Next-generation sequencing of centenarians to identify genetic variants predisposing to human longevity

Dissertation

In fulfillment of the requirements for the degree

"Dr. rer. nat."

of the Faculty of Mathematics and Natural Sciences

at Christian-Albrechts-Universität zu Kiel

Submitted by

Nandini Badarinarayan

Research Group for Healthy Ageing

Institute of Clinical Molecular Biology

Kiel, July 2014

First referee: Prof. Dr. Tal Dagan

Second referee: Prof. Dr. Almut Nebel

Date of oral examination: 24th October 2014

Signed:

Prof. Dr. Wolfgang Duschl (Dean)

“Always keep your smile. That's how I explain my long life.”- Jeanne Calment (who had the longest confirmed human lifespan in history, living to the age of 122 years, 164 days)

For my family

Table of Contents

List of figures

No.

List of tables

No.

List of abbreviations

1000G	1000 genomes project
AD	Alzheimer's disease
BGI	Beijing genomics institute
BWA	Burrows-Wheeler aligner
CCA	Case-control analysis
CCDS	Consensus coding sequences database
CNAG	Centre Nacional d'Anàlisi Genòmica
CRG	Centre de Regulació Genòmica
dbSNP	Single nucleotide polymorphism database
DNA	Deoxyribonucleic acid
EDTA	Ethylenediaminetetraacetic acid
ePCR	Emulsion PCR
ESP	NHLBI GO exome sequencing project
GA	Genome analyzer
GATK	Genome analysis toolkit
GEHA	Genetics of healthy ageing
GWAS	Genome-wide association studies
GWS	Genome-wide significance
HAGR	Human ageing genomic resources
HGMD	Human gene mutation database
het/hom	Heterozygous/homozygous variants
HWE	Hardy-Weinberg equilibrium
ICMB	Institute of clinical and molecular biology
IGV	Intergrative genomics viewer
LLI	Long-lived individuals
LMP	Long mate-pair
MADT	Study of middle-aged Danish twins
MAF	Minor allele frequency
MALDI-TOF	Matrix-assisted laser desorption/ionization time-of-flight
mapQV	Mapping quality value
NCBI	National center for biotechnology information
NECS	New England centenarians study
NGS	Next-generation sequencing
NHLBI	National heart, lung, and blood institute
OCS	Okinawa centenarian study
PCR	Polymerase chain reaction
PE	Paired-end
Polyphen	Polymorphism phenotyping
RNA	Ribonucleic acid
ROS	Reactive oxygen species
SICS	Southern Italian centenarian study
SIFT	Sorts intolerant from tolerant
SNAP	Screening for non-acceptable polymorphisms
SNP	Single nucleotide polymorphisms
SNV	Single nucleotide variant
SOLiD	Sequencing by oligonucleotide ligation and detection
SU.VI.MAX	Supplementation en vitamines et mineraux antioxydants study
SVM	Support vector machine

TE	Tris-EDTA
Ti/Tv	Transition/transversion ratio
UCSC	University of California, Santa Cruz
UTR	Untranslated region
WES	Whole exome sequencing
WGS	Whole genome sequencing

1 Introduction

1.1 Longevity phenotype

Longevity is often defined as a complex, polygenic multifactorial phenotype that involves survival to an exceptional age such as 90 years or older, or the potential to survive beyond the species-specific average age at death (Murabito et al. 2012). This definition involves not only the individual's ability to achieve old age but also population-level mortality, measured in this case by mean age at death of a population (or life expectancy) (De Benedictis and Franceschi 2006). Life expectancy is the average number of years that a person at a specific age can expect to live, assuming that age-specific mortality levels remain constant (Oeppen and Vaupel 2002). Over the past two centuries, in developed countries improvements in standard of living and health care have resulted in a significant increase in life expectancy at a steady pace in both males and females. For example, this can be illustrated well in Germany's recent history. During the separation of Germany, mortality in East Germany was comparatively higher than West Germany. However, post-reunification (1989-1990), mortality in East Germany declined among the oldest-old, largely due to improved medical, social, and economic improvements even for the elderly (Oeppen and Vaupel 2002). The average life expectancy in 75 to 85 years in developed countries (Oeppen and Vaupel 2002; Christensen et al. 2006). As there has been a linear acceleration in life expectancy since the 1900s, life expectancy trajectories do not appear to be approaching a maximum. Recent studies have shown that there is an increase in the number of elderly populations as the occurrence of age-related diseases is significantly declining with increase in lifespan (Oeppen and Vaupel 2002; Manton et al. 2006).

The longevity phenotype, without the consideration of health and physical or cognitive function, reflects the overall lifespan. Therefore, it is a heterogeneous phenotype that is influenced by genetic factors as well as by non-genetic factors such as healthcare, nutrition and lifestyle (Murabito et al. 2012). It is commonly accepted that genetic variation explains around 30% of the variability in adult human lifespan and that environmental factors contribute to the remaining 70% in the average-lived populations. However, in populations with more exceptional survivors, the genetic contribution to lifespan may be a lot higher (McGue et al. 1993; Herskind et al. 1996; Willcox et al. 2008).

Longevity studies focus on long-lived individuals (LLI), that is, people surviving to the 95th percentile (and beyond) of their respective birth cohort-specific age distributions (Gudmundsson et al. 2000). In Germany, it is 95 years for females and 92 years for males according to the Human Mortality Database (http://www.mortality.org/). In 1980, it was proposed that one has to markedly delay both morbidity and disability (compression of morbidity hypothesis) towards the end of life in order to survive to a 100 years (Fries 1980). Centenarians or individuals aged 100 years or more, exceed the average human life expectancy by 20 to 25 years, live mostly in good health and show a rapid decline towards the end of their life (Hitt et al. 1999). They are considered as models for successful or healthy ageing as centenarians represent a unique population with a remarkable capability to escape or postpone major age-related diseases until their mid-nineties (Franceschi and Bonafè 2003; Engberg et al. 2009). Healthy ageing is defined as a combination of old age and health, that is, absence of diseases and disabilities along with high physical and cognitive functional capacity and in additional, being socially active (Rowe and Kahn 1997). Longevity studies mainly focus on lifespan, whereas healthy ageing concentrates on healthspan. However, both lifespan and healthspan are closely related as LLI, who live exceptionally long also tend live in good health for most of their lives (Brooks-Wilson 2013). The gradual increase in the number of centenarians in developed countries at a rate of 8% per year is largely attributed to environmental factors (such as improvements in lifestyle and health care) that led to a steady decline in early and late-life mortality (Vaupel 1995; Kirkwood 2008). The frequency of centenarians in the global population is approximately 1 in 10,000 persons. The number of centenarians in the world is expected to increase from 316,600 in 2011 to 3.2 million in 2050 (United Nations Population Fund, 2012).

A new subpopulation of extraordinarily LLI has arisen within the centenarian population, called *supercentenarians* - people aged 110 years or more. Supercentenarians have emerged consistently from the 1970s and the numbers have been growing since as shown in Figure 1-1 (Robine and Vaupel 2001). Until now Jeanne Louise Calment, a French supercentenarian has been confirmed and validated to have the longest human lifespan in history, living to the age of 122 years and 164 days (Robine and Allard 1998). As of July 2014, there are 70 validated living supercentenarians, 65 females and 5 males (Young and Coles 2014). Supercentenarians appear to be more phenotypically homogeneous with respect to morbidity and function than centenarians (Schoenhofen et al. 2006; Sebastiani and Perls 2012). Lethal diseases such as cardiovascular disease and stroke were found to be less common in supercentenarians than in centenarians (Schoenhofen et al. 2006).

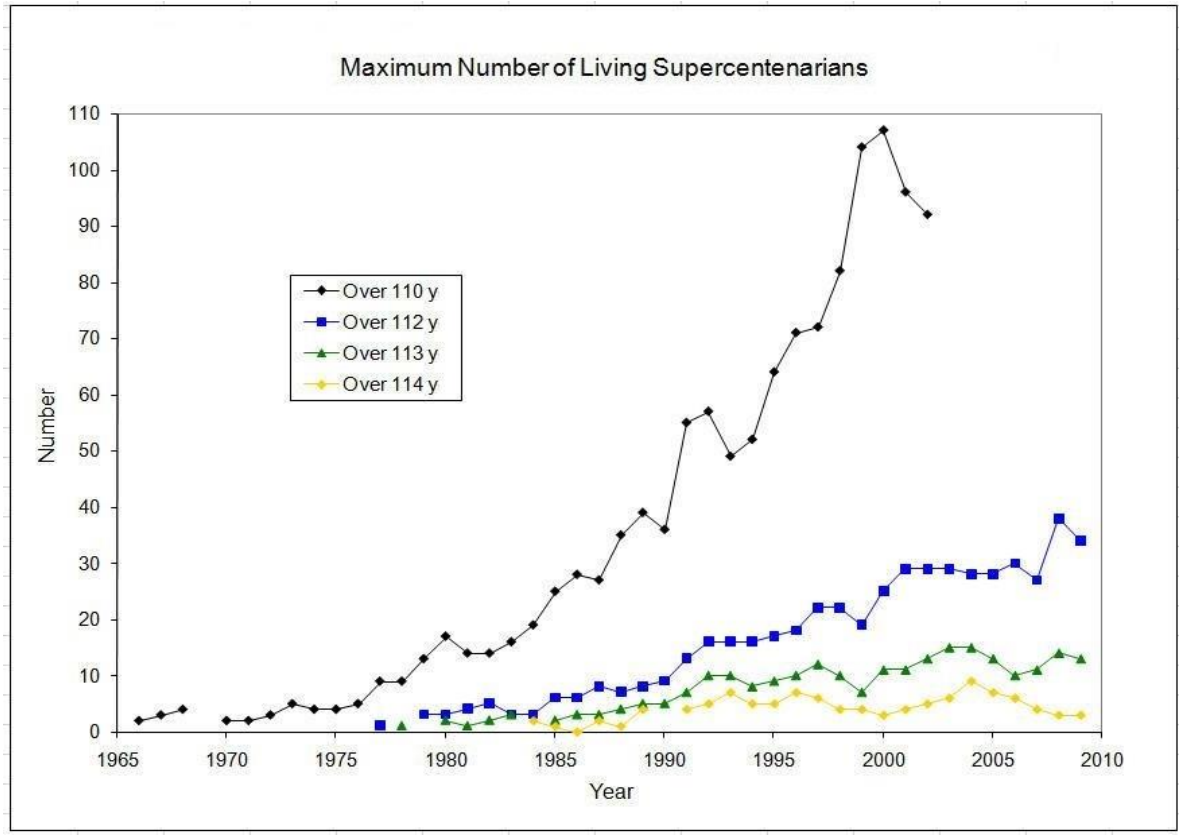

Figure 1-1: N**umber of living supercentenarians**: Number of validated living supercentenarians increasing over the years (Figure from http://www.grg.org).

It was hypothesized that LLI have a low number of disease associated variants and/or an increase in the number of protective variants that may delay age-related diseases, thereby leading to a longer, healthier life (Perls and Terry 2003). Studies have shown that centenarians either delay or escape age-associated diseases such as heart disease, stroke, diabetes and Alzheimer's disease until very late in life, often past eighty years of age or later (Evert et al. 2003). It was also seen that most of the centenarians, in spite of the presence of diseases, delayed disability until the mean age of 93 years, which indicates that genetic influence enables the LLI to remain independent for a long time (Hitt et al. 1999). A longevity genome-wide association study (GWAS) reported that LLI share the same number of risk alleles for age-related diseases such as coronary artery disease and type 2 diabetes compared with younger controls from the same population (Beekman et al. 2010). More recently, Sebastiani and co-workers published that LLI carried a similar number of disease variants, when compared to the Venter (Levy et al. 2007) and Watson (Wheeler et al. 2008) genomes, and yet survived to the most extreme ages (Sebastiani et al. 2011). It was then concluded that the number of disease-associated variants are not lower in LLI as proposed earlier, but there is an enrichment for longevity-associated variants that may resist the damaging effects of disease variants and offer protection from various age-related

diseases (Sebastiani and Perls 2012) (see Figure 1-2). These studies support the existence of buffering mechanisms, which indicates that the 'longevity-enabling' genes may act to buffer the deleterious effects of genes causing age-related diseases. Hence, the frequencies of deleterious genotypes might be even higher among LLI because their protective genotype allow the disease-related genes to accumulate with extreme lifespan (Bergman et al. 2007). As they are exceptional survivors, studies on centenarians and supercentenarians can help discover common and rare genetic variants predisposing to extreme longevity, and thus gain a better insight into the genetic basis for human longevity.

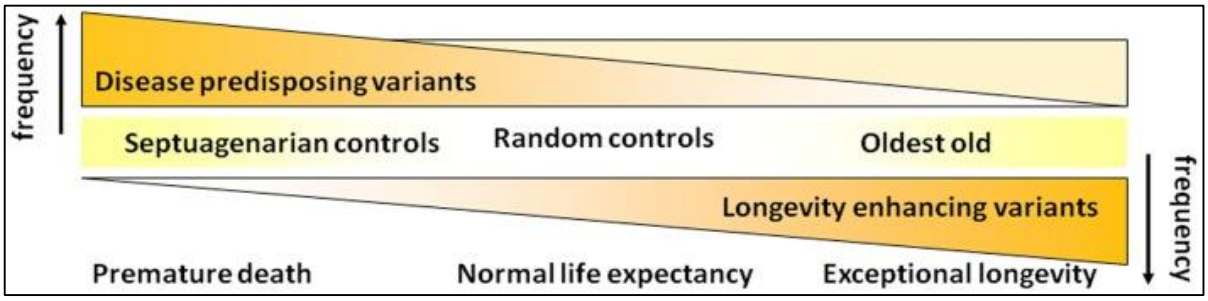

Figure 1-2: **Disease and longevity variants**: Prevalence of disease and longevity-associated variants with increasing age (Figure from Sebastiani and Perls 2012) .

1.2 Genetic epidemiology of human longevity

Studying centenarians, family-based cohorts or population-based cohorts, can help to identify variants that are enriched or deprived with age and, thereby have a higher or lower frequency in the population strata of increasing age (Tazearslan et al. 2012). As presented in Table 1-1 twin studies have shown that longevity can be inherited and the contribution of the genetic variation is about 25 to 30% (Herskind et al. 1996; vB Hjelmborg et al. 2006). Further, family studies on centenarians in different populations have suggested a significant genetic contribution to exceptional longevity (Abbott et al. 1974; Gudmundsson et al. 2000; Perls et al. 2002). The genetic influence on longevity from population-based studies is approximately 15 to 25% (Kerber et al. 2001; Mitchell et al. 2001; Murabito et al. 2012) and ranges from 20 to 30% in twin registers (McGue et al. 1993; Herskind et al. 1996). A study has been reported that African-Americans have a lower heritability rate than Europeans or Caribbean-Hispanic populations, which shows that genetic influences on longevity can vary by ethnicity (Lee et al. 2004).

Many studies have shown that longevity clusters within families and that there is an increase in the genetic effect after the age of 60 (Gavrilova et al. 1998; Perls et al. 2000; vB Hjelmborg et al. 2006; Sebastiani et al. 2013). A study was conducted with the U.S. 1900 birth cohort in the New England centenarian study (NECS) to analyse the survival rate of siblings of centenarians. It was observed that male and female siblings of centenarians have a greater probability of surviving to the age of 100: males 16.95-fold (95% CI, 10.84–23.07) and females 8.22-fold (95% CI, 6.55–9.90), when compared to siblings of those with average life expectancy (Perls et al. 2002). A population-based study in Iceland showed that the first generation relatives of LLI are twice as likely to survive to the same age as compared to controls (Gudmundsson et al. 2000). The immediate generation of relatives of Jeanne Calment were shown to have a 10-fold or higher probability of living to 80 years or more compared with a control family of average age (Robine and Allard 1998). Furthermore, offsprings of centenarians are comparatively healthy, with a marked delay in lethal age-related diseases such as Alzheimer's disease, cancer and cardiovascular disease (Atzmon et al. 2005; Schoenmaker et al. 2006; Sebastiani et al. 2013).

