Genetic Variation and Human Longevity

Association studies of DNA repair, GH/IGF-1/insulin signaling, oxidative stress and classical candidate genes

Mette Soerensen, M.Sc.

This review has been accepted as a thesis together with 4 previously published papers by University of Southern Denmark on the 25th of January 2012 and defended on the 29th of March 2012.

Tutor(s): Lene Christiansen, Kaare Christensen, Tinna Stevnsner & Vilhelm A. Bohr

Official opponents: Hélène Blanché-Koch & Mette Nyegaard

Correspondence: Mette Soerensen, Danish Aging Research Center, Epidemiology, Institute of Public Health, University of Southern Denmark, J.B. Winsløws Vej 9B, 5000 Odense C, Denmark or Department of Clinical Genetics, Odense University Hospital, Sdr. Boulevard 29, 5000 Odense C, Denmark.

E-mail: msoerensen@health.sdu.dk

Dan Med J 2012;59(5):B4454

1. This PhD thesis was based on four manuscripts, which have been published as shown below:

Paper I: Soerensen M, Dato S, Tan Q, Thinggaard M, Kleindorp R, Beekman M, Jacobsen R, Suchiman HED, de Craen AJM, Westendorp RGJ, Schreiber S, Stevnsner T, Bohr VA, Slagboom PE, Nebel A, Vaupel JW, Christensen K, McGue M, Christiansen L., "Human longevity and variation in GH/IGF-1/insulin signaling, DNA damage signaling and repair and pro/antioxidant pathway genes: Cross sectional and longitudinal studies", Experimental Gerontology. 2012 Mar 3.

Paper II: Soerensen M, Dato S, Tan Q, Thinggaard M, Kleindorp R, Beekman M, Suchiman HED, Jacobsen R, McGue M, Stevnsner T, Bohr VA, de Craen AJM, Westendorp RGJ, Schreiber S, Slagboom PE, Nebel A, Vaupel JW, Christensen K, Christiansen L, ”Evidence from case-control and longitudinal studies supports associations of genetic variation in *APOE*, *CETP*, and *IL6* with human longevity”, Age (Dordr). 2012 Jan 12.

Paper III: Soerensen M, Dato S, Christensen K, McGue M, Stevnsner T, Bohr VA, Christiansen L., ”Replication of an association of variation in the *FOXO3A* gene with human longevity using both case-control and longitudinal data”, Aging Cell. 2010 Dec; 9(6): 1010-7.

Paper IV: Soerensen M, Thinggaard M, Nygaard M, Dato S, Tan Q, Hjelmborg J, Andersen-Ranberg K, Stevnsner T, Bohr VA, Kimura M, Aviv A, Christensen K, Christiansen L., ”Genetic variation in *TERT* and *TERC* and human leukocyte telomere length and longevity: a cross-sectional and longitudinal analysis”, Aging Cell. 2012 Apr; 11(2): 223-227.

2. INTRODUCTION

Since the 1950s the mortality rate among the elderly in the developed countries has declined dramatically, causing an increase in the number of the oldest-old (1). Seen both from a health care perspective and a socioeconomic perspective this increase makes the study of human aging and longevity increasingly important. Twin studies have indicated a genetic component of the inter-individual variation in aging phenotypes, as well as in longevity itself. Hence, in order to obtain a better understanding of human aging and longevity, the underlying genetic and molecular processes need to be elucidated.

2.1. Biological aspects of aging and longevity

It is an inevitable fact that all human beings will age and die. But why do we age? What is aging seen from a biological point of view and which processes lead to aging? To give an exhaustive answer to these questions is out of scope of this thesis, however a few points concerning the biological aspects of aging will be given. In several animal species the aging process is initiated at some point after maturity and the reproduction phase by a decline in the physiological functions necessary for survival; in humans this is characterized by physiological changes such as reduction in muscle strength, loss of bone mass, changes in the cardiovascular system, loss of elasticity in the lungs and changes in hormone signaling (2-6). Moreover, human aging is often accompanied by age-related diseases such as cardiovascular diseases, atherosclerosis, dementia, type 2 diabetes, Alzheimer's, osteoporosis and cancers. At the cellular level aging is among other things characterized by a decreased rate of cell division, a change in gene expression and a change in the response to intra- and extracellular stimuli. Furthermore, vital cellular components accumulate damage during aging; mutations and lesions accumulate in the genome and the DNA becomes less stable, while modified and damaged proteins and lipoproteins arise (7). These age-related physiological changes and the biological, molecular and biochemical mechanisms which might be their basis, have challenged researchers for decades; why do the biological functions normally ensuring the homeostasis of the body slow down? How do these changes in body functioning affect aging phenotypes and the occurrence of age-related diseases? And at a population level; why do we age so differently, i.e. why do some people age with well preserved physical and cognitive functioning, while others do not? Why do some people live to extreme ages, while others do not? The questions are many and the explanations put forward comprehensive.

Biological theories of aging

Several theories aiming to explain the occurrence of biological changes during aging have been proposed. The theories of profound importance for the object and strategy of this PhD study will be explained shortly in the following.

First of all, evolutionary biologists have argued that aging occurs post-reproductively, since there is no longer a need to withstand the physiological functions of the body necessary for reproduction, i.e. aging is a default state occurring after an organism has fulfilled the requirements of natural selection (8). Following this thought there should be no selection for genetic variants during aging which promote long life. However, as it was pointed out, due to the lack of selection, detrimental late-acting genetic variations might accumulate in the elderly populations (8). Later the idea of antagonistic pleiotropy (9) extended this by suggesting that the harmful genetic variations accumulating during old age had been selected during the earlier reproductive phase of life due to a positive effect on reproduction. It was, furthermore, proposed that there appears to be a trade-off between reproduction and longevity in the disposable soma theory (10). Based on the idea that under pressure of natural selection there is a trade-off between maintaining the soma (non-reproductive tissue) of an organism and reproduction, i.e. the resources invested in maintenance only need to be sufficient to keep the body in good condition until reproduction has occurred, it was deduced that to maintain the soma after reproduction (to become long-lived) will require investment in maintenance and will, consequently, come with a cost in reproduction. Thus, the resources invested in either reproduction or longevity determines the lifespan of an organism. One consequence of this idea is that due to the limited investment in maintenance, damage will accumulate throughout lifespan and consequently lead to aging. Accordingly, in addition to the detrimental variations mentioned above, longevity could be regulated by genes controlling processes counteracting damage accumulation (11).

A number of biological theories of aging focus on specific biological processes affecting this damage accumulation. One theory is the free radical theory of aging (12) concerning the harmful effects of reactive oxygen species (ROS). ROS are produced primarily in the mitochondria of the cells, where the last steps of oxidative phosphorylation take place. Oxidative phosphorylation is the process by which energy released by oxidation of nutrients is converted into ATP (adenosine triphosphate), the main energy 'currency' of the cell. In the final step of oxidative phosphorylation, oxygen is the final electron acceptor, however, about 2-3% of the oxygen is reduced insufficiently giving rise to ROS (13). Furthermore, ROS are generated in other cellular compartments in a variety of cellular processes, in which oxidation takes place and also, ROS can be generated by exogenous sources such as UV radiation. ROS can oxidize and damage nucleic acids, proteins and cellular membranes and this damage is believed to reduce the cellular function, ultimately manifesting itself at the organ and body level and consequently contributing to aging (14).

Another theory is the theory of an age-related increase in genetic instability. It is based on the fact that more DNA mutations are found in cells from old than from young donors (15, 16), i.e. damage appears to accumulate as mutations with age. These damages might be introduced by spontaneous decay or replication errors in the DNA, by ROS or by external sources such as ionizing radiation (17, 18). Different DNA repair mechanisms generally ensure the removal of the different types of DNA damage before they are converted to mutations, however, it is predicted that a change in these DNA repair mechanisms with age contributes to the age-associated accumulation of DNA damage (19). An age-related accumulation of damage is suspected to increase the instability of the DNA, obviously affecting a wide range of processes. The theory is supported by the existence of human premature aging syndromes such as Rothmund-Thomson and Werner syndromes, which are caused by mutations in DNA repair protein encoding genes (reviewed in (20)), and among other things are characterized by cataracts, skin alterations and short stature, and for the latter also diabetes and atherosclerosis. In addition, impaired DNA repair has been associated with several age-related diseases including Alzheimer's disease (21).

Also related to genetic instability is the theory of telomere deterioration. Telomeres are the regions at the ends of chromosomes which protect the chromosomes from deterioration, hereby maintaining genetic stability. Telomeres consist of TTAGGG repeats which in human cells with an active telomerase complex are added to the ends of the chromosomes by the catalytic subunit tert and the RNA template subunit terc. However, in human cells in culture, telomeres are known to shorten with every cell cycle (reviewed in (22)), until reaching a critical length at which point the cell enters cellular senescence (reviewed in (23)). Hence, telomeres have been suggested to be a sort of "clock" which eventually prevents the cell from dividing (24) and thus ceases its function. In cross-sectional studies of human blood cells, telomere length has been reported to be inversely related to the age of the individual (25, 26) and telomere length has been associated with increased risk of age-related diseases and mortality (reviewed in (27, 28)). Furthermore, it was recently shown in a study of twin pairs that short leukocyte telomere length (LTL) predicts an early death (29).

Lastly, some theories deal with the functions of organ systems essential for survival, essential in the sense that they regulate other body systems and/or ensure the communication and adaptation of these body systems to internal and external stimuli and therefore ensure an optimal functional state of the body for reproduction and survival. The neuroendocrine theory of aging (30, 31) suggests that aging occurs due to age-related changes in the hypothalamo-pituitary-adrenal (HPA) axis constituted of complex hormonal signal interactions between the hypothalamus, the pituitary gland and the adrenal glands. These communications affect a wide range of processes including development, growth and reproduction, as well as stress responses, i.e. the ability to maintain the steady state of the body despite constant changes in the environment. One of the signaling networks of the HPA axis is the growth hormone 1/insulin-like growth factor 1/insulin (GH/IGF-1/INS) signaling pathway, which, as it will be described below, is one of the major candidate pathways of human longevity. Finally, the neuroendocrine-immuno theory of aging (32) focuses on the role of the immune system and its interaction and integration with the neuroendocrine system in the aging process.

2.2. GENETIC EPIDEMIOLOGICAL STUDIES

As it can be inferred from the above, one major interest in the aging research field is the genetic influence on the aging process; what is the genetic contribution to variation in aging phenotypes and lifespan and which genes and biological processes play a role? The investigation of the connection between phenotype and genotype variation is conducted by different means; diverse types of genetic variation have been explored and different study approaches have been applied.

Intuitively we are familiar with the existence of genetic variation; that is we are aware of the abundant variation in human phenotypes both within a population and between populations and that phenotypic similarity appears to cluster among related individuals. The human genome is constituted of approximately $3*10^9$ base pairs (bps) divided on 23 chromosome pairs, thus giving plenty of room for variation. Briefly, large genetic variations such as chromosomal rearrangements or translocations are generally not considered relevant to a polygenic complex state such as human longevity, whereas variations of smaller size such as copy number variations, tandem repeats, insertions/deletions of single nucleotides and inter-individual variation in the individual nucleotides (single nucleotide polymorphisms (SNPs)) (33) are considered relevant. SNPs are highly investigated in genetic association studies; i.e. the inspection of differences in allele frequencies between population groups e.g. diseased and controls. It has been estimated that approximately 0.1-0.5% of the human genome is polymorphic, i.e. there is approximately $5\text{-}15*10^6$ SNPs in the human genome or 1 for every 200-1,000 nucleotides. These estimates were recently confirmed in the first publication on whole-genome sequencing of several individuals (179 individuals from four different populations); $14.4*10^6$ SNPs were identified (34)).

Study designs

When investigating the connection between phenotypic and genotypic variation, researchers have in general applied two types of studies: linkage studies and association studies. In short, linkage studies are the search for disease (phenotype) causing genes in related individuals i.e. collections of affected sib pairs or pedigrees covering several generations with a number of affected individuals. Exploring such individuals enable the measurement of frequency with which known genetic loci markers segregate together through meiosis from one generation to the next. If the markers tend to segregate together more often than expected by chance, they are linked. The linkage analysis then investigates the co-segregation of markers and the phenotype of interest. In this way the rough location of the phenotype causing gene, relative to the known genetic markers, can be deduced. The linkage studies have in general been successful in determining monogenic diseases such as cystic fibrosis (35) and Huntington disease (36). Association studies are the investigation of the association of specific genetic variants with a given phenotype; for instance the comparison of allele frequencies between a group of affected individuals (cases) and a group of unaffected individuals (controls), or for example the association of an allele with a continuous trait such as cognitive function. The individuals under study might be related or, as is often the case, unrelated. Association studies have in general been applied for identifying genes involved in polygenic complex disorders such as type 2 diabetes and in longevity (37, 38). One type of study design often applied in association studies is the case-control study. These studies are rather fast and less expensive to conduct than the prospective cohort studies (see below) and they are very suitable for rare diseases. However, a disadvantage is that they are difficult to design, especially the issue of choosing a proper control group is critical, i.e. bias is potentially introduced due to differences in characteristics of cases and controls (cohort effects). Another type of study design is the prospective cohort study; a population of individuals is enrolled and is followed prospectively with repeated collection of phenotype data, enabling longitudinal data analyses. Two advantages of this study design are the avoidance of the cohort effect and the possibility to investigate several phenotypes. Still, the disadvantage is that they are labor-intensive, time-consuming and expensive to construct (39).

That it is possible to identify common phenotype causing genetic variations in unrelated individuals is based on the knowledge about human evolution and the spreading from sub-Saharan Africa to the rest of the globe (40), that is the relatedness of human beings is reflected in our genome. One such reflection is the principle of linkage disequilibrium (LD), which is the non-random association of alleles in the genome, i.e. that some combinations of alleles (haplotypes) are more frequent in a population than would be expected by chance. LD is influenced greatly by recombination processes; it is known that the recombination rate is not equally distributed over the genome and that some regions have limited recombination and that loci in these regions will have higher degrees of LD (41). The loci with high degree of LD will segregate together from one generation to the next and consequently it is possible to investigate the variation in a given stretch (‘block’) of the genome with high LD by investigating a few variations in the block. The principle of LD can nowadays easily be exploited due to the HapMap consortium; the exhaustive collection of genotype frequency data for several different human populations, which is freely accessible via the HapMap consortium webpage (http://hapmap.ncbi.nlm.nih.gov/index.html.en). In this way data on the degree of LD in a given genomic region can be explored, and SNPs covering the majority of the genetic variation in that genomic region (so-called tagging SNPs) can be identified, hence enabling a thorough investigation of the genomic region. Obviously the HapMap populations are not completely similar to the researcher’s own study population, still the emergence of this database facilitates the process of genetic association studies and makes it simpler to compare studies.

Finally, the appearance of the HapMap consortium and the development of genome-wide genotyping techniques have introduced a difference in study concept within genetic association studies; the genome-wide association study (GWAS) ‘scanning’ the genome for association versus the candidate study based on an a priori hypothesis of for instance biological function of a gene or perhaps of a functional effect of the variant under study, e.g. an amino acid substitution. This difference in concept leads to profound differences in how to handle the data and consequently in how to interpret the results. A major aspect of this is the issue of multiple testing; an issue which becomes even more challenging with the recent progress in next generation sequencing techniques.

2.3. GENETIC ASPECTS OF AGING AND LONGEVITY

Until the emergence of the tagging SNP approach and the GWAS, genetic association studies of human longevity were conducted as candidate association studies of one or a few variations in one or a few genes. As it will be described in the Discussion of Materials and Methods section, this PhD project was initiated by a thorough literature and database search to identify candidate variations, genes and pathways of human aging and longevity. These candidates have traditionally been identified by different means.

Biological model systems of aging and longevity

As compared to studies of humans, studies of animal models have one major advantage when investigating the genetic contribution to aging and longevity: a population of genetically uniform organisms can be genetically manipulated, for instance by knocking out or over-expressing a gene and the consequences of such manipulation can be examined. Moreover, opposed to humans the environment of the model organisms can be controlled and since the animals generally have short lifespans it is easier to perform longitudinal studies. Frequently applied animal models are the common fruit fly *Drosophila melanogasta*, the roundworm *Caenorhabditis elegan*s and the house mouse *Mus musculus*. Based on major biological theories of aging, numerous candidate genes have been investigated in these animal models, some of which have shown effects on lifespan. Classical examples are the animal models knocked out in genes taking part in the GH/IGF-1/INS signaling pathway. *Caenorhabditis elegans* with mutations in *daf-2* (homologue to the igf-1/ins receptors) bypass dauer formation and become long-lived (42). Extended lifespan was also observed in *Drosophila melanogasta* mutated in the *daf-2* homologue gene (43) as well as in female mice mutated in the *IGF1R* gene (44). Contrary, knocking out the *INSR* gene in mice resulted in mice with decreased lifespan (45) while increased lifespan was observed if knocking out the *INSR* gene in adipose tissue only (46), illustrating that things often become more complex when investigating mammals. In any case, it appears that reduced GH/IGF-1/INS signaling extends lifespan in the model organisms. On the whole, these model systems have pointed to genes involved in maintenance and repair mechanisms, in metabolism and in anti-oxidant activities (47). Nevertheless, the biology and importantly the life circumstances of these animal models do clearly not resemble human beings, still they might point to some conserved mechanisms of aging and longevity and might for that reason be used for hypothesis generation.

Finally, for the exploration of the cellular changes occurring during aging and the molecular basis for such changes, cellular model systems have been employed; budding yeast *Saccharomyces cerevisiae* is a commonly applied system and so is human cells in culture (23, 48). Special kinds of human cells important for the study of the cellular mechanisms are cells derived from humans affected by premature aging syndromes such as Hutchinson–Gilford progeria, Werner syndrome and Rothmund-Thomson syndrome (20, 49).

Genetic studies of human longevity

It seems intuitively acceptable to state that living to very high ages runs in families; most of us probably know old siblings still going strong, maybe despite ‘unfortunate’ habits like smoking. Indeed it has been shown that siblings of centenarians have higher chances of becoming centenarians themselves compared to other members of their birth cohort, and that the survival advantage of family members of long-lived individuals is lifelong (50). The genetic component of longevity has been investigated in twin studies making use of the genetic similarity of dizygotic and monozygotic twins; it has been estimated that 15-25% of the variation in human lifespan is caused by genetic differences (51, 52). Moreover, this genetic contribution to lifespan appears to be minimal before age 65 and most profound from age 85 (53). Hence, to identify genetic variants which influence longevity it appears reasonable to study the oldest-old.

For identification of the genetic variants that influence human longevity, some researchers have used family-based studies, for example studies of long-lived siblings (e.g. (54)), enabling linkage analysis as described above. Overall the main advantage of conducting family-based studies is that the potential bias introduced by differences in environment between individuals is reduced. One disadvantage might, though, be that the genetic variations identified may be rare in the general population, i.e. they may be unique for the group of related individuals investigated. However, case-control association studies have by far been the rule, often comparing frequencies of genetic variations in a group of oldest-old to the frequencies in a younger (often middle-aged) control group. Here cohort effects, that is differences in characteristics (which have arisen over time (generations)) of the oldest-olds and the controls can be considered an issue. Such differences could for instance have been mediated by the improvement in living standard and health care occurring over the last century. Moreover, the case-control study is based on the assumptions of a constant effect of the genetic variation (i.e. the effect does not depend on the birth cohort) and of similar initial frequencies of the genetic variation (that is similar frequencies in the control group and in the hypothetical cohort of the oldest-old, i.e. when the oldest-old had the same age as the controls). These assumptions are not always true (55). One special version of the case-control study, in part avoiding these problems, is the comparisons of oldest-old individuals, their off-spring and the genetically unrelated spouses of the off-spring (e.g. (56)). In this setup it is investigated whether certain genetic variations are enriched or reduced in frequency among the oldest-old and their off-spring opposed to the spouses (serving as the control group). The advantage of comparing the off-spring of the oldest-old to the controls as opposed to comparing the oldest-old to controls, as it is done in a conventional case-control study, is that the potential bias due to cohort effect is smaller. Anyway, in order to substantiate that initial findings of the case-control studies are not spurious, replication studies of initial findings in additional study populations have become the rule in genetic epidemiological studies of human longevity. One issue complicating this is however, that similar findings for the exact same polymorphism might not necessarily be found in different populations; some investigations indicate that even comparing European populations might be difficult, especially if comparing northern and southern European populations (57, 58). Lastly, prospective cohort studies have also been conducted for following oldest-old individuals from inclusion to death, examples being the Leiden 85-plus cohort (59) and the Danish 1905 birth cohort (60) investigated in this thesis. Furthermore, some prospective cohorts have been established for studying specific age-related diseases, e.g. the Framingham heart study (61), the Cardiovascular Health study (62) and the Copenhagen City Heart Study (63). Still, these prospective cohorts are fewer, probably due to the cost in time and money for constructing them. The major advantage of these cohorts is the avoidance of the cohort-effect bias, although the disadvantage is that a given association identified might potentially be age-span specific, i.e. only relevant for the age-span of the cohort investigated.

