz	22	6	114	0.006409777	0.00391702
CR46082	1183	6	90	0.006409777	0.004108884
CG9733	1184	5	145	0.005341481	0.006619868
CG31028	1183	5	139	0.005341481	0.006345942
Pex7	660	5	136	0.005341481	0.004672936
CG5339	534	5	114	0.005341481	0.005204586
CG31030	1183	5	79	0.005341481	0.002714426
wda	1097	5	76	0.005341481	0.002611346
CR45350	535	4	148	0.004273185	0.005085254
Phm	535	4	116	0.004273185	0.003985739
CR44046	1182	4	115	0.004273185	0.00525024
put	989	4	91	0.004273185	0.003126744
Ppi1	1183	4	91	0.004273185	0.003126744
CG5597	534	4	87	0.004273185	0.002989304
CG11388	535	4	80	0.004273185	0.003652341
CG30178	535	4	77	0.004273185	0.003515378
Axud1	30	4	75	0.004273185	0.002576987
CG17660	22	4	70	0.004273185	0.003195798
CG1983	1184	4	69	0.004273185	0.003150144
CG15529	1183	4	67	0.004273185	0.002302108
CG15530	1184	4	67	0.004273185	0.003058836
CG16787	535	4	59	0.004273185	0.002693601
RpS8	1180	4	58	0.004273185	0.00199287
CG3163	535	4	53	0.004273185	0.001821071
Cog7	1180	4	50	0.004273185	0.002282713
cpb	22	4	33	0.004273185	0.001133874
CG4882	534	4	31	0.004273185	0.001065154
CG34133	1180	3	95	0.003204889	0.004337155
boss	1131	3	92	0.003204889	0.004200192
HSPC300	535	3	91	0.003204889	0.004154538
CG13562	534	3	90	0.003204889	0.004108884
CG9737	1184	3	84	0.003204889	0.002886225

CG18404	1184	3	79	0.003204889	0.002714426
CG9870	30	3	71	0.003204889	0.003241453
EloA	1097	3	63	0.003204889	0.002164669
CG43750	30	3	61	0.003204889	0.002095949
CG13827	1097	3	61	0.003204889	0.002095949
Arr2	660	3	60	0.003204889	0.002739256
h	660	3	59	0.003204889	0.002693601
Mlc2	1183	3	55	0.003204889	0.002510984
CG15531	1184	3	55	0.003204889	0.002510984
CG45072	1183	3	54	0.003204889	0.00246533
Sry-alpha	1182	3	53	0.003204889	0.001821071
Sry-beta	1182	3	53	0.003204889	0.002419676
CG4434	1097	3	39	0.003204889	0.001340033
CG17712	22	3	36	0.003204889	0.001236954
CG15515	1180	3	36	0.003204889	0.001643553
Gr22f	22	3	34	0.003204889	0.001552245
Capa	1180	3	31	0.003204889	0.001415282
CG7943	1182	3	31	0.003204889	0.001065154
CG34213	535	3	27	0.003204889	0.001232665
Rab5	30	2	117	0.002136592	0.004020099
CG31029	1183	2	89	0.002136592	0.004063229
Atxn7	30	2	83	0.002136592	0.003789304
CG4324	535	2	63	0.002136592	0.002164669
CG34299	1184	2	63	0.002136592	0.002164669
Lpt	534	2	61	0.002136592	0.00278491
RpS7	1184	2	61	0.002136592	0.00278491
CG4449	1097	2	58	0.002136592	0.002647947
Sox14	535	2	52	0.002136592	0.002374022
CG7946	1182	2	52	0.002136592	0.002374022
CG9682	1184	2	50	0.002136592	0.001717991
Fmo-1	535	2	45	0.002136592	0.001546192
dgt1	1180	2	45	0.002136592	0.001546192
Bet5	1184	2	38	0.002136592	0.001305673
mRpL48	22	2	33	0.002136592	0.001133874
CG17083	1097	2	33	0.002136592	0.001506591
CG6763	1097	2	33	0.002136592	0.001506591
CG17652	22	2	31	0.002136592	0.001065154
CR31032	1180	2	30	0.002136592	0.001369628
CR46111	1182	2	28	0.002136592	0.000962075
Rrp40	22	2	24	0.002136592	0.000824636
spn-A	1182	2	24	0.002136592	0.001095702
CecB	1184	2	24	0.002136592	0.000824636

Upf3	534	2	18	0.002136592	0.000821777
Cec-Psi1	1184	2	18	0.002136592	0.000618477
CR44976	22	2	16	0.002136592	0.000730468
CG2812	534	2	16	0.002136592	0.000730468
E(spl)mbeta-HLH	1131	2	16	0.002136592	0.000549757
CG17650	22	2	15	0.002136592	0.000684814
l(3)L1231	989	1	159	0.001068296	0.007259028
CG3609	30	1	54	0.001068296	0.00246533
CG31937	22	1	40	0.001068296	0.00182617
CG43448	1184	1	40	0.001068296	0.00182617
Cdc16	1097	1	37	0.001068296	0.001689208
alpha-Man-Ic	1180	1	37	0.001068296	0.001271313
CG30172	535	1	32	0.001068296	0.001099514
janA	1182	1	30	0.001068296	0.001369628
Eogt	30	1	29	0.001068296	0.001323974
mir-1009	535	1	29	0.001068296	0.000996435
Unc-89	535	1	29	0.001068296	0.001323974
Chi	535	1	28	0.001068296	0.001278319
CG43116	1131	1	28	0.001068296	0.000962075
Tpi	1183	1	27	0.001068296	0.001232665
Pask	535	1	25	0.001068296	0.001141357
CG3597	30	1	24	0.001068296	0.000824636
CG14540	1131	1	24	0.001068296	0.000824636
CR44526	660	1	23	0.001068296	0.000790276
CG15522	1180	1	23	0.001068296	0.000790276
CG7824	1180	1	23	0.001068296	0.000790276
CG15514	1180	1	22	0.001068296	0.000755916
CecC	1184	1	22	0.001068296	0.001004394
DCP1	534	1	21	0.001068296	0.000721556
Taldo	534	1	21	0.001068296	0.00095874
CR46112	1183	1	18	0.001068296	0.000618477
CG13566	535	1	16	0.001068296	0.000549757
CG3860	535	1	16	0.001068296	0.000549757
ZIPIC	1182	1	16	0.001068296	0.000730468
CecA2	1184	1	15	0.001068296	0.000684814
E(spl)mdelta-HLH	1131	1	14	0.001068296	0.00063916
CG3557	30	1	13	0.001068296	0.000593505
CR45926	534	1	13	0.001068296	0.000593505
Orcokinin	534	1	13	0.001068296	0.000593505
E(spl)malpha-BFM	1131	1	12	0.001068296	0.000547851