Exceptional survival: centenarians	Sibling survival probability
New England Centenarian Study, likelihood of achieving age 100 (Perls et al. 2002)	Women 8-fold; men 17-fold
Okinawa Centenarian Study likelihood of achieving age 90 (Willcox et al. 2006a)	Women 2.6-fold; men 5.4-fold
Twin registries and population based samples	**Heritability**
Twin registries (McGue et al. 1993; Herskind et al. 1996)	20% to 30%
Old Order Amish (Mitchell et al. 2001)	25%
Utah Population Database (Kerber et al. 2001)	15%
Framingham Heart Study (Newman et al. 2012)	16%
Medicare recipients, New York City (Lee et al. 2004)	
European ancestry	26%
African-American	4%
Caribbean-Hispanic	29%

Table 1-1: **Genetic contribution to longevity:** Genetic contribution to longevity from different studies (Murabito et al. 2012).

In 1977, Kirkwood proposed the disposable soma theory, which states that longevity requires investments in somatic maintainance and therefore, the resources available for reproduction is reduced (Kirkwood 1977). The hypothesis indicates that women exhaust resources with repeated pregnancies that would otherwise be available for maintainance and repair of the body (Kirkwood and Rose 1991). Evolutionary theories predict a trade-off between fertility and longevity, where

the chances of a higher survival comes at a cost of lower reproduction, which implies that individuals who bear fewer offsprings may live longer than those who have more number of offsprings (Westendorp and Kirkwood 1998; Kirkwood 2005; Mukhopadhyay and Tissenbaum 2007; Mitteldorf 2010; Tabatabaie et al. 2011). It has also been observed that the population of centenarians is mostly dominated by females, constituting over 85% (Max Planck Institute for Demographic Research, 2003). The reasons for such dominance could involve a number of factors (social, cultural, environmental, biological and genetic), although at present, they are not completely understood. Moreover, females are more resistant to age-related diseases as compared to men. Estrogen could have a protective role for cardiovascular diseases in females due to its effective serum lipids (Waldron 1995) and males are more exposed to infections, leading to the immunocompetence hypothesis, which suggests a significantly low effect of testosterone on immunity (Crimmins and Finch 2006). Behavioural factors such as smoking and alcohol consumption ('risky behaviour', formerly predominant in males) has also been proposed to explain the gender difference in mortality (Wallace 1996). Cigarette smoking increases the risk of several serious diseases such as lung cancer, which affect men more than females, but this cannot be a primary factor, as it has been shown that the sex difference is consistent among non-smokers as well (Waldron 1983; Waldron 1993). Although it is not clearly known what proportion of the gap reflects behavioural and biological factors and how much is due to genetic influences, we commonly observe a higher number of female centenarians in a population compared to males. However, the fewer men who reach the age of 100 are usually more healthy than a 100-year-old female (Franceschi et al. 2000).

1.3 Genetic influences on longevity

1.3.1 Findings in model organisms

Studies in model organisms can be used to explore the genetic effect in longevity and have provided abundant evidence for molecular pathways such as metabolism, anti-oxidant activities, and maintenance and repair mechanisms that extend lifespan nearly tenfold (Ayyadevara et al. 2008; Kuningas et al. 2008). The common genetic models used in longevity research are *Saccharomyces cerevisiae* (baker's yeast), *Canenorhabditis elegans* (round worm), *Drosophila melanogaster* (fruit fly) and *Mus musculus* (common mouse) (Christensen et al. 2006). Selected model organisms and their genetic findings are listed in Supplementary Table 12-1. Model organisms are useful to study the genetic variants associated with longevity as they exhibit short

lifespan along with the ability to control their environment and manipulate the genotype. In addition, there is a considerable genetic homology between humans and model organisms.

Model organism	No. of longevity-associated genes
Saccharomyces cerevisiae	825
Canenorhabditis elegans	741
Drosophila melanogaster	140
Mus musculus	112

Table 1-2: **Number of longevity-associated genes**: Number of longevity-associated genes in model organisms listed in GenAge database (Tacutu et al. 2013).

Today, many databases and tools for the biology and genetics of ageing and longevity are freely available, such as the Human ageing genomic resources (HAGR) (Tacutu et al. 2013), which hosts a variety of curated databases of candidate genes associated with longevity in humans and in model organisms. The public database GenAge (de Magalhães and Toussaint 2004) provides a comprehensive overview of the genetics of human ageing and longevity by incorporating findings from model organisms to humans (see Table 1-2). A number of pathways and associated genes contributing to longevity that have been revealed by genetic manipulations in model organisms are also available in the NetAge database (Tacutu et al. 2010) (see Table 1-3).

Pathway	No. of longevity genes
Insulin signaling	39
mTOR signalling	19
Focal adhesion	28
Adherens junction	10
Wnt signaling	17
Notch signaling	5
DNA repair	36

Table 1-3: **Number of longevity-genes reported in NetAge**: Number of longevity genes associated with various pathways identified from model organisms and reported in the NetAge database (Tacutu et al. 2010) .

With respect to longevity research, the most important study in *C. elegans* showed that the insulin-like signaling pathway regulates lifespan and metabolism in the round worm. The first evidence for genetic effect on lifespan reported in *C. elegans* referred to the gene *age-1* (Friedman and Johnson 1988; Kuningas et al. 2008). Here, it was shown that alteration in genes

such as *daf-2* and *age-1* were able to bypass the dauer formation, thereby increasing the lifespan. The same was later observed in *D. melanogaster* (Giannakou and Partridge 2007) as well as in female *M. musculus,* where mutations in the *daf-2* homologue gene increase lifespan and also increase stress and starvation resistance. However, in the case of male knock-out mouse models, mutations in *daf-2* led to a decrease in lifespan, insulin resistance and diabetes, while increase in lifespan was observed in fat-specific insulin receptor knockout mice (Holzenberger et al. 2003). The results clearly demonstrated that vertebrates are more complex to study but altogether, evidence shows that a reduced insulin signaling system increases life span in model systems.

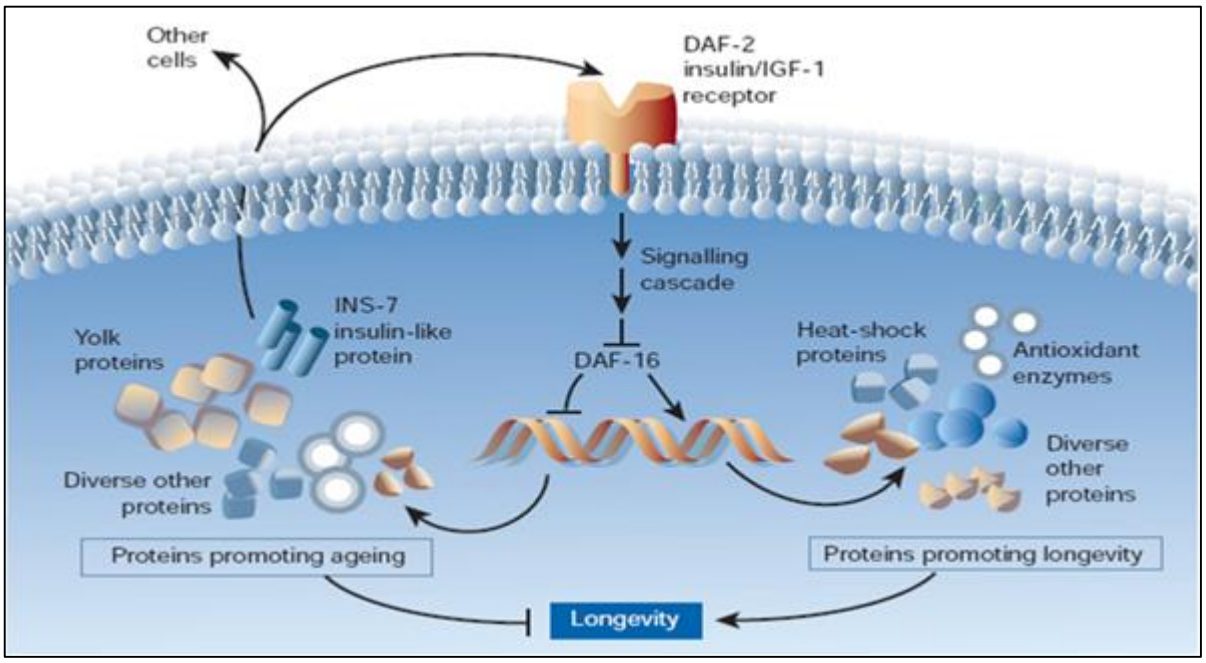

Figure 1-3: **Lifespan regulation in *C. elegans***: *daf-2, daf-16* have been identified as important genes in regulating lifespan in *C. elegans.* A high *daf-16* expression promotes longevity in *C.elegans* (Gems and McElwee 2003).

Lifespan extension caused by *daf-2* mutations is dependent upon the presence of the dauer formation protein *daf-16*, an orthologue of mammalian forkhead box O (FOXO) transcription factors. Mutations in *daf-16* suppress all phenotype of *daf-2* and *age-1* mutants, including lifespan extension, dauer arrest and reduced fertility (Lin et al. 1997). Overexpression of *daf-16* increases lifespan by approximately 20%, whereas a loss of function allele shortens it (Henderson and Johnson 2001) (see Figure 1-3). In *D. melanogaster*, overexpression of dFOXO (homologue for DAF-16) similarly extends lifespan (Hwangbo et al. 2004). It has also been observed that mice lacking the insulin receptor or the insulin-like growth factor receptor-1 live longer than wild-type mice, which implies that active FOXO transcription factors support mammalian longevity (Carter and Brunet 2007). The mammalian target of rapamycin (mTOR) is an evolutionarily conserved nutrient-sensing protein kinase that plays a crucial role in cell growth,

proliferation and metabolism (Kuningas et al. 2008). Various studies in model organisms such as yeast (Kaeberlein et al. 2005), roundworms (Vellai et al. 2003) and flies (Kapahi et al. 2004) have shown that inhibition of TOR has resulted in an increase in lifespan. Studies have also shown that caloric restriction inhibits mTOR activity, leading to longer lifespan, thus illustrates an association between nutrient intake and longevity (Christensen et al. 2006).

Genetic studies from model organisms provide a framework for understanding the factors by which life span is determined. However, investigation of human homologues of these genes would certainly be more complicated because longevity and susceptibility to disease-associated ageing in human is influenced by multiple genes, as well as by the environment.

1.3.2 Findings in humans

The genetic contribution to the variation in adult human lifespan is approximately 25 to 30% (Herskind et al. 1996; vB Hjelmborg et al. 2006) and is likely to be driven by several genes, each of which has small effects (Kirkwood 2005). Because of its high complexity and strong environmental influences, so far the number of validated genes influencing human longevity is limited (Deelen et al. 2013). Longevity genes that have been previously explored in model organisms have eventually driven the search for longevity genes in humans.

Based on previous findings in model organisms, genes involved in insulin-IGF1 signaling and the mTOR pathway (Lin et al. 1997; Lamming and Sabatini 2011) can be considered strong candidates for human longevity. However, very few validated findings have been reported (Bonafè et al. 2003; Suh et al. 2008; Albani et al. 2011; Passtoors et al. 2013) and most of the initial findings have not been replicated in different study populations. Other candidate genes that have been reported to influence human longevity are genes involved in antioxidant activity (superoxidase dismutases) (Soerensen et al. 2009), Cholesteryl ester transfer protein (*CETP*) (Barzilai et al. 2003), and Sirtuin (*SIRT3*) (Rose et al. 2003; Bellizzi et al. 2005). However, they also proved to be difficult to replicate in independent longevity populations (Nebel et al. 2005; Lescai et al. 2009; Di Cianni et al. 2013; Gentschew et al. 2013). So far, only two genes listed below, where the genetic variation influences human longevity, have been consistently replicated in various populations (Deelen et al. 2013).

i) APOE: apolipoprotein E gene is involved in the regulation of lipoproteins (Kervinen et al. 1994; Schächter et al. 1994; Deelen et al. 2011; Nebel et al. 2011). The *APOE* ε4 allele, which is

found significantly less in centenarians, confers a higher risk of Alzheimer's disease (AD) and cardiovascular disease, while ε2 is enriched in LLI and is associated with a beneficial lipid profile and may offer a protective effect for Alzheimer's disease and cardiovascular disease (Schächter et al. 1994; Christensen et al. 2006; Bennet et al. 2007).

ii) FOXO3A: forkhead box O3A gene acts as a transcription factor for multiple genes and is involved in processes such as cellular stability mechanisms and stress response. Variations in the *FOXO3A* gene revealed stronger effects in the centenarians compared to younger controls (Suh et al. 2008; Willcox et al. 2008; Anselmi et al. 2009; Flachsbart et al. 2009; Li et al. 2009c; Soerensen et al. 2010).

The most common study designs applied to identify genetic variants involved in human longevity are linkage analysis, longitudinal cohort design or case-control association studies. In longevity research, linkage analysis with long-lived sib-pairs measures the frequency of shared alleles that occur more often than expected between two sibs with the same phenotype, which then indicates that a gene involved with longevity is located in a region nearby (Willcox et al. 2006b). Linkage analysis has the advantage of being robust to population stratification (differences in allele frequencies between subpopulations due to ancestry, ethnicity or geography differences). Contradictory results have been reported with small scale genome-wide linkage studies conducted with a small sample size (Puca et al. 2001; Reed et al. 2004; Boyden and Kunkel 2010). However, recently, the largest longevity linkage analysis to date was performed in the Genetics of Healthy Aging (GEHA) in Europe study with 2,118 nonagenarian European sib-pairs. As a result, four regions (14q11.2, 17q12-q22, 19p13.3-p13.11 and 19q13.11-q13.32) that showed linkage with longevity were reported. Fine mapping of these linkage regions identified a variant (rs4420638) near *APOE* at the 19q locus as significantly associated with longevity (Beekman et al. 2013). Overall, this method has been proved disadvantageous in the case of human longevity research, primarily due to the lack of availability of multi-generational DNA or long-lived sib-pairs. Large sample size is required to achieve the statistical power needed to identify genetic regions involved in longevity (Risch and Merikangas 1996; Christensen et al. 2006).

Longitudinal studies are based on a population of individuals enrolled and followed over time to collect periodically phenotype data that includes interviews, physical and cognitive tests and collection of biological samples. In Denmark, a longitudinal study of the Danish 1905 cohort (ages 92 to 93) was established, which was a unique opportunity to investigate the genetic

contribution to human longevity (Nybo et al. 2003). The Danish 1905 cohort is special as the selection from birth to the age of 92-93, and selection from age 92-93 to 100 is 1 in 20 individuals (Soerensen 2012). Though the study design is less susceptible to biases in the control group in comparison to case-control studies, it is expensive and time consuming to follow a large group of people over a very long period of time to collect the data concerning longevity studies (Christensen et al. 2006). On the other hand, a cross-sectional approach is dependent on the assumption that there is no secular change in the frequency of the observed gene variant. The assumption was explored in a study (Lewis and Brunner 2004), and it was concluded that this hypothesis can be debatable, as gene frequency differs in populations and gene-environment interactions exist. The limitation could be overcome by conducting long-term follow-up or longitudinal studies to ensure verifiable results. Therefore, in spite of the logistical challenge, it is reasonable to carry out follow-up studies on LLI given the high mortality rate at advanced ages.