Finally, when planning and performing a genetic association study of human longevity there are some principles of population genetics which are important to consider. It appears that longevity is, as it is the case for several of the aging phenotypes, a polygenic trait (a trait influenced by several genes) and therefore we can expect that the effect of each of the individual genetic variants is small and that it is the combined effect of all the variations which contributes to the phenotype (64). Moreover, based on the common disease: common variation hypothesis (65), we can expect the variations to be found in humans with considerable frequency. These ideas have generally been dominant during recent years of genetic epidemiological research on human longevity and they are very suitable for the tagging SNP approach. However, it has been suggested that longevity might be influenced by numerous rare variants and that each of these variants have a larger effect (66), a hypothesis known as the common disease: rare variation hypothesis. For investigating such rare variants one needs to study an exceptionally large number of individuals and possibly needs to perform sequencing for identification of new variants. This will become possible with the technology of next generation sequencing.

Findings from genetic association studies of human longevity

Before the emergence of the tagging SNP approach, researchers often investigated one or a few variations in biologically plausible candidate genes. Often non-synomonous coding variations or other variations with putative functional effects (e.g. located in transcription factor binding sites) were explored. These findings indicated variations in genes involved in insulin signaling (e.g. growth hormone 1 (*GH1*) (67)), antioxidant activity (e.g. superoxide dismutases (*SOD1* and *SOD2*) (68)), maintenance and repair mechanisms (e.g. werner (*WRN*) (69) and heat shock protein 1 A (*HSPA1A*) (70)) and lipoprotein metabolism (e.g. apolipoprotein E (*APOE*) (71)) to influence human longevity. The majority of the initial findings proved difficult to replicate in additional study populations, i.e. finding common genetic variants associated with lifespan turned out to be a difficult task. Actually only one variation, namely the *APOE* ε haplotype, was repeatedly found to pose an effect on longevity.

With the appearance of the HapMap database, the tagging SNP method became an attractive approach due to the possibility to cover almost all of the common genetic variation in a given genomic region. This is different to studying one variation, where one can evaluate the effect of a single SNP, still a lack of association cannot disregard association of other variations in the gene and hence a relevance of variation in the gene as such. By use of the tagging SNP approach new candidate genes were put forward; for example the *FOXO3A* (72), *FOXO1A* (73), and *AKT1* (74) genes of the GH/IGF-1/INS signaling pathway; *APOC3* (75) involved in lipoprotein metabolism, *EXO1* taking part in DNA repair (76) and *DUSP6*, *NALP1* and *PERP* (77) affecting cellular proliferation and differentiation, apoptosis, and p53, respectively. Of these genes *FOXO3A* has so far been the only one showing replication in several study populations. Finally, to date three genome-wide association studies on human longevity have been published; two studies point only to the *APOE* gene (78, 79) and one study points to the *MINPP1* gene, where the gene product is involved in the regulation of cellular proliferation (80). In general a 'lack' of novel findings in GWA studies can be due to too small sample size and thus aging consortia are presently gathering samples. Moreover, since the tagging SNP approach covers common variations only, rare variations, perhaps the causal ones, cannot be captured by this technique.

To summarize, based on the biological theories of aging, the investigations using the biological model systems of aging and longevity and the genetic association studies in humans, a number of biological pathways can be considered as candidate pathways of human longevity. First of all, due to the associations of *APOE* and *FOXO3A* variation observed in humans and the investigations in the model systems, lipoprotein metabolism and the GH/IGF-1/INS signaling pathway must be considered relevant. Moreover, primarily based on the biological theories of aging and the investigations in the model systems, antioxidant, DNA repair, mitochondria, cell cycle regulation, and immune response pathways must be considered potential candidates (39).

2.4. THE AIM OF THE PHD PROJECT

The overall aim of this PhD project was to investigate the association of human longevity with sequence variations in a large number of candidate genes in order to identify genes and gene variations involved in human longevity. Due to the comprehensive number of longevity candidate genes, the genes investigated were limited to three very relevant and promising candidate pathways: the DNA damage signaling and repair, GH/IGF-1/INS signaling and pro-/antioxidant pathways. Moreover, a few genes not belonging to the core functions of these pathways, but which were commonly suggested in the literature, were included.

This objective was implemented by the execution of several sub-projects ultimately resulting in the following manuscripts:

"Common Genetic Variation in the GH/IGF-1/Insulin Signaling, DNA Damage Signaling and Repair and Pro-/Antioxidant Pathways is Associated with Human Longevity" (Paper I)

"Evidence from Case-control and Longitudinal Studies Supports Associations of Genetic Variation in *APOE*, *CETP* and *IL6* with Human Longevity" (Paper II)

Furthermore, during the PhD study several new studies, suggesting specific candidate variations and genes to be associated with longevity and related phenotypes, impelled us to include replication studies of these findings as part of the PhD project. This work resulted in the following papers/manuscripts:

"Replication of an association of variation in the *FOXO3A* gene with human longevity using both case-control and longitudinal data" (Paper III)

"Genetic variation in *TERT* and *TERC* and human leukocyte telomere length and longevity; a cross sectional and longitudinal analysis" (Paper IV)

3. DISCUSSION OF MATERIALS AND METHODS

The details on the majority of the materials and methods of relevance to this PhD project are described in Papers I-IV, hence the following sections will mainly be a discussion of the choice of materials and methods, as well as elaborations on issues for which there were no room in the papers.

For the main genotype-generation of the PhD project we chose to use the GoldenGate platform (Illumina Inc), since it is very suitable for a candidate gene study like the one performed here. Alternatively one might have chosen a genome-wide SNP array (covering the genome by tagging SNPs) and extracted the data for the specific genes. The main reason for not applying that option was financial. Moreover, if we had used a genome-wide SNP array the genetic coverage of the specific genes investigated here might not have been as good as by using the GoldenGate platform. By the use of the GoldenGate platform we genotyped the main study population, while additional study populations were genotyped by other methods for the replication studies of the initial GoldenGate findings and for the study of SNPs associated with telomere length.

3.1. STUDY POPULATIONS AND STUDY DESIGNS

The main study population

The unique nature of our cohorts enabled us to pursue the major aim of the study by employing two types of study designs: the widely used case-control study for investigating the genetic association with survival from middle age to old age and the less applied longitudinal study for exploring the genetic association with survival during old age. Hence, these two approaches investigate two different aspects of longevity, and, as it will be described in the Results and Discussions section, the results found using the two study designs did not always show concordance.

For the case-control study a middle-aged control group of 800 individuals was randomly selected from the Study of Middle Aged Danish Twins (MADT). Among these 800 individuals were no twin pairs, since only one twin from each pair was included. MADT was initiated in 1998 by random selection of 2,640 intact twin pairs from 22 consecutive birth years (1931-1952) via the Danish Central Person Registry (81). As the case group 1,200 oldest-old individuals were randomly selected from the Danish 1905 birth cohort study (1905 cohort), a survey of the entire 1905 birth cohort started in 1998, when the birth cohort members were 92-93 years of age (60). In general the advantage of using these two nation-wide cohorts gathered in a rather genetically homogenous country like Denmark is that we can presume a low degree of heterogeneity. As it will be described in section 3.3, this was indeed observed when estimating the pairwise identity-by-state (IBS) using the GoldenGate data, i.e. population stratification can generally be disregarded. However, as is the case for any case-control study the comparison of for instance allele frequencies between the MADT and 1905 cohort individuals is prone to bias introduced by cohort effects. Therefore, we chose to perform replication studies of novel SNP findings in oldest-old and middle-aged Germans.

To extend our studies we also applied the longitudinal study design, in which the cohort effects are avoided. The Danish 1905 birth cohort study is quite unique since it is a comprehensive nation-wide study of an entire birth cohort alive at ages 92-93. Such nation-wide cohort studies are rare and are only possible in countries like the Scandinavian countries with wide-ranging registration systems. However, one issue must be considered when evaluating the findings based on this cohort; there is a certain degree of selection bias for the genotyped individuals. The complete birth cohort was contacted in 1998 when 3,600 individuals born in 1905 were still alive. Of these, 2,262 chose to participate and 1,651 gave blood samples. In a subsequent analysis of demographic characteristics, the participants were found to be similar to the nonparticipants, except that there were more males and residents from rural areas among the participants. Moreover, despite no difference in hospitalization, the immediate death rate within the first six months after intake was higher among the non-participants, indicating that terminal illness was a likely reason for nonparticipation (60). Of specific interest to the study presented here, the 1,651 individuals giving blood had to be cognitively functioning to a certain extent, since blood samples were not taken without prior consent. A comparison of the mean mini-mental state examination (MMSE) at intake in 1998 shows a significantly lower mean cognitive score among participants not giving blood than among participants giving blood (data not shown). Moreover, of specific relevance to the longitudinal survival analysis, a calculation of the mean survival time from intake in 1998 to death shows a significantly longer mean survival time for participants giving blood compared to participants not giving blood (data not shown). Such selection bias might affect the survival estimates in absolute numbers (such as risk differences), while the relative difference (such as relative risks) between genotype groups should remain unaffected. Additionally, problems might arise if selection bias is also present with respect to exposure (SNP genotypes). This might be the case for SNPs associated with mortality; fewer individuals holding a mortality allele could have entered the study and given blood, as compared to individuals holding the longevity allele. This kind of potential selection bias might also cause a problem in the case-control study, since the allele frequencies of mortality alleles in the 1905 participants giving blood might be different than among participants not giving blood. Still, for obvious reasons, we cannot check whether selection bias with respect to exposure is present.

Despite these shortcomings, the 1905 cohort must be considered a good and unique opportunity to investigate the genetic contribution to survival during the ninth decade of life. Emphasizing this, the selection from birth to ages 92-93 was similar (1 in 20 individuals) to the selection from ages 92-93 to age 100 for the 1905 birth cohort. Furthermore, by the time of initiation of the PhD project the cohort was nearly extinct, which is advantageous since censoring can be avoided. However, the findings obtained do in principal only account for the 92+ individuals, and therefore in order to inspect for similar effects on survival prior to this age we also inspected novel SNP findings in our case-control data.

Additional study populations

The additional study populations are described in the papers. In short, further Danish cohorts were investigated in the replication study of Paper IV. Of these cohorts the LSADT (the Longitudinal Study of Aging Danish Twins (81)) and the UT cohort (the Unilever Twin Cohort Study (82)) are rather similar in design to the MADT cohort, while the Danish Longitudinal Centenarians Study (DLCS) (83) is quite similar to the 1905 cohort. Hence, the same issues of bias as described above must be considered relevant for these cohorts. In Papers I and II we included replication data from German and Dutch samples. Contrary to the Danish cohorts these study populations were not nation-wide since the participants were not recruited based on a single national registry. The German case-control samples were identified via local registry offices within the different geographic regions of Germany (84) while the Dutch cohort of 85 year olds was inhabitants of the city of Leiden in the Netherlands (59). Therefore, the participants might possibly be somehow less representative of their nations than the Danish cohorts.

3.2. SELECTION OF THE GENETIC VARIANTS

As the original aim of this project was to investigate genetic variation in candidate genes and the association with human aging and longevity, the study was initiated with a thorough literature and database search in order to choose which genes and polymorphisms to explore. One pitfall in this regard is how to choose which databases to use, i.e. some databases appeared obvious (e.g. NCBI), whereas the validity of newer databases seemed more difficult to judge. With the exception of the animal models, I did, however, always use more than one database for the different searches and at all times accordance between databases was regarded as an advantage. A list of all the databases applied can be found in Appendix 1 of this thesis. Furthermore, as for all literature and database mining the searches were up-to-date at the time of execution, still, this might change rapidly thereafter. This was especially true for the development in the releases in the HapMap database during the time of performing the PhD project.

Briefly, a systematic search for candidate pathways, genes and variations was first conducted in the NCBI databases by employing the search terms ‘human longevity’, ‘human aging’, ‘premature aging syndromes’ and ‘age-related disease’ (the latter including specific states such as ‘age-related cognitive decline’ and ‘myocardial infarction’). As already mentioned, the search was conducted at a point in time, when the use of the tagging SNP approach was rather new, i.e. the candidates identified were primarily based on studies of single variations, especially with regard to longevity. Moreover, a systematic search for animal models of aging and longevity was made using NCBI. Publicly available databases put forward by researchers in the field were also consulted and so were non-public databases available to me via collaborations in research consortia (e.g. www.lifespannetwork.nl). Consulting these databases verified the large part of the candidates identified via the NCBI databases. Based on these searches we decided to choose a pathway-based approach, i.e. to focus on DNA damage signaling/DNA repair, the GH/IGF-1/Insulin signaling pathways and pro-/and antioxidants.

Next, the genes covering the core of the three biological pathways had to be chosen. To define a biological pathway is a difficult task; it depends greatly on the level of detail (e.g. should all subunits of a candidate protein complex be included?) and on the width of the pathway (e.g. how does one distinguish the core function of the pathway and the related sub-pathways?). Hence evaluation of the importance of the pathway components can easily become somehow subjective. In any case, via our combined knowledge and thorough mining of several databases, a total of 152 core candidate genes were chosen. In addition, 16 candidate genes, which were not part of the core of these pathways, yet which were commonly discussed in the literature (e.g. *APOE*), were included. In order to ensure proper gene IDs, the Human Gene Nomenclature Committee webpage was checked. A list of the genes accordingly chosen is shown in Appendix 1, while Figures 1-3 on the following pages show the definitions of the three pathways.

Subsequently, the specific chromosomal regions composing the 168 genes were identified via the NCBI and UCSC databases. In addition to the encoding regions the 5,000 bp upstream and 1,000 bp downstream regions were added in order to investigate the genetic variation in the regulatory regions. These regions may not cover the entire regulatory regions of the genes, as these might be further away. Nevertheless, the identification of all the specific regulatory regions of the 168 genes would have been beyond the time frame of the project. Finally, we identified the potentially functional SNPs in each gene region, i.e. non-synonymous SNPs, SNPs located in potential splice or transcription factor binding sites and SNPs potentially inducing frame shifts or nonsense-mediated mRNA decay.

The tagging SNPs in the 168 chromosomal regions were ascertained for the CEU cohort via the HapMap consortium database and were analysed in the HaploView software with the appliance of pairwise-tagging between SNPs (with a minimum LD of r^2 >0.8, see note 1 under References) and exclusion of SNPs having a minor allele frequency (MAF) in the CEU cohort of less than 5%. Furthermore, for each gene region SNPs reported by Illumina Inc. to perform poorly on the GoldenGate platform were excluded. Lastly, due to the length of the PCR products in GoldenGate, the minimum distance between the SNPs was set to 60 bp. In this way 1,536 SNPs were chosen for genotyping.

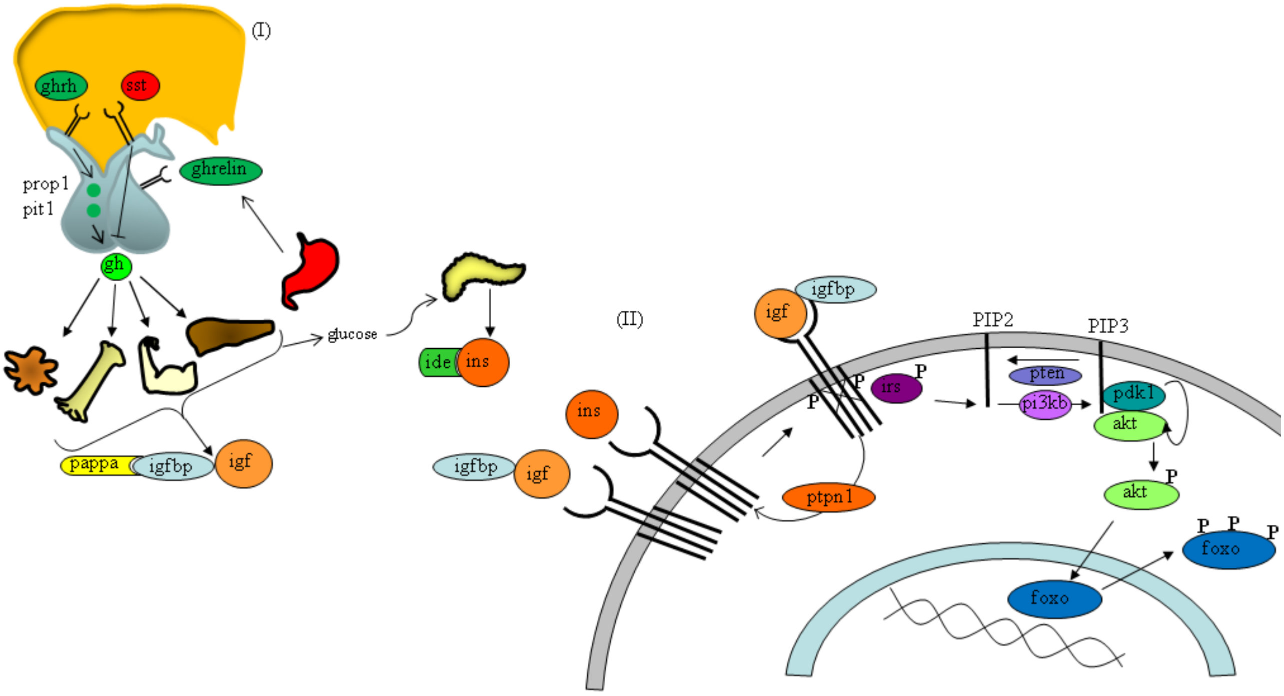


Figure 1: The GH/IGF-1/Insulin pathway

I) The growth hormone releasing hormone (ghrh) is produced in the hypothalamus and is transported to the cell membranes of the somatotroph cells in the pituitary gland, where it binds to its receptor, hereby inducing the transcription factors pit1 and prop1 which leads to gh production. This activity is opposed by somatostatin (sst) and its receptor. Gh is secreted to the serum and in its target organs (primarily the liver, but also others, e.g. fat cells, bone and muscle) gh binds to its receptor and induces signaling cascades such as the mapk/erk and jak-stat pathways, the latter inducing IGF expression. Igf is secreted and in the serum it interacts with igf binding protein (igfbp), affecting both igf level and activity. The level of igfbp is regulated by cleavage by pappa. Several gh feedback mechanisms exist; one is grehlin secreted from the stomach which binds its receptor and stimulates gh secretion. In the liver gh among numerous affects also stimulates glucose synthesis, in turn affecting insulin (ins) production from the islets of Langerhans of the pancreas. In the blood, the level of insulin is affected by the insulin degrading enzyme (ide). (the top left of the figure is adapted from http://edrv.endojournals.org/cgi/content-nw/full/23/5/623/F2 (for free download))

II) In the target cell membrane ins binds to its receptor, and igf binds either to the igf or the insulin receptor. This binding induces auto phosphorylation between receptor dimers, inducing either the pi3kb/akt or the MAPK pathways. In the pi3kb/akt pathway, the insulin receptor substrate (irs) is phosphorylated, which in turn activates pi3kb, then phosphorylating phosphatidylinositol-4,5-biphosphate (PIP2) leading to PIP3. This phosphorylation is counteracted by the pten phosphatase. The akt and pdk1 bind to PIP3 and pdk1 phosphorylates akt. The activated akt phosphorylates different targets including mtor and foxo. Phosphorylation of foxo leads to its displacement from the nucleus to the cytoplasm, hereby inhibiting foxo induced transcription of target genes.