Sry-delta	1182	1	9	0.001068296	0.000309238
CG3803	535	1	8	0.001068296	0.000365234
CG6511	660	0	91	0	0.003126744
SrpRbeta	660	0	41	0	0.001871825
CG45073	1183	0	39	0	0.001340033
Galphas	534	0	38	0	0.001734862
CecA1	1184	0	34	0	0.001168234
fzr2	534	0	31	0	0.001065154
CG42557	1180	0	29	0	0.000996435
CG42558	1180	0	29	0	0.001323974
CG3907	535	0	26	0	0.001187011
CR45037	1131	0	26	0	0.001187011
CG15517	1180	0	26	0	0.001187011
CG34317	1182	0	26	0	0.001187011
CG7950	1182	0	26	0	0.000893355
CG3065	535	0	22	0	0.000755916
CG3735	534	0	21	0	0.00095874
TM4SF	534	0	21	0	0.000721556
CG15526	1182	0	21	0	0.00095874
alpha-catenin-related	535	0	20	0	0.000687196
CG17648	22	0	19	0	0.000652837
Cp18	660	0	18	0	0.000821777
E(spl)mgamma-HLH	1131	0	18	0	0.000618477
RpL32	1182	0	18	0	0.000821777
SERCA	534	0	17	0	0.000584117
thoc5	535	0	17	0	0.000584117
janB	1182	0	16	0	0.000549757
RpS28a	1182	0	16	0	0.000549757
Anp	1184	0	16	0	0.000730468
CR44262	1184	0	15	0	0.000515397
Cec2	1184	0	14	0	0.00063916
CG15390	30	0	13	0	0.000593505
CG32022	660	0	13	0	0.000446678
ocn	1182	0	12	0	0.000412318
His4r	989	0	11	0	0.000377958
PHDP	534	0	10	0	0.000456543
Jon99Ci	1180	0	10	0	0.000456543
CR44982	30	0	7	0	0.000240519
or	535	0	6	0	0.000206159
CR44045	1182	0	6	0	0.000273926
CG44094	989	0	5	0	0.000228271

CR45039	1131	0	5	0	0.000228271
CR45038	1131	0	4	0	0.000137439
CG43965	660	0	3	0	0.000136963
Jon99Ciii	1180	0	3	0	0.000136963
snoRNA:Psi18S-1377c	1184	0	3	0	0.000103079
snoRNA:Psi28S-2626	1184	0	3	0	0.000136963
snoRNA:Or-CD8	1180	0	2	0	6.87196E-05
snoRNA:Psi28S-2149	1184	0	2	0	9.13085E-05
CR46153	989	0	1	0	4.56543E-05
snoRNA:Me28S-A2564	1180	0	1	0	4.56543E-05
snoRNA:Psi18S-1377a	1184	0	1	0	3.43598E-05
snoRNA:Psi18S-1377b	1184	0	1	0	4.56543E-05
snoRNA:Psi18S-1377d	1184	0	1	0	4.56543E-05
snoRNA:Psi18S-1377e	1184	0	1	0	3.43598E-05
Jon99Cii	1180	0	0	0	0
tRNA:CR31023	1184	0	0	0	0
tRNA:CR31383	1184	0	0	0	0

Table S4.4 Novel bins in extended Synthetic GWAS-based network harbouring significant SNPs in DGRP GWAS dataset.

Novel bins in extended Synthetic GWAS-based network harbouring significant SNPs in DGRP GWAS dataset	Chromosome	Significant SNP	P value	SNP harbouring gene
21	2L	1632386	5.90E-08	
	2L	1632388	3.74E-07	
34	2L	2712044	2.56E-05	*CG31690*
46	2L	3618373	2.22E-05	
47	2L	3752571	2.35E-07	*CG10019*
	2L	3746990	1.14E-05	*CG10019*
56	2L	4401879	2.36E-05	*CG33003*
89	2L	7048386	1.59E-05	*milt*
91	2L	7258591	1.54E-05	*Wnt4*
	2L	7252498	2.70E-05	*CG13786*
123	2L	9826990	1.67E-05	*nAChRalpha6*
126	2L	10070707	6.77E-06	*CG44153*
	2L	10068812	9.41E-06	*CG44153*

129	2L	10315909	1.55E-05	*CG4972*
311	2R	1885907	2.08E-05	*Src42A*
340	2R	4308355	7.86E-06	
	2R	4308343	8.41E-06	
529	2R	19425848	2.26E-05	*bw*
562	3L	824939	1.82E-05	
576	3L	1966180	7.14E-06	*SCOT*
618	3L	5319539	1.12E-05	
	3L	5320154	1.46E-05	
	3L	5320475	2.15E-05	
	3L	5320661	2.42E-05	
622	3L	5636181	2.69E-06	*Blimp-1*
668	3L	9310689	2.78E-05	
670	3L	9507749	2.97E-06	*CG33700*
698	3L	11687861	2.56E-05	*CG11652*
699	3L	11792808	5.37E-06	*CG10361*
	3L	11792799	2.19E-05	*CG10361*
729	3L	14162655	2.93E-05	
736	3L	14781414	1.45E-06	*Tdrd3*
	3L	14778725	3.50E-06	*bmm*
	3L	14778027	3.71E-06	*bmm*
	3L	14780164	4.00E-06	*Tdrd3*
774	3L	17762728	1.13E-05	
778	3L	18140585	6.51E-06	
787	3L	18810814	4.36E-06	*CG14073*
788	3L	18934159	1.06E-05	*CG32204*
873	3R	1174299	1.79E-05	*mtd*
1045	3R	14921157	1.15E-05	*ATPsynD*
1063	3R	16347543	2.12E-05	
	3R	16347541	3.06E-05	
1091	3R	18577501	5.63E-06	*Usp12-46, CG7029*
1097	3R	19071977	5.64E-07	*CG4467*
	3R	19082073	2.52E-05	*wda*
1120	3R	20944700	9.10E-06	*CG31108, CG31510*
1131	3R	21833264	1.49E-05	
	3R	21833263	1.62E-05	
1132	3R	21913681	1.04E-05	*dysf*
1152	3R	23482833	9.26E-06	*Mlc1*
1168	3R	24748071	1.19E-05	*Doa*
	3R	24748001	1.73E-05	*Doa*
1173	3R	25189263	1.05E-05	
1178	3R	25562159	5.27E-06	*CG7601*
	3R	25562054	1.33E-05	*CG7601*
	3R	25526160	2.94E-05	