Alternatives to linkage analysis and longitudinal studies, where variants with small effects can be detected, include association studies or case-control studies. Association studies are designed to compare the allele or genotype frequencies of genetic variants in LLI (cases) with younger controls. In longevity research, case-control association studies are by far the most common approach implemented. Candidate gene studies have pointed to genes involved in lipid metabolism and insulin signaling as well-verified longevity influencing loci such as *APOE* (Schächter et al. 1994; Blanché et al. 2001; Bennet et al. 2007) and *FOXO3A* (Willcox et al. 2008; Anselmi et al. 2009; Flachsbart et al. 2009; Li et al. 2009c) . It is very important to choose an appropriate control individuals in such a way that they differ from cases with regard to phenotype only and match them as much as possible to other variables, such as gender, ethnicity and ancestry, to avoid false-positive findings (Nebel and Schreiber 2005). Further, in order to confirm initial findings of case-control studies, replication in additional longevity populations is now a common practice followed in human longevity research (Soerensen 2012). The most widely tested genetic markers in association studies are bi-allelic single-nucleotide variants (SNVs). SNVs constitute a single base change in the DNA sequence but they can also occur at a very low-frequency (minor allele frequency (MAF<1%). Single nucleotide polymorphisms (SNPs) are SNVs that occur in the general population usually with a defined minor allele frequency (e.g. MAF>5%). They are mostly validated in different populations and are included in the single nucleotide polymorphism database (dbSNP). As SNVs are often observed in only one or few individuals, some of them are not well characterized and, thus not validated in dbSNP. Some of the SNVs located in the coding regions could alter the protein by an amino acid

substitution (nonsynonymous variants), and are thus more likely to affect the gene function (Kenny and Bustamante 2011; Pavlopoulos et al. 2013).

Apart from candidate gene studies, hypothesis-free approaches such as genome-wide association studies (GWAS) have also been applied to study genetic variation in human longevity. GWAS involve rapidly scanning and prioritizing markers across the genome and eventually identify genetic variations associated with longevity. These studies should have been a favourable method to determine new genetic variants involved in longevity, as GWAS previously has shown to be successful for the discovery of novel genes involved in many common complex conditions such as Crohn's disease and inflammatory bowel disease (Klein et al. 2005; Duerr et al. 2006; Wellcome Trust Case Control Consortium 2007; Manolio et al. 2008; Franke et al. 2010). However, six longevity GWAS studies have been conducted to date (Newman et al. 2010; Deelen et al. 2011; Malovini et al. 2011; Nebel et al. 2011; Walter et al. 2011; Sebastiani et al. 2012) (Supplementary Table 12-2) and only variants in or near *APOE* have achieved genome-wide significance (GWS: generally $p \leq 5x10^{-8}$) for human longevity (Deelen et al. 2011; Nebel et al. 2011). The reason for such limited success is probably due to a combination of factors, including the heterogeneity of the phenotype, the influence of environmental factors that vary widely across populations, and mostly, the small sample size used for longevity GWAS. Many successful GWAS with replicated signals usually have a large sample size, sometimes more than 10,000 (Murabito et al. 2012). Therefore, to increase study power to detect new association signals for the longevity phenotype, GWAS are increasing the sample size through meta-analysis (Deelen et al. 2013). Recently, a GWAS meta-analysis was performed with 7,729 LLI of European descent (85 years and above) and 16,121 younger controls (less than 65 years), followed by replication in an additional set of 13,060 LLI and 61,156 controls (Deelen et al. 2014). In this study, a novel locus on chromosome 5q33.3 that associates with survival beyond 90 years (OR=1.10, $P=1.74\times10^{-8}$) was identified and replicated. The minor allele of the lead SNV (rs2149954), located in an intergenic region, is thought to promote human longevity by lowering the risk of mortality owing to stroke and non-cardiovascular causes (Deelen et al. 2014). Many candidate genes have been reported to be involved in human longevity, but very few have been confirmed to influence exceptional survival to old age despite the wide range of study designs utilized (Schächter et al. 1994; Puca et al. 2001; Willcox et al. 2008; Anselmi et al. 2009; Flachsbart et al. 2009; Pawlikowska et al. 2009; Deelen et al. 2011; Nebel et al. 2011; Bell et al. 2012; Passtoors et al. 2012; Passtoors et al. 2013). Given the complex phenotype, longevity is assumed to be determined by a combination of many genetic variants with rather small effects (Yashin et al.

2010). The inconsistency in findings from different studies can be a consequence of various factors such as sample size selection and poor replication approaches. Underpowered studies can result in false-positive findings that rightfully fail to replicate (Brooks-Wilson 2013). Furthermore, the association between variations in genes for longevity can be population-specific or can have a gene-environmental component or both, which is why meta-analysis studies, where different populations well matched for ethnicity with different genetic background are combined, may obscure true signals (Brooks-Wilson 2013). Today, with the use of recent technological advances, next-generation sequencing (NGS) can act as a powerful tool to identify associations between genetic variants and human longevity (de Magalhães et al. 2010).

1.4 Next-generation sequencing to detect variants associated with human longevity

Next-generation sequencing (NGS) or second-generation sequencing is considered the state-of-the-art sequencing technology and was motivated by the first generation sequencing of genomes using Sanger sequencing, which is time-consuming and expensive. The principle behind next-generation sequencing is to randomly fragment DNA into shorter pieces and then construct a DNA library, which is then sequenced at a high coverage followed by mapping to a reference genome of the species; on the other hand, reads can also be assembled *de novo*. The read length from NGS is lower when compared with Sanger sequencing, thus generating a large amount of data faster and more cost effectively (de Magalhães et al. 2010).

In 2005, 454 Life Science introduced the first commercial machine (GS20, Roche) and since then many other sequencing technologies from Illumina and Applied Biosystems have become popular (Liu et al. 2012). Comparison between the two most popular technologies, SOLiD and Illumina have been listed in Table 1-4. The cost of sequencing has decreased from several million dollars to less than $5,000 for one genome and continues to decrease. In January 2013, Archon Genomics X Prize announced a $10 million grand prize competition for the team that reaches $1,000 per genome for sequencing 100 human genomes in a month to an efficiency of 1 error per 1,000,000 bases, with 98% completeness along with identification of structural variations (Kedes and Campany 2011). However, the competition was eventually cancelled due to the immense price drop in NGS (GenomeWeb 2013). In January 2014, Illumina launched HiSeq X Ten Sequencer,

which promises the first $1,000 genome at 30x coverage (Illumina 2014). This makes whole genome sequencing more feasible to study genetic variation in humans.

	Illumina	SOLiD v4 (Sequencing by Oligonucleotide Ligation and Detection)
Method	Sequencing by synthesis	Ligation and two-base coding
Read length	50 to 300 bp	50 + 35 bp or 50 + 50 bp
Accuracy	98%	99.9%
Reads per run	3 billion	1.2 to 1.4 million
Time per run	3 to 10 days	1 to 2 weeks
Costs per million bases	$0.05 to $0.15	$0.13
Advantage	High throughput	Accuracy
Disadvantage	Equipment can be very expensive. It requires high concentrations of DNA.	It is slower than other methods.

Table 1-4: **Comparison of sequencing technologies**: Comparison of SOLiD and Illumina sequencing technologies that have been implemented in this project (Liu et al. 2012)

Next-generation sequencing represents new opportunities for the investigation of this complex phenotype as whole genome and exome sequencing of exceptionally old individuals provides a very high level of resolution of variant discovery to understand the genetic basis of human longevity. Detection of variants with a functional impact may help interpret why LLI delay or escape age-related diseases. Sequencing can also help to detect low-frequency (MAF≤10%) and rare (MAF<1%) variants, in comparison to GWAS that focuses on common variants (MAF>10%) (de Magalhães et al. 2010). However, to detect rare variants, large sample sizes are required to distinguish true genetic signals and to avoid false-positive results. Some studies have suggested

that low-frequency variants play an important role in the genetic architecture of the studied longevity phenotype and also contribute to the missing heritability (Vaupel 2004; Chan et al. 2014). Also, low-frequency variants may go undetected in a GWAS study as the statistical power to such variants with MAF≤10% is much lower.

Until now, whole genome sequencing data for one female and one male supercentenarian (Sebastiani et al. 2011) and whole genome sequencing of a centenarian twin pair and a middle aged monozygotic twin pair have been reported (Ye et al. 2013). In 2011, Sebastiani *et al.* showed that the number of known disease-associated variants in centenarians was similar to that of other control genomes sequenced to date, indicating that exceptional lifespan may not be due to the absence of known disease-associated variants. A recent study in 2013 by Ye *et al.* reported a small number (eight) of somatic variations detected in blood by whole genome sequencing of a centenarian twin pair and middle age monozygotic twin pair by two independent next-generation sequencing platforms (Illumina and Complete Genomics). The study concluded that, by using two independent NGS platforms, somatic single nucleotide substitutions can be detected (although stochastic somatic variation occurring in less than 20% of cells will go undetected), and that accumulation of mutations is not necessarily a result of a long-lived life.

1.5 Research objectives

There is growing interest among researchers in the complex trait of longevity; one of the primary reasons might be the gradual increase in life expectancy and the growing percentage of centenarians in the world (Vaupel 2010). It is now hypothesized that LLI carry a similar number of disease variants as the general population (Beekman et al. 2010; Sebastiani and Perls 2012) and that is a selection for longevity-associated variants, which may not only resist the damaging effects of disease variants but also offer protection (Bergman et al. 2007; Beekman et al. 2010; Sebastiani and Perls 2012). The state-of-the-art technology of next-generation sequencing provides a new tool to generate a large amount of data to unravel the genetic mechanisms of exceptional lifespan.

In this project, we combine two innovative genetic platforms (next-generation sequencing plus high-throughput genotyping technologies) with contemporary study designs (case-control association studies) and statistical methods to identify new variants that may influence human

longevity. To reach this goal, we carried out whole genome and whole exome sequencing of six centenarians (108 to 114 years) of European origin (four Germans, one French and one Spanish) on a SOLiD and Illumina platform. As it would be very cost intensive to genotype all variants identified, SNVs were selected and prioritized for typing in large study populations, where frequencies of the selected variant were compared between LLI and younger controls to infer genetic influence of longevity. SNVs were selected based on two different approaches.

Method 1: SNVs that may have a functional impact (MAF 1 to 50%)

Method 1 was carried out in collaboration with the Centre de Regulació Genòmica (CRG) and the Centre Nacional d'Anàlisi Genòmica (CNAG), Spain. Here, we performed whole genome sequencing of four centenarians (two Germans, one French, one Spanish) using the SOLiD technology, and exome sequencing using the Illumina technology. Due to their potentially functional relevance based on amino-acid substitutions, exonic variants were selected for subsequent genotyping by combining SOLiD SNVs calls with Illumina exome SNV calls. Variant frequencies were annotated using the 1000 Genomes data (1000G) and the NHLBI Exome Sequencing Project (ESP) database. Variants that were present in at least two centenarians with significantly different frequencies compared with the 1000G and ESP databases ($p<0.05$), and that were found by PhyloP to be conserved, constituted a list of 116 potentially functional SNVs. These 116 variants were selected for further genotyping in our German study population followed by replication experiment of relevant association signals in additional French and Danish longevity samples. The detailed variant selection criteria have been explained in section 2.4.1.

Method 2: Low-frequency variants with functional impact (MAF≤10%)

Method 2 was carried out in collaboration with the Institute of Medical Informatics and Statistics (IMIS), Kiel. Here, we used the same sequencing data of the four centenarians (two Germans, one French and one Spanish). Furthermore, two more German centenarians were exome-sequenced using the Illumina technology. The approach focused on selecting exonic low-frequency variants that showed an intersection of exonic SOLiD and Illumina variant calls with a MAF≤10% in the 1000G and ESP database. Variants with MAF≤10% were chosen mainly because they would have been missed out in the previous longevity GWAS study due to lack of power. Furthermore, low-frequency variants should be enriched for functional variants (Casals et al. 2013). The effect of SNVs was then determined by eight different prediction tools, where each variant was assigned a score of -1 or 1, whereby -1 indicates ‘no effect’ and 1 implies ‘effect’ for each tool. SNVs that showed a top score were selected for further analysis. Furthermore, known longevity genes and

pathways identified from various model organisms from the NetAge database (Tacutu et al. 2010) were used as filter masks for variant selection. Also, GWAS hit regions that were used as target regions for variant selection, were generated using our previous longevity GWAS data described elsewhere (Nebel et al. 2011). Each SNV of interest was visualized manually using the Integrative Genomics Viewer (IGV) in order to select good quality and true variants. This comprised a list of 51 potentially functional SNVs, which were chosen for further genotyping and replication experiment of relevant association signals in independent longevity samples. The detailed SNV selection criteria are explained in section 2.4.2.

The high depth of sequencing achieved by combining SNV calls generated with the Illumina and SOLiD technology allowed us to mine the data confidently for variants of interest. To further substantiate our initial findings, selected SNVs were tested for association by means of direct genotyping in German LLI (n = 1,610, age range: 95 to 110 years including a centenarian subset n = 745) and younger controls (n = 1,104, age range: 60 to75 years) matched for ancestry, gender and geographical origin within Germany. The relevant association signals from the German population were followed up by replication in different independent longevity populations. The French longevity sample comprised 1,269 LLI (age range: 90 to115 years) and 1,834 younger controls (age range: 35 to 61 years). The Danish longevity population consisted of 910 LLI (age range: 94 to101 years) and 760 younger controls (age range: 60 to 72 years). The American population comprised 352 LLI (age range: 90 to 114 years) and 365 younger controls (age range: 0 to 35 years) and the Italian population constituted 489 LLI (age range: 90 to 108 years) and 480 younger controls (age range: 18 to 48 years).

2 Materials

2.1 Enzymes, kits and instruments

Enzymes, kits and instruments

ABI TaqMan assays	Life Technologies Corporation, Foster City, CA
Affymetrix 6.0 array	Affymetrix Inc., Santa Clara, CA
Agilent Bioanalyzer 2100	Agilent Technologies, Germany
Covaris™ system	Life Technologies Corporation, Foster City, CA
End-Polishing Enzyme 1 and 2	Applied Biosystems Inc.; Foster City, CA, USA
Illumina Genome Analyser	Illumina, Inc. San Diego, CA, USA
Illumina Hi-Seq system	Illumina, Inc. San Diego, CA, USA
Illumina OmniExpress BeadChip kit	Illumina Inc., San Diego, CA
Illumina paired-end sequencing kit	Illumina, Inc. San Diego, CA, USA
Invisorb Blood Giga Kit	Invitek; Berlin, Germany
iPLEX™ Mass ARRAY kit	Sequenom, San Diego, CA
Klenow enzyme	Illumina, Inc. San Diego, CA, USA
NimbleGen Human Exome Array	Roche NimbleGen Systems GmbH, Germany
Proteinase K	Molecular Research Center; Cincinnati, USA
PureLink™ columns	Applied Biosystems Inc.; Foster City, CA, USA
QIAquick PCR Purification Kit	Qiagen, Hamburg, Germany
SOLiD™ Library Column Purification Kit	Applied Biosystems Inc.; Foster City, CA, USA
SOLiD™ Long Mate-Paired Library	Applied Biosystems Inc.; Foster City, CA, USA
SOLiD™ Paired-End Library	Applied Biosystems Inc.; Foster City, CA, USA
SOLiD™ 4 system	Applied Biosystems Inc.; Foster City, CA, USA
SureSelect Human All Exon kit	Agilent Technologies, Germany
T4 DNA polymerase	Illumina, Inc. San Diego, CA, USA
TaqMan-Assays	Applied Biosystems; Weiterstadt, Germany
TaqMan Universal PCR Master Mix	Applied Biosystems; Weiterstadt, Germany

2.2 Online databases and software

Databases and software

1000 Genomes Project	http://www.1000genomes.org/
Affymetrix Genotyper Console	Affymetrix Inc., Santa Clara, CA
Annovar	http://www.openbioinformatics.org/annovar/
BedTools	http://code.google.com/p/bedtools/
Bioscope	Applied Biosystems Inc.; Foster City, CA, USA
BWA	http://bio-bwa.sourceforge.net/
FastQC	http://www.bioinformatics.babraham.ac.uk/
Genome Analysis Toolkit	http://www.broadinstitute.org/gatk/
Integrative Genomics Viewer	http://www.broadinstitute.org/igv/
Illumina GenomeStudio	Illumina, Inc. San Diego, CA, USA
MutPred	http://mutpred.mutdb.org/
NCBI dbSNP	http://www.ncbi.nlm.nih.gov/SNP/
NetAge	http://netage-project.org/
NHLBI Exome Sequencing Project	http://evs.gs.washington.edu/EVS/
Picard	http://picard.sourceforge.net/
PMut	http://mmb2.pcb.ub.es:8080/PMut/
PLINK	http://pngu.mgh.harvard.edu/~purcell/plink/
PhyloP	http://compgen.bscb.cornell.edu/phyloP/
Polyphen-2	http://genetics.bwh.harvard.edu/pph2/
PS Power and Sample Size Program	http://biostat.mc.vanderbilt.edu/twiki/bin/view/
SAMtools	http://Samtools.sourceforge.net/
SIFT	http://sift.jcvi.org/
SNAP	https://www.rostlab.org/services/snap/
SnpActs	http://snpacts.ikmb.uni-kiel.de/
SNPs&GO	http://snps-and-go.biocomp.unibo.it/snps-and-go/
UCSC	http://genome.ucsc.edu/

3 Methods

3.1 Sequencing

The different studies (Method 1 and Method 2) were performed in collaboration with the CRG and CNAG, Spain, scientists at the Institute of Clinical Molecular Biology, Kiel (ICMB - Prof. Dr. Andre Franke, Prof. Dr. Almut Nebel and PD Dr. Friederike Flachsbart) and scientists at IMIS, Kiel (Prof. Dr. Michael Krawczak and Dr. Amke Caliebe).