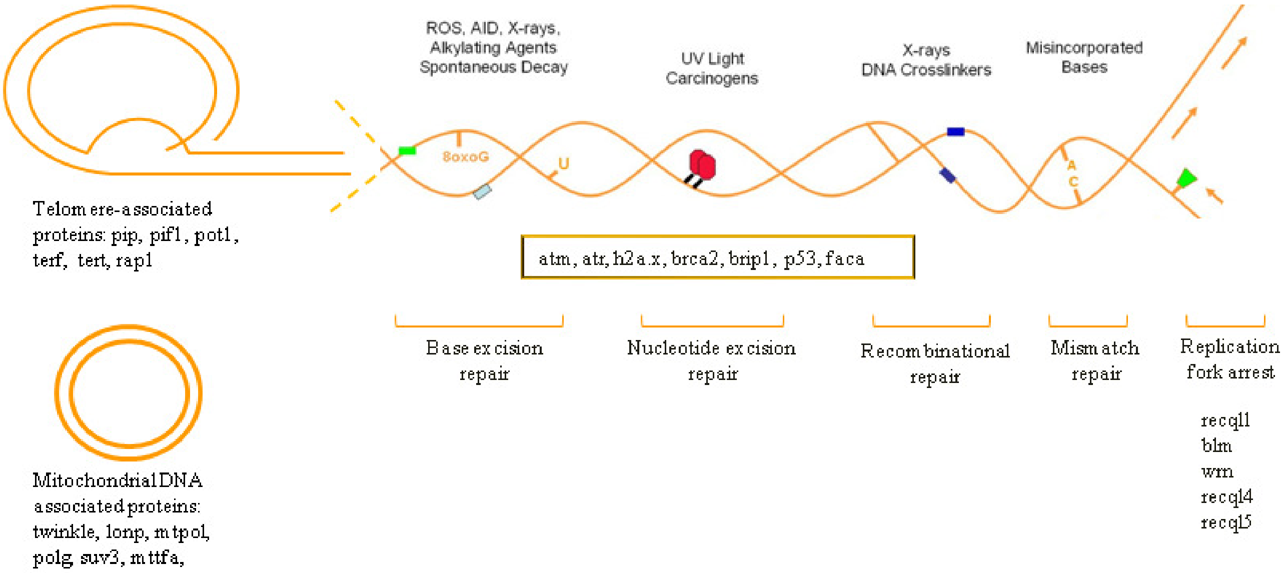

ROS, AID, X-rays, Alkylating Agents Spontaneous Decay
UV Light Carcinogens
X-rays DNA Crosslinkers
Misincorporated Bases
8oxoG
U
A
C
Telomere-associated proteins: pip, pif1, pot1, terf, tert, rap1
atm, atr, h2a.x, brca2, brip1, p53, faca
Base excision repair
Nucleotide excision repair
Recombinational repair
Mismatch repair
Replication fork arrest
recql1
blm
wrn
recql4
recql5
Mitochondrial DNA associated proteins: twinkle, lonp, mtpol, polg, suv3, mttfa,

Figure 2A: DNA stability, damage and repair.

In addition to the DNA damage signaling and repair, we included some genes encoding proteins affecting the stability of telomeres and mitochondrial DNA (left). DNA damage is in general introduced by different sources (top middle) and is detected by different proteins including atm and atr (box middle) signaling to different DNA repair pathways.

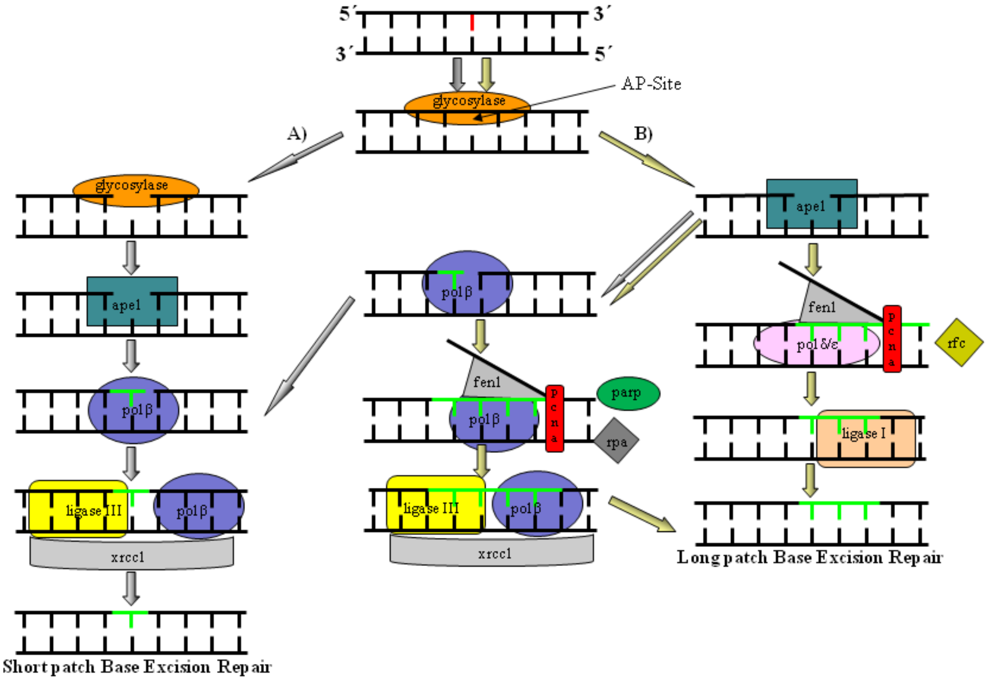


Figure 2B: Base excision repair (BER).

BER is composed of short patch BER (A), where 1 nucleotide is replaced, and long patch BER (B), where 2-7 nucleotides are replaced. A glycosylase removes the damaged base and (A (grey arrows)) ape1 and pnk modify the abasic site (AP site), while pol β mediates DNA synthesis and ligase III seals the gap while xrcc1 stabilizes. (B (yellow arrows)) pol β, pol δ or pol ε mediates pcna stimulated DNA synthesis, fen1 cleaves the single stranded flap structure, and ligase I or III seals the gap.

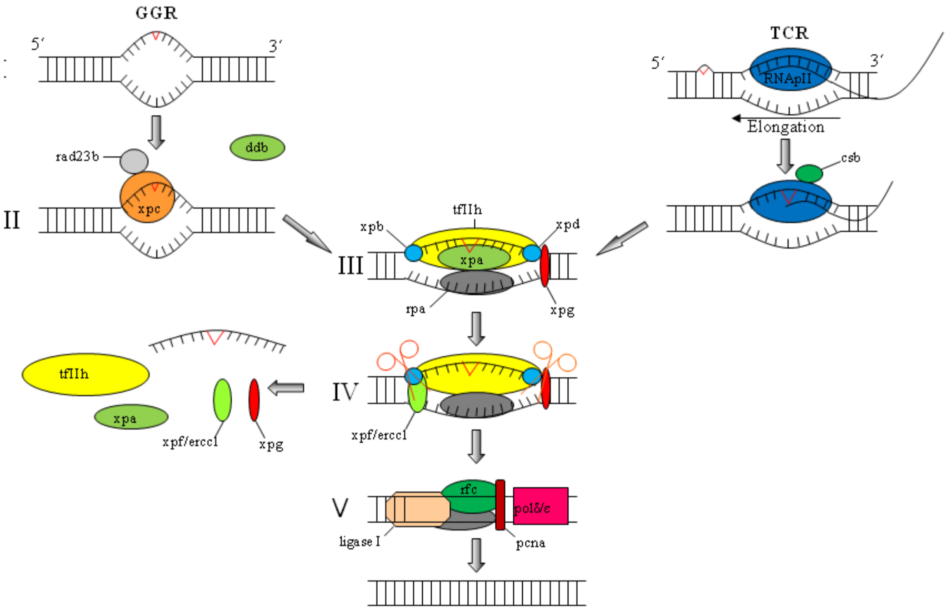


Figure 2C: Nucleotide excision repair (NER).

NER is composed of global genomic NER (GG-NER) and transcription coupled NER (TC-NER). (I-II) xpc-rad23b detects the DNA damage, probably assisted by ddb. In TC-NER the elongating RNA polymerase II is blocked by DNA damage in the transcribed strand. csb probably modulates the RNApII-DNA interface and recruits repair proteins. (III) the tfIIh complex (containing xpb and xpd) is recruited and opens the DNA helix for repair, while xpa binds and verifies the damage. rpa binds to the DNA strand complementary to the damage patch and interacts with xpa, hereby stabilizing the protein repair complex. IV) xpf/ercc1 and xpg cut the single stranded DNA containing the damage. (V) pol δ/ε mediates DNA synthesis and ligase I seals the gap.

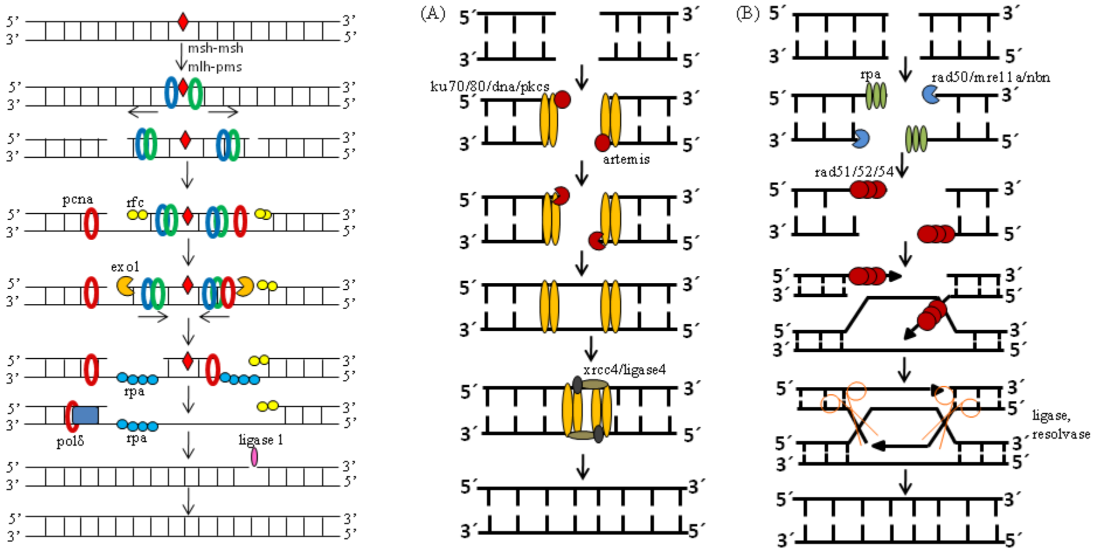


Figure 2D: Mismatch repair (MMR) (left) and Recombinational repair (RCR) (right)

The msh-msh dimer recognizes the mismatch and recruits the mlh-pms dimer and they nick the newly synthesized strand holding the mismatch. Pnca and rcf bind to stabilize. Exo1 is activated and degrades the strand with the mismatch, while rpa binds the single strand. Polδ fills the gap and ligase 1 seals the gap. RCR is divided into (A) Non-homologous end joining (NHEJ) and (B) Homologous recombinatonal repair (HRR). (A) The ku70-ku80-dna-pkcs binds the ends and catalyses the rejoining of the ends of the broken strands. Artemis trims the ends of the DNA strands before rejoining. Xrcc4/ligase4 seals the break. (B) The rad50/mre11a/nbn protein complex digests the damaged ends, rpa binds and rad51, 52 and 54 ”sense” homology between the damaged and undamaged (sister chromatide) helix. Homologous recombination takes place.

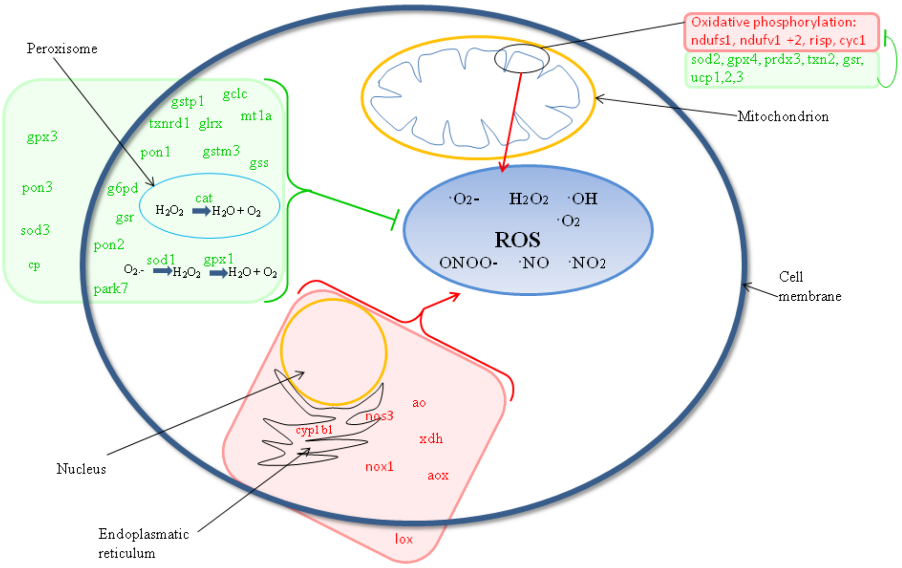


Figure 3: Pro- and antioxidants.

Genes encoding proteins holding direct or indirect pro- or antioxidant effects. Prooxidants are written in red and antioxidants in green. The primary cellular locali zations of the proteins are indicated. The antioxidant reactions catalyzed by catalase, superoxide dismutase and glutathione peroxidase are shown as examples

3.3. GENERATION OF GENOTYPE DATA AND QUALITY CONTROL

In the studies presented in this thesis, DNA was purified from either blood spot cards or whole blood samples and genotypes were determined by either GoldenGate (Illumina Inc., USA), Sequenom MassARRAY iPLEX®Gold (Sequenom Inc., USA) or TaqMan allelic discrimination technologies (Applied Biosystems, USA).

GoldenGate genotyping assay data

For the GoldenGate genotyping of the 1,536 SNPs, DNA was purified from 10 year old blood spot cards; 3 punches from each card were used with the QIAamp DNA Mini and Micro Kits (Qiagen) and purified either by hand or automatically using a QIAcube (Qiagen). Since DNA purified from blood spot cards are often more degraded than DNA purified from whole blood samples, the applicability of the DNA on the GoldenGate platform was initially tested in a small pilot study of 32 blood spot card samples. Since this test appeared satisfactory, DNA was purified from the 2,000 blood spot cards and used for GoldenGate genotyping. The genotyping was outsourced to Aros Biotechnology (Denmark), while data quality control, data cleaning and data validation were performed by us. Moreover, during the data cleanup (see below) we were in contact with colleagues also applying the GoldenGate technology, and our genotype data appeared similar to theirs.

After genotyping the 2,000 samples, the quality of the genotype data was first inspected in the GenomeStudio Genotyping Module (Illumina Inc.) by examining the call rates of the samples, the call rate quality (the so-called GenScore) and by applying the standard quality control procedure of the internal controls of the GoldenGate platform. The latter checks the individual steps of the genotyping procedure: allele specific extension, PCR, hybridizations etc. The results of these check-ups are shown in Appendix 2 of this thesis; they indicated that the genotype data were acceptable. In addition we performed verification experiments. For checking the intra-plate and the inter-plate reproducibility of the system, 24 and 48 samples were included in duplicate. When our cluster definitions (see below) were applied, the data showed an intra-plate reproducibility of 99.4% (using 24 samples) and an inter-plate reproducibility of 96.8% (using 48 samples). Moreover, the 800 MADT samples were re-genotyped for 4 SNPs using TaqMan allelic discrimination technology (as described below), showing a reproducibility of 94.6%. So overall, considering the use of blood spot cards as opposed to full blood samples, the quality of the genotype data was found to be acceptable.

One key issue of epidemiological concern regarding the quality of the genotype data is obviously information bias, i.e. if some individuals are placed in the wrong genotype group it can affect the findings obtained in the subsequent association studies. The blood spot cards of the MADT and 1905 cohorts were collected in the same year, however we experienced that the blood spot cards from the oldest-old individuals tended to be less soaked with blood than the blood spot cards from the middle-aged. This difference might cause a difference in DNA concentration and, hence, in the signal intensity in the GoldenGate. Therefore, to define the SNP clusters we chose to group all the 2,000 samples as one group. First the raw data were re-clustered using the top-120 samples with a call rate above 97.8%. Subsequently, 175 samples with a call rate below 90% were excluded, leaving data on 1,089 oldest-old and 736 middle-aged individuals. SNPs with a call frequency below 90% were excluded, while 254 SNPs with a call frequency between 90-95% were manually checked using the so-called ‘grey zone criteria’ recommended by Illumina Inc.: SNP clusters being close together (score <2.3), clusters with low intensity (score <0.2 or >0.8), clusters having a heterozygote cluster shifting towards a homozygote cluster (score <0.13), SNPs with excess heterozygosity (score <-0.3 or >0.2) and SNPs located on the X chromosome. After this cleaning a total of 142 SNPs were excluded. One issue of this grey zone cleaning is that it can become somehow non-objective due to the subjective evaluation of the SNP plots carried out by the data cleaner. In any case, we chose to perform the grey zone cleaning according to the recommended parameters and then after association analyses to go back to the data and check the quality of the plots of the associated SNPs. Finally, as mentioned above, replication studies of novel findings were conducted in additional German and Dutch samples.

Lastly, before initiating data analysis, the homogeneity of the study population was inspected by calculating the pairwise identity-by-state (IBS) in the Plink software, i.e. the similarity in genotypes of all individuals one to one. All individuals were assigned to the same cluster with a mean IBS of 0.7345 (SE = $1.4*10^{-5}$), an IBS of 1 is complete similarity, that is the study population appeared homogenous. A plot illustrating the similarity of the study population is shown in Appendix 2. It must, however, be mentioned that this IBS estimation is most reliable when genome-wide data are used.

Additional genotyping

For testing the inter-method reproducibility, the 800 MADT samples (genotyped by GoldenGate) were re-genotyped by TaqMan allelic discrimination technology using the Fast or Standard protocols on a StepOne Real Time PCR instrument (Applied Biosystems). The additional samples genotyped for Paper IV were genotyped in the same way, although here both whole blood and blood spot card samples were used. The blood spot card samples in general showed slightly lower intensity.

Finally, for replication of novel findings in Papers I and II, the German and the Dutch whole blood samples were genotyped by Sequenom MassARRAY iPLEX®Gold technology. SNPs that were not compatible with the iPLEX system were genotyped by TaqMan SNP genotyping assays (Applied Biosystems), the German samples via a homemade automated platform (85). The data quality control was done by our collaborators.

3.4. STATISTICAL ANALYSIS METHODS APPLIED

The genetic association studies were performed either as case-control analyses by comparing genotype data in the MADT and 1905 cohorts (GoldenGate data), in the German middle-aged and oldest-old (Papers I and II) or in different age groups (Paper IV), or they were carried out as longitudinal analyses of follow-up survival data of the oldest-old; for the 1905 cohort (GoldenGate data) by regression analyses and for the Dutch replication cohort (Papers I and II) by Cox regression.

Case-control analyses

The case-control analyses were performed either by comparing allele and genotype frequencies between cases and controls by simple χ^2-statistics (Plink software) or by regression analysis using the generalized linear model (R software), the latter in order to adjust for gender. Moreover, haplotype analyses were conducted for 'scanning' the gene regions for combined effects of single-SNPs: a sliding window of 3 SNPs (along their physical position) was applied in Plink. In general, the haplotype-findings supported the findings observed at the single-SNP level, and hence did not bring additional major findings. The case-control analyses were generally performed for both genders combined, as well as separately for the two genders, the reason for the latter being that some studies have indicated gender specific association of SNPs with longevity (86-88). As it will be described in the Results and Discussions section, we did observe gender-specific associations. Furthermore, to consider all biologically relevant genotype models, three genotype models were applied for the genotype analyses: dominant and recessive models, as well as an additive model for the explorative studies (Papers I and II) and an assumption free model for replication studies (Papers III and IV). On the whole there was concordance between the models and significance was not observed in the dominant or recessive models if the estimates for the assumption free or additive models were insignificant. Finally, in Paper II the set-based association test in Plink was used in order to explore the gene as the unit of exploration.

Longitudinal analyses

Of the 1,089 1905 cohort members, for which the GoldenGate genotyping assay data remained after data-cleaning, 14 individuals were alive by the 1 January 2010 (the date of survival update used). To enable investigation of the survival during old age by performing regression analyses, I imputed the remaining life expectancy for these 14 individuals using the www.mortality.org database holding cohort mortality data for the Danish population (based on data from the Statistics Denmark). The advantage of conducting regression analysis as opposed to Cox regression is that it enables the estimation of the quantitative effect (the reduced/increased time lived) for individuals holding the rare allele of an associated SNP. However, if doing Cox regression on the GoldenGate dataset more or less the same SNPs were found to be the most significant, indicating that the same findings could be captured using either procedure (data not shown). I chose to generate two survival variables. First the number of days lived per individual, i.e. the exact number of days lived from intake-date in 1998 to death (for the 14 individuals still alive the imputed date of death), was investigated for application in linear regression. Tests of the assumptions of normal distribution and equal variance (Shapiro-Wilk and Breusch-Pagan/Cook-Weisberg tests in Stata 11) did not show normal distribution, therefore the number of days lived were transformed to the square root for a better fit; these values were designated the Number_of_days_lived variable. As the second survival variable the 1,089 individuals were divided into two groups depending on their time of death: early = living from intake in 1998 to 31 December 2000 (i.e. living to maximum age 95) and late = living to minimum 1 January 2001 (i.e. living to minimum age 95). This variable was termed the Early_late death variable. The reasoning for this variable was that the selection in survival to ages 92-93 was similar to the selection in survival from ages 92-93 to 100 for the 1905 birth cohort, and consequently by dividing the oldest-old into the early and late death groups, the genetic contribution to this selection could be explored. The reason for making the cut-off by New Year 2000/2001 was that the distribution of survival times from ages 92-93 and onwards was clearly right skewed (data not shown); by 31 December 2000 approximately half of the males and one third of the females in the cohort had died. Both survival variables were analysed in Plink by linear (Number_of_days_lived) and logistic (Early_late) regression, adjusting for the potential confounders sex and age or stratifying by sex while adjusting for age. As for the case-control analyses, additive, recessive and dominant genotype models were applied.