1199	3R	27244251	1.36E-05	*Gprk2*
	3R	27245123	2.80E-05	*Gprk2*
1231	X	604933	8.31E-06	*sdk*
	X	601759	1.98E-05	*sdk*
1261	X	2997707	1.66E-05	*kirre*
	X	2997709	1.81E-05	*kirre*
1485	X	20940365	4.95E-06	*bves*

Table S4.5 Novel bins in extended DGRP GWAS-based network harbouring significant SNPs in Synthetic GWAS dataset.

Novel bins in extended DGRP GWAS-based network harbouring significant SNPs in Synthetic GWAS dataset	Chromosome	Significant SNP	D value	SNP harbouring gene
11	2L	826143	7.913430734	*dock*
	2L	830057	8.139679995	*dock*
	2L	837910	8.316410473	*drongo*
	2L	835407	8.387239025	*drongo*
12	2L	952524	8.079787385	*CG4341*
	2L	931835	8.14332228	*CG4341*
	2L	927174	8.813519969	*CG4341*
	2L	916065	9.470033426	*GluRIIC*
14	2L	1105961	7.996222372	*mtRNApol*
	2L	1099941	8.208956173	*CG4629*
	2L	1014551	9.237906397	*IA-2*
15	2L	1123177	8.667261632	*Pino*
	2L	1155301	7.901374128	*capt*
	2L	1131236	7.983946876	*CG4552*
	2L	1178569	7.996034724	*CG4896*
	2L	1151212	8.559932329	*Vps29*
	2L	1148233	8.930734981	*l(2)10685*
17	2L	1283879	8.135716428	*robo3*
24	2L	1904413	7.948992874	*CG7337*
	2L	1911695	8.045435095	*CG7337*
	2L	1889232	8.156835797	*CG7337*
	2L	1917529	8.373709195	*CG7337*
	2L	1874447	8.406752111	*CG31663*
	2L	1901024	8.468424042	*CG7337*
	2L	1915485	8.73900548	*CG7337*
25	2L	1929405	8.026218568	*CG7337*
	2L	1934894	8.187206607	*CG7337*
	2L	1936961	8.241995114	*CG7337*

	2L	1920199	8.405161345	*CG7337*
	2L	2005275	7.984234108	*CG33543*
	2L	1929405	8.026218568	*CG7337*
	2L	1911695	8.045435095	*CG7337*
	2L	1950350	8.117007768	*erm*
	2L	1889232	8.156835797	*CG7337*
	2L	1934894	8.187206607	*CG7337*
	2L	1936961	8.241995114	*CG7337*
	2L	1955401	8.338093138	*erm*
	2L	1917529	8.373709195	*CG7337*
	2L	1961205	8.397049762	*erm*
	2L	1920199	8.405161345	*CG7337*
	2L	1988867	8.449057083	*Npc2a*
	2L	1901024	8.468424042	*CG7337*
	2L	1945334	8.610575689	*CG15357*
	2L	1958746	8.714808353	*erm*
	2L	1915485	8.73900548	*CG7337*
	2L	1998001	9.087747357	*CG33543*
	2L	1987099	9.168572049	*Got2*
	2L	1976747	9.326228485	*CG15356*
	2L	1985412	9.433835942	*Got2*
	2L	1978896	9.738595965	*CG15356*
	2L	1983091	9.806585669	*CG7289*
	2L	1981203	9.820670657	*CG15362*
	2L	1998001	9.087747357	*CG33543*
26	2L	2005275	7.984234108	*CG33543*
	2L	2026220	8.030508805	*CG4238*
	2L	2071535	8.113756722	*dpr3*
	2L	2022829	8.179535278	*CG4238*
	2L	2029663	8.345478443	*CG4238*
	2L	2039069	8.541166134	*Su(dx)*
	2L	2046036	8.764865447	*Kebab*
	2L	2012579	8.774584291	*CG4238*
	2L	2068318	8.789641927	*dpr3*
	2L	2015464	9.03661616	*CG4238*
	2L	2018025	9.201295403	*CG4238*
	2L	2009049	9.551478447	*Nplp4*
	2L	2071535	8.113756722	*dpr3*
	2L	2068318	8.789641927	*dpr3*
28	2L	2168185	8.284752379	*aop*
	2L	2173842	8.367070889	*aop*
	2L	2164521	8.594890484	*aop*
	2L	2160218	8.869791186	*aop*
	2L	2210836	7.956956677	*CG15385*