3.1.1 Study participants

Whole genome and exome sequencing of the six centenarians (two German males and females plus one French female and one Spanish female, Table 3-1) were performed with SOLiD (genome) and Illumina (genome and exome) technologies. All the samples belonged to the cohort born between 1880 and 1990; the blood samples were taken in the year 2004.When the blood sample were taken, the German female was 106 years old. She was raised in the city; her former job was that of a craftswoman and she had one daughter. The German male was 108 years old and was from the country side. His former job was that of a craftsman. He had a history of smoking for over 40 years: on average two cigarettes per day. The French female was over 114 years old, who had a history of smoking most of her life: on average two cigarettes per day. The Spanish female was over 110 years old. The other two samples, sequenced at Beijing Genomics Institute (BGI) in China, were a German female, 108 years old, and a German male, 106 years old, neither of whom had any history of smoking. All participants and/or their legally authorized representatives took part in the written informed consent process, as required by the Institutional Review Boards/ local medical ethical committees of all participating countries before starting the study.

Sample no.	Country of origin	Gender	Age (years)
(1)	German	Female	108
(2)	German	Male	109
(3)	French	Female	>114
(4)	Spanish	Female	>110
(5)	German	Female	108
(6)	German	Male	106

Table 3-1: Study participants used for whole genome and exome sequencing

DNA extraction

Peripheral blood was obtained from all subjects and used for DNA extraction. For the French sample, DNA was obtained from a cell line. The DNA was extracted from frozen blood samples using the 'Blood Giga kit' from Invitek™ (Berlin, Germany) following the manufacturer's protocol. Here, the whole blood sample was lysed in an optimized lysis buffer and proteins were degraded during the lysis with Proteinase K. The DNA was precipitated with the addition of 96% ethanol containing solution followed by washing and final elution and, lastly, resuspended in low salt buffer for subsequent downstream applications. For the French female sample, DNA was extracted from a lymphoblastoid cell line pellet using the classical phenol chloroform method, followed by an isopropanol DNA precipitation and an ethanol wash. DNA was resuspended in TE 10:1.

3.1.2 SOLiD technology

Whole genome sequencing

Whole genome sequencing was performed using the SOLiD™ 4 system (Applied Biosystems, Foster City, CA) at the sequencing facility in ICMB, Kiel. The SOLiD system implements 2-base encoding based on sequencing by ligation (Metzker 2010). Four different libraries were constructed per individual: one paired-end library [50 + 35 bp (SOLiD™ Paired-End Library Construction Kit)] and three genomic mate-pair libraries [50 + 50 bp (SOLiD™ Long Mate-Paired Library Construction Kit)]. The library preparation (Applied Biosystems SOLiD™ 4 System 2010) in principle consists of four steps: fragmentation of DNA using a Covaris™ system, end repair of fragmented DNA using specific enzymes (End Polishing Enzyme 1 and End Polishing Enzyme 2), ligation of specific adapter sequence to ends of the fragment and finally, library amplification. After amplification, the libraries were purified with the SOLiD™ Library Column Purification Kit and measured by quantitative PCR (Applied Biosystems SOLiD™ 4 System 2010).

Emulsion PCR (ePCR) was carried out to generate clonal bead populations for all the libraries before sequencing as shown in Figure 3-1(Metzker 2010). Clonal amplification takes place in an emulsion that consists of droplets of an aqueous phase, which includes PCR components (template, primers, DNA polymerase, and DNA Beads). Once the ePCR is complete, the beads were bonded covalently to the glass slide followed by subsequent rounds of ligation with different labelled probes that are eight bases in length (Supplementary Figure 12-1).

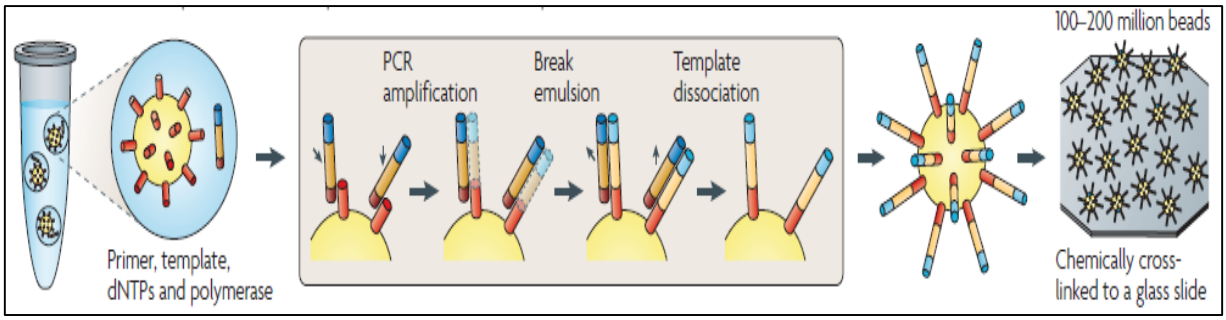

Figure 3-1: **Emulsion PCR for SOLiD™ sequencing**: Emulsion PCR is performed to generate clonal bead populations in microreactors containing a reaction mixture of an oil–aqueous emulsion to capture bead–DNA complexes into single aqueous droplets. PCR amplification is performed within these droplets to create beads containing several thousand copies of the same template sequence. Each bead is then chemically attached to the surface of a glass slide (Figure from Metzker 2010).

The sequences generated were determined in SOLiD specific 'colour space', representing the first two bases of the dinucleotide as shown in Figure 3-2. Due to colour space, each base was determined independently twice, which allows a high accuracy in sequencing, and distinction between true variants and sequencing errors (Metzker 2010).

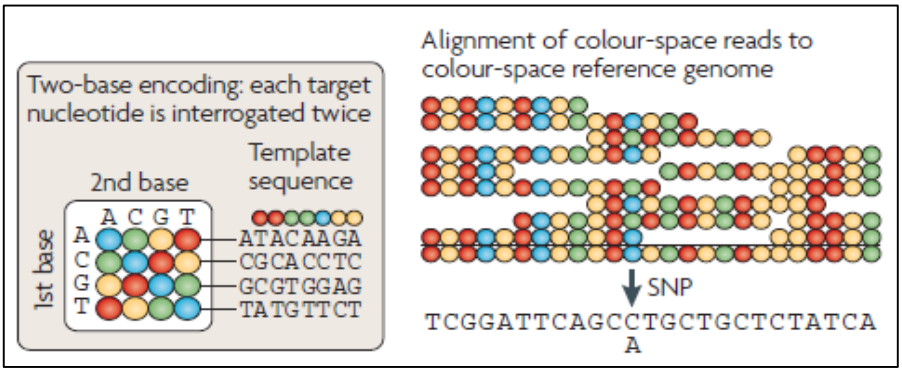

Figure 3-2: **Colour-space reference for SOLiD™ sequencing**: For decoding the data, each colour indicates two bases in which the second base of each dinucleotide unit constitutes the first base of the following dinucleotide. Because each base is interrogated twice it is possible to determine particular bases were at those positions, this can finally lead to interpret the whole sequence. The colour-space reads are aligned to a colour-space reference sequence to decode the DNA sequence (Figure from Metzker 2010).

3.1.3 Illumina technology

Whole genome sequencing

Whole genome sequencing using the Illumina Genome Analyser or Hi-Seq machines was performed at the Centre for Genomic Regulation (CRG) and Centre Nacional d'Anàlisi Genòmica (CNAG), Spain.

Illumina sequencing is based on the principle of sequencing by synthesis (Metzker 2010).The library generation was prepared for paired-end sequencing using the 'PE-102-1001-paired-end sequencing sample prep kit'. For the genomic DNA library preparation, the steps involved were fragmentation of the sample to generate desired size range of less than 800 bp, end repair using specific enzymes (T4 DNA polymerase and Klenow enzyme), adenylation of DNA ends, ligation of specific adaptors and PCR amplification to enrich the fragments that have adapter molecules on both ends (Illumina Inc. 2011). Finally, the library was purified and then quantitated to create optimum cluster densities across every lane prior to seeding clusters on a flow cell. Before the libraries are ready for sequencing, single molecules were amplified in a flow cell to generate clusters by bridge amplification as presented in Figure 3-3 (Metzker 2010).

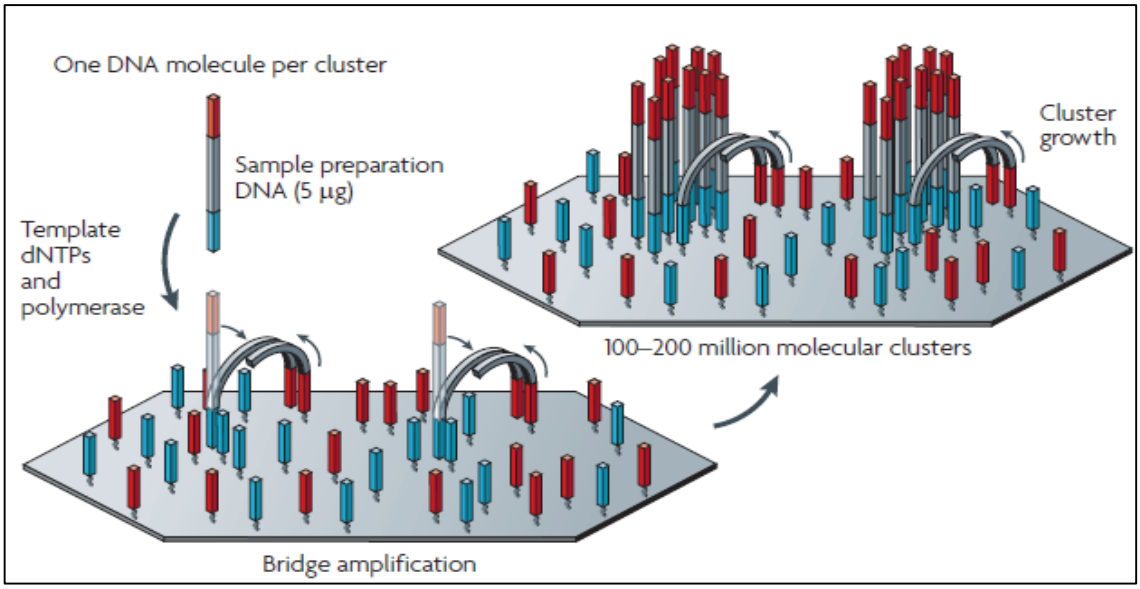

Figure 3-3: **Bridge amplification for Illumina sequencing**: Unlabeled nucleotides and enzymes are added to initiate solid-phase bridge amplification. It is composed of two primary steps: initial priming and extending of the single-stranded template, and bridge amplification of the immobilized template with immediately adjacent primers to form clusters (Figure from Metzker 2010).

The sequencing is then performed by the inclusion of four fluorescently-labelled nucleotides followed by high resolution imaging of the entire flow cell (Supplementary Figure 12-2). The clusters were stimulated by a laser to show a colour, to identify the newly added base after each round of synthesis. This approach results in high accuracy in sequencing by reducing sequence-context specific errors and enabling robust base calling across the genome.

Whole exome sequencing

The whole exome sequencing of four centenarians (German male and female, French female and Spanish female) was performed using the Illumina Genome Analyser II machines at the CRG, Spain. Two additional German samples were sequenced at BGI, China.

Exome sequencing, also known as *targeted exome capture*, is to selectively sequence the protein coding regions of the genome. The sample input was created using an exome capturing array, in this case, Agilent SureSelect Human All Exon 50 Mb capture (for samples sequenced at the CRG) and NimbleGen 2.1M Human Exome Array (for samples sequenced at the BGI), and sequenced on Illumina GAII machines according to standard protocols.

The SureSelect Human All Exon kit (SureSelect Target Enrichment Kit 2010) design covered more than 50 Mb of the human genome: 1.22% of human genomic regions corresponding to the NCBI's Consensus CDS database (CCDS) and more than 300 additional human non-coding RNAs. The library preparation was similar to that of the Illumina paired-end sequencing sample prep kit (PE-102-1001) as described above. The resulting DNA library was purified using the QIAquick PCR Purification Kit, amplified by PCR and assessed for quality and quantity with Agilent 2100 Bioanalyzer (SureSelect Target Enrichment Kit 2010). The resuspended DNA was then hybridized with biotinylated RNA library 'baits' (Supplementary Figure 12-3) of SureSelect All Exon capture library, according to the standard Agilent SureSelect Target Enrichment protocol. SureSelect magnetic beads that were prepared for the isolation of the exonic DNA were washed thoroughly, eluted, purified, re-amplified and finally checked for fragment quality.

The NimbleGen 2.1M Human Exome Array covered 34 Mb of the human genome and captured around 180,000 coding exons and 551 miRNA exons (Roche NimbleGen Inc 2009). Genomic DNA was randomly fragmented by nebulization to an average size of 500bp, and a pair of linkers was ligated to both ends of DNA fragments (Ellinghaus et al. 2013). Fragmented DNA of each individual was hybridized to NimbleGen 2.1M Human Exome Array. Exome-enriched DNA fragments were eluted from the array and were amplified by PCR, followed by random ligation of DNA fragments. The ligated long exon-enriched DNA was sheared to 200bp on average, and then the fragments were ligated with Illumina compatible adaptors and subjected to library preparation and sequencing (Roche NimbleGen Inc 2009). The exome libraries were then prepared for cluster generation and was sequenced on the Illumina Genome Analyzer II following the manufacturer's instructions.