The quantitative effects of the associated SNPs were estimated for Number_of_days_lived-associated SNPs by regression analysis of the untransformed survival variable for same sex and same age (93 years) individuals, thereby obtaining differences in mean survival time between the homozygotes for the common allele and the individuals holding the rare allele. For Early_late-associated SNPs the risks of being alive by 1 January 2001 were calculated for all genotype groups of an associated SNP using the odds ratio (ORs) attained in the logistic regression and risk differences were obtained for individuals holding the rare allele (with the homozygotes for the common allele as reference). Finally, when performing linear regression analysis it is possible to estimate the proportion of variation in longevity which can be ascribed to the SNPs identified. Hence, the coefficient of determination (R^2) was calculated by regression analysis of the residuals from the regression analysis of the Number-of-days-lived variable and all the associated SNPs (adjusted for age and gender), thus giving R^2 of the combined effect of the SNPs.

Finally, in Paper II the set-based association test for the survival variables was also applied.

Cox regression was carried out in three instances. First, in Paper III a sex-adjusted, left-truncated Cox proportional hazards model was used for investigating survival during old age for the 1,089 1905 cohort members, since the survival variables had not yet been completed. Moreover, Cox regression was also performed in Paper IV, where the non-extinct UT and LSADT cohorts were explored, i.e. censoring was necessary when analyzing these individuals. In Paper IV the survival of the 1905 cohort individuals was also analysed by Cox regression for consistency in analysis. In all cases the fulfillment of the proportional hazard assumption was initially evaluated using Schoenfeld residuals and conducting an Aalen linear hazard model. If not fulfilled with respect to sex, sex-stratified analyses were performed and in case of a change in effect with age, an extended Cox model (splitting up effect into age spans which the Aalen model supported) was conducted. Finally, the Dutch replication data of Papers I and II was also analysed by Cox regression either adjusting for or stratifying by sex. For the Dutch replication data an additive model was used and only in cases where an inspection of Kaplan-Meier plots indicated a recessive or dominant model, such a model was applied.

Analysis of telomere length

Linear regression analysis was also performed with respect to leukocyte telomere length (LTL) in the replication study of Paper IV. Again the assumptions of normal distribution and equal variance were initially tested and the regression analysis was adjusted for the two confounders gender and age at blood sampling. In the same paper, haplotype-based association studies were performed with the survival and LTL variables using the Thesias software (89).

Correction for multiple testing

To sum up, the association analyses of genotype data and longevity data were carried out at several levels: three gender groups (both genders combined and males and females separately), three genotype levels (allele, genotype and haplotype) and at the genotype level three models were assumed. Obviously application of these various tests increases the risk of false positive findings. However, I believe that there were sound a priori arguments for doing the tests, and correction for multiple testing was conducted when relevant. Correction was performed by the permutation approach (max(T) permutation mode set at 10,000 permutations) in the Plink or the R software. However, due to the design of the software we could only correct within each test. Therefore, in the papers we included corrections by Bonferroni (see note 2 under References), Bonferroni Step-down (Holm) (see note 3 under References) and/or Benjamini and Hochberg False Discovery Rate (see note 4 under References) in the Discussion sections. We discussed the consequences of applying the different types of correction methods, e.g. the overly conservative Bonferroni vs. the less conservative False Discovery Rate. One major issue of these tests is that they assume independency between the tests, which is clearly not the case here, and moreover assume independency between SNPs which was not always found to be the case when calculating LD in Plink.

3.5. FUNCTIONAL GENOMICS – MOLECULAR EFFECTS OF THE GENETIC VARIATION; QPCR EXPERIMENTS

As a consequence of the central dogma of molecular biology, the variation observed at the nucleotide level might affect the gene product. Accordingly a SNP in a coding region may introduce an amino acid substitution (a non-synonymous SNP), possibly affecting the activity of the encoded protein, while a SNP in a non-coding region might introduce alternative splicing, affect nonsense-mediated decay of mRNAs or affect the binding of transcription factors, all possibly affecting the level of RNA/protein and the protein activity level. Therefore, in order to investigate the molecular basis of a SNP, functional studies can be carried out.

In this PhD project I have initiated gene expression experiments of some of the genes holding SNPs found to be associated with human longevity e.g. *H2AFX*, *INS*, *RAD52*, *NTHL1*, *RAD23B*, *CETP* and *IL6*. From these studies, data on *IL6* expression have so far been included in a manuscript (Paper IV). I applied approximately 200 whole blood samples collected in PAXgene Blood RNA tubes (Qiagen) during the last wave of sample collection for the MADT cohort (2009-2011). RNA was isolated, the integrity and concentration were inspected (using the RNA 6000 Nano Kit and a Bioanalyser 2100 (Agilent Technologies, US)), and reverse transcription was performed using the High-capacity cDNA Reverse Transcription kit (Applied Biosystems, US). We chose to inspect the gene expression of the candidate genes by comparative ΔΔCt analysis; using duplex reactions of FAM- and VIC-labeled TaqMan gene expression assays (Applied Biosystems, US) for the candidate gene and for an endogenous control gene, respectively. When performing such duplex reactions, the experiment procedure for each candidate gene must be carefully validated; first experiments were carried out checking for equal efficiency of the two assays in single and in duplex reactions (with varying input amounts of cDNA), secondly the dynamic range of the duplex reaction (i.e. the cDNA dilution range where the ΔCt does not vary) was determined. Based on these validation experiments, a dilution of cDNA in the middle of the dynamic range was chosen, and the thresholds of the assays were noted and applied in the subsequent experiments. The real-time PCR reactions were run in triplicates under standard conditions on the StepOnePlus real-time PCR system (Applied Biosystems).

4. RESULTS AND DISCUSSIONS

The detailed results and discussions are given in Papers I-IV, thus the following will be short overviews with additional elaborations in case of novel findings of relevance to the studies.

4.1. STUDIES OF PATHWAY GENES AND CLASSICAL CANDIDATE GENES

Papers I and II hold the findings from the association studies of the GoldenGate genotyping assay data and longevity. In Paper I the data for the 148 genes covering the DNA damage signaling and DNA repair, GH/IGF-1/insulin and pro-/antioxidant pathways are investigated, while Paper II explores genetic variation in 16 commonly discussed candidate genes (including APOE) which are not part of the core function of the three pathways. The association studies were performed at the single-SNP and haplotype levels and for the 16 classical candidate genes also at the gene-level.

Common Genetic Variation in the GH/IGF-1/Insulin Signaling, DNA Damage Signaling and Repair and Pro-/Antioxidant Pathways is Associated with Human Longevity (Paper I)

The aim of this study was to investigate the association of human longevity with common genetic variation in the genes composing three major candidate pathways of longevity: DNA-damage signaling and repair, GH/IGF-1/Insulin signaling and pro-/antioxidant processes. Altogether data on 1,273 SNPs in 148 genes in 1,089 oldest-old and 736 middle-aged Danes were available after data cleaning, which makes this study the largest of its kind to date.

In general more SNPs were found to be associated with longevity than would have been expected simply by chance. In the case-control study 1 SNP in the pro-/antioxidant gene *GSR*, 1 SNP in each of the GH/IGF-1/INS genes *INS*, *KL*, *GHRHR*, *GHSR* and *IGF2R* and 1 SNP in each of the DNA-damage/repair genes *RAD52*, *WRN*, *POLB*, *RAD23B*, *NTHL1*, *XRCC1* and *XRCC5* were found to be associated with longevity after correction for multiple testing.

These findings were supported by haplotype-based analyses. The *INS*, *GHSR*, *IGF2R*, *RAD52*, *WRN*, *NTHL1* and *XRCC1* SNPs showed tendencies of the same direction of effect in the longitudinal study of the 1,089 oldest-old (P<0.11) and the same was true for the *INS*, *RAD52* and *NTHL1* SNPs in the German case-control samples (P<0.11). In all the replication studies the SNPs were first inspected at the same genotype level (allelic or genotypic association) and in the same gender group (both genders combined, males separately or females separately) in which the SNP was initially found to pass correction for multiple testing. However, the SNPs were also inspected in other gender groups and at the other genotype level, consequently not all of the replications mentioned can be considered true replications. However, I believe that showing the same tendency in a different gender group and/or at a different genotype level is supportive of an association when a similar effect was evident in this different gender group and/or at this different genotype level in the initial analysis. Table 1, on the next page, depicts the concordance between the initial case-control findings and the German case-control samples and the Danish longitudinal study. Furthermore it specifies whether the minor variant showed a negative or a positive effect on longevity.

In the longitudinal study of the 1,089 oldest-old Danes, 1 SNP in each of the pro-/antioxidant genes *TXNRD1* and *XDH*, 1 SNP in the GH/IGF-1/Insulin gene *GHRL*, 1 SNP in each of the DNA-damage/repair genes *MHL1* and *H2AFX* and 2 SNPs in the DNA-damage/repair gene *XRCC5* were found to be associated with longevity after correction for multiple testing. In general these associations were more gender-specific than those observed in the case-control study, possibly indicating that the genetic impact becomes more gender-specific at advanced ages. The *XDH* and *GHRL* SNPs showed the same direction of effect in the Danish case-control study, while the *TXNRD1*, *H2AFX* and 1 of the *XRCC5* SNPs posed similar effects in the Dutch prospective cohort. The findings are summarized in Table 2 on the next page.

The strength of performing regression analyses as opposed to Cox regression in the longitudinal analyses is that it enables the estimation of the quantitative effects of the associated SNPs as well as the coefficient of determination, i.e. the amount of variation in longevity in the oldest-old which can be ascribed to the SNPs. When estimating the quantitative effects of the SNPs, rather profound effects were observed: for the SNPs associated with the Early_late variable, a maximum effect was seen for rs26802 (*GHRL*) CC males having a 52.9% higher chance of being alive after 1 January 2001 (as compared to AA males). For the SNPs that were associated with the Number_of_days_lived variable, a maximum effect of 1.66 additional years of survival was seen for rs705649 (*XRCC5*) GG males (as compared to rs705649 AA males with a mean survival time of 2.15 years). Calculation of the coefficient of determination (R^2) for the Number_of_days_lived variable and the 7 SNPs identified in the longitudinal analyses indicated that 4.99% and 20.5% of the variation in longevity for females and males, respectively, could be ascribed to the SNPs. The higher percentage for males might indicate that the genetic impact in old age may be most pronounced in males, at least for the SNPs investigated here.

However, one important point to keep in mind when evaluating the findings above is that the inspection in the longitudinal data of the SNPs initially identified in the case-control study and vice versa is not a replication as such; the case-control analysis investigates survival from middle age to old age, whereas the longitudinal analysis explores survival during old age. Since it cannot necessarily be expected that the effects of all SNPs are constant over the entire age span (e.g. some SNPs might be important for survival in the younger elderly, but may loose their importance at the highest ages), a lack of concordance could simply mean that the effect is relevant in one part of old age but not in other parts, while concordance indicates an effect throughout old age.

Moreover, Tables 1 and 2 show that primarily longevity variants are seen, i.e. the rare alleles are enriched in the oldest-old as compared to the middle-aged (case-control analysis) or they are associated with increased survival in the oldest-old (longitudinal analysis), that is the majority of the rare alleles of the SNPs has a positive effect on longevity. This observation is in line with a recent study which showed that oldest-old individuals carry the same frequencies of a number of age-related disease risk (minor) alleles as controls, indicating that long life in these individuals was not due to a decreased burden of negatively affecting minor alleles (90).

Finally, if considering the biological roles of the genes holding the associated SNPs, some interesting reflections can be made. The GH/IGF-1/INS genes in general belong to the upper part of the pathway (the synthesis/secretion of growth hormone and signaling to the target cells) and not the lower part of the pathway (the intracellular signaling ultimately affecting foxo), suggesting that this upper part of the pathway poses the most profound effect on longevity. Still, as it will be explained in Section 4.2., if zooming in on *FOXO3A*, we do replicate the associations previously reported by others, although no associations were found for *FOXO3A* when analyzing all the GH/IGF-1/INS SNPs together (Paper I). This illustrates that when correcting for multiple testing in Paper I we might overlook findings (induce false negatives), while at the same time we obviously decrease the risk of false positives. With respect to DNA repair, the majority of the associated SNPs belong to the base excision repair or the recombinational repair pathways, indicating that variation in the repair of DNA lesions induced by for instance ROS or ionizing radiation appears to have effects on longevity. For the latter pathway, SNPs in both pro- and antioxidants were observed.

Pathway	Gene	SNP	Concordance between Danish and German studies	Concordance between case-control and longitudinal studies	Longevity or mortality minor variant
Pro-/ antioxidants	GSR	rs1002149			
GH/IGF-1/ INS	INS	rs3842755	Yes^{3}	Yes^{1}	Longevity
	KL	rs1207362			
	GHRHR	rs2267723			
	GHSR	rs572169		Yes^{5}	Mortality
	IGF2R	rs9456497		Yes^{5}	Longevity
DNA repair	RAD52	rs11571461	Yes^{2}	Yes^{2}	Longevity
	WRN	rs13251813		Yes^{1}	Longevity
	POLB	rs2953983			
	RAD23B	rs1805329			
	NTHL1	rs3211994	Yes^{6}	Yes^{1}	Longevity
	XRCC1♀	rs25487		Yes^{5}	Mortality
	XRCC5♂/♀	rs11685387			

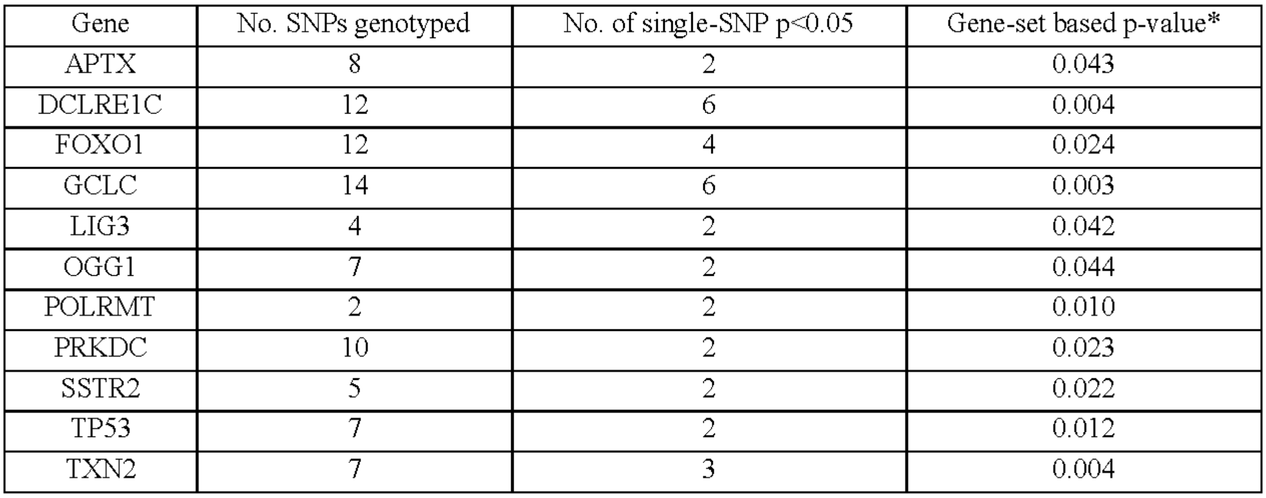
Table 1: Concordance between the initial case-control findings and the German case-control samples and the Danish longitudinal study.

♂: separately for males, ♀: separately for females, 1: significant in the same gender group and at the same genotype level (allelic or genotypic association), 2: as 1 yet with tendency (p<0.11) significance, 3: significant in the same gender group, however at a different genotype level, 4: as 3 yet with tendency, 5: significant at the same genotype level, however in different gender group, 6: as 5 yet with tendency.

Pathway	Gene	SNP	Concordance between Danish and Dutch studies	Concordance between longitudinal and case-control studies	Longevity or mortality minor allele
Pro-/ antioxidants	TXNRD1	rs10047589	Yes^{4}		Longevity
	XDH	rs207444		Yes^{1}	Longevity
GH/IGF-1/ INS	GHRL♂	rs26802		Yes^{6}	Longevity
DNA repair	MLH1♂	rs13320360			
	H2AFX♀	rs2509049	Yes^{5}	No^{2} / Yes^{5}	(Mortality)
	XRCC5♂	rs705649			
	XRCC5♂	rs828910	Yes^{7}		Longevity

Table 2: Concordance between the initial longitudinal findings and the Dutch prospective cohort and the Danish case-control study

♂: separately for males, ♀: separately for females, 1: significant in the same gender group and at the same genotype level (allelic or genotypic association), 2: as yet with tendency (p<0.11) significance, 3: significant in the same gender group, however at a different genotype level, 4: as 3 yet with tendency , 5: significant at the same genotype level, however in different gender group, 6: as 5 yet with tendency , 7: significant in different gender group and different genotype model.



Gene	No. SNPs genotyped	No. of single-SNP p<0.05	Gene-set based p-value*
APTX	8	2	0.043
DCLRE1C	12	6	0.004
FOXO1	12	4	0.024
GCLC	14	6	0.003
LIG3	4	2	0.042
OGG1	7	2	0.044
POLRMT	2	2	0.010
PRKDC	10	2	0.023
SSTR2	5	2	0.022
TP53	7	2	0.012
TXN2	7	3	0.004

Table 3: Gene-based analysis of allele frequency case–control comparisons for both genders combined.

*: The p-values are corrected for the number of SNPs tested within each gene.

Gene-based analysis of pathway genes (unpublished data)

In Paper I the single-SNP and 3-SNP haplotypes in each candidate gene were the units of exploration. However, seen from a biological perspective, it must be considered relevant also to investigate larger units; the gene as a whole is interesting because it encodes the functional unit (the protein) and, moreover, the pathway is relevant since it is the collective functions of the gene products. Therefore we have now initiated gene-based and pathway-based analyses; the former using the set-test in Plink, the latter is carried out applying sum statistics (91, 92).

The gene-based analyses are not performed in order to obtain further support of the associations discovered at the single-SNP level, but rather to find genes in which the combined group of SNPs defines an association. This is exemplified in Table 3 (top of page) listing additional genes found by gene-based analysis. Such candidate genes can be explored further in additional study populations.

Evidence from Case-control and Longitudinal Studies Supports Associations of Genetic Variation in APOE, CETP and IL6 with Human Longevity (Paper II)

In addition to the pathway genes, 16 frequently proposed candidate genes not being part of the core of any of the three pathways were included in the GoldenGate Genotyping Assay: *APOE*, *ACE*, *CETP*, *HFE*, *IL6*, *IL6R*, *MTHFR*, *TGFB1*, *APOA4*, *APOC3*; *SIRTs 1*, *3* and *6*; and *HSPAs 1A*, *1L* and *14*. Surprisingly, with the exception of *SIRT1* and *SIRT3*, no studies employing the tagging SNP approach have previously been published for these genes.

When the 102 tagging SNPs were investigated by the use of the case-control approach variation in the two lipoprotein metabolism related genes *APOE* and *CETP* was found to be disadvantageous and advantageous for longevity, respectively. Haplotype- and gene-based analyses confirmed these associations, while the gene-based analysis also pointed to the heat-shock protein *HSAP14* as a longevity gene. The effects of the *APOE* and *CETP* SNPs were confirmed in the German case-control samples, although for the *CETP* SNP it was at the genotype level. The *CETP* SNP also posed an effect on survival during old age (p = 0.04-0.05) in the Danish oldest-old. The *APOE* SNP turned out simply to be a proxy for rs429358 defining the well-known *APOE* ε haplotype, since adjusting for rs429358 eliminated the effect. The effects are summarized in Table 4 (top of next page).