	2L	2168185	8.284752379	*aop*
	2L	2173842	8.367070889	*aop*
	2L	2216709	8.398047541	*CG15386, CG42371*
	2L	2164521	8.594890484	*aop*
	2L	2231172	8.615514941	*Sec24CD*
	2L	2160218	8.869791186	*aop*
	2L	2222232	8.891994166	*papi*
	2L	2182110	8.942315968	*CR43751*
	2L	2206990	8.991297737	*CG33124*
	2L	2187134	9.471822997	*CR44066*
	2L	2195103	10.01293322	*CG34172*
	2L	2191209	10.19746239	*CG10874, CR44065*
31	2L	2407844	8.372823989	*VGlut*
	2L	2435403	8.076461663	*dpp*
	2L	2450708	8.130006578	*dpp*
	2L	2407844	8.372823989	*VGlut*
	2L	2431174	9.122685221	*dpp*
	2L	2414822	9.321642147	*CG18641*
	2L	2419104	9.435656194	*CG34447*
	2L	2421989	10.32301113	*CG9886*
93	2L	7363970	8.137811972	*Wnt10*
	2L	7367097	8.179864852	*Wnt10*
128	2L	10216704	7.992298701	*Npc1a*
166	2L	13259566	8.338614244	*CG31730*
187	2L	14959571	9.614838315	*CG42313*
	2L	14969191	9.641481299	*CG42313*
	2L	14912792	9.79416925	*CG42313*
	2L	14965929	9.82393229	*CG42313*
	2L	14921880	10.29006228	*CG42313*
	2L	14926230	10.31176121	*CG42313, CG3491*
	2L	14917849	10.3233364	*CG42313*
	2L	14954267	10.33806256	*CG42313*
	2L	14949702	10.54090147	*CG42313*
	2L	14939003	10.66965032	*CG42313*
	2L	14932967	10.86810467	*CG42313*
188	2L	15024079	8.73447932	*GABA-B-R1*
	2L	14993422	8.793673011	*mol*
	2L	15027203	9.345127787	*GABA-B-R1, CG33310*
	2L	15006259	9.48730524	*DCTN5-p25*
	2L	15009715	9.543385878	*l(2)35Bc*
	2L	14959571	9.614838315	*CG42313*
	2L	14969191	9.641481299	*CG42313*
	2L	15029307	9.722862963	*GABA-B-R1, CG33310*
	2L	15020521	9.761576829	*GABA-B-R1*

	2L	14981938	9.779007515	*mol*
	2L	14912792	9.79416925	*CG42313*
	2L	14965929	9.82393229	*CG42313*
	2L	15034311	9.843897127	*GABA-B-R1*
	2L	14979114	10.26286969	*mol*
	2L	14921880	10.29006228	*CG42313*
	2L	14926230	10.31176121	*CG42313, CG3491*
	2L	14917849	10.3233364	*CG42313*
	2L	14954267	10.33806256	*CG42313*
	2L	14976077	10.46813555	*mol*
	2L	15017629	10.47348805	*GABA-B-R1*
	2L	14949702	10.54090147	*CG42313*
	2L	14939003	10.66965032	*CG42313*
	2L	14932967	10.86810467	*CG42313*
	2L	15038696	11.23563303	*CIAPIN1*
189	2L	15076131	8.259623256	*CG15270*
	2L	15087196	8.371539528	*CG15270*
	2L	15080877	8.531753341	*CG15270*
	2L	15090266	9.466825628	*CG15270*
	2L	15065225	10.10835162	*solo, vas, vig*
	2L	15051047	10.2379874	*ck*
	2L	15045278	10.93336965	*ck*
207	2L	16633684	7.914845957	*CG42389*
	2L	16583833	8.017132063	*CG42389*
	2L	16647072	8.031415715	*CG42389*
	2L	16565774	8.066606913	*CG42389*
	2L	16581210	8.137045149	*CG42389*
	2L	16562212	8.204347712	*CG42389*
	2L	16534143	8.233518282	*CR45354*
	2L	16586830	8.255059157	*CG42389*
	2L	16506016	8.381684825	*Tpr2*
	2L	16530492	8.387006659	*CG5953*
	2L	16610301	8.449556615	*CG42389*
	2L	16522776	8.706708717	*CG5953*
	2L	16637297	8.726073526	*CG42389*
	2L	16526499	8.849867261	*CG5953*
	2L	16508145	8.920175504	*CG5953*
	2L	16504874	8.996306533	*Tpr2*
	2L	16519489	9.02442891	*CG5953*
	2L	16659290	9.239070028	*CG42389*
	2L	16642641	10.22936235	*CG42389*
209	2L	16633684	7.914845957	*CG42389*
	2L	16583833	8.017132063	*CG42389*
	2L	16647072	8.031415715	*CG42389*

	2L	16565774	8.066606913	*CG42389*
	2L	16581210	8.137045149	*CG42389*
	2L	16562212	8.204347712	*CG42389*
	2L	16586830	8.255059157	*CG42389*
	2L	16610301	8.449556615	*CG42389*
	2L	16709477	8.604740457	*CG31808*
	2L	16637297	8.726073526	*CG42389*
	2L	16705625	8.809591385	*CG31808*
	2L	16664927	9.134490514	*Trpgamma*
	2L	16659290	9.239070028	*CG42389*
	2L	16715915	9.339220455	*CG31808, Cyt-c-d*
	2L	16712926	9.361004466	*CG31808*
	2L	16698691	9.396219953	*grp*
	2L	16692565	9.485458298	*grp*
	2L	16684100	9.776822465	*grp*
	2L	16675692	9.884971888	*Trpgamma*
	2L	16670176	10.10933892	*Trpgamma*
	2L	16642641	10.22936235	*CG42389*
	2L	16702028	10.82083259	*CR43670*
218	2L	17401169	9.477977562	*CLIP-190*
	2L	17363745	9.494387876	*CG31804*
	2L	17396361	10.47869367	*CLIP-190*
	2L	17370596	10.84661996	*Lrch*
	2L	17409708	10.99985303	*Rpb11*
	2L	17405130	11.06097347	*CLIP-190*
	2L	17428917	11.09336575	*Dif*
	2L	17425463	11.2363135	*Dif*
	2L	17393434	11.34936673	*CLIP-190*
	2L	17445865	11.37230587	*dl*
	2L	17431306	11.53367196	*Dif*
	2L	17390408	11.53844491	*CLIP-190*
	2L	17375012	11.54956483	*Lrch*
	2L	17415410	11.55721837	*Dif*
	2L	17378258	11.60644747	*Lrch*
	2L	17419440	11.68953153	*Dif*
	2L	17422294	11.7236514	*Dif*
	2L	17442938	11.78471038	*dl*
	2L	17435322	11.81837315	*CG5043*
	2L	17383125	12.07172352	*Lrch*
	2L	17387106	12.3382856	*CLIP-190*
	2L	17438871	12.98333043	*dl*
233	2L	18620751	8.375597724	*MESR3*
	2L	18548323	8.950408659	*Pde11*
	2L	18534021	9.024866934	*Pde11*