3.2 Mapping and variant calling

For the SOLiD™ sequencing data, primary analysis (image analysis and base calling) and secondary analysis (mapping, calling of single nucleotide variants) were performed with Bioscope™ software v1.2 using default parameters. Bioscope implements repetitive mapping with a seed length (i.e. number of bases used to find accurate matching positions in the reference) of 25 to get more uniquely matched reads and allowing two color space mismatches including a penalty score of -2 for extension (User Guide: Applied Biosystems 2010). All sequenced data were mapped against human genome hg19 reference. Depth and breadth of sequence coverage was calculated with BedTools package v2.12 (Quinlan and Hall 2010). Variant calling was performed with diBayes algorithm using medium stringency settings from Bioscope™ v1.2 (Applied Biosystems), mpileup from sequence alignment/map tools (SAMtools) v.0.1.8 (parameters: |-q 15|-Q 20| (Li et al. 2009b) and Genome Analysis Toolkit-GATK v1.2 (parameters: |-baqGOP 30|-dcov 1000|-mbq 10|-mmq 10) (McKenna et al. 2010). The parameters in ‘q’ or ‘mmq’ is the minimum mapping quality and ‘Q’ or ‘mbq’ is the minimum base quality to be considered. Minimum mapping quality signifies minimal quality mapping filter, for example, mapping quality with zero signifies non-uniquely mapped reads and it is recommended to filter out the ambiguously mapped reads. The minimum base quality sets a threshold for a given base so that the user has the option to omit reads with low quality during variant calling. The recommended values from both the tools were applied (Li et al. 2009b; McKenna et al. 2010). As suggested by GATK (http://www.broadinstitute.org/gatk/gatkdocs), the gap open penalty for base alignment quality (baqGOP) is usually 30 for whole genome call sets. The parameter ‘dcov’ or downsample to coverage helps to get rid of excessive coverage above a certain depth, because having such additional data is not that useful and imposes unreasonable computational costs (McKenna et al. 2010). Variant calls for each sample were subsequently merged for further analysis using GATK CombineVariants, where all calls from different files in .vcf format are combined into one file (Supplementary Figure 12-4a).

For whole genome and exome sequences generated with Illumina technology, image analysis and base calling were performed by the Illumina Genome Analyzer’s pipeline v1.3 with default parameters. The quality of the raw sequence data was first checked with FastQC package v0.9 (Andrews 2010) that provides an overview (for example, duplication levels, overrepresented sequences, GC content per base/sequence), summarizes the data in graphs and tables and exports the results to an HTML based report. Burrows-Wheeler Aligner (BWA v0.5.9) (Li and Durbin 2009) was implemented to align sequences against the indexed hg19 human genome reference

(parameters: |-q 15| -l 5000|-t 8). BWA algorithm benefits from high alignment accuracy, supports gapped alignments for both paired-end and single-end reads and also unmapped reads are automatically assigned a mapping quality of zero (Hatem et al. 2013). The '-q' parameter in BWA indicates that the reads are trimmed at a position when quality starts to decrease below the set threshold and '-l' takes the first 5000 sub-sequence as seed length. The '-t' parameter specifies the number of threads to use, which in this case is 8 (http://bio-bwa.sourceforge.net/bwa.shtml). Duplicates were removed that were present in the reads due to amplification biases in PCR and optical duplicates (Illumina software mistakenly identifies a single cluster as two or more clusters) using Picard's 'MarkDuplicates' v1.55 (Picard 2009). To prevent false positive variants at the end of sequencing reads, and to obtain accurate scores on variant calls, local realignment around indels and quality score recalibration were performed using GATK v2.2. To improve the quality of the data, duplicates and non-uniquely mapped reads were removed by SAMtools v0.1.18 (Petersen 2014). PCR duplicates can be problematic for variant calling, since some alleles can be overrepresented as they share the same sequences and same alignment position. Sequence coverage and statistics were calculated with BedTools package v2.12. Variant calling was performed with SAMtools (parameters: |-q 15|-Q 20) and GATK (parameters: |-dcov 1000 |-mbq 10|-mmq 10), where the results were consequently merged for subsequent analysis using GATK's CombineVariants (Supplementary Figure 12-4b and (Petersen et al. 2014)).

For functional annotation of the variants generated from SOLiD and Illumina technology, an internal single nucleotide variation (SNV) categorization package snpActs (http://snpacts.ikmb.uni-kiel.de/) and Annovar (Wang et al. 2010) were implemented. All known variant positions were matched to the National Center for Biotechnology Information database of SNPs (NCBI's dbSNP) build 135 (Sherry et al. 2001; Bethesda 2010). snpActs, developed by Björn Stade from ICMB, is a database-driven toolset that scans different gene annotations to identify and annotate SNVs in functional elements (Petersen 2014). snpActs enables the user to filter the variant list using special rules (e.g.: coding SNVs), allele frequency (e.g.: 1000G/ESP database) or target regions of interest (e.g.: candidate genes list, GWAS hit regions). To differentiate between sequencing artefacts and true variants, variants were visualized using the Integrative Genomics Viewer (IGV v2.1.24) (Thorvaldsdóttir et al. 2013).

For additional assessment of variant quality, the genotype concordance was computed for all samples. Three samples were genotyped using the Illumina OmniExpress 700k array (Illumina Inc., San Diego, CA) containing approximately 700,000 SNVs. Genotype calling was done with

GenomeStudio software according to the protocol provided by Illumina. The other two samples sequenced at the BGI were genotyped using Affymetrix 6.0 (Affymetrix Inc., Santa Clara, CA), which contains more than 906,600 SNVs. The genotype calling was performed with the Affymetrix Genotyper Console v4.0, using the default quality control thresholds. The calling for all array data was initially performed using the genome build hg18 but later converted to hg19 coordinates using the liftOver tool provided by the UCSC Genome Bioinformatics Group (http://genome.ucsc.edu/cgi-bin/hgLiftOver). The genotype concordance was then calculated with snpActs.

3.3 Selection of variants

3.3.1 Method 1: SNVs that may have functional impact

SNVs (MAF: 1% to 50%) for Method 1 were selected by Daniel Trujillano, a Ph.D. student at the CRG in Spain. For this method, the variants were selected from the exonic SNVs detected in whole genome SOLiD and Illumina exome-sequencing data of the four centenarians, as shown in Figure 3-4.

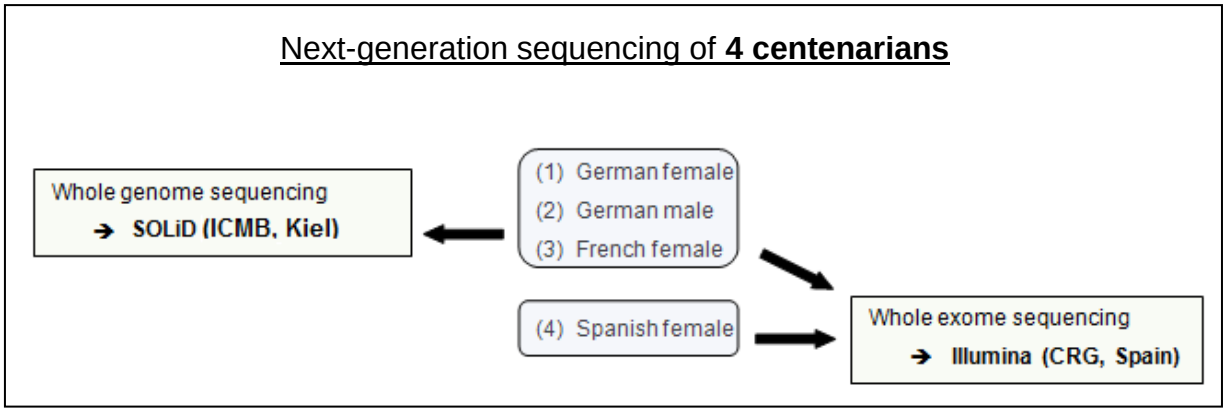

Figure 3-4: **Individuals sequenced on SOLiD and Illumina technologies for Method 1**: Four individuals sequenced with SOLiD and Illumina technologies to select SNVs that may have a functional impact.

The exome pipeline implemented by CRG has been presented in Supplementary Figure 12-5. The resulting alignments were used as input for three different variant prediction tools, namely GATK (McKenna et al. 2010), mpileup (Li et al. 2009b) and SHORE (Schneeberger et al. 2009). The three independent SNV predictions were subsequently quality filtered using GATK's VariantFiltration and intersected with GATK's CombineVariants. Functional annotation of all variants was performed using Annovar, providing a comparison of predicted variants with NCBI's dbSNP 132, 1000G and NHLBI Exome Sequencing Project (ESP), as well as multiple

estimates of the impact of amino acid substitution on the structure and function of proteins. The 1000G data constitutes 38 million variants constructed using a combination of low-coverage whole-genome and exome sequencing of 1,092 individuals from 14 populations spread across Europe, East Asia, sub-Saharan Africa and America (Abecasis et al. 2012). The ESP database comprises individuals from a number of large-scale National Heart, Lung, and Blood Institute (NHLBI) cohorts: 2,203 African-Americans and 4,300 European-Americans unrelated individuals, totalling 6,503 individuals. The database contains approximately 3 million variants.

Due to the potentially functional impact of amino acid substitutions, exonic variants were selected for subsequent genotyping by combining exonic SNVs detected in the whole genome SOLiD data with the Illumina exome SNV calls. Variant frequencies were annotated using the 1000G and ESP database. Further, p-values were calculated with an in-house script at the CRG institute that compared allelic frequencies, taking into consideration the sample size. Variants that were present in *(i)* at least two samples with significantly different MAFs to the 1000G or ESP databases and *(ii)* found to be conserved by PhyloP were selected for further investigation.

The above selection of variants constituted a list of 116 potentially functional SNVs that were chosen for further genotyping in our German population (n = 1,614 LLI including a centenarian subset n = 748; younger controls n = 1,104). Seven SNVs that showed a significant association signal in the German longevity sample were typed for replication in the French (n = 1,269 LLI and 1,834 younger controls) and Danish (n = 910 LLI and 760 younger controls) longevity samples.

3.3.2 Method 2: Low-frequency variants with functional impact

For Method 2, sequencing data generated from the SOLiD and Illumina platforms for the same four individuals in Method 1, plus an additional two centenarians exome-sequenced with the Illumina technology were used as presented in Figure 3-5.

Low-frequency variants (MAF≤10%) that show an intersection between the two platforms (SOLiD and Illumina) were selected. The effect of change in amino acid substitution for all SNVs was evaluated with eight different prediction tools. Furthermore, known longevity genes and pathways identified in various model organisms listed in the NetAge database (Tacutu et al. 2010) and longevity GWAS hit regions (Nebel et al. 2011) were used as filter masks for variant

selection. The scores for the prediction tools for all SNVs were computed by Carolin Knecht, a Ph.D. student at the IMIS, University of Kiel.

The variant generated with the SOLiD and Illumina technology for all six samples were annotated using snpActs (http://snpacts.ikmb.uni-kiel.de/) and Annovar (Wang et al. 2010). All known variant positions were matched to the NCBI's dbSNP 135. The variants that showed an intersection of both the technologies were chosen for further investigation. As SOLiD and Illumina employ different techniques, it is expected that the intersection would represent true-positive variants among large proportions of putative false-positive calls and sequencing artefacts, that are randomly distributed over the genome (Ratan et al. 2013).

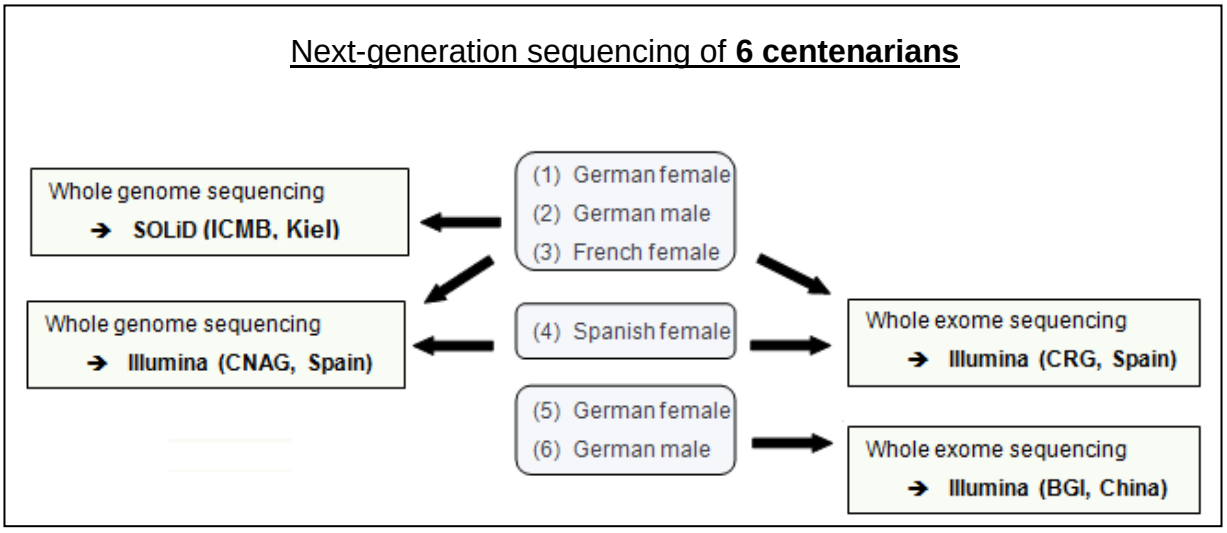

Figure 3-5: **Individuals sequenced on SOLiD and Illumina technologies for Method 2**: Six individuals sequenced on SOLiD and Illumina sequencing platforms to select low-frequency variants that may have a functional impact.

The variant list was filtered using rules enabled in snpActs (http://snpacts.ikmb.uni-kiel.de/), where only coding regions such as missense, nonsense, cancel-start, read-through and splice-sites were selected [cancel-start: changes a start codon of the mRNA inhibiting transcription; read-through: changes the stop codon to a codon for an amino acid]. The list was further refined by retaining variants, which showed a MAF≤10% against the 1000G and ESP database. Low-frequency variants with MAF≤10% were selected, mainly because these low-frequency variants might have been missed in the previous longevity GWAS studies (Nebel et al. 2011), since the statistical power to detect such variants with small effects in GWAS is very low (Chan et al. 2014). Also, many studies suggest that variants that influence longevity are likely to be low-frequency variants with large effects sizes (Vaupel 2010; Garagnani et al. 2014). The low-

frequency variants with large effect sizes would definitely have a functional consequence as well, particularly for nucleotide changes that affect protein function (Casals et al. 2013)

The functional impact of SNVs were predicted by eight different tools: Grantham (Grantham, 1974); PMut (Ferrer-Costa, et al., 2005); Screening for Non-acceptable Polymorphisms (SNAP) (Bromberg and Rost, 2007); Sorting Intolerant From Tolerant (SIFT) (Kumar, et al., 2009); SNPs&GO (Calabrese, et al., 2009); MutPred (Li, et al., 2009); Polymorphism Phenotyping (Polyphen-2) (Adzhubei, et al., 2010); and PhyloP (Pollard, et al., 2010). All these different tools implement different algorithms to make predictions regarding the functionality of mutated proteins. The basic idea behind the implementation of all the different tools is to take advantage of their possible complementary performance at classifying functionally relevant variants, thereby help prioritizing target SNVs. The input information here was protein sequence, amino acid substitution or UniProt IDs. The tools were implemented using Perl routines or, when possible, by batch queries. For each variant, the binary decisions of each tool were aggregated by summation, whereby -1 indicates 'no effect' and 1 implies 'effect' for each tool. The 'top-scoring' SNVs are variants that showed an effect for seven or eight out of all eight tools and these SNVs were selected for further analysis.