In the longitudinal analysis of survival during old age only one SNP (in the *IL6* gene) showed a p-value corrected for multiple testing below 0.10, a finding which was supported in the gene-based analysis. The positive effect of the *IL6* SNP on longevity was confirmed in the Dutch replication cohort. The effects are summarized in Table 5 (on the next page). Finally, as the *IL6* SNP was located in potential transcription factor binding sites, qPCR experiments were conducted, still no difference in *IL6* expression was observed between SNP genotypes.

Considering that 102 SNPs were studied, one might object that rather few associations in these commonly suggested candidate genes were observed. However, the genes were previously suggested as candidates mainly due to published studies of single variants, hence our study underlines the importance of examining the variation in the entire gene region (and not just single variants) when evaluating the association of a candidate gene with longevity.

Gene	SNP	OR (Danes)	OR (Germans)	Danish longitudinal study
APOE	rs769449	0.700	0.430	NS
CETP	rs9923854	1.948	1.32 (heterozygotes) 1.76 (rare homozygotes)	OR (Early_late) = 1.380 β-coef. (Number_of_days_lived) = 2.114

Table 4: Effects of the *APOE* and *CETP* SNPs. (OR: odd ratio NS: non-significant)

Gene	SNP	Additional years of survival for AA/AC (Danes)	HR (Dutch)	Danish case-control study
IL6	rs2069827	Males: 0.77 (mean survival CC = 2.91) Females: 0.74 (mean survival CC = 3.74)	0.74	NS

Table 5: Effects of the *IL6* SNP. (HR: hazard ratio, NS: non-significant)

4.2. STUDIES PERFORMED IN RESPONSE TO NOVEL ASSOCIATIONS PUBLISHED BY OTHER RESEARCHERS

The studies of Papers III and IV were carried out due to new and interesting findings regarding the *FOXO3A* and *TERT* genes published by others during the execution of the PhD study. Both genes had been chosen for genotyping in the PhD project, hence we felt impelled to investigate the reported associations in our study populations.

Replication of an association of variation in the FOXO3A gene with human longevity using both case-control and longitudinal data (Paper III)

In September 2008 Willcox et al. (72) published a case-control study of tagging SNPs in 5 genes of the GH/IGF-1/Insulin signaling pathway and their association with longevity in Hawaiian males of Japanese origin. The major findings were the associations of three SNPs in the *FOXO3A* transcription factor encoding gene with longevity, i.e. increased frequencies of the rare alleles were observed in the oldest-old as compared to controls. Advantageous effects of *FOXO3A* SNPs were subsequently confirmed in other case-control studies using study populations of Italian, German, Chinese and American (European descent) origin (88, 93, 94, and 74).

Since, *FOXO3A* had been chosen for investigation in this PhD project we used our data to perform a replication study of 15 *FOXO3A* tagging SNPs in the 1,089 oldest-old and 736 middle-aged Danes. Overall our case-control analyses confirmed the effects of the previously published SNPs as well as showed positive effects of novel SNPs: for 14 SNPs the frequencies of the rare alleles and genotypes tended to be higher in the oldest-old than in the middle-aged controls, while for 1 SNP it was lower. The effects appeared most pronounced in males where 5 and 8 SNPs showed association with longevity at the allele and genotype levels, respectively, while for both genders combined, 5 SNPs showed association with longevity at the genotype level and no associations were seen in females. These findings were supported by haplotype analyses. After correction for multiple testing only estimates at the genotype level remained significant and this was primarily in males. Lastly, as Flachsbart et al. 2009 (94) had reported the effects of the *FOXO3A* SNPs to be more pronounced in centenarians than in nonagenarians, we repeated the analyses separately for the centenarians of the 1905 cohort. We were, however, not able to replicate the increased effect, probably due to our small sample size of centenarian males (N=30). The replications found from our case-control analyses are summarized in Table 6, on the next page.

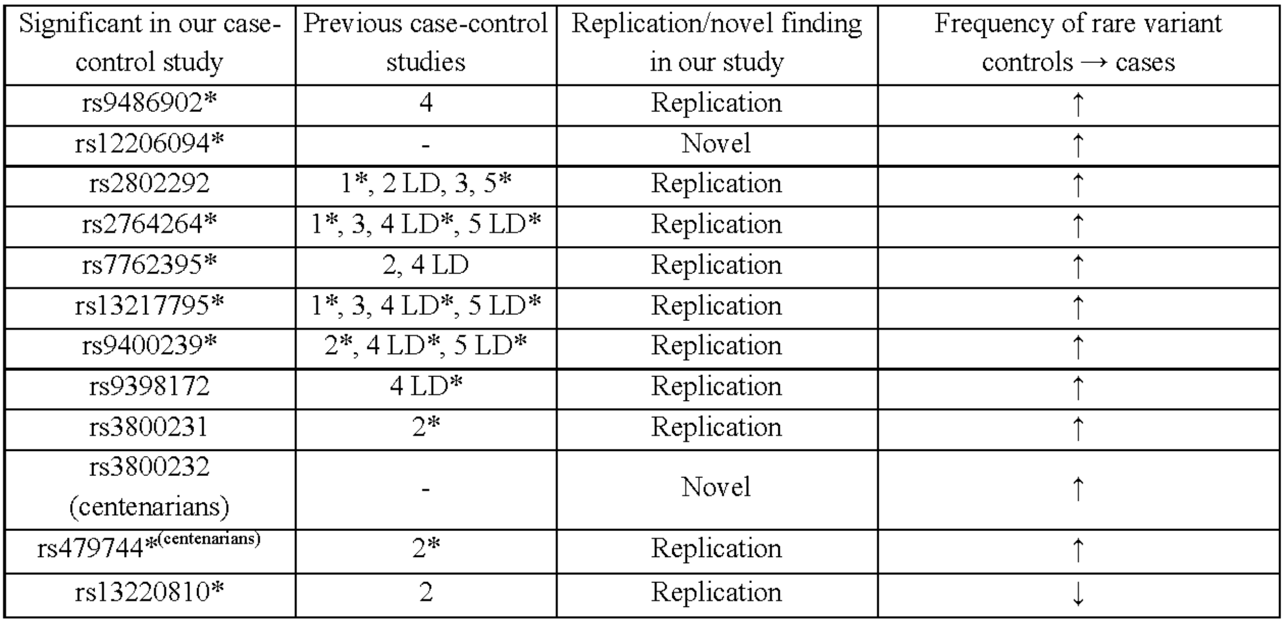
The fact that we observed the associations to be most pronounced in males is interesting considering that we observed no associations for females despite that our sample size of females was approximately 2.5 times that of males. Male-specific effects were also observed for the Italian cohort (93). The original paper by Willcox et al. (72) did report the effects for males, but no females were inspected and hence no conclusion regarding a gender-specific effect can de drawn. Moreover, bearing in mind that our sample size is at least as large as those used in the previous studies, it appears peculiar that our estimates only remain significant after correction for multiple testing at the genotype level. However, in general an assumption of a recessive model for the genotype frequency data showed the lowest p-values, indicating that the association effects in our study population are most prominent for the rare homozygote individuals, and, hence, that the association effects might be 'masked' when analyzing the allele frequencies (that is the heterozygotes might 'dilute' the association effect).

Retrospectively, when considering the number of case-control studies published on *FOXO3A* SNPs and longevity, it would be interesting to perform a meta-analysis of all studies. Such a study would, in addition to possibly verifying the associations, also be useful in order to evaluate the size of the effects of the SNPs (ORs).

We also conducted longitudinal analyses of survival in the oldest-old Danes. At the time of our study only one previous publication had applied this study design to investigate *FOXO3A* SNPs and longevity: a study of the Leiden 85+ cohort (73), still here haplotypes and not single-SNPs were explored. Our longitudinal analyses did not show the same pronounced effects as seen in our case-control analyses. rs10499051 showed a positive effect on survival during old age for females and both genders combined (HR<1), while rs7762395 and rs9486902 displayed negative effects on survival for males (HRs>1). These findings were contradictory to the case-control study where rs10499051 did not have a significant effect, while rs7762395 and rs9486902 showed positive effects. For the latter SNPs this might, though, indicate an antagonistic pleiotropic effect, i.e. the SNPs have positive effects on survival from middle age to old age yet negative effects on survival during old age. Nevertheless, none of these longitudinal findings remained significant after correction for multiple testing (Bonferroni correction) which makes it hard to conclude, especially considering that this study was the first study published holding single-*FOXO3A*-SNP- and follow-up survival data.

However, at the time of publication of our paper, another paper including Cox regression analysis of *FOXO3A* SNP data and longevity was published: a population of 727 Han Chinese individuals aged 92+ in 1998 and living to a minimum age of 100 years in the period 1998-2008 were investigated (95). The analyses were adjusted for several potential confounders, making it somewhat difficult to compare directly to our analysis which was corrected only for gender and age. In any case, 2 of the 3 investigated *FOXO3A* tagging SNPs showed a positive effect on survival during old age (P<0.01 and P=0.01-0.05, respectively). One of these SNPs (rs2802292) was also included in our study, where it did, though, not display significance. Nonetheless, more studies applying follow-up survival data are needed in order to conclude on the effects of *FOXO3A* SNPs on survival during old age.

Finally, a very exciting aspect of the Zeng et al. paper (95) was their examination of a combined effect of *FOXO3A* and *FOXO1A* SNPs on survival; they found that positive effects of *FOXO3A* SNPs and negative effects of *FOXO1A* SNPs observed in the study population compensated one another. We are presently performing a similar study of *FOXO1A*, *FOXO3A* and *FOXO4* SNPs genotyped by the GoldenGate procedure during this PhD project. Lastly, we are also conducting gene-based analysis of the *FOXO* genes applying a recently proposed method (96).



Significant in our case-control study	Previous case-control studies	Replication/novel finding in our study	Frequency of rare variant controls → cases
rs9486902*	4	Replication	↑
rs12206094*	-	Novel	↑
rs2802292	1*, 2 LD, 3, 5*	Replication	↑
rs2764264*	1*, 3, 4 LD*, 5 LD*	Replication	↑
rs7762395*	2, 4 LD	Replication	↑
rs13217795*	1*, 3, 4 LD*, 5 LD*	Replication	↑
rs9400239*	2*, 4 LD*, 5 LD*	Replication	↑
rs9398172	4 LD*	Replication	↑
rs3800231	2*	Replication	↑
rs3800232 (centenarians)	-	Novel	↑
rs479744*$^{\text{(centenarians)}}$	2*	Replication	↑
rs13220810*	2	Replication	↓

Table 6: Replication study of the associations between FOXO3A SNPs and longevity

*: significant after correction for multiple testing, LD: previous study investigates SNP being in LD (r^2>0.8) with our SNP (according to HapMap cohort data), 1) Willcox et al. 2008 (72); 2) Flachsbart et al. 2009 (94), 3) Anselmi et al. 2009 (93), 4) Pawlikowska et al. 2009 (74) and 5) Li et al. 2009 (88).

Genetic variation in TERT and TERC and human leukocyte telomere length and longevity; a cross-sectional and longitudinal analysis (Paper IV)

One major biological theory of aging is the effect of telomere deterioration. Telomerase is the reverse transcriptase which adds telomeres to the ends of chromosomes and hence, genetic variations in the genes encoding the telomerase subunits are obvious candidates for association with variation in telomere length as well as in longevity. In 2010 two GWAS of LTL reported association of the minor alleles of SNPs in the *TERC* gene (encoding the RNA template subunit of the telomerase) with reduced LTL in British and US (European descent) populations (97, 98). The association of one of the SNPs was subsequently confirmed in Han Chinese (99). At the same time a candidate gene study was published showing association of variation in *TERT* (encoding the catalytic subunit of the telomerase) with both LTL and longevity in Ashkenazi Jews (56). *TERT* had been chosen for investigation in this PhD study (as part of the DNA damage signaling/DNA repair pathway), and therefore I decided to thoroughly explore the reported associations in our study population. In order to obtain more power additional members of some of our cohorts for which LTL had previously been determined by Southern Blotting were included. These individuals were genotyped for the 2 *TERC* and 4 *TERT* candidate SNPs and analyses were carried out with respect to both LTL and longevity.

Despite a much larger sample size we did not replicate the findings seen by Atzmon et al. (56) neither regarding LTL nor longevity. Moreover, the study of 11 *TERT* tagging SNPs showed no association. The data analyses with respect to longevity were performed using both the case-control/cross sectional and longitudinal approaches.

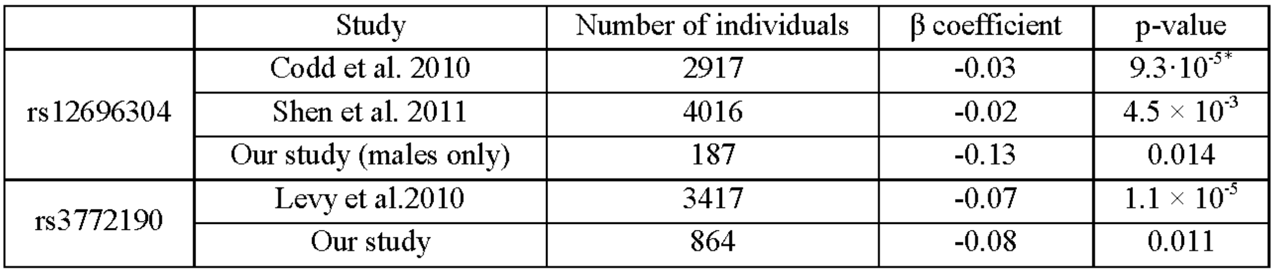
On the contrary, we replicated the associations of the minor alleles of two *TERC* SNPs with reduced LTL. The results of our study and of the previous studies are shown in Table 7 for comparison. When initially testing the assumption of the linear regression analysis, we observed an interaction between gender and rs12696304, and a subsequent sex-stratified analysis showed a significant effect in males, but not in females. One might presume that a test of compliance of the regression model was not conducted in the original GWA studies (97, 98) due to the vast number of SNPs tested, hence pointing to the strength of doing thorough replication studies in which such tests can be performed.

Finally, we investigated the association of the two *TERC* SNPs with longevity using both cross-sectional and longitudinal approaches. The most important findings were seen in the longitudinal analyses where an effect was observed for individuals above age 80: the heterozygotes for rs3772190 showed reduced survival: HR = 1.31, p = 0.009, while the rare homozygotes AA displayed the same but non-significant tendency: HR = 1.32, p = 0.112. Combining the two groups in a dominant model showed: HR = 1.31 p = 0.006. In the view of the theory of telomere deterioration, this is an interesting finding since the rare variant of rs3772190 is associated with both reduced survival during old age and shorter LTL.

Taken together, considering that only two *TERC* SNPs (which we actually estimated to be in profound LD) were explored, an interesting aspect of this study is that variation in the *TERC* gene was found to be associated with both LTL and longevity. On the contrary, 14 SNPs were investigated in *TERT*, and no associations were observed. This indicates that variation in the template for telomere synthesis (terc), not the catalytic subunit (tert), is of relevance for variation in LTL and longevity.

5. CONCLUDING REMARKS

The overall aim of this PhD study was to investigate the association of human longevity with sequence variations in a large number of candidate genes in order to identify genes and gene variations involved in human longevity. Consequently the present PhD thesis makes a contribution to the knowledge about such factors, first of all by exploring tagging SNPs in the genes composing three of the major candidate pathways of human longevity: DNA damage signaling and repair, GH/IGF-1/INS signaling and pro-/antioxidants, and moreover by examining the variation in frequently suggested candidate genes including the replication of specific SNPs. Due to strong a priori hypotheses it was intuitively expected to observe associations, however, the identification of SNPs that influence human longevity is a difficult task, mainly due to issues such as the modest effect sizes of the individual SNPs, statistical power, and the need of replication of initial findings in additional study populations. This is underlined by the fact that despite years of research, so far only the *APOE* and *FOXO3A*



	Study	Number of individuals	β coefficient	p-value
rs12696304	Codd et al. 2010	2917	-0.03	$9.3 \cdot 10^{-5*}$
	Shen et al. 2011	4016	-0.02	4.5×10^{-3}
	Our study (males only)	187	-0.13	0.014
rs3772190	Levy et al.2010	3417	-0.07	1.1×10^{-5}
	Our study	864	-0.08	0.011

Table 7: Replication study of the associations between TERC SNPs and leukocyte telomere length.

*: Replication was performed in 3 additional replication cohorts all showing similar beta coefficients. In a meta-analysis of all the cohorts (9,492 individuals) the p-value was $3.72 \cdot 10^{-14}$.

genes have been convincingly established as longevity-associated genes.

The genetic variation was explored in two homogenous nation-wide cohorts of middle-aged and oldest-old Danes enabling the use of the frequently applied case-control study design, as well as the infrequent employed longitudinal study design of survival during the last part of life. We did not observe the same initial results when applying these two approaches which indicate that the genetic contribution is not constant over the entire age span.

In Paper I we identified variations in genes of all three candidate pathways, some of which showed concordance in the replication populations, with SNPs in the *INS*, *RAD52* and *NTHL1* genes showing the highest degree of replication. The majority of the rare alleles of the identified SNPs were longevity variants, not mortality variants, suggesting that at least in our study population, longevity is primarily affected by positively acting minor alleles. Moreover, since our cohort of oldest-old was nearly extinct, we were able to perform the longitudinal analyses as regression analyses which enabled the estimation of the quantitative effects of the SNPs as well as the coefficient of determination; both indicated rather considerable effects. Hence, new candidates of longevity are put forward for future investigation by others. An interesting aspect of Paper I is the sex-specific effects found in the longitudinal study as compared to the case-control study. Sex-specific effects have been reported by others with respect to longevity (e.g. (95, 100)), as well as regarding aging-related phenotypes (e.g. (101, 102)). The sex-specific effects are suggested to be based on a complex interplay between genetic factors and gender-specific environmental differences. Moreover, gender-specific differences in the occurrence of age-related diseases, such as cardiovascular diseases, have been proposed to contribute to the sex-specific effects on longevity (86). In any case, the observed sex-specific effects point to the environmental influence on longevity and to the interaction between genetic variation and environmental variation, aspects which until now are somehow unexplored in the literature. Finally, the coefficient of determination was found to be larger for males than for females, indicating that the genetic contribution to longevity might be larger for males. Such effects have been suggested for Italian cohorts (87, 103), however it needs to be examined further in our twin cohorts to enable conclusion regarding Danes.

In addition to the pathway candidates, the present PhD thesis contributes to the knowledge about genetic factors influencing human longevity by the investigation of individual frequently suggested candidate genes. In Paper II we explored the 16 classical candidate genes of which 14 genes surprisingly had not previously been explored by a tagging SNP approach. The tagging SNP approach must be considered a stronger approach than the investigation of a single or a few variations due to the coverage of the common genetic variation in the entire gene. We observed association with human longevity of SNPs in the *APOE*, *CETP* and *IL6* genes, but in none of the remaining 13 genes, suggesting that at least in our study population these genes appear not to be relevant for longevity. Lastly, we performed replication studies of SNPs in the *FOXO3A*, *TERT* and *TERC* genes. These studies exemplify that the manner in which the polymorphisms under study are chosen very much influences the conclusions drawn, i.e. when testing many SNPs we might overlook findings (introduce false negatives); when analyzing all the pathway genes together (Paper I) no findings were seen for the *FOXO3A* and *TERT* genes, still, when studying *FOXO3A* separately, we replicated the previously published case-control associations (Paper III). These case-control findings support *FOXO3A* as a persuasive longevity gene.

Future Research and Perspectives

The pathway approach applied in Paper I is a recent strategy (74, 76, 77) which is very reasonable from a biological point of view, since the candidate variations are chosen based on the biological functions of the gene products. An interesting application of the pathway-approach is obviously the pathway-based analysis which evaluates the combined association of the entire biological pathway. Our preliminary analyses indicate associations with longevity of the DNA damage signaling and repair pathway and of the GH/IGF-1/INS pathway, but not of the pro-/antioxidants. In connection with this, considering that most of the SNPs of the DNA repair pathway found to be associated with longevity belonged to certain sub-pathways, it would be exciting to explore such sub-pathways separately. Another appealing aspect is the gene-based analysis and the investigation of joint effects of genes in a given pathway. Both the pathway- and gene-based analyses are compelling since the dependency between SNPs and genes can be taken into account, which seen from a biological perspective is self-evidently relevant. Until now we have examined the data set by applying the simple set-based gene test in Plink (section 3.1), however the use of more advanced methods (such as (96)) would indeed be attractive.