	2L	18617300	9.314780069	*MESR3*
	2L	18543661	9.395924029	*Pde11*
	2L	18594055	9.697563697	*CG15160*
	2L	18573351	9.844553872	*Pde11*
	2L	18590039	10.34254036	*Pde11*
	2L	18568994	10.35461386	*Pde11*
	2L	18538240	10.44167447	*Pde11*
	2L	18552571	10.44975269	*Pde11*
	2L	18541301	10.45568606	*Pde11*
	2L	18577286	10.47107399	*Pde11*
	2L	18586552	10.47228858	*Pde11*
	2L	18580705	10.68932593	*Pde11*
	2L	18562331	10.9343188	*Pde11*
	2L	18603107	11.12889517	*CG10413*
	2L	18555514	11.34336166	*Pde11*
	2L	18622621	11.37846888	*MESR3*
	2L	18625093	11.43628597	*MESR3*
	2L	18659607	11.87661316	*MESR3*
	2L	18627588	12.1689362	*MESR3*
	2L	18655428	13.12263328	*MESR3*
	2L	18607167	13.30907581	*CG10333*
	2L	18631460	13.43371042	*MESR3*
	2L	18652001	13.7995579	*MESR3, Cyp310a1*
	2L	18647569	13.8801458	*MESR3*
	2L	18611181	14.7508588	*Atac2*
	2L	18642477	15.06951324	*MESR3*
243	2L	19396892	9.772517141	*CG17544*
	2L	19372230	10.23676715	*CG17349*
	2L	19392368	10.94372431	*CG17544*
	2L	19401404	11.02817897	*Pax*
	2L	19393967	11.07647656	*CG17544*
	2L	19425513	11.49652551	*Pax, CG16771, CG13085*
	2L	19388437	11.65474048	*CG17549*
	2L	19429195	11.85641717	*CG16771, CG13085*
	2L	19385426	11.87857343	*fon*
	2L	19419520	12.09199289	*Pax,lectin-37Db*
	2L	19413784	12.16729873	*Pax*
	2L	19382434	12.1918126	*CG17350*
	2L	19433505	12.43549977	*Rab9*
	2L	19405437	12.53110015	*Pax*
	2L	19410347	12.55913781	*Pax*
	2L	19344262	12.61859912	*dnt*
	2L	19436514	12.65827922	*CG10237*
	2L	19339596	12.69074402	*dnt*

	2L	19363030	12.71358295	*dnt*
	2L	19439754	13.55662644	*CG10237*
244	2L	19525002	12.29698328	*CG10132*
	2L	19436514	12.65827922	*CG10237*
	2L	19494518	13.35544371	*swm*
	2L	19439754	13.55662644	*CG10237*
	2L	19513743	13.57857952	*CG10186*
	2L	19506716	13.94984239	*CG10188*
	2L	19449969	14.29068081	*Top2*
249	2L	19818677	8.902116029	*sick*
	2L	19928031	9.149767796	*sick*
	2L	19814784	9.261677958	*sick*
	2L	19906914	9.641595094	*sick*
	2L	19925189	9.747072175	*sick*
	2L	19812082	9.755517426	*sick*
	2L	19900890	9.930997273	*sick*
	2L	19903534	10.04485792	*sick*
	2L	19897884	10.15571853	*sick*
	2L	19809561	10.23326875	*sick*
	2L	19910321	10.25247139	*sick*
	2L	19806642	10.61980518	*sick*
	2L	19854256	10.66798106	*sick*
	2L	19821795	10.66835826	*sick*
	2L	19914368	10.70181493	*sick*
	2L	19922727	10.7178283	*sick*
	2L	19857961	10.79666133	*sick*
	2L	19920292	10.82418858	*sick, CR43828*
	2L	19866821	10.87971247	*sick*
	2L	19845828	10.94343847	*sick*
	2L	19917081	10.95307879	*sick*
	2L	19800945	11.05154803	*sick*
	2L	19849450	11.05198489	*sick*
	2L	19954020	11.07239713	*sick*
	2L	19824302	11.12220092	*sick*
	2L	19803776	11.17299903	*sick*
	2L	19892551	11.21784385	*sick*
	2L	19945313	11.24101283	*sick*
	2L	19942015	11.28627009	*sick*
	2L	19895268	11.28760575	*sick*
	2L	19931878	11.40450655	*sick*
	2L	19851815	11.41550141	*sick*
	2L	19949356	11.51230404	*sick*
	2L	19839189	11.52165276	*sick*
	2L	19886638	11.54470816	*sick*

	2L	19883874	11.56354874	*sick*
	2L	19889044	11.61822355	*sick*
	2L	19871268	11.63935837	*sick*
	2L	19843393	11.64379763	*sick*
	2L	19827684	11.74625239	*sick*
	2L	19874296	11.93298412	*sick*
	2L	19878201	11.99381723	*sick*
	2L	19881730	12.04142042	*sick*
	2L	19797922	12.0832266	*sick*
	2L	19835274	12.19556981	*sick*
	2L	19831689	12.89080648	*sick*
	2L	19934731	13.30270361	*sick*
	2L	19938543	13.62534782	*sick*
261	2L	20801263	10.57057931	*CG9328*
	2L	20809509	11.23260707	*CG9328*
	2L	20807250	11.30978097	*CG9328*
	2L	20865777	11.31392457	*CG9338*
	2L	20860823	11.50655728	*CG9336*
	2L	20869266	11.57073299	*CG31675*
	2L	20835780	11.60666898	*Oseg5*
	2L	20841257	11.65581679	*CG31676*
	2L	20803816	11.7654823	*CG9328*
	2L	20856550	11.80156294	*twit*
	2L	20815775	11.91397464	*CG33322*
	2L	20829765	11.96127357	*CG31673*
	2L	20825566	12.18072446	*CG9331*
	2L	20908231	12.25537046	*sky, CG43739*
	2L	20847306	12.43572922	*CG31676*
	2L	20912924	12.58707497	*sky, CG43739*
	2L	20900300	12.6804374	*sky, CG43739*
	2L	20884904	12.79996614	*sky, CG43739*
	2L	20889472	12.81151924	*sky, CG43739*
	2L	20880910	12.83822097	*sky, CG43739*
	2L	20852626	12.84289714	*twit*
	2L	20798471	13.25879876	*CG9328*
	2L	20874913	13.37732227	*sky, CG43739, CR44981*
266	2L	21217693	12.62281546	*Atg18b, CG8679*
	2L	21240816	12.68856184	*Hr39*
	2L	21231505	12.83034877	*CG8677*
	2L	21253149	12.85846706	*Hr39*
	2L	21263836	13.51529453	*CG8671*
	2L	21276328	13.61969003	*CG8671*
267	2L	21319064	10.02811035	*crc*
	2L	21334900	10.64621471	*dimm*