Known longevity genes and pathways identified from various model organisms listed in the NetAge database (Tacutu et al. 2010) were used as filter masks to prioritize the variant list. Insulin pathway and mTOR pathway are the most interesting conserved pathways identified using animal models that influence longevity (Barzilai et al. 2012). Over 50 longevity genes that involved pathways such as insulin and the mTOR pathway were included. GWAS hit regions were also used as target regions for selection of SNVs. The longevity GWAS data used has been described in detail elsewhere (Nebel et al. 2011). It comprises 664,472 autosomal SNPs in 763 LLI (mean age: 99.7 years) and 1,085 controls (mean age: 60.2 years) from Germany. Variants were 'clumped together' using the clumping algorithm (parameters: $p1 < 0.001$, $p2 < 0.01$, $r2 > 0.8$, 200kb distance) of PLINK (Purcell et al. 2007), starting from the most significant SNV and moving to the less significant, to generate a list of LD-dependent associated genomic regions. These associated 'clumped' regions were used as target regions for variant selection.

Different lists were then generated based on various selection criteria: variants identified based on target regions (longevity pathway list, GWAS associated hit regions) and top-scoring SNVs computed from the different prediction tools. Furthermore, variants present in four or more

individuals and that showed an 'effect' for at least five prediction tools, as well as SNVs present in five to six individuals, were selected for further analysis. These variants were of interest because they were found in four or more of the sequenced centenarians and had a MAF≤10% or no MAF in the 1000G or ESP database, which is not expected by chance.

Each SNV of interest was then visualized manually using the Integrative Genomics Viewer (IGV) in order to select good quality and true variants, which resulted in a list of 51 potentially functional SNVs that were chosen for further genotyping in the German LLI population set and a subsequent replication experiment in the Danish, Italian and American longevity samples.

3.4 Genotyping and replication

The selected SNVs were genotyped at the genotyping facility in ICMB, Kiel by the iPLEX™ Mass ARRAY technology (Sequenom, San Diego, CA) and the ABI TaqMan® technology (Life Technologies Corporation, Foster City, CA) (see Table 3-1).

Technology	iPLEX™ (Jurinke et al. 2002)	TaqMan® (Livak 2003)
Company	Sequenom http://www.sequenom.com/	Applied Biosystems http://www.appliedbiosystems.com/
Assay type	Primer extension	5' exonuclease/PCR
Pros	High sample throughput	Simplicity, very high reliability
Cons	Maintenance intensive equipment	Single plex, custom assays expensive

Table 3-2: **Genotyping platforms**: Summary of genotyping technologies used in this project (Ragoussis 2006).

3.4.1 Sequenom technology

Genotyping using Sequenom MassARRAY iPLEX™ platform was performed according to the manufacturer's protocol. The Sequenom method is based on a label-free primer extension chemistry that produces allele-specific extension products (Jurinke et al. 2002). After PCR amplification of the allele-specific fragment, MALDI-TOF spectrometry was employed to

analyse the extended primer, permitting precise determination of the size of products generated, which can be converted into genotype information. Genotype calls were produced by automated allele calling with the help of software provided by Sequenom.

3.4.2 TaqMan technology

Genotyping was performed using TaqMan® Genotyping Assays (Applied Biosystems) according to the manufacturer's protocol. TaqMan is based on the principle of using the 5' exonuclease activity of Taq polymerase, which employs a combination of PCR and competitive hybridization (Livak 2003). At the end of the PCR reaction, normalised intensities of the fluorescent signals were plotted on a scatter plot using a clustering algorithm in the data analysis software provided by Applied Biosystems, thus determining the genotypes (Ragoussis 2006).

3.4.3 Study population

The samples used for genotyping were the German LLI matched for ancestry, gender and geographical origin within Germany. Significant association signals in the German samples were then typed for replication in independent longevity samples from the French, Danish, Italian and American study populations. Blood samples and DNA obtained from the study participants were isolated by using standard methods.

German sample

The German 'case' population consists of 1,610 unrelated subjects, (age range: 95 to 110 years). The samples cover about 27% males (n = 435) and 73% (n = 1,175) females. The subjects were selected and recruited from different regions of Germany, based on data available from local registry offices. The subjects were contacted by a letter with a questionnaire and a blood sampling kit. A complete summary of the socio-economic conditions, quality of life and health status of the subjects was recorded with the help of a questionnaire (Nebel et al. 2005). The younger control population comprised 1,104 unrelated individuals (age range: 60 to 75 years), covering 26% males (n = 283) and 74% females (n = 821). The participants were recruited from different geographic regions of Germany and were all of German ancestry. Although regional genetic differences in the population structure within Germany are classified as very low (Steffens et al. 2006), care was taken to match controls to cases for ancestry, gender and geographical origin within Germany to avoid false-positive association signals due to biases such as population stratification (Flachsbart et al. 2009).

Independent longevity samples for the replication experiment

Replication of the significant association signals obtained from the German longevity sample was carried out in independent longevity samples from France, Denmark and Italy. The French replication sample comprised 1,269 LLI from different regions throughout France (Île-de-France, Northeast, Northwest, Southeast and Southwest). These individuals were matched for gender and geographical origin with 1,834 healthy unrelated younger controls (age range: 35 to 61 years) (Blanché et al. 2001). Controls were selected from two sample sets: *i*) unrelated young European people born in France (Île-de-France, Northeast, Northwest, Southeast, Southwest) (Blanché et al. 2001); and *ii*) unrelated French subjects who participated in the Supplementation in Vitamins and Mineral Antioxidants (SU.VI.MAX) study (Hercberg et al. 1998). The age range for cases was 90-115 years, covering 18% (n = 232) males and 82 % (n = 1037) females. For the younger controls, the age range was 35 to 61 years, including 40% (n = 730) males and 60% (n = 1104) females.

The Danish cohort consisted of 910 LLI selected from four nation-wide birth cohort studies: the 1905 Birth Cohort Study (Nybo et al. 2001); the 1910 Birth Cohort Study (Christensen et al. 2013); the 1911-12 Birth Cohort Study (Robine et al. 2010); and the 1915 Birth Cohort Study (Christensen et al. 2013). The 760 younger controls were randomly selected from the Study of Middle-Aged Danish Twins (MADT), which was initiated in 1998 and includes twins born from 1931 to 1952. The control group includes only one twin from each twin pair (Soerensen et al. 2010). The age range for cases was 94 to 101 years, covering 30% (n = 273) males and 70 % (n = 637) females. For the controls, the age range was 60 to 72 years, covering 40% (n = 301) males and 60% (n = 459) females.

The Italian longevity sample comprised 489 LLI (age range: 90 to 114 years) and 480 unrelated younger controls (age range: 18 to 48 years) geographically matched. The participants were obtained from the Southern Italian Centenarian Study (SICS), where they were recruited from regions of Southern Italy east of Naples. SICS LLIs were thoroughly investigated for demographic and clinical characteristics and they were enrolled by Associazione Longevita (Anselmi et al. 2009; Malovini et al. 2011).

The American population comprised 352 Caucasian LLI (age range: 90 to 114 years) recruited through Elixir Pharmaceuticals, Beth Israel Deaconess Medical Center, Children's Hospital of Boston and Boston University Medical Center. Individuals were recruited by various methods

such as institutional websites and organizations involved with the ageing community. A set of 365 young unrelated controls (age range: 0 to 35years), self-identified as 'Caucasian' and less than 35 years of age, were obtained from several anonymous sources in the U.S.A. To avoid genetic stratification, only those controls were selected who best matched the cases with respect to genetic background (Geesaman et al. 2003).

All participants and/or their legally authorized representatives took part in the written informed consent process, as required by the Institutional Review Boards/ local medical ethical committees of all participating countries before starting the study.

3.4.4 Statistical analysis

Allele-based single marker case–control analyses (CCA) and odds ratio (OR) statistics were calculated with χ^2 statistics, using the open-source analysis toolset PLINK v.1.07 (Purcell et al. 2007). P-values less than 0.05 were considered nominally significant. All SNVs were tested for Hardy–Weinberg equilibrium (HWE) in controls before inclusion in the analyses ($P_{HWE} > 0.001$), using PLINK. Power and sample sizes were calculated with the PS Power and Sample Size Program (Dupont and Plummer 1990), applying a significance level of 0.05.

4 Results

4.1 Mapping, coverage and variant calling

4.1.1 SOLiD technology

Whole genome sequencing

The sequencing for the three centenarians ((1) German male, (2) German female and (3) French female) was performed on SOLiD™ 4 system (Applied Biosystems, Foster City, CA) using one paired-end library and three genomic mate-pair libraries. This produced 3,009,170,818 reads for the German female, 2,891,350,241 reads for the German male and 3,109,676,528 reads for the French female. Reads were mapped with Bioscope™ (Applied Biosystems) to the human genome reference hg19. More than 65% of reads were mapped for all three samples with over 90% of genome coverage. The mapping statistics are presented in Table 4-1.

	(1) German female 108 years	(2) German male 109 years	(3) French female >114 years
No. of reads	3,007,757,220	2,890,132,319	3,109,676,528
No. of reads aligned	1,976,859,186	2,023,852,602	1,974,842,082
Covered base positions at 1x (%)	93.34	92.53	93.24
Covered base positions at 8x (%)	84.66	89.77	77.12
Covered base positions at 20x (%)	59.64	78.35	47.19
Average coverage (mean)	43.73	41.51	44.44

Table 4-1: **Mapping statistics for SOLiD sequencing data**: Mapped sequences and coverage depth across the genome for data generated by SOLiD technology.

SNVs were called using diBayes from Bioscope™ (Applied Biosystems), SAMtools (Li et al. 2009a) and GATK (McKenna et al. 2010). The variants from all three callers were combined for each sample. This resulted in 3,264,816 SNVs for the German female, 3,923,324 for the German male and 2,695,673 for the French female as shown in Table 4-2. For functional annotation, snpActs (http://snpacts.ikmb.uni-kiel.de/) was implemented to determine the distribution of the all variants and to give an overview of coding and non-coding SNVs, cancel-start, read-through, and transition (Ti)/transversion (Tv) ratio (The Ti/Tv ratio is generally used to evaluate the quality of variant calls: reported to be between 2 to 2.2 for variants in whole genome and 2.8 to 3.0 in the coding region respectively (DePristo et al. 2011)).

	(1) German female 108 years	(2) German male 109 years	(3) French female >114 years
Total SNVs	3,264,816	3,923,324	2,695,673
known SNVs (dbSNP 135)	2,539,458 (74.46%)	3,143,342 (80.50%)	2,225,357 (81.72%)
novel SNVs	871,207 (25.54%)	761,059 (19.50%)	497,906 (18.28%)
heterozygous/homozygous ratio (het/hom)	1.41	1.66	1.09
transition/transversion ratio (Ti/Tv)	2.11	2.11	1.98

Table 4-2: **SNV distribution for SOLiD genome sequences**: Summary of SNV distribution for all three genomes generated by SOLiD technology.

About 75 to 80% of the SNVs present in all three samples were already known and reported in the dbSNP (135) database and the remaining were novel. The ratio of heterozygous variants to homozygous variants is 1.09 to 1.66, which matches the published range of 1.2 to 1.7 for European ethnicity (Levy et al. 2007; Sebastiani et al. 2011). The Ti/Tv ratio was found to be 2.11 for both the German male and the German female and 1.98, for the French female (Table 4-2), which is according to the expected Ti/Tv ratio for whole genome sequencing (DePristo et al. 2011). The distribution of the variants was as follows: non-coding SNVs (introns and intergenic) comprised around 92%; approximately 7.3% of the variants were located in untranslated region (5' UTR, 3'UTR and UTR-Splice sites); and around 0.7% constituted coding SNVs (synonymous, missense and nonsense SNVs) for all three individuals (Supplementary Table 12-3).

For additional assessment of SNV quality, genotype concordance was computed for all three subjects using the Illumina OmniExpress Chip 700k array, where more than 94% concordance was observed as presented in Figure 4-1.

	No. of SNVs on Illumina OmniExpress Array	Overlapping SNVs	Genotype concordance (%)
(1) German female	710,663	285,648	96.5
(2) German male	710,746	328,816	98.3
(3) French female	710,038	251,325	94.6

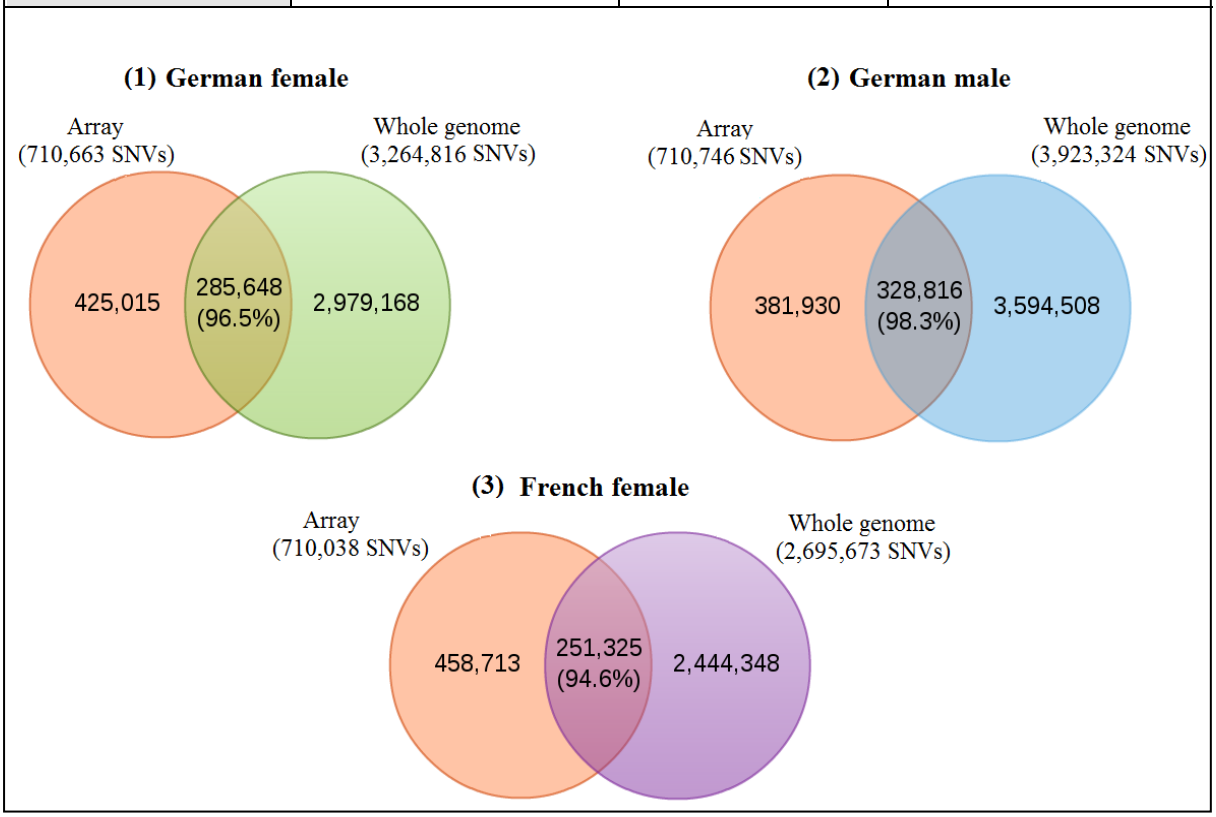

Figure 4-1: **Genotype concordance for SOLiD sequencing data**: Genotype concordance of whole genome SOLiD sequencing data compared with the Illumina OmniExpress array; (1) German female (96.5%), (2) German male (98.3%) and (3) French female (94.6%).

4.1.2 Illumina technology

Whole genome sequencing

The whole genome sequencing for the same three centenarians ((1) German female, (2) German male and (3) French female) plus one (4) Spanish female was performed on Illumina Genome Analyzer or Hi-Seq machines using the 'PE-102-1001-paired-end sequencing sample prep kit' at the CNAG, Spain. This produced: 1,027,097,060 reads for the German female, 1,143,873,254 reads for the German male, 850,231,574 reads for the French female; and 1,053,944,714 for the Spanish female. Reads were mapped with BWA to the human genome reference hg19. More than 90% of reads were mapped for all three samples as shown in Table 4-3.