In addition to the abovementioned, the future plans for the use of the GoldenGate data set is to investigate the association with aging-related phenotypes, i.e. to explore the association of the genetic variation in the candidate genes with predictors of mortality in old age such as cognitive and physical impairment, which among other phenotypes have been found to predict mortality in the 1905 cohort (104). Besides the survival data applied in the PhD project, both the 1905 and the MADT surveys included a home-based interview focusing on health issues (e.g. self-reported diseases and self-reported health) and assessment of functional and cognitive abilities using validated tests (e.g. grip strength, activities of daily living (ADL) and MMSE). Moreover, follow-up waves have been carried out for the 1905 cohort in 2001, 2003 and 2005, and for the MADT in 2009-2011, giving rise to longitudinal data on the aging-related phenotypes. Finally, due to the registration system in Denmark, it is possible to apply for additional information such as disease diagnosis, hospital admission and cause of death. We have already applied for and received data from the Danish Cancer Registry on the cohort members genotyped and these data will be employed. For the studies of aging-related phenotypes we plan to perform the analyses at the single-SNP level, but also at the gene and pathway levels. It will indeed be interesting to see whether we find overlap in the genes found to be associated with longevity and the genes associated with the aging-related phenotypes.

Finally, the overall goal of the genetic epidemiological research of human longevity is to understand which biological processes, genes and gene variations influence the inter-individual variation in survival during old age, i.e. which genetic variants contribute to an extreme long lifespan for some people? In my opinion this study is important, not with the aim of prolonging human lifespan but rather in order to get an increased knowledge about the inevitably intertwined aging process and hopefully in the long run enable better prevention, recognition and treatment of age-related diseases and disabilities in order to improve quality of life in old age. So how do we achieve this? It will undoubtedly call for collaboration: to continue the gathering of samples from different study populations in order to increase statistical power, to continue the conduction of replication studies of initial findings obtained from either candidate studies or GWAS studies, to continuously develop and improve the statistical methods, including the handling of studies of gene effects, gene-gene interactions and pathway effects, to continue constructing prospective cohort surveys and family studies holding data on both aging-related phenotypes and survival, and to continue the studies of other genetic aspects such as epigenetics, the roles of the non-coding part of the genome (e.g. micro RNA) and the studies of additional types of genetic variation. The latter was illustrated recently in a study showing association of copy number variations with human longevity (105). Lastly, the next generation sequencing technique opens new possibilities for the identification of novel genetic variants and for new applications such as mRNA sequencing for the investigation of gene expression. Still, with next generation sequencing also follows challenges in data handling and correction for multiple testing. Therefore, the questions and tasks are still many, however by a many-faceted cooperative approach we will hopefully get closer to the answers.

6. SUMMARY

The overall aim of the PhD project was to elucidate the association of human longevity with genetic variation in major candidate genes and pathways of longevity. Based on a thorough literature and database search we chose to apply a pathway approach; to explore variation in genes composing the DNA damage signaling, DNA repair, GH/IGF-1/Insulin signaling and pro-/antioxidant pathways. In addition, 16 genes which did not belong to the core of either pathway, however recurrently regarded as candidate genes of longevity (e.g. *APOE*), were included. In this way a total of 168 genes were selected for investigation. We decided to explore the genetic variation in the form of single nucleotide polymorphisms (SNPs), a highly investigated type of genetic variation. SNPs having potential functional impact (e.g. affecting binding of transcription factors) were identified, so were specific SNPs in the candidate genes previously published to be associated with human longevity. To cover the majority of the common genetic variation in the 168 gene regions (encoding regions plus 5,000 bp upstream and 1,000 downstream) we applied the tagging SNP approach via the HapMap Consortium. Consequently 1,536 SNPs were selected.

The majority of the previous publications on genetic variation and human longevity had employed a case-control study design, e.g. comparing centenarians to middle-aged controls. This type of study design is somehow prone to bias introduced by for instance cohort effects, i.e. differences in characteristics of cases and controls, a kind of bias which is avoided when a prospective cohort is under study. Therefore, we chose to investigate 1,200 individuals of the Danish 1905 birth cohort, which have been followed since 1998 when the members were 92-93 years old. The genetic contribution to human longevity has been estimated to be most profound during the late part of life, thus these oldest-old individuals are excellent for investigating such effect. The follow-up survival data enabled performance of longitudinal analysis, which is quite unique in the field of genetic epidemiology of human longevity. Since the cohort was nearly extinct when initiating the PhD study we were able to conduct the longitudinal analyses as regression analyses enabling the estimation of the quantitative effects of the associated SNPs. However, this study explores the genetic contribution to survival during the ninth decade of life, hence, in order to investigate the genetic contribution to survival in younger elderly we also included 800 individuals of the Study of Middle-aged Danish twins (MADT). MADT was initiated in 1998 by random selection of 2,640 intact twin pairs from 22 consecutive birth years (1931-1952) via the Danish Central Person Registry. Only one twin from each twin pair was included in the PhD study. The inclusion of these 800 middle-aged individuals enabled the performance of case-control analysis. Consequently DNA was purified from 2,000 blood samples, genotyping was carried out via the GoldenGate genotyping platform (Illumina Inc.), quality control and data cleaning were conducted, leading to genotype data on 1,394 SNPs in 1,089 oldest-old and 736 middle-aged individuals.

The genotype data were analysed at the single-SNP and haplotype levels, and for the 16 candidate genes also at the gene level. The analyses of the data set verified the association of a few of the 16 candidate genes not being part of the candidate pathways under study; SNPs in the *APOE*, *CETP* and *IL6* genes, while the analyses of the 152 pathway-related genes pointed to new candidate genes of human longevity; especially SNPs in the *INS*, *RAD52* and *NTHL1* genes appeared promising. As part of these investigations, replication studies of the single-SNP level findings were conducted in German case-control samples of 1,613 oldest-old (ages 95-110) and 1,104 middle-aged individuals and in a Dutch prospective cohort of 563 oldest-old (age 85+). The Dutch oldest-old have been followed for approximately the same number of years as the Danish oldest-old. Interesting aspects of the study were that the majority of the rare alleles of the identified SNPs were longevity variants, not mortality variants, indicating that at least in our study population, longevity is primarily affected by positively acting minor alleles. Moreover, we observed sex-specific differences in the association of the genetic variation with survival during old age.

Furthermore, the genotype data generated were used for a number of replication studies on variation in the *FOXO3A*, *TERT* and *TERC* genes. These studies were performed in response to new data being published on the association of genetic variation in the genes with longevity (*FOXO3A* and *TERT*) and with telomere length (*TERT* and *TERC*). Our studies verified a role of *TERC* in human telomere length and of F*OXO3A* in human longevity (survival from middle age to old age), while a novel role of *TERC* in human longevity was found.

Finally, in addition to the literature and database searches, the genotype data generation and the data analyses mentioned here, RNA purification and qPCR experiments have been initiated in order to investigate gene expression of some of the genes holding SNPs found to be associated with human longevity. Data on one of these genes (*IL6*) have been included in a manuscript.

7. REFERENCES

1. Vaupel JW, Carey JR, Christensen K, Johnson TE, Yashin AI, Holm NV, Iachine IA, Kannisto V, Khazaeli AA, Liedo P, Longo VD, Zeng Y, Manton KG, and Curtsinger JW (1998) Biodemographic trajectories of longevity. Science 280 (5365):855-860

2. Chan GK and Duque G (2002) Age-related bone loss: old bone, new facts. Gerontology 48 (2):62-71

3. Grounds MD (2002) Reasons for the degeneration of ageing skeletal muscle: a central role for IGF-1 signalling. Biogerontology 3 (1-2):19-24

4. Lakatta EG (2000) Cardiovascular aging in health. Clin Geriatr Med 16 (3):419-444

5. Roth GS (1995) Changes in tissue responsiveness to hormones and neurotransmitters during aging. Exp Gerontol 30 (3-4):361-368

6. Taylor BJ and Johnson BD (2010) The pulmonary circulation and exercise responses in the elderly. Semin Respir Crit Care Med 31 (5):528-538

7. 'Molecular Biology of Aging' (2008) Molecular Biology og Aging. Cold Spring Harbor Laboratory Press,

8. Medawar PB (1952) An Unsolved Problem in Biology. Lewis, London

9. Williams GC (1957) Pleiotropy, natural selection and the evolution of senescence. Evolution 11:398-411

10. Kirkwood TB (1977) Evolution of ageing. Nature 270 (5635):301-304

11. Kirkwood TB and Austad SN (2000) Why do we age? Nature 408 (6809):233-238

12. Harman D (1956) Aging: a theory based on free radical and radiation chemistry. J Gerontol 11 (3):298-300

13. Valko M, Izakovic M, Mazur M, Rhodes CJ, and Telser J (2004) Role of oxygen radicals in DNA damage and cancer incidence. Mol Cell Biochem 266 (1-2):37-56

14. Harman D (1991) The aging process: major risk factor for disease and death. Proc Natl Acad Sci U S A 88 (12):5360-5363

15. Finette BA, Sullivan LM, O'Neill JP, Nicklas JA, Vacek PM, and Albertini RJ (1994) Determination of hprt mutant frequencies in T-lymphocytes from a healthy pediatric population: statistical comparison between newborn, children and adult mutant frequencies, cloning efficiency and age. Mutat Res 308 (2):223-231

16. Hamilton ML, Van RH, Drake JA, Yang H, Guo ZM, Kewitt K, Walter CA, and Richardson A (2001) Does oxidative damage to DNA increase with age? Proc Natl Acad Sci U S A 98 (18):10469-10474

17. Bertram JS (2000) The molecular biology of cancer. Mol Aspects Med 21 (6):167-223

18. Lindahl T (1993) Instability and decay of the primary structure of DNA. Nature 362 (6422):709-715

19. de Souza-Pinto NC, Hogue BA, and Bohr VA (2001) DNA repair and aging in mouse liver: 8-oxodG glycosylase activity increase in mitochondrial but not in nuclear extracts. Free Radic Biol Med 30 (8):916-923

20. Bohr VA (2008) Rising from the RecQ-age: the role of human RecQ helicases in genome maintenance. Trends Biochem Sci 33 (12):609-620

21. Weissman L, Jo DG, Sorensen MM, de Souza-Pinto NC, Markesbery WR, Mattson MP, and Bohr VA (2007) Defective DNA base excision repair in brain from individuals with Alzheimer's disease and amnestic mild cognitive impairment. Nucleic Acids Res 35 (16):5545-5555

22. Wong JM and Collins K (2003) Telomere maintenance and disease. Lancet 362 (9388):983-988

23. Campisi J and d'Adda de Fagagna F. (2007) Cellular senescence: when bad things happen to good cells. Nat Rev Mol Cell Biol 8 (9):729-740

24. Harley CB (1991) Telomere loss: mitotic clock or genetic time bomb? Mutat Res 256 (2-6):271-282

25. Lindsey J, McGill NI, Lindsey LA, Green DK, and Cooke HJ (1991) In vivo loss of telomeric repeats with age in humans. Mutat Res 256 (1):45-48

26. Slagboom PE, Droog S, and Boomsma DI (1994) Genetic determination of telomere size in humans: a twin study of three age groups. Am J Hum Genet 55 (5):876-882

27. Njajou OT, Cawthon RM, Damcott CM, Wu SH, Ott S, Garant MJ, Blackburn EH, Mitchell BD, Shuldiner AR, and Hsueh WC (2007) Telomere length is paternally inherited and is associated with parental lifespan. Proc Natl Acad Sci U S A 104 (29):12135-12139

28. Shay JW and Woodring WE (2008) Telomeres and Telomerase in Aging and Cancer. In: Guarente L.P., Partridge L., Wallace D.C. (eds) Molecular Biology of Aging. Cold Spring Harbor Laboratory Press, pp 575-597

29. Kimura M, Hjelmborg JV, Gardner JP, Bathum L, Brimacombe M, Lu X, Christiansen L, Vaupel JW, Aviv A, and Christensen K (2008) Telomere length and mortality: a study of leukocytes in elderly Danish twins. Am J Epidemiol 167 (7):799-806

30. Finch CE (2007) The Biology of Human Longevity. Academic Press, Burlington, MA, USA,

31. Finch CE (1976) The regulation of physiological changes during mammalian aging. Q Rev Biol 51 (1):49-83

32. Franceschi C, Valensin S, BonaFe M, Paolisso G, Yashin AI, Monti D, and De BG (2000) The network and the remodeling theories of aging: historical background and new perspectives. Exp Gerontol 35 (6-7):879-896

33. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N et al (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453 (7191):56-64

34. 1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. Nature 467 (7319):1061-1073

35. Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, and Tsui LC (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245 (4922):1073-1080

36. Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY, and . (1983) A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306 (5940):234-238

37. Ahlqvist E, Ahluwalia TS, and Groop L (2011) Genetics of type 2 diabetes. Clin Chem 57 (2):241-254

38. Slagboom PE, Beekman M, Passtoors WM, Deelen J, Vaarhorst AA, Boer JM, van den Akker EB, van HD, de Craen AJ, Maier AB, Rozing M, Mooijaart SP, Heijmans BT, and Westendorp RG (2011) Genomics of human longevity. Philos Trans R Soc Lond B Biol Sci 366 (1561):35-42

39. Christensen K, Johnson TE, and Vaupel JW (2006) The quest for genetic determinants of human longevity: challenges and insights. Nat Rev Genet 7 (6):436-448

40. McEvoy BP, Powell JE, Goddard ME, and Visscher PM (2011) Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. Genome Res 21 (6):821-829

41. Nachman MW (2002) Variation in recombination rate across the genome: evidence and implications. Curr Opin Genet Dev 12 (6):657-663

42. 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):461-464

43. Giannakou ME and Partridge L (2007) Role of insulin-like signalling in Drosophila lifespan. Trends Biochem Sci 32 (4):180-188

44. Holzenberger M, Dupont J, Ducos B, Leneuve P, Geloen A, Even PC, Cervera P, and Le BY (2003) IGF-1 receptor regulates lifespan and resistance to oxidative stress in mice. Nature 421 (6919):182-187

45. Okamoto H and Accili D (2003) In vivo mutagenesis of the insulin receptor. J Biol Chem 278 (31):28359-28362

46. Bluher M, Kahn BB, and Kahn CR (2003) Extended longevity in mice lacking the insulin receptor in adipose tissue. Science 299 (5606):572-574

47. Kuningas M, Mooijaart SP, van HD, Zwaan BJ, Slagboom PE, and Westendorp RG (2008) Genes encoding longevity: from model organisms to humans. Aging Cell 7 (2):270-280

48. Gershon H and Gershon D (2000) The budding yeast, Saccharomyces cerevisiae, as a model for aging research: a critical review. Mech Ageing Dev 120 (1-3):1-22

49. Pereira S, Bourgeois P, Navarro C, Esteves-Vieira V, Cau P, De Sandre-Giovannoli A, and Levy N (2008) HGPS and related premature aging disorders: from genomic identification to the first therapeutic approaches. Mech Ageing Dev 129 (7-8):449-459

50. Perls TT, 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. Proc Natl Acad Sci U S A 99 (12):8442-8447

51. Herskind AM, McGue M, Holm NV, Sorensen TI, Harvald B, and Vaupel JW (1996) The heritability of human longevity: a population-based study of 2872 Danish twin pairs born 1870-1900. Hum Genet 97 (3):319-323

52. Ljungquist B, Berg S, Lanke J, McClearn GE, and Pedersen NL (1998) The effect of genetic factors for longevity: a comparison of identical and fraternal twins in the Swedish Twin Registry. J Gerontol A Biol Sci Med Sci 53 (6):M441-M446

53. Hjelmborg JVb, Iachine I, Skytthe A, Vaupel JW, McGue M, Koskenvuo M, Kaprio J, Pedersen NL, and Christensen K (2006) Genetic influence on human lifespan and longevity. Human Genetics 119 (3):312-321

54. Geesaman BJ, Benson E, Brewster SJ, Kunkel LM, Blanche H, Thomas G, Perls TT, Daly MJ, and Puca AA (2003) Haplotype-based identification of a microsomal transfer protein marker associated with the human lifespan. Proc Natl Acad Sci U S A 100 (24):14115-14120

55. Lewis SJ and Brunner EJ (2004) Methodological problems in genetic association studies of longevity--the apolipoprotein E gene as an example. Int J Epidemiol 33 (5):962-970

56. Atzmon G, Cho M, Cawthon RM, Budagov T, Katz M, Yang X, Siegel G, Bergman A, Huffman DM, Schechter CB, Wright WE, Shay JW, Barzilai N, Govindaraju DR, and Suh Y (2010) Genetic variation in human telomerase is associated with telomere length in Ashkemazi centenarians. Proc Natl Acad Sci U S A 107,1:1710-1717

57. Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M et al (2008) Correlation between genetic and geographic structure in Europe. Curr Biol 18 (16):1241-1248

58. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, and Bustamante CD (2008) Genes mirror geography within Europe. Nature 456 (7218):98-101

59. Bootsma-van der Wiel A, Van Exel E, de Craen AJ, Gussekloo J, Lagaay AM, Knook DL, and Westendorp RG (2002) A high response is not essential to prevent selection bias: results from the Leiden 85-plus study. J Clin Epidemiol 55 (11):1119-1125

60. Nybo H, Gaist D, Jeune B, Bathum L, McGue M, Vaupel JW, and Christensen K (2001) The Danish 1905 cohort: a genetic-epidemiological nationwide survey. J Aging Health 13 (1):32-46

61. DAWBER TR, MEADORS GF, and MOORE FE, Jr. (1951) Epidemiological approaches to heart disease: the Framingham Study. Am J Public Health Nations Health 41 (3):279-281

62. Fried LP, Borhani NO, Enright P, Furberg CD, Gardin JM, Kronmal RA, Kuller LH, Manolio TA, Mittelmark MB, Newman A, and . (1991) The Cardiovascular Health Study: design and rationale. Ann Epidemiol 1 (3):263-276

63. Nordestgaard BG and Tybjaerg-Hansen A (1999) Susceptibility mutations for ischemic heart disease. Curr Atheroscler Rep 1 (2):108-114

64. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI et al (2009) Finding the missing heritability of complex diseases. Nature 461 (7265):747-753

65. Risch N and Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273 (5281):1516-1517

66. Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69 (1):124-137

67. van Heemst D, Beekman M, Mooijaart SP, Heijmans BT, Brandt BW, Zwaan BJ, Slagboom PE, and Westendorp RG (2005) Reduced insulin/IGF-1 signalling and human longevity. Aging Cell 4 (2):79-85

68. Yashin AI, De BG, Vaupel JW, Tan Q, Andreev KF, Iachine IA, BonaFe M, Valensin S, De LM, Carotenuto L, and Franceschi C (2000) Genes and longevity: lessons from studies of centenarians. J Gerontol A Biol Sci Med Sci 55 (7):B319-B328

69. Castro E, Edland SD, Lee L, Ogburn CE, Deeb SS, Brown G, Panduro A, Riestra R, Tilvis R, Louhija J, Penttinen R, Erkkola R, Wang L, Martin GM, and Oshima J (2000) Polymorphisms at the Werner locus: II. 1074Leu/Phe, 1367Cys/Arg, longevity, and atherosclerosis. Am J Med Genet 95 (4):374-380

70. Altomare K, Greco V, Bellizzi D, Berardelli M, Dato S, DeRango F, Garasto S, Rose G, Feraco E, Mari V, Passarino G, Franceschi C, and De BG (2003) The allele (A)(-110) in the promoter region of the HSP70-1 gene is unfavorable to longevity in women. Biogerontology 4 (4):215-220

71. Schachter F, Faure-Delanef L, Guenot F, Rouger H, Froguel P, Lesueur-Ginot L, and Cohen D (1994) Genetic associations with human longevity at the APOE and ACE loci. Nat Genet 6 (1):29-32

72. Willcox BJ, Donlon T.A., He Q., Chen R., Grove J.S., Yano K., Masaki K.H., Wilcox D.C., Rodriguez B., and Curb J.D. (2008) FOXO3A genotype is strongly associated with human longevity. Proceedings of the National Academy of Sciences 105 (37):13987-13992

73. Kuningas M, Magi R, Westendorp RG, Slagboom PE, Remm M, and van HD (2007) Haplotypes in the human Foxo1a and Foxo3a genes; impact on disease and mortality at old age. Eur J Hum Genet 15 (3):294-301

74. Pawlikowska L, Hu D, Huntsman S, Sung A, Chu C, Chen J, Joyner AH, Schork NJ, Hsueh WC, Reiner AP, Psaty BM, Atzmon G, Barzilai N, Cummings SR, Browner WS, Kwok PY, and Ziv E (2009) Association of common genetic variation in the insulin/IGF1 signaling pathway with human longevity. Aging Cell 8 (4):460-472