	2L	21301206	10.65940745	*Mondo*
	2L	21312214	11.05279055	*crc*
	2L	21307093	11.32135662	*Mondo*
	2L	21295670	12.2690877	*Mondo, Gr39a*
	2L	21291132	13.01304247	*Mondo, Gr39a*
	2L	21282927	13.24188425	*Cyp6t2Psi*
	2L	21263836	13.51529453	*CG8671*
	2L	21276328	13.61969003	*CG8671*
342	2R	4479321	7.911827459	*CG14752*
	2R	4457000	8.034438204	*Cyp6a15Psi*
343	2R	4511904	8.080954426	*Cirl*
	2R	4516154	8.293785626	*CG14749*
	2R	4505735	8.316558277	*Cirl*
	2R	4488792	8.485888632	*rgr*
	2R	4496559	8.628184125	*rgr*
	2R	4500423	8.692729392	*CG8642*
	2R	4492774	8.744648683	*rgr*
344	2R	4644568	7.934034645	*CG8740*
	2R	4582765	8.254366073	*CG8586*
349	2R	5031590	8.400764798	*Prp38*
358	2R	5733814	8.093695411	*Orc6*
	2R	5719683	8.176431727	*hebe*
	2R	5739624	8.185300157	*PCB*
	2R	5685806	8.270568616	*CG1688*
	2R	5708487	8.279119383	*CG34033*
	2R	5667360	8.344410123	*CG1688*
	2R	5745390	8.41754215	*PCB*
	2R	5677198	8.531761519	*CG1688, CR44208*
	2R	5663224	8.626809511	*CG1688, CR44207*
	2R	5706873	8.93496721	*CG1648*
	2R	5723585	8.936302964	*hebe*
	2R	5716443	8.979694244	*hebe*
	2R	5712273	8.984813152	*dila*
	2R	5673096	9.004005551	*CG1688*
	2R	5759407	9.050150775	*cbx, CG18446*
	2R	5729152	9.442057217	*Vamp7*
	2R	5658089	9.457556526	*CG1688*
359	2R	5824288	8.139613978	*Mef2*
	2R	5833457	8.202627946	*Mef2*
	2R	5819950	8.405902734	*Mef2*
	2R	5828828	8.485392842	*Mef2*
	2R	5845421	8.684809801	*Mef2*
	2R	5759407	9.050150775	*cbx, CG18446*
	2R	5773718	9.178210057	*CG1513*

	2R	5811302	9.214694542	*Mef2*
	2R	5837098	9.251193774	*Mef2*
	2R	5815948	9.326837515	*Mef2*
	2R	5782050	9.769793811	*CG30007*
	2R	5806472	9.838125595	*Mef2*
	2R	5789089	10.19560362	*CG1441*
	2R	5794612	10.69997677	*FMRFa*
	2R	5799770	10.94460218	*Etf-QO*
361	2R	5978445	8.347401584	*CG2269*
	2R	5992617	8.525233016	*14-3-3zeta*
	2R	5984444	8.549790598	*Jra*
	2R	6002190	8.617669454	*Pfk*
	2R	5924574	8.651113943	*oys*
	2R	5951919	9.093808231	*CG2292*
	2R	5941647	9.393171004	*magu*
	2R	5946505	9.411175958	*magu*
	2R	5969110	9.903767415	*egr*
	2R	5960383	10.17791354	*CG1371*
	2R	5955652	10.48311845	*Cdc2rk*
	2R	5974318	11.00134627	*CG2269, sut4*
364	2R	6207149	8.021101025	*Ndg*
	2R	6282722	8.030936444	*CG42732*
	2R	6280007	8.104777245	*CG42732*
	2R	6285415	8.167426022	*CG42732*
	2R	6176817	8.389592586	*CAP*
	2R	6294935	8.47461854	*CG42732, CG12898*
	2R	6172930	8.598260512	*CAP*
	2R	6250772	8.636715323	*CG42732, Gr47a*
	2R	6296707	8.680105991	*CG42732, CG33477*
	2R	6203252	8.896988427	*Ndg*
	2R	6180745	8.995117815	*CAP*
	2R	6229900	9.018407532	*CG42732*
	2R	6169222	9.022196302	*CAP*
	2R	6260162	9.022531389	*CG42732*
	2R	6214618	9.078388577	*CG42732*
	2R	6233532	9.078936166	*CG42732*
	2R	6226379	9.140896745	*CG42732*
	2R	6248379	9.23817985	*CG42732*
	2R	6245099	9.298597555	*CG42732*
	2R	6223429	9.383556115	*CG42732*
	2R	6219343	9.439792379	*CG42732*
	2R	6256992	9.671348417	*CG42732*
	2R	6159065	9.681507906	*CAP*
	2R	6198098	9.719766457	*CG12909*