	(1) German female 108 years	(2) German male 109 years	(3) French female >114 years	(4) Spanish female >110 years
No. of reads	1,027,097,060	1,143,873,254	850,231,574	1,053,944,714
No. of reads aligned	940,941,134	1,039,187,976	780,832,784	967,729,186
Covered base positions at 1x (%)	92.81	91.80	92.77	93.09
Covered base positions at 8x (%)	91.04	90.18	90.78	91.05
Covered base positions at 20x (%)	79.55	81.87	73.27	77.19
Average coverage	30.58	33.07	25.52	28.27

Table 4-3: **Mapping statistics for Illumina sequencing data:** Mapped sequences and coverage depth across the genome for data generated by Illumina whole genome sequencing.

SNVs were called using SAMtools (Li et al. 2009a) and GATK (McKenna et al. 2010). The variants from both callers were combined for each sample to give 4,013,012 for the German female; 4,071,554 for the German male; 4,022,164 for the French female and 4,081,702 for the Spanish female. The SNV distribution for all four genomes is presented in Table 4-4. For functional annotation of the variants, snpActs (http://snpacts.ikmb.uni-kiel.de/) was implemented to determine the distribution of the SNVs.

	(1) German female 108 years	(2) German male 109 years	(3) French female >114 years	(4) Spanish female >110 years
Total SNVs	4,013,012	4,071,554	4,022,164	4,081,702
known SNVs (dbSNP 135)	3,673,475 (91.54%)	3,698,341 (90.9%)	3,678,686 (91.5%)	3,733,683 (91.5%)
novel SNVs	339,537 (8.46%)	373,213 (9.1%)	343,478 (8.5%)	348,019 (8.5%)
het/hom ratio	1.77	1.72	1.75	1.82
ti/tv ratio	2.08	2.06	2.01	2.1

Table 4-4: **SNV distribution for Illumina genome sequences:** Summary of SNV distribution for all four genomes generated by Illumina whole genome sequencing.

About 90% of the SNVs present in all four samples were known and reported in dbSNP and the remaining were novel. The ratio of heterozygous variants to homozygous variants is 1.72 to 1.82, which is slightly higher when compared to other genomes (Levy et al. 2007). As shown in Table 4-4, the Ti/Tv ratio was found to be 2.01 to 2.21 for all four genomes, which meets the expected value according to published data (DePristo et al. 2011). Non-coding variants (introns and intergenic) comprised around 92%, approximately 7.3% of the variants were located in untranslated region (5' UTR, 3'UTR and UTR-Splice sites) and the remaining 0.7% constituted coding SNVs (synonymous, missense and nonsense SNVs) (Supplementary Table 12-4).

	No. of SNVs on Illumina OmniExpress Array	Overlapping SNVs	Genotype concordance (%)
(1) German female	710,663	341,373	99.7
(2) German male	710,746	344,369	99.8
(3) French female	710,038	342,427	99.8

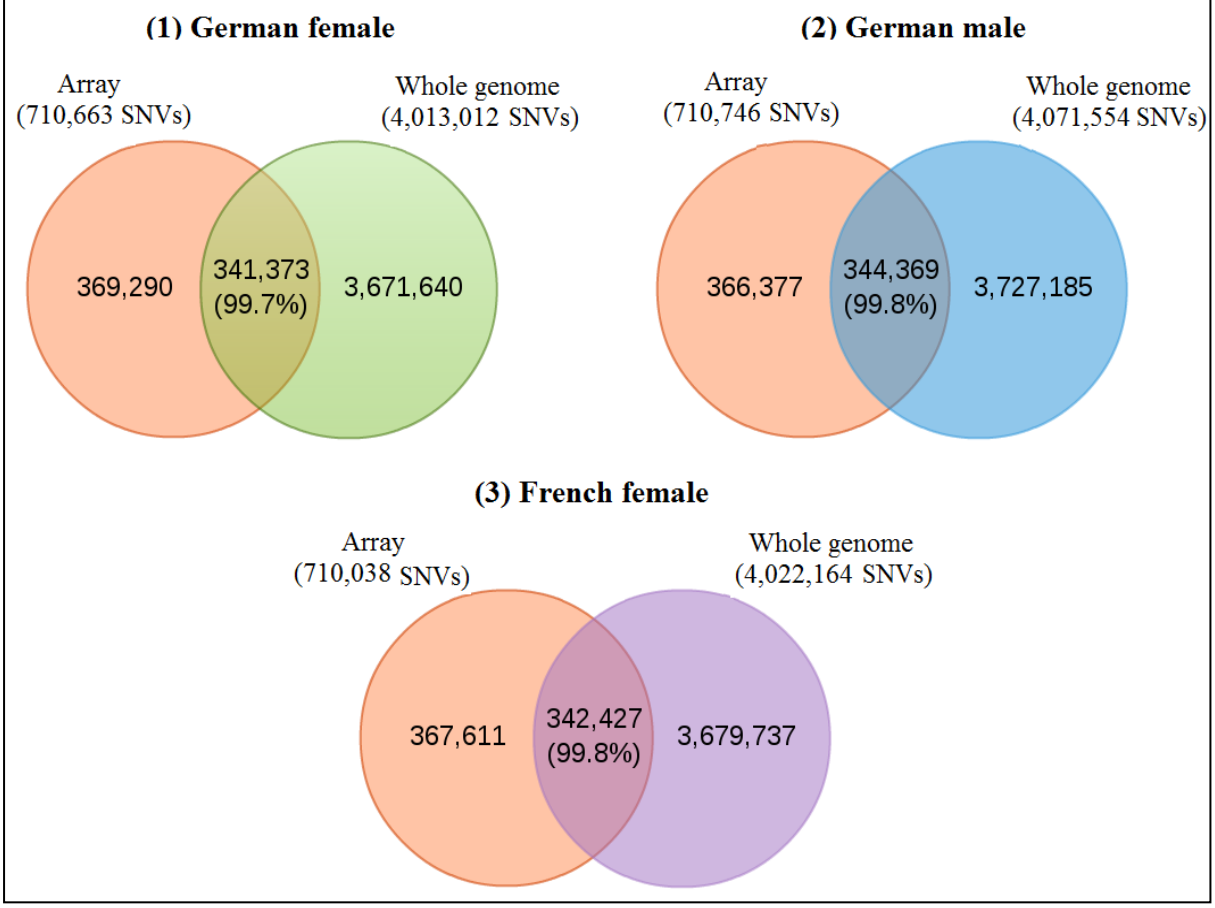

Figure 4-2: **Genotype concordance for Illumina sequencing data**: Genotype concordance of whole genome Illumina sequencing data compared with the Illumina OmniExpress array; (1) German female (99.7%), (2) German male (99.8%), and (3) French female (99.8%).

For additional assessment of SNV quality, the genotype concordance was computed using snpActs (http://snpacts.ikmb.uni-kiel.de/) for three subjects with the Illumina OmniExpress Chip 700k array, where more than 99% concordance was observed as presented in Figure 4-2.

Whole exome sequencing

The previous four centenarians ((1) German female, (2) German male, (3) French female and (4) Spanish female) and two additional German centenarians ((5) German male and (6) female) were sequenced using Agilent SureSelect Human All Exon target enrichment kit at CRG, Spain and NimbleGen 2.1M Human Exome enrichment kit at BGI, China on Illumina GAII machines according to standard protocols. Reads were mapped with BWA to the human genome reference hg19. More than 90% of reads were mapped for all six samples. The exome sequencing coverage and mapping statistics are shown in Table 4-5.

	(1) German female 108 years	(2) German male 109 years	(3) French female >114 years	(4) Spanish female >110 years	(5) German female 108 years	(6) German male 106 years
No. of reads	89,613,010	90,724,560	90,704,224	86,744,490	40,962,578	43,761,129
No. of reads aligned	87,572,479	88,671,162	89,373,419	85,610,982	36,866,320	39,270,973
Covered base positions at 1x (%)	90.95	91.14	91.43	91.1	97.77	97.32
Covered base positions at 8x (%)	78.9	79.07	79.36	79.76	82.92	83.55
Covered base positions at 20x (%)	67.86	67.96	68.35	68.72	54.66	59.76
Average coverage	73.2	75.1	72.1	70.1	27.7	29.7

Table 4-5: **Mapping statistics for exome sequencing data:** Mapped sequences and coverage depth for Illumina exome sequencing data.

The variants were called using SAMtools (Li et al. 2009a) and GATK (McKenna et al. 2010) and variants enriched in targeted regions (i.e. coding regions of the genome) were selected. The SNVs from both callers were combined for each sample to give 18,456 to 27,178 SNVs for all six centenarians as shown in Table 4-6.

	(1) German female 108 years	(2) German male 109 years	(3) French female >114 years	(4) Spanish female >110 years	(5) German female 108 years	(6) German male 106 years
Total SNVs	26,223	26,790	26,767	27,178	18,456	18,481
known SNVs	25,112 (95.76%)	25,712 (95.97%)	25,647 (95.81%)	26,085 (95.97%)	17,531 (94.98%)	16,182 (95.21%)
novel SNVs	1,111 (4.24%)	1,078 (4.03%)	1,120 (4.19%)	1,093 (4.03%)	925 (5.02%)	885 (4.78%)
het/hom ratio	1.61	1.63	1.68	1.68	1.50	1.50
ti/tv ratio	2.68	2.64	2.74	2.71	2.87	2.85

Table 4-6: **SNV distribution for Illumina exome sequences:** Summary of SNV distribution for all six exomes sequenced with the Illumina technology.

About 95% of the SNVs present in all six samples were already reported in the dbSNP 135 database and the remaining were novel. The ratio of heterozygous variants to homozygous variants is 1.5 to 1.68 and the Ti/Tv ratio was found to be 2.6 to 2.8 for all six exomes, which is the expected ratio (DePristo et al. 2011) (Table 4-4). Non-coding SNVs (introns and intergenic) comprised around 25 to 31%, approximately 3 to 4% of the variants were located in untranslated region (5' UTR, 3'UTR and UTR-Splice sites) and the remaining 65 to 70% constituted coding SNVs (synonymous, missense and nonsense SNVs) (Supplementary Table 12-5).

For additional assessment of SNV quality, the genotype concordance was computed using snpActs (http://snpacts.ikmb.uni-kiel.de/) by comparing the sequencing data of the three centenarians ((1) German female, (2) German male, (3) French female) with the Illumina OmniExpress Chip 700k array. The genotype concordance for the Spanish exome sequencing

sample was calculated by CRG, Spain using the Illumina 2.5M array and was reported to be 85%. For the two additional samples ((5) German male and (6) female) sequenced at the BGI, the Affymetrix 6.0 array was used to calculate the genotype concordance with snpActs (http://snpacts.ikmb.uni-kiel.de/). The genotype concordance results are presented in Figure 4-3.

	Array	Overlapping SNVs	Genotype concordance (%)
(1) German female	Illumina OmniExpress (700k)	9,084	99.5
(2) German male	Illumina OmniExpress (700k)	9,226	99.46
(3) French female	Illumina OmniExpress (700k)	9,217	99.48
(5) German female	Affymetrix 6.0 (900k)	4,553	99.01
(6) German male	Affymetrix 6.0 (900k)	4,781	99.19

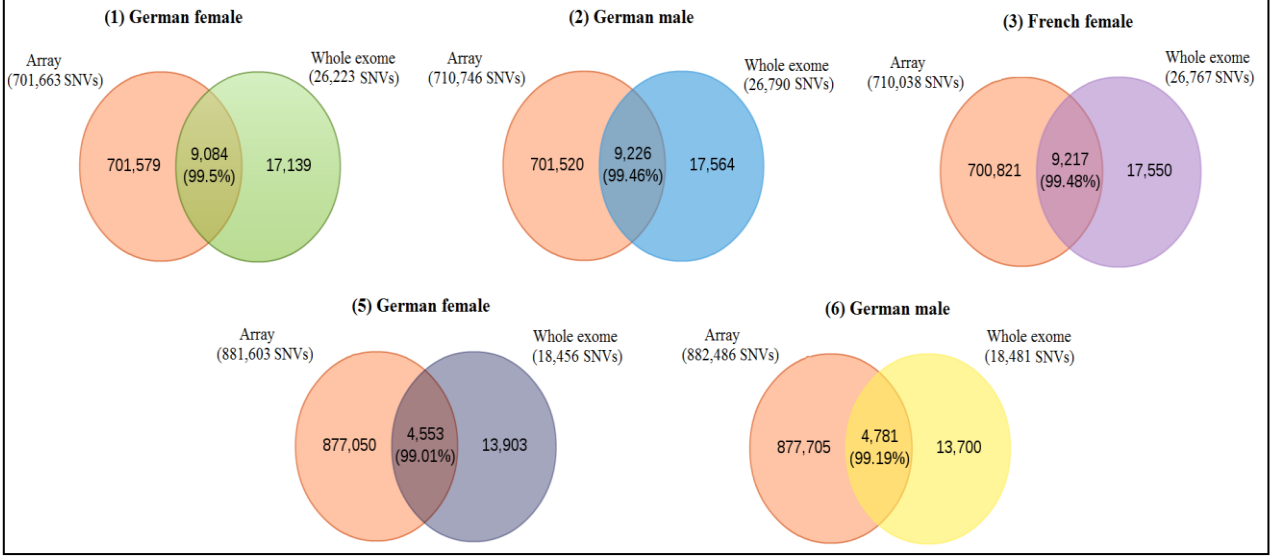

Figure 4-3: **Genotype concordance for Illumina exome sequencing data**: Genotype concordance of Illumina whole exome sequencing data compared with the array data; (1) German female (99.5%), (2) German male (99.46%), (3) French female (99.48%), (4) German female (99.01%) and (5) German male (99.19%).

4.2 Method 1: SNVs that may have a functional impact

4.2.1 Selection of variants

The SNVs with MAF 1% to 50% for further genotyping were selected by Daniel Trujillano from CRG in Spain. For this method, variants generated from SOLiD whole genome sequencing and Illumina exome sequencing of four centenarians were used ((1) German male and (2) female, (3) French female and (4) Spanish female). Reads were aligned to hg19 human genome reference with BWA, where 93.41% of all targeted bases were covered by at least 20 reads, obtaining an

average coverage of 50x for all four subjects. The resulting alignments from the exome sequencing were used as input for three different variant calling tools, namely GATK (McKenna et al. 2010), mpileup (Li et al. 2009b) and SHORE (Schneeberger et al. 2009). A genotype concordance of 85% was observed by the CRG for the Spanish sample when compared with the Illumina Omni 2.5M array. For the other three individuals, over 95% genotype concordance was observed when compared with the Illumina OmniExpress Array using snpActs (http://snpacts.ikmb.uni-kiel.de/). Combining the target-enriched variants (which are coding regions of the genome), 65,826 variants were found to be common among all four individuals. As shown in Figure 4-4, 97.4% of the variants were already known and reported in dbSNP 132 and the remaining variants were novel. The Ti/Tv ratio was found to be 2.48 for known variants and 2.25 for novel variants, which meets the expected value according to DePristo (DePristo et al. 2011).