75. Atzmon G, Rincon M, Schechter CB, Shuldiner AR, Lipton RB, Bergman A, and Barzilai N (2006) Lipoprotein genotype and conserved pathway for exceptional longevity in humans. PLoS Biol 4 (4):e113

76. Nebel A, Flachsbart F, Till A, Caliebe A, Blanche H, Arlt A, Hasler R, Jacobs G, Kleindorp R, Franke A, Shen B, Nikolaus S, Krawczak M, Rosenstiel P, and Schreiber S (2009) A functional EXO1 promoter variant is associated with prolonged life expectancy in centenarians. Mech Ageing Dev 130 (10):691-699

77. Flachsbart F, Franke A, Kleindorp R, Caliebe A, Blanche H, Schreiber S, and Nebel A (2010) Investigation of genetic susceptibility factors for human longevity - A targeted nonsynonymous SNP study. Mutat Res 694:9-13

78. Deelen J, Beekman M, Uh HW, Helmer Q, Kuningas M, Christiansen L, Kremer D et al (2011) Genome-wide association study identifies a single major locus contributing to survival into old age; the APOE locus revisited. Aging Cell

79. Nebel A, Kleindorp R, Caliebe A, Nothnagel M, Blanche H, Junge O, Wittig M, Ellinghaus D, Flachsbart F, Wichmann HE, Meitinger T, Nikolaus S, Franke A, Krawczak M, Lathrop M, and Schreiber S (2011) A genome-wide association study confirms APOE as the major gene influencing survival in long-lived individuals. Mech Ageing Dev

80. Newman AB, Walter S, Lunetta KL, Garcia ME, Slagboom PE, Christensen K, Arnold AM et al (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. J Gerontol A Biol Sci Med Sci 65 (5):478-487

81. Skytthe A, Kyvik K, Holm NV, Vaupel JW, and Christensen K (2002) The Danish Twin Registry: 127 birth cohorts of twins. Twin Res 5 (5):352-357

82. Gunn DA, Rexbye H, Griffiths CE, Murray PG, Fereday A, Catt SD, Tomlin CC, Strongitharm BH, Perrett DI, Catt M, Mayes AE, Messenger AG, Green MR, van der OF, Vaupel JW, and Christensen K (2009) Why some women look young for their age. PLoS One 4 (12):e8021

83. Andersen-Ranberg K, Schroll M, and Jeune B (2001) Healthy centenarians do not exist, but autonomous centenarians do: a population-based study of morbidity among Danish centenarians. J Am Geriatr Soc 49 (7):900-908

84. Nebel A, Croucher PJ, Stiegeler R, Nikolaus S, Krawczak M, and Schreiber S (2005) No association between microsomal triglyc-

eride transfer protein (MTP) haplotype and longevity in humans. Proc Natl Acad Sci U S A 102 (22):7906-7909

85. Hampe J, Wollstein A, Lu T, Frevel HJ, Will M, Manaster C, and Schreiber S (2001) An integrated system for high throughput TaqMan based SNP genotyping. Bioinformatics 17 (7):654-655

86. Candore G, Balistreri CR, Listi F, Grimaldi MP, Vasto S, Colonna-Romano G, Franceschi C, Lio D, Caselli G, and Caruso C (2006) Immunogenetics, gender, and longevity. Ann N Y Acad Sci 1089:516-537

87. Franceschi C, Motta L, Valensin S, Rapisarda R, Franzone A, Berardelli M, Motta M et al (2000) Do men and women follow different trajectories to reach extreme longevity? Italian Multicenter Study on Centenarians (IMUSCE). Aging (Milano) 12 (2):77-84

88. Li Y, Wang WJ, Cao H, Lu J, Wu C, Hu FY, Guo J, Zhao L, Yang F, Zhang YX, Li W, Zheng GY, Cui H, Chen X, Zhu Z, He H, Dong B, Mo X, Zeng Y, and Tian XL (2009) Genetic association of FOXO1A and FOXO3A with longevity trait in Han Chinese populations. Hum Mol Genet 18 (24):4897-4904

89. Tregouet DA and Garelle V (2007) A new JAVA interface implementation of THESIAS: testing haplotype effects in association studies. Bioinformatics 23 (8):1038-1039

90. Beekman M, Nederstigt C, Suchiman HE, Kremer D, van der BR, Lakenberg N, Alemayehu WG, de Craen AJ, Westendorp RG, Boomsma DI, de Geus EJ, Houwing-Duistermaat JJ, Heijmans BT, and Slagboom PE (2010) Genome-wide association study (GWAS)-identified disease risk alleles do not compromise human longevity. Proc Natl Acad Sci U S A 107 (42):18046-18049

91. Ott J and Hoh J (2003) Set association analysis of SNP case-control and microarray data. J Comput Biol 10 (3-4):569-574

92. Wille A, Hoh J, and Ott J (2003) Sum statistics for the joint detection of multiple disease loci in case-control association studies with SNP markers. Genet Epidemiol 25 (4):350-359

93. Anselmi CV, Malovini A, Roncarati R, Novelli V, Villa F, Condorelli G, Bellazzi R, and Puca AA (2009) Association of the FOXO3A locus with extreme longevity in a southern Italian centenarian study. Rejuvenation Res 12 (2):95-104

94. Flachsbart F, Caliebe A, Kleindorp R, Blanche H, von Eller-Eberstein H, Nikolaus S, Schreiber S, and Nebel A (2009) Association of FOXO3A variation with human longevity confirmed in German centenarians. Proc Natl Acad Sci U S A 106 (8):2700-2705

95. Zeng Y, Cheng L, Chen H, Cao H, Hauser ER, Liu Y, Xiao Z, Tan Q, Tian XL, and Vaupel JW (2010) Effects of FOXO genotypes on longevity: a biodemographic analysis. J Gerontol A Biol Sci Med Sci 65 (12):1285-1299

96. Li MX, Gui HS, Kwan JS, and Sham PC (2011) GATES: a rapid and powerful gene-based association test using extended Simes procedure. Am J Hum Genet 88 (3):283-293

97. Codd V, Mangino M, van der HP, Braund PS, Kaiser M, Beveridge AJ, Rafelt S et al (2010) Common variants near TERC are associated with mean telomere length. Nat Genet 42 (3):197-199

98. Levy D, Neuhausen SL, Hunt SC, Kimura M, Hwang SJ, Chen W, Bis JC et al (2010) Genome-wide association identifies OBFC1 as a locus involved in human leukocyte telomere biology. Proc Natl Acad Sci U S A 107 (20):9293-9298

99. Shen Q, Zhang Z, Yu L, Cao L, Zhou D, Kan M, Li B, Zhang D, He L, and Liu Y (2011) Common variants near TERC are associated with leukocyte telomere length in the Chinese Han population. Eur J Hum Genet

100. Hong MG, Reynolds C, Gatz M, Johansson B, Palmer JC, Gu HF, Blennow K, Kehoe PG, de FU, Pedersen NL, and Prince JA (2008) Evidence that the gene encoding insulin degrading enzyme influences human lifespan. Hum Mol Genet 17 (15):2370-2378

101. Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, Myers RH, Pericak-Vance MA, Risch N, and van Duijn CM (1997) Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. JAMA 278 (16):1349-1356

102. Lehmann DJ, Refsum H, Nurk E, Warden DR, Tell GS, Vollset SE, Engedal K, Nygaard HA, and Smith AD (2006) Apolipoprotein E epsilon4 and impaired episodic memory in community-dwelling elderly people: a marked sex difference. The Hordaland Health Study. J Neurol Neurosurg Psychiatry 77 (8):902-908

103. 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. Eur J Hum Genet Ahead of print

104. Nybo H, Petersen HC, Gaist D, Jeune B, Andersen K, McGue M, Vaupel JW, and Christensen K (2003) Predictors of mortality in 2,249 nonagenarians--the Danish 1905-Cohort Survey. J Am Geriatr Soc 51 (10):1365-1373

105. Kuningas M, Estrada K, Hsu YH, Nandakumar K, Uitterlinden AG, Lunetta KL, van Duijn CM, Karasik D, Hofman A, Murabito J, Rivadeneira F, Kiel DP, and Tiemeier H (2011) Large common deletions associate with mortality at old age. Hum Mol Genet

106. LEWONTIN RC (1964) The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics 49 (1):49-67

NOTES:

1: $r^2 = D^2$/product of the frequencies of all the alleles at two loci. D is an estimate describing the degree of LD for two loci, i.e. the difference between the observed and expected (by chance) frequencies of the haplotype. An r^2 value of 1 is complete dependency between loci, while a value of 0 is complete independency (106).

2: Bonferroni correction: corrected p-value = uncorrected p-value * number of tests

3 Bonferroni Step-down (Holm) correction: corrected p-value = uncorrected p-value * rank. Rank: p-values are ranked from the smallest p-value to the largest, the rank of the smallest p-value is the number of tests performed, the rank of the second smallest p-value is the number of tests performed minus 1, etc.

4: Benjamini and Hochberg False Discovery Rate: corrected p-value = uncorrected p-value * n/n-rank. n: number of tests performed. Rank: p-values are ranked from the largest to the smallest, the largest is left uncorrected, the second largest p-value has rank 1, the third largest rank 2 etc.

ABBREVIATIONS:

1905 cohort: The Danish 1905 birth cohort study
BER: base excision repair
bp: base pair
cDNA: complementary DNA
CEU: the CEPH cohort (Utah residents of northern and western European ancestry, collected by the Centre d'Etude du Polymorphisme Humain)
Ct: cycle threshold
DLCS: the Danish Longitudinal Centenarians Study
DNA: deoxyribonucleic acid
GH/IGF-1/INS pathway: growth hormone 1/ insulin-like growth factor 1 / insulin pathway
GWAS: genome-wide association study
HPA: hypothalamo-pituitary-adrenal (axis)
HR: hazard rate
IBS: identity-by-state
LD: linkage disequilibrium
LSADT: the Longitudinal Study of Aging Danish Twins
LTL: leukocyte telomere length
MADT: the Study of Middle Aged Danish Twins
MAF: minor allele frequency
MMR: mismatch repair
MMSE: mini-mental state examination
mRNA: messenger ribonucleic acid
NCBI: National Center for Biotechnology Information
NER: nucleotide excision repair
NS: non-significant
OR: odds ratio
PCR: polymerase chain reaction
r2: estimate for the degree of LD between loci
R^2: coefficient of determination
RCR: recombinational repair
RD: risk difference
ROS: reactive oxygen species
RR: relative risk
SE: standard error
SNP: single nucleotide polymorphism
UCSC: University of California Santa Cruz

UT: Unilever twin cohort study
The full names of the genes investigated in this PhD project are listed in Appendix 1 of this thesis

8. APPENDIXES

APPENDIX 1:

Databases applied for the literature and database searches

1) Searching for genes with an affect on lifespan of the model organisms:
http://www.ncbi.nlm.nih.gov/sites/entrez

2) Searching for variants and genes associated with human longevity, aging, premature aging syndromes and/or age-related disease:
http://www.ncbi.nlm.nih.gov/sites/entrez, /OMIM, /dbGaP
http://geneticassociationdb.nih.gov
http://genomics.senescence.inf/genes
http://www.hgvbaseg2p.org/phenotypemethod/list
http://www.lifespannetwork.nl

3) Searching for genes constituting the pathways:
http://www.ncbi.nlm.nih.gov/sites/entrez
http://www.biocarta.com
http://www.reactome.org
http://www.genome.utah.edu/genesnps
http://www.genome.jp/kegg/pathway.html
http://www.genome.utah.edu/genesnps/

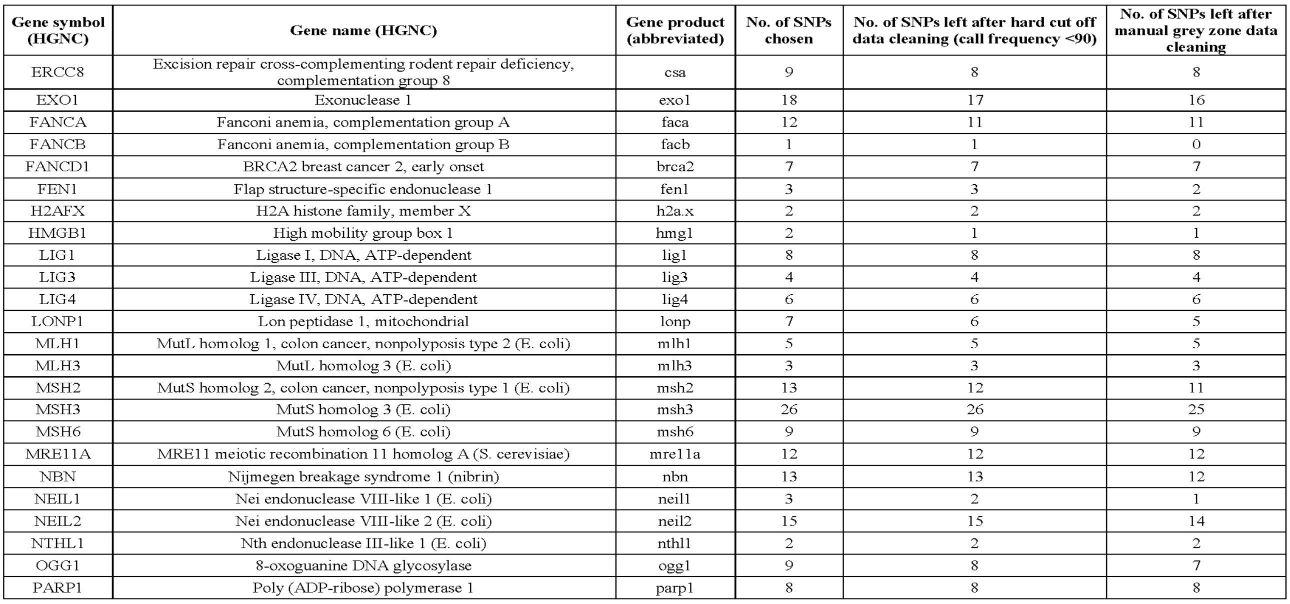
4) Confirming gene IDs and gene chromosomal positions:
http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
http://www.genenames.org/
http://genome.ucsc.edu/

5) Searching for functional SNPs:
http://www.ncbi.nlm.nih.gov/SNP/
http://genome.ucsc.edu/
http://www.genome.utah.edu/genesnps/
http://snpper.chip.org/
http://www.snps3d.org/
http://variome.kobic.re.kr/SNPatPromoter/, /SNP2NMD, /FESD
http://fastsnp.ibms.sinica.edu.tw./pages/input_CandidateGeneSe

arch.jsp
http://manticore.niehs.nih.gov/snpfunc.htm

The IDs of all SNPs were checked by linking back to the http://www.ncbi.nlm.nih.gov/SNP/ database.

Lists of SNPs and genes investigated in this PhD project (see the following pages):



Gene symbol (HGNC)	Gene name (HGNC)	Gene product (abbreviated)	No. of SNPs chosen	No. of SNPs left after hard cut off data cleaning (call frequency <90)	No. of SNPs left after manual grey zone data cleaning
ERCC8	Excision repair cross-complementing rodent repair deficiency, complementation group 8	csa	9	8	8
EXO1	Exonuclease 1	exo1	18	17	16
FANCA	Fanconi anemia, complementation group A	faca	12	11	11
FANCB	Fanconi anemia, complementation group B	facb	1	1	0
FANCD1	BRCA2 breast cancer 2, early onset	brca2	7	7	7
FEN1	Flap structure-specific endonuclease 1	fen1	3	3	2
H2AFX	H2A histone family, member X	h2a.x	2	2	2
HMGB1	High mobility group box 1	hmg1	2	1	1
LIG1	Ligase I, DNA, ATP-dependent	lig1	8	8	8
LIG3	Ligase III, DNA, ATP-dependent	lig3	4	4	4
LIG4	Ligase IV, DNA, ATP-dependent	lig4	6	6	6
LONP1	Lon peptidase 1, mitochondrial	lonp	7	6	5
MLH1	MutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)	mlh1	5	5	5
MLH3	MutL homolog 3 (E. coli)	mlh3	3	3	3
MSH2	MutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)	msh2	13	12	11
MSH3	MutS homolog 3 (E. coli)	msh3	26	26	25
MSH6	MutS homolog 6 (E. coli)	msh6	9	9	9
MRE11A	MRE11 meiotic recombination 11 homolog A (S. cerevisiae)	mre11a	12	12	12
NBN	Nijmegen breakage syndrome 1 (nibrin)	nbn	13	13	12
NEIL1	Nei endonuclease VIII-like 1 (E. coli)	neil1	3	2	1
NEIL2	Nei endonuclease VIII-like 2 (E. coli)	neil2	15	15	14
NTHL1	Nth endonuclease III-like 1 (E. coli)	nthl1	2	2	2
OGG1	8-oxoguanine DNA glycosylase	ogg1	9	8	7
PARP1	Poly (ADP-ribose) polymerase 1	parp1	8	8	8

DNA damage signaling and DNA repair:

(80 genes chosen, 77 genes passed data cleaning. 653 SNPs chosen, 592 SNPs passed data cleaning)

Gene symbol (HGNC)	Gene name (HGNC)	Gene product (abbreviated)	No. of SNPs chosen	No. of SNPs left after hard cut off data cleaning (call frequency <90)	No. of SNPs left after manual grey zone data cleaning
ACD	Adrenocortical dysplasia homolog (mouse)	pip1	2	2	2
APEX1	APEX nuclease (multifunctional DNA repair enzyme) 1	ape1	3	2	2
APTX	Aprataxin	aprataxin	8	8	8
ATM	Ataxia telangiectasia mutated	atm	8	8	8
ATR	Ataxia telangiectasia and Rad3 related	atr	9	9	8
BLM	Bloom syndrome, RecQ helicase-like	blm (recq2)	16	16	15
BRIP1	BRCA1 interacting protein C-terminal helicase 1	facj	13	13	13
C10orf2	Chromosome 10 open reading frame 2	twinkle, mitochondrial	2	2	1
DCLRE1C	DNA cross-link repair 1C	artemis	14	12	12
DDB1	Damage-specific DNA binding protein 1	uv-ddb1	4	4	3
DDB2	Damage-specific DNA binding protein 2	uv-ddb2	7	6	6
ERCC1	Excision repair cross-complementing rodent repair deficiency, complementation group 1	ercc1	6	6	6
ERCC2	Excision repair cross-complementing rodent repair deficiency, complementation group 2	xpd	10	10	9
ERCC3	Excision repair cross-complementing rodent repair deficiency, complementation group 3	xpb	5	4	4
ERCC4	Excision repair cross-complementing rodent repair deficiency, complementation group 4	xpf	4	4	4
ERCC5	Excision repair cross-complementing rodent repair deficiency, complementation group 5	xpg	13	13	13
ERCC6	Excision repair cross-complementing rodent repair deficiency, complementation group 6	csb	13	13	11

Gene symbol (HGNC)	Gene name (HGNC)	Gene product (abbreviated)	No. of SNPs chosen	No. of SNPs left after hard cut off data cleaning (call frequency <90)	No. of SNPs left after manual grey zone data cleaning
PCNA	Proliferating cell nuclear antigen	pcna	4	4	4
PIF1	PIF1 5'-to-3' DNA helicase homolog (S. cerevisiae)	pif1	3	3	3
PMS1	Postmeiotic segregation increased 1 (S. cerevisiae)	pms1	7	7	6
PMS2	Postmeiotic segregation increased 2 (S. cerevisiae)	pms2	6	6	5
POLB	Polymerase (DNA directed), beta	polb	3	3	3
POLD1	Polymerase (DNA directed), delta 1, catalytic subunit 125kDa	pold	7	6	6
POLE	Polymerase (DNA directed), epsilon	pole	11	11	10
POLG	Polymerase (DNA directed), gamma	polg	7	7	6
POLRMT	Polymerase (RNA) mitochondrial (DNA directed)	mtrpol	2	2	2
POT1	Protection of telomeres 1 homolog (S. pombe)	pot1	6	6	5
PNKP	Polynucleotide kinase 3'-phosphatase	pnk	2	1	1
PRKDC	Protein kinase, DNA-activated, catalytic polypeptide	dna-pkcs	10	10	10
RAD23B	RAD23 homolog B (S. cerevisiae)	rad23b	15	15	14
RAD50	RAD50 homolog (S. cerevisiae)	rad50	4	4	3
RAD51	RAD51 homolog (S. cerevisiae	rad51	5	5	5
RAD52	RAD52 homolog (S. cerevisiae)	rad52	12	11	11
RAD54L	RAD54-like (S. cerevisiae)	rad54	7	7	7
RECQL1	RecQ protein-like (DNA helicase Q1-like)	recq1	17	17	17
RECQL4	RecQ protein-like 4	recq4	4	3	2
RECQL5	RecQ protein-like 5	recq5	3	2	1
RFC1	Replication factor C (activator 1) 1, 145kDa	rfc	10	10	10
RPA1	Replication protein A1, 70kDa	rpa	16	15	15
SUPV3L1	Suppressor of var1, 3-like 1 (S. cerevisiae)	suv3	4	4	4
TERF1	Telomeric repeat binding factor (NIMA-interacting) 1	terf1	6	5	5
TERF2	Telomeric repeat binding factor 2	terf2	8	8	8