	2R	6185338	9.781438438	*CAP*
	2R	6236914	9.82730856	*CG42732*
	2R	6241450	10.36507958	*CG42732, CG12907*
	2R	6266541	10.40542062	*CG42732, Ir47a*
	2R	6264173	10.49405527	*CG42732, Ir47b*
	2R	6188427	10.71251724	*CAP*
	2R	6195132	10.91332588	*JhI-1*
	2R	6191565	11.18797107	*CAP*
366	2R	6334187	8.020877605	*Galphao*
	2R	6362626	8.184129821	*whd*
	2R	6402198	8.513201452	*lola*
	2R	6396694	8.768630428	*lola*
369	2R	6631659	8.06337521	*CG33144*
540	2R	20262675	8.192767981	*bs*
	2R	20268811	8.226784362	*mAChR-A*
	2R	20255160	8.431539493	*bs, CR44811*
	2R	20249400	8.771286522	*bs, CR44811*
	2R	20318927	8.89625056	*Pgam5-2*
	2R	20237112	8.95416988	*bs*
	2R	20258303	9.056489502	*bs*
	2R	20275182	9.095427714	*mAChR-A*
	2R	20280808	9.581279777	*Slik*
	2R	20310579	9.679717994	*prom*
	2R	20314951	9.805264192	*prom*
	2R	20289603	9.904039251	*Slik, Rpn8*
	2R	20300562	10.05570806	*CG45068, CG45069*
	2R	20307414	10.19834748	*prom*
	2R	20297228	10.24690081	*SerT*
	2R	20304878	10.42217698	*prom*
541	2R	20367109	8.461292163	*CG13579*
	2R	20352800	8.583582238	*CG13579*
	2R	20371688	8.676782793	*CG13579*
	2R	20361364	8.832082187	*CG13579, CG3492*
	2R	20318927	8.89625056	*Pgam5-2*
	2R	20364396	8.941530822	*CG13579*
	2R	20400512	9.095130262	*Letm1*
	2R	20381381	9.098004564	*CG13590*
	2R	20356578	9.240450931	*CG13579*
	2R	20349022	9.354203756	*CG13579*
	2R	20378157	9.425973249	*CG13579*
	2R	20388490	9.472135098	*Prosalpha4T2*
	2R	20344983	9.551221517	*CG13579*
	2R	20339431	9.684010631	*CG4563*
542	2R	20400512	9.095130262	*Letm1*

	2R	20422474	9.304204989	*CG4622*
	2R	20410940	9.356308865	*CG4612*
	2R	20405592	9.387205425	*Ir60c*
	2R	20439282	9.734759559	*CG4622, ITP*
	2R	20432522	9.738821019	*CG4622, ITP*
	2R	20418172	9.781457021	*Ir60e*
	2R	20484017	10.44811643	*pio*
	2R	20471821	10.52611413	*pio*
	2R	20459321	10.61491257	*Fcp1, CG3511*
	2R	20475989	10.66633371	*pio*
543	2R	20527263	8.403006673	*CG3640*
	2R	20508727	8.641716644	*CG13594*
	2R	20512939	8.89497285	*CG13594*
	2R	20517386	9.063847917	*CG13594*
	2R	20503679	9.550149224	*CG13594*
	2R	20493569	9.683279442	*CG4707*
	2R	20484017	10.44811643	*pio*
	2R	20471821	10.52611413	*pio*
	2R	20475989	10.66633371	*pio*
544	2R	20610503	8.173945534	*Mid1*
	2R	20604295	8.221945207	*Mid1*
	2R	20623818	8.405656896	*Mid1*
	2R	20637353	9.159156717	*Usp15-31*
545	2R	20637353	9.159156717	*Usp15-31*
	2R	20703653	9.174322254	*Dll*
	2R	20667705	9.37511708	*Lcp9*
	2R	20682012	9.557757391	*CR43257*
	2R	20714399	9.672662733	*Dll*
546	2R	20781528	8.398746174	*NaCP60E*
	2R	20768289	8.617764651	*CG44247*
	2R	20757872	8.819680511	*Atf-2*
	2R	20703653	9.174322254	*Dll*
	2R	20714399	9.672662733	*Dll*
547	2R	20781528	8.398746174	*NaCP60E*
	2R	20811348	8.557563769	*pain*
	2R	20831763	8.934341339	*CG15861*
	2R	20822788	9.265029309	*CG30427*
	2R	20896979	9.584750485	*zip*
	2R	20868369	9.660600445	*emp*
	2R	20883043	9.805000775	*zip*
548	2R	20936282	8.445072864	*gsb-n*
	2R	20930062	8.725627463	*gsb-n*
	2R	20922725	8.781822343	*Nplp1*
	2R	20950781	8.898137587	*gsb*

	2R	20916983	9.256021923	*uzip*
	2R	20910311	9.573993017	*uzip*
	2R	20906158	9.581787398	*uzip*
	2R	20896979	9.584750485	*zip*
549	2R	20999197	8.307430746	*lov*
	2R	21022346	8.368120951	*lov*
	2R	21026518	8.45171267	*lov, CG43106*
	2R	20962613	8.452644726	*gol*
	2R	21011096	8.969266308	*lov*
	2R	21002964	9.330791873	*lov*
	2R	21006399	9.381893174	*lov*
	2R	21018708	9.516199853	*lov*
550	2R	21077475	8.608352879	*CG9380*
	2R	21096574	8.810341072	*CG9380*
	2R	21072792	8.864909148	*CG9380*
	2R	21084101	9.183423115	*CG9380*
	2R	21079766	9.205515185	*CG9380*
	2R	21088898	9.425762385	*CG9380*
	2R	21092965	9.563802455	*CG9380*
	2R	21100988	9.767904737	*CG9380*
607	3L	4411303	8.365550335	*DOR*
610	3L	4652769	8.10617432	*axo*
	3L	4655227	8.956726024	*axo*
	3L	4665781	9.182704337	*axo*
	3L	4658252	9.387576332	*axo*
	3L	4661203	9.47773455	*axo*
	3L	4652769	8.10617432	*axo*
	3L	4655227	8.956726024	*axo*
	3L	4665781	9.182704337	*axo*
	3L	4658252	9.387576332	*axo*
	3L	4661203	9.47773455	*axo*
613	3L	4995774	7.967347575	*Con, CG32232*
614	3L	4995774	7.967347575	*Con, CG32232*
616	3L	5184159	7.919026141	*shep*
	3L	5193437	8.048705646	*shep*
617	3L	5184159	7.919026141	*shep*
	3L	5193437	8.048705646	*shep*
619	3L	5434425	8.006141367	*DIP-delta*
	3L	5430091	8.149984978	*DIP-delta*
	3L	5426490	8.179270556	*DIP-delta*
620	3L	5508026	8.078125601	*Lkr*
	3L	5514406	8.528357509	*Lkr*
621	3L	5554181	7.923059898	*sinu*
	3L	5508026	8.078125601	*Lkr*