Gene symbol (HGNC)	Gene name (HGNC)	Gene product (abbreviated)	No. of SNPs chosen	No. of SNPs left after hard cut off data cleaning (call frequency <90)	No. of SNPs left after manual grey zone data cleaning
TERF2IP	Telomeric repeat binding factor 2, interacting protein	rap1	4	4	4
TERC	Telomerase RNA component	terc	1	1	0
TERT	Telomerase reverse transcriptase	tert	16	15	11
TFAM	Transcription factor A, mitochondrial	mttfa	8	8	8
TP53	Tumor protein p53	p53	9	9	7
UNG	Uracil-DNA glycosylase	ung	5	5	5
XPA	Xeroderma pigmentosum, complementation group A	xpa	9	9	8
XPC	Xeroderma pigmentosum, complementation group C	xpc	9	9	8
XRCC1	X-ray repair complementing defective repair in Chinese hamster cells 1	xrcc1	12	12	12
XRRC4	X-ray repair complementing defective repair in Chinese hamster cells 1	xrcc4	18	17	17
XRCC5	X-ray repair complementing defective repair in Chinese hamster cells 1	ku80	20	20	20
XRCC6	X-ray repair complementing defective repair in Chinese hamster cells 1	ku70	3	2	0
YBX1	Y box binding protein 1	ybx1	8	8	8
WRN	Werner syndrome, RecQ helicase-like	wrn (recq3)	18	18	17
Total			**653**	**630**	**592**

GH/IGF-1/Insulin signaling:

33 genes chosen. 420 SNPs chosen, 370 SNPs passed data cleaning

Gene symbol (HGNC)	Gene name (HGNC)	Gene product (abbreviated)	No. of SNPs chosen	No. of SNPs left after hard cut off data cleaning (call frequency <90)	No. of SNPs left after manual grey zone data cleaning
AKT1	v-akt murine thymoma viral oncogene homolog 1	akt	8	8	6
FOXO1	Forkhead box O1	foxo1a	12	12	12
FOXO3	Forkhead box O3	foxo3a	15	15	15
FOXO4	Forkhead box O4	foxo4	2	2	2
GH1	Growth hormone 1	gh	5	5	3
GHR	Growth hormone receptor	ghr	22	22	20
GHRH	Growth hormone releasing hormone	ghrh	3	3	2
GHRHR	Growth hormone releasing hormone receptor	ghrhr	12	12	11
GHRL	Ghrelin/obestatin prepropeptide	ghrelin	10	10	10
GHSR	Growth hormone secretagogue receptor	ghsr	7	7	4
IDE	Insulin-degrading enzyme	ide	12	11	10
IGF1	Insulin-like growth factor 1 (somatomedin C)	igf1	14	14	13
IGF1R	Insulin-like growth factor 1 receptor	igf1r	63	61	57
IGF2	Insulin-like growth factor 2 (somatomedin A)	igf2	16	14	10
IGF2R	Insulin-like growth factor 2 receptor	igf2r	30	28	28
IGFALS	Insulin-like growth factor binding protein , acid labile subunit	igfals	3	3	1
IGFBP2	Insulin-like growth factor binding protein 2, 36kDa	igfbp2	12	12	12
IGFBP3	Insulin-like growth factor binding protein 3	igfbp3	10	9	9
IGF2BP2	Insulin-like growth factor 2 mRNA binding protein 2	igf2bp2	16	16	15
INS	Insulin	ins	2	2	2
INSR	Insulin receptor	insr	32	30	29
IRS1	Insulin receptor substrate 1	irs1	13	12	10
IRS2	Insulin receptor substrate 2	irs2	14	14	14
KL	Klotho	kl	20	20	19
PAPPA	Pregnancy-associated plasma protein A, pappalysin 1	pappa	20	20	18
PDPK1	3-phosphoinositide dependent protein kinase-1	pdk1	2	2	1
PIK3CB	Phosphoinositide-3-kinase, catalytic, beta polypeptide	pi3kb	4	4	2
POU1F1	POU class 1 homeobox 1	pit1	7	7	7
PROP1	PROP paired-like homeobox 1	prop1	6	5	5

Gene symbol (HGNC)	Gene name (HGNC)	Gene product (abbreviated)	No. of SNPs chosen	No. of SNPs left after hard cut off data cleaning (call frequency <90)	No. of SNPs left after manual grey zone data cleaning
PTEN	Phosphatase and tensin homolog	pten	7	7	7
PTPN1	Protein tyrosine phosphatase, non-receptor type 1	ptp1b	9	8	7
SST	Somatostatin	sst	6	5	4
SSTR2	Somatostatin receptor 2	ss2r	6	6	5
Total			**420**	**406**	**370**

Pro-/antioxidants:

39 genes chosen, 38 genes passed data cleaning. 342 SNPs chosen, 311 SNPs passed data cleaning.

Gene symbol (HGNC)	Gene name (HGNC)	Gene product (abbreviated)	No. of SNPs chosen	No. of SNPs left after hard cut off data cleaning (call frequency <90)	No. of SNPs left after manual grey zone data cleaning
ACOX1	Acyl-CoA oxidase 1, palmitoyl	aox	11	11	10
AOX1	Aldehyde oxidase 1	ao	25	25	24
CAT	Catalase	cat	15	15	14
CP	Ceruloplasmin (ferroxidase)	cp	16	15	13
CYP1B1	Cytochrome P450, family 1, subfamily B, polypeptide 1	cyp1b1	9	9	7
CYC1	Cytochrome c-1	cyc1, mitochondrial	2	2	2
G6PD	Glucose-6-phosphate dehydrogenase	g6pd	3	2	2
GCLC	Glutamate-cysteine ligase, catalytic subunit	gclc	16	14	14
GLRX	Glutaredoxin (thioltransferase)	glrx	9	9	8
GPX1	Glutathione peroxidase 1	gpx1	1	0	0
GPX3	Glutathione peroxidase 3 (plasma)	gpx3	11	11	11
GPX4	Glutathione peroxidase 4 (phospholipid hydroperoxidase)	gpx4	4	4	4
GSR	Glutathione reductase	gr	9	8	8
GSS	Glutathione synthetase	gshs	8	8	8
GSTM3	Glutathione S-transferase mu 3 (brain)	gstm3	4	4	3
GSTP1	Glutathione S-transferase pi 1	gstp1	5	5	4
LOX	lysyl oxidase	lox	1	1	1
MT1A	Metallothionein 1A	mt1a	3	3	3
NDUFS1	NADH dehydrogenase (ubiquinone) Fe-S protein 1, 75kDa (NADH-coenzyme Q reductase)	ndufs1	6	5	5
NDUFV1	NADH dehydrogenase (ubiquinone) flavoprotein 1, 51kDa	ndufv1	2	2	1
NDUFV2	NADH dehydrogenase (ubiquinone) flavoprotein 2, 24kDa	ndufv2	7	7	7
NOS3	Nitric oxide synthase 3 (endothelial cell)	nos3	14	13	13
NOX1	NADPH oxidase 1	nox1	8	8	8
PARK7	Parkinson protein 7	park7	7	7	7
PON1	Paraoxonase 1	pon1	28	26	25
PON2	Paraoxonase 2	pon2	13	13	12
PON3	Paraoxonase 3	pon3	6	6	6
PRDX3	Peroxiredoxin 3	prx3	6	6	6
SOD1	Superoxide dismutase 1, soluble	sod1	3	3	3
SOD2	Superoxide dismutase 2, mitochondrial	sod2	5	5	5
SOD3	Superoxide dismutase 3, extracellular	sod3	8	7	6
SRXN1	Sulfiredoxin 1	srxn1	5	5	5
TXN2	Thioredoxin 2	txn2	8	7	7
TXNRD1	Thioredoxin reductase 1	tr	17	16	15
UCP1	Uncoupling protein 1 (mitochondrial, proton carrier)	ucp1	11	11	10
UCP2	Uncoupling protein 2 (mitochondrial, proton carrier)	ucp2	4	4	4
UCP3	Uncoupling protein 3 (mitochondrial, proton carrier)	ucp3	4	4	4
UQCRFS1	Ubiquinol-cytochrome c reductase, Rieske iron-sulfur polypeptide 1	risp	4	4	4
XDH	Xanthine dehydrogenase	xdh	24	24	22
Total			**342**	**329**	**311**

Classical candidate genes:

16 genes chosen. 112 SNPs chosen, 103 SNPs passed data cleaning

Gene symbol (HGNC)	Gene name (HGNC)	Gene product (abbreviated)	No. of SNPs chosen	No. of SNPs left after hard cut off data cleaning (call frequency <90)	No. of SNPs left after manual grey zone data cleaning
ACE	Angiotensin I converting enzyme (peptidyl-dipeptidase A) 1	ace	11	11	11
APOA4	Apolipoprotein A-IV	apoa4	1	1	1
APOC3	Apolipoprotein C-III	apoc3	3	2	2
APOE	Apolipoprotein E	apoe	4	3	3
CETP	Cholesteryl ester transfer protein, plasma	cetp	21	21	20
HFE	Hemochromatosis	hfe	9	8	6
HSPA1L	Heat shock 70kDa protein 1-like	hsp70-1L	4	4	4
HSPA1A	Heat shock 70kDa protein 1A	hsp70-2	2	1	1
HSPA14	Heat shock 70kDa protein 14	hsp70-L1	6	6	6
IL6	Interleukin 6 (interferon, beta 2)	il6	5	4	3
IL6R	Interleukin 6 receptor	il6r	10	9	8
MTHFR	Methylenetetrahydrofolate reductase (NAD(P)H)	mthfr	14	14	13
SIRT1	Sirtuin 1	sirt1	6	5	5
SIRT3	Sirtuin 3	sirt3	10	10	10
SIRT6	Sirtuin 6	sirt6	4	4	3
TGFB1	Transforming growth factor, beta 1	tgfb1	9	9	7
Total			**119**	**112**	**103**

APPENDIX 2: GOLDENGATE GENOTYPE DATA:

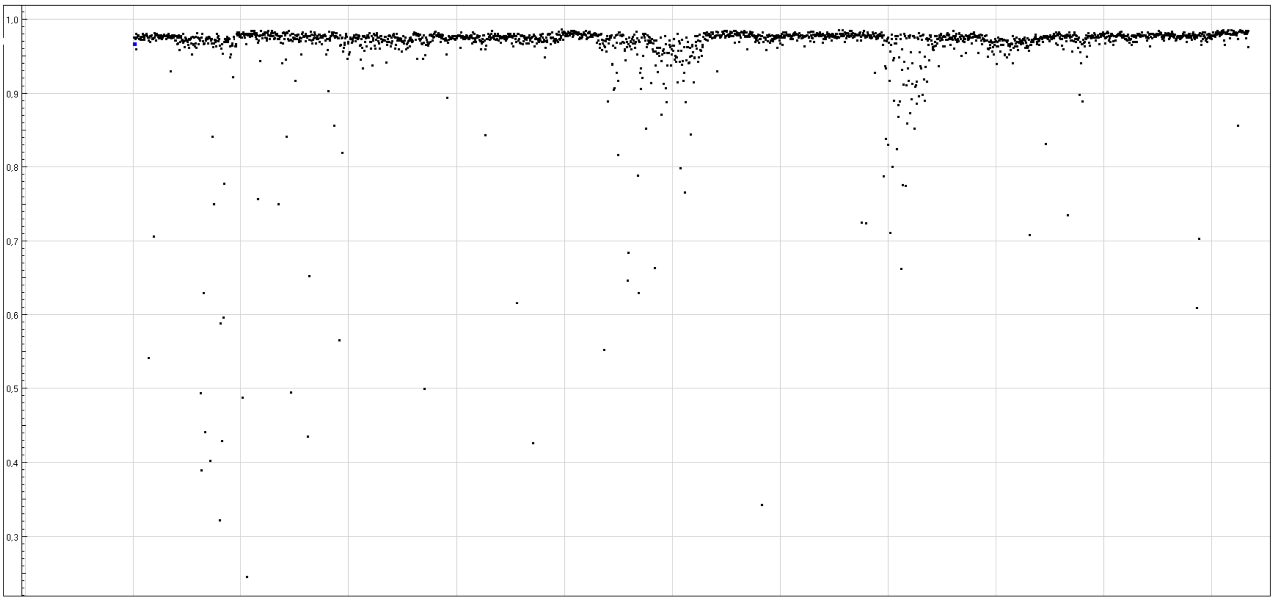


Figure A1: Plot of raw data (the default clustering of GenomeStudio is applied). Each dot is a sample and the call rate of each sample is indicated on the y-axis.

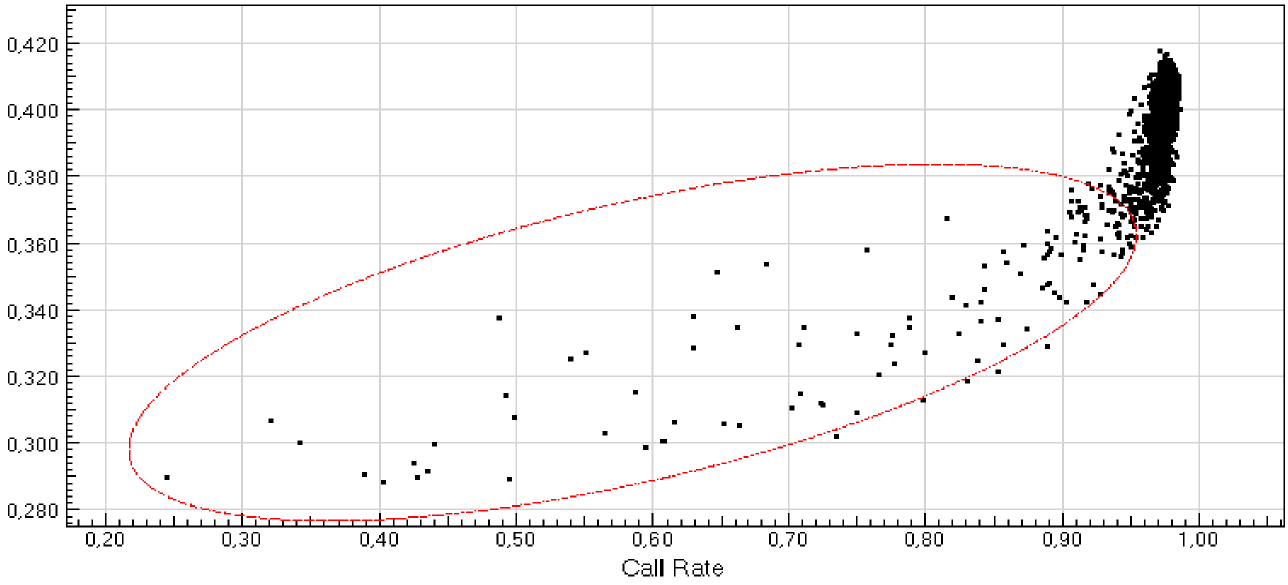


Figure A2: Call rate of raw data vs. GenScores. The GenScore is a quality parameter generated by GenomeStudio describing the quality of the genotypes obtained for each sample. The circle indicates unusable samples. Notice that the majority of the samples clusters together with high call rates and high GenScores.

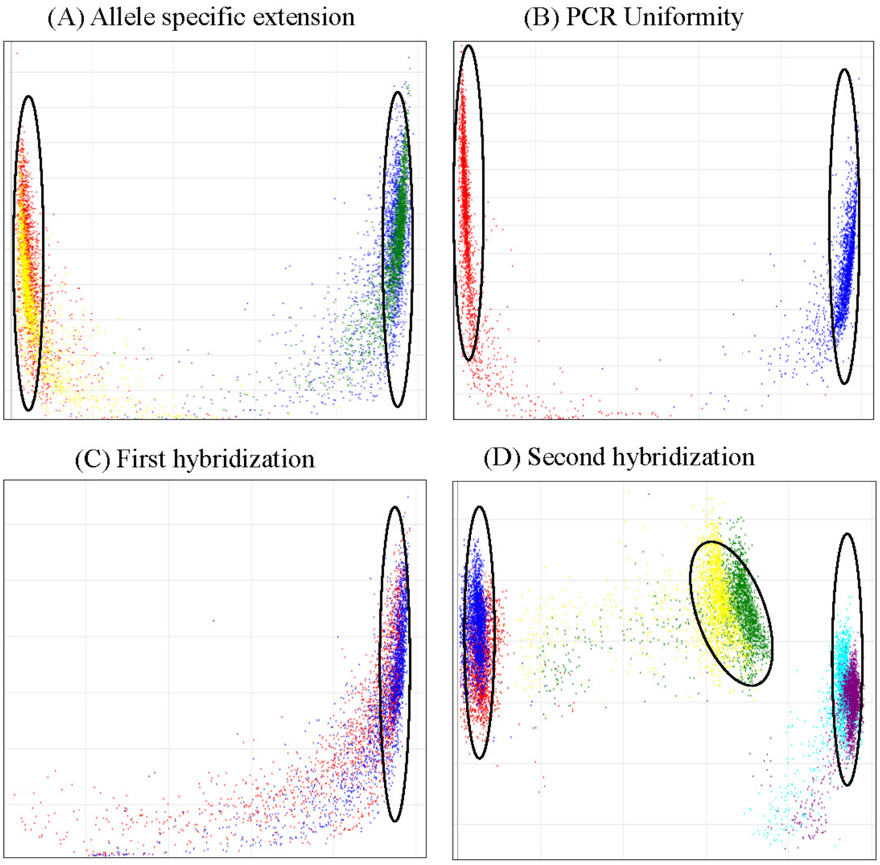


Figure A3: Internal quality controls of the GoldenGate Steps. All plots have the theta dimension on the x-axis (shows the separation of signal from the allele specific probes) and R (intensity of probe signal) is on the y-axis. The circles indicate approximate location in case of optimal genotyping.

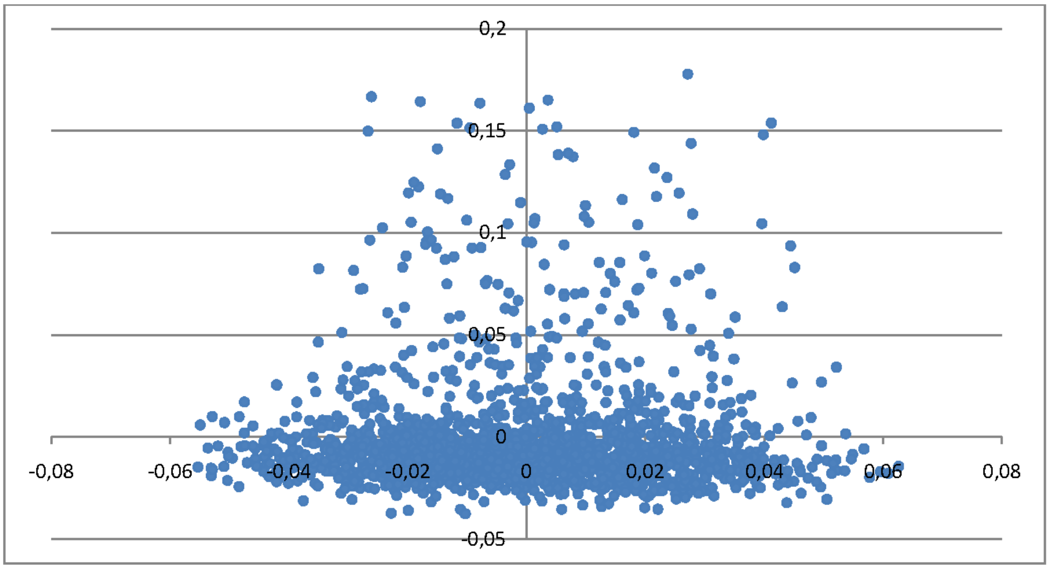


Figure A4: The homogeneity of the main study population (individuals remaining after data cleaning). Identical-by-state plotted in two dimensions. Each dot is an individual. Notice how the majority of the samples clusters together at the bottom.