	3L	5514406	8.528357509	*Lkr*
625	3L	5890819	8.020681031	*CG5592*
656	3L	8330319	7.989422084	*Cpr66Cb*
	3L	8376901	8.132212625	*ImpE1*
	3L	8382712	8.188048465	*ImpE1*
	3L	8335478	8.256094523	*DNApol-alpha50, CG7083*
	3L	8373245	8.43583655	*ImpE1*
	3L	8341064	8.493413831	*GAPcenA*
	3L	8394690	8.561538683	*CG7120*
	3L	8353190	8.664767933	*Idh*
	3L	8344746	8.784729323	*GAPcenA*
	3L	8348301	8.817704078	*CG17352*
	3L	8390378	9.028586235	*CG7120*
657	3L	8406515	7.918677933	*Exo70, mtrm*
	3L	8401872	8.284220426	*Oseg1*
658	3L	8557247	7.942725669	*rhea*
	3L	8560432	7.959301534	*rhea*
	3L	8555270	7.989406124	*rhea*
	3L	8563681	8.078552947	*rhea*
	3L	8538373	8.09233012	*rhea*
	3L	8526183	8.230693706	*foi*
	3L	8523530	8.278826054	*foi*
	3L	8520362	8.28402805	*foi*
	3L	8514769	8.391871142	*GstO2*
	3L	8552986	8.424453929	*rhea*
	3L	8530411	8.518915607	*ergic53*
	3L	8557247	7.942725669	*rhea*
	3L	8555270	7.989406124	*rhea*
	3L	8538373	8.09233012	*rhea*
	3L	8552986	8.424453929	*rhea*
659	3L	8560432	7.959301534	*rhea*
	3L	8563681	8.078552947	*rhea*
	3L	8571048	8.127538028	*CG43078*
	3L	8568834	8.432503473	*CG6638*
	3L	8630889	8.636979223	*Zasp66*
	3L	8633981	8.668593147	*Cpr66D*
	3L	8638486	9.41115768	*CG13305*
	3L	8636312	9.539676587	*Cpr66D*
667	3L	9275428	8.962785021	*GluRIB*
	3L	9275428	8.962785021	*GluRIB*
673	3L	9776740	7.980860934	*CG8177*
	3L	9705980	8.270482595	*SH3PX1*
	3L	9769339	8.359399417	*CG8177*
	3L	9701052	8.539011487	*CG16711*

	3L	9716652	8.571594906	*defl*
700	3L	11872520	8.108926634	*CG44837, Sprn*
701	3L	11872520	8.108926634	*CG44837, Sprn*
732	3L	14407896	7.985191979	*bbg*
733	3L	14407896	7.985191979	*bbg*
764	3L	17032285	8.029365359	*scaf6*
	3L	16960379	9.984936617	*Nc73EF*
790	3L	19064896	7.925997862	*Mkp3*
	3L	19060381	8.65286812	*Mkp3*
	3L	19054504	8.931791805	*CG3797*
792	3L	19182883	7.975482599	*fz2*
882	3R	1918978	8.042751691	*Gasp*
883	3R	1918978	8.042751691	*Gasp*
988	3R	10342379	8.117181822	*Pde6*
1082	3R	17780676	7.914143692	*Eip93F*
	3R	17840803	8.095620087	*CG6332*
	3R	17874306	8.160191078	*how*
	3R	17850805	8.613944922	*CG6028*
	3R	17846079	8.709548583	*Mitofilin*
	3R	17843065	8.791370954	*CG6439*
1098	3R	19136450	8.104903395	*pnt*
	3R	19136450	8.104903395	*pnt*
1099	3R	19286990	8.303204481	*CG4374*
1100	3R	19286990	8.303204481	*CG4374*
	3R	19309845	8.443308541	*Ir94g*
1105	3R	19717639	8.413817534	*tbrd-1*
1130	3R	21750347	8.033025284	*CCAP-R*
	3R	21750347	8.033025284	*CCAP-R*
1180	3R	25758236	8.40783802	*Cog7*
	3R	25757489	8.556456688	*Cog7*
1187	3R	26297991	8.135591915	*PH4alphaEFB*
1305	X	6450844	7.993819816	*CG34417*
	X	6524040	8.004217577	*pigs*
	X	6445958	8.190374429	*CG34417*
	X	6506212	8.447558626	*pigs*
	X	6440925	8.775819635	*CG34417*
	X	6437894	8.805067599	*CG34417*
	X	6518856	9.121203709	*pigs*
	X	6514002	9.277793286	*pigs*
1353	X	10426959	9.29046566	*spri*
	X	10433992	9.870449686	*spri*
1354	X	10426959	9.29046566	*spri*
	X	10433992	9.870449686	*spri*
1421	X	15831668	8.087824816	*Stim*

Table S5.1 Average SNP count for all sequences recorded under each of the 49 selected TFs.

TF name	**Total number of SNPs counted across all TFBS sequences recorded**	**Average number of SNPs per sequence**
abd-A	37	0.860465116
Abd-B	49	1.96
Antp	18	1.125
ap	69	4.928571429
bcd	143	2.6
bin	48	1.411764706
br-Z1	32	2.285714286
Br-Z2	45	2.045454545
Br-Z3	43	2.529411765
cad	12	0.923076923
da	23	1.4375
dl	112	2.285714286
Dref	80	2.857142857
EcR	39	2.071428571
en	44	1.517241379
eve	25	1.923076923
exd	22	1.466666667
ey	43	2.866666667
fkh	30	2.727272727
ftz	108	2.29787234
grh	31	2.583333333
hb	203	2.136842105
jing	18	1.5
kni	111	2.846153846
Kr	68	1.511111111
Mad	96	1.5
Med	41	1.366666667
pan	87	2.289473684
pho	26	2.6

pnr	38	2.714285714
pnt	23	1.4375
prd	16	1.333333333
sd	67	3.045454545
sna	37	2.916666667
srp	61	2.44
Su(H)	40	1.904761905
tin	40	1.666666667
tll	71	1.918918919
Trl	64	1.361702128
ttk	21	1.75
twi	35	2.1875
Ubx	101	1.463768116
usp	26	2.363636364
vfl	15	1.5
vnd	8	0.615384615
vvl	54	3.857142857
z	69	1.769230769
zen	54	2.347826087