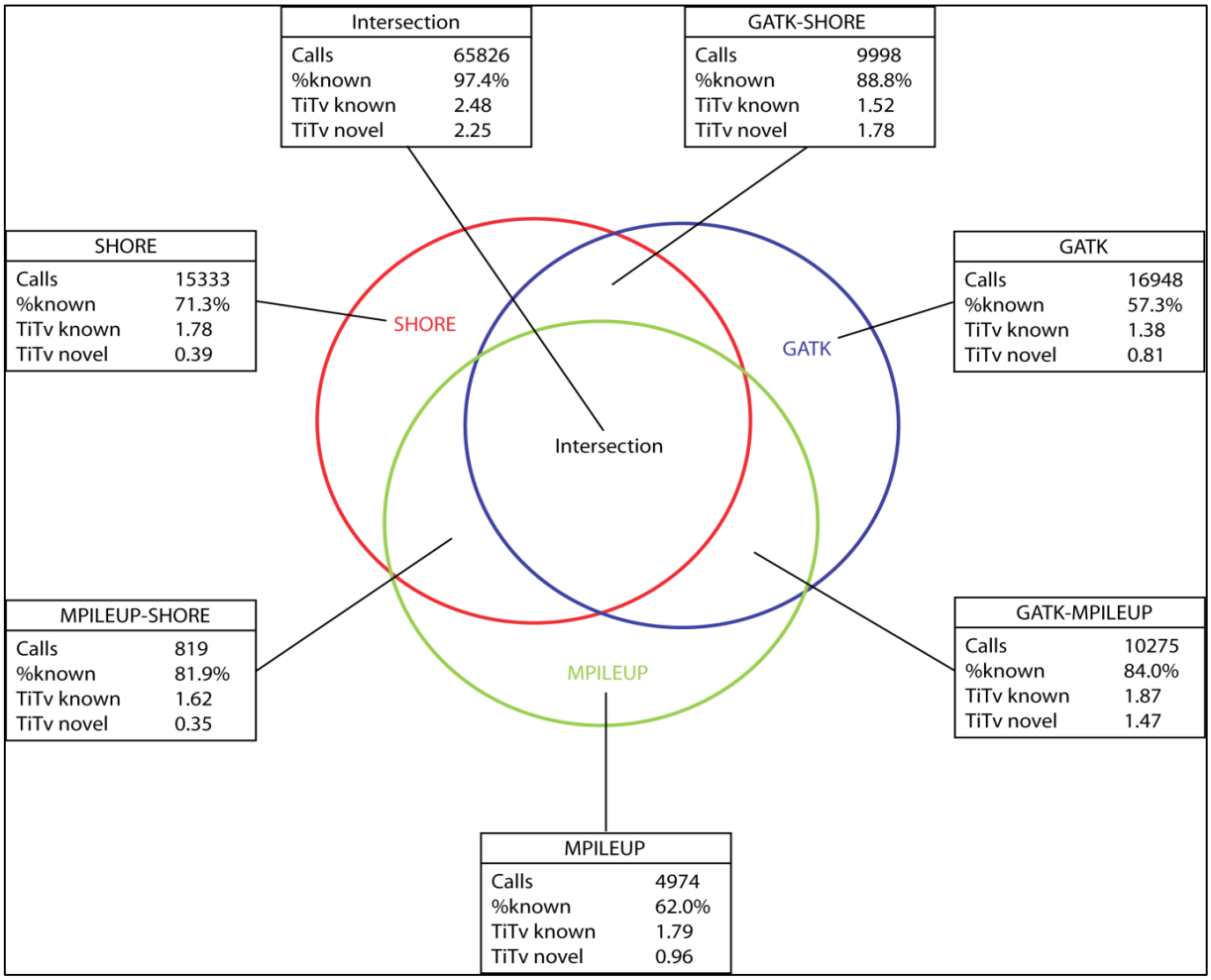

Figure 4-4: **Variant calling metrics for four exomes generated by CRG, Spain**: Three different variant calling tools, namely GATK, mpileup and SHORE were implemented for the exome sequencing data ((1) German male and (2) female, (3) French female and (4) Spanish female) and the subsequent results from all three variant callers were combined to obtain 65,826 variants (Figure by CRG, Spain).

The selection criteria for variants chosen for this approach have been outlined in Figure 4-5. The exome Illumina variant calls were then combined with exonic SNVs (including splice-sites and untranslated exonic regions) detected in whole genome SOLiD SNV calls to give a total of 66,658 SNVs. To filter the variant list further, SNVs that were present in at least two samples that had significantly different MAFs with respect to 1000G and ESP databases, and were found to be conserved by PhyloP were selected for further investigation. This constituted a list of 116 potentially functional variants listed in Supplementary Table 12-6. These 116 SNVs were chosen for further genotyping in our German LLI (*n* = 1,610) and younger controls (*n* = 1,104). The significant SNVs were subsequently tested for replication in the French (*n* = 1,269 LLI and 1,834 younger controls) and Danish (*n* = 910 LLI and 760 younger controls) longevity samples.

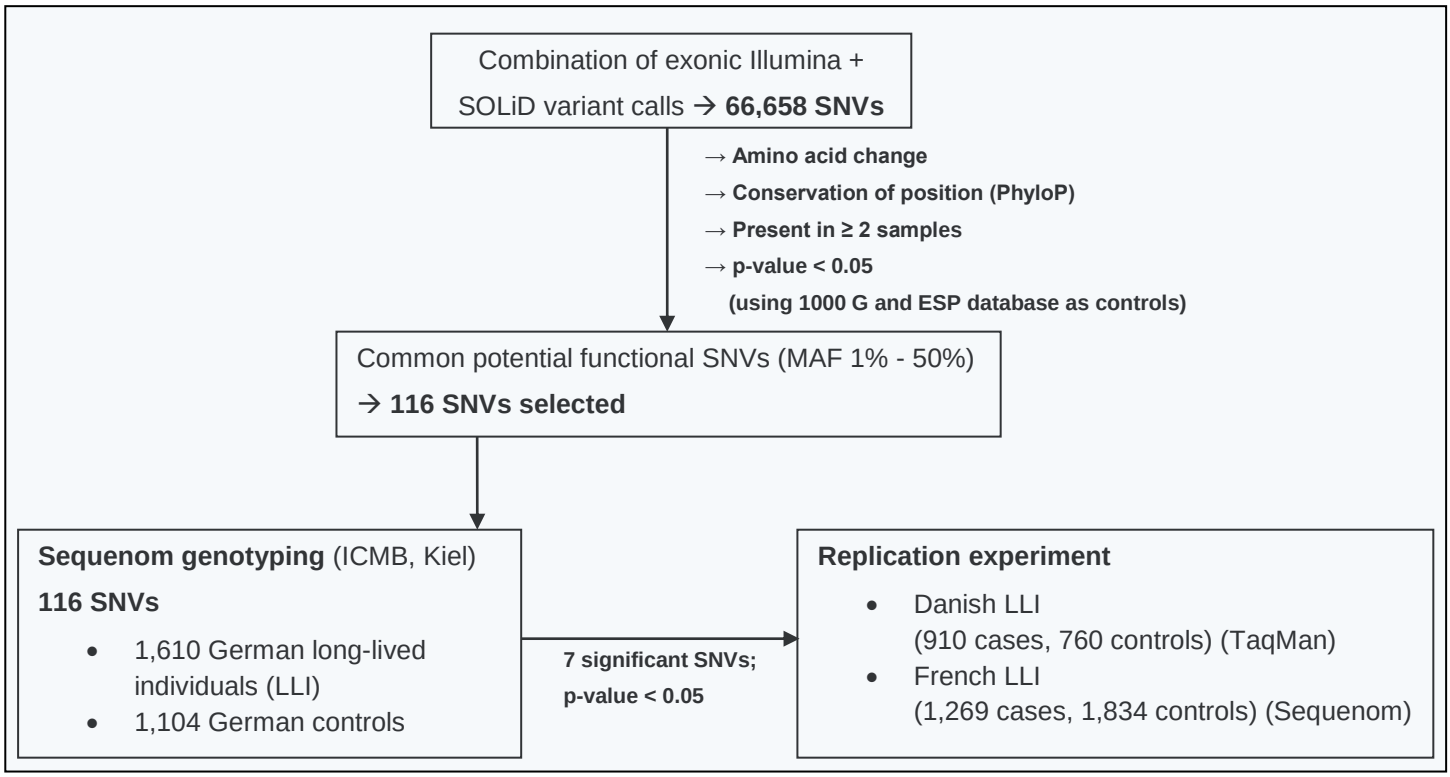

Figure 4-5: **Selection of variants for Method 1**: selection of variants from four sequenced centenarians ((1) German female and (2) male, (3) French female and (4) Spanish female) to detect SNVs that may have a functional impact.

4.2.2 Analysis of selected SNVs

After combining the exome Illumina SNV calls with exonic SNVs detected in whole genome SOLiD data, a total of 116 SNVs was selected for subsequent genotyping and association testing. German LLI [*n* = 1,610; age range: 95-110 years, including a centenarian subset (*n* = 745)] were compared to younger controls (*n* = 1,104; age range: 60-75 years) matched for ancestry, gender and geographical origin within Germany (Supplementary Table 12-7).

Out of 116 SNVs, 109 variants were genotyped with Sequenom technology and 1 variant failed the assay design. The remaining six SNVs were typed with TaqMan. Seven SNVs showed a significant association signal with a p-value of less than 0.05 in either the whole sample (see Table 4-6) or the centenarian subset (see Table 4-7).

Chr.	dbSNP ID	Gene	MAF cases *n* = 1,610	MAF controls *n* = 1,104	P_{CCA}	OR	95% CI
1	rs10927851	FBLIM1	0.293 (C)	0.342	0.0002	0.80	0.7037-0.8961
14	rs3093921	PARP2	0.036 (G)	0.021	0.0020	1.71	1.215-2.432
5	rs61757629	NAIP	0.031 (T)	0.018	0.0050	1.68	1.166-2.432
13	rs35719359	PCCA	0.064 (G)	0.050	0.0277	1.30	1.029-1.657
6	rs17054318	PLEKHG1	0.035 (C)	0.047	0.0306	0.74	0.5628-0.9732
11	rs34108746	PRG3	0.062 (G)	0.074	0.0909	0.83	0.6686-1.03
11	rs78489201	TNKS1BP1	0.066 (G)	0.075	0.1883	0.87	0.7004-1.073

Chr.: chromosome id
MAF: minor allele frequency
P_{CCA}: p-value obtained from an allele-based case–control comparison, using a $\chi 2$-test with 1 degree of freedom
OR: Odds ratio for attaining old age with the minor allele in controls as reference allele
95% CI: 95% confidence interval for OR

Table 4-6: **Longevity association statistics in German LLI for seven SNVs**: Association statistics for seven SNVs with potential functional impact in German LLI (n=1,610) and younger controls (n=1,104).

Chr.	dbSNP ID	Gene	MAF cases *n* = 745	MAF controls *n* = 1,104	P_{CCA}	OR	95% CI
1	rs10927851	FBLIM1	0.289 (C)	0.342	0.0009	0.78	0.6745-0.9046
14	rs3093921	PARP2	0.028 (G)	0.021	0.1758	1.34	0.8756-2.054
5	rs61757629	NAIP	0.032 (T)	0.019	0.0155	1.69	1.099-2.578
13	rs35719359	PCCA	0.062 (G)	0.050	0.0963	1.27	0.9573-1.69
6	rs17054318	PLEKHG1	0.036 (C)	0.047	0.1170	0.76	0.5408-1.072
11	rs34108746	PRG3	0.053 (G)	0.074	0.0102	0.69	0.5226-0.9179
11	rs78489201	TNKS1BP1	0.055 (G)	0.075	0.0202	0.72	0.5471-0.9513

For abbreviations see legend to Table 4-6

Table 4-7: **Longevity association statistics in German centenarian subgroup for seven SNVs:** Association statistics for seven SNVs in German centenarians subset (n=745) and younger controls (n=1,104).

The seven SNVs that showed an association in either the whole German sample or centenarian subset were investigated for replication in two independent LLI populations. The French

replication samples comprised 1,269 LLI (age range: 90-115 years) and 1,834 younger controls (age range: 35-61 years). The Danish samples consisted of 910 LLI (age range: 94-100 years) and 760 younger controls (age range: 60-72 years). No significant association was observed in the French replication sample for the seven selected SNVs as seen in Table 4-8.

Chr.	dbSNP ID	Gene	MAF cases *n* = 1,269	MAF controls *n* = 1,834	P_{CCA}	OR	95% CI
1	rs10927851	FBLIM1	0.291 (C)	0.305	0.2271	0.93	0.8358-1.044
14	rs3093921	PARP2	0.018 (G)	0.022	0.2899	0.82	0.5662-1.186
5	rs61757629	NAIP	0.036 (T)	0.031	0.3100	1.16	0.8729-1.533
13	rs35719359	PCCA	0.050 (G)	0.046	0.3708	1.11	0.88-1.409
6	rs17054318	PLEKHG1	0.049 (C)	0.046	0.5780	1.07	0.8434-1.357
11	rs34108746	PRG3	0.084 (G)	0.075	0.1553	1.15	0.9499-1.38
11	rs78489201	TNKS1BP1	0.084 (G)	0.076	0.1865	1.13	0.9412-1.365

For abbreviations see legend to Table 4-6

Table 4-8: **Longevity association statistics for replication in French LLI for seven SNVs:** Association statistics for seven SNVs in French LLI (n=1,269) and younger controls (n=1,834).

Chr.	dbSNP ID	Gene	MAF cases *n* = 910	MAF controls *n* = 760	P_{CCA}	OR	95% CI
1	rs10927851	FBLIM1	0.308 (C)	0.271	**0.0298**	1.20	1.018-1.381
14	rs3093921	PARP2	0.027 (G)	0.035	0.1485	0.75	0.4993-1.112
5	rs61757629	NAIP	0.022 (T)	0.024	0.6122	0.88	0.5612-1.406
13	rs35719359	PCCA	0.070 (G)	0.078	0.3940	0.89	0.6841-1.161
6	rs17054318	PLEKHG1	0.040 (C)	0.042	0.8471	0.97	0.6848-1.365
11	rs34108746	PRG3	0.082 (G)	0.080	0.8151	1.03	0.7972-1.334
11	rs78489201	TNKS1BP1	0.093 (G)	0.081	0.2489	1.15	0.9026-1.484

For abbreviations see legend to Table 4-6

Table 4-9: **Longevity association statistics for replication in Danish LLI for seven SNVs:** Association statistics for seven SNVs in Danish LLI (n=910) and younger controls (n=760).

For the Danish replication experiment (see Table 4-9), one SNV (rs10927851) showed a nominally significant P_{CCA} of 0.0298, but this cannot be considered a positive replication as the frequency difference between cases and controls is opposite when compared with the German LLI sample as shown in Table 4-11.

Chr.	dbSNP ID	Gene	MAF cases *n* = 2,179	MAF controls *n* = 2,594	P_{CCA}	OR	95% CI
1	rs10927851	FBLIM1	0.298 (C)	0.296	0.8916	1.01	0.92-1.101
14	rs3093921	PARP2	0.021 (G)	0.026	0.1958	0.84	0.6396-1.096
5	rs61757629	NAIP	0.029 (T)	0.029	0.8293	1.03	0.8083-1.304
13	rs35719359	PCCA	0.059 (G)	0.055	0.4141	1.08	0.9031-1.281
6	rs17054318	PLEKHG1	0.046 (C)	0.044	0.8623	1.02	0.8376-1.236
11	rs34108746	PRG3	0.083 (G)	0.076	0.1779	1.11	0.9542-1.288
11	rs78489201	TNKS1BP1	0.088 (G)	0.077	**0.0503**	1.16	0.9997-1.341

For abbreviations see legend to Table 4-6

Table 4-10: **Longevity association statistics for the combined analysis in French and Danish LLI:** Association statistics for seven SNVs in combined French and Danish LLI (n=2,179) and younger controls (n=2,594).

Both the French and Danish data were combined to increase power; after combining the data, one SNV (rs78489201) showed a marginally significant association (P_{CCA}=0.05) (see Table 4-10). However, rs78489201 could not be considered a positive replication as again, the frequency difference between cases and controls is contrary to the frequency distribution in the German sample (see Table 4-11).

Chr.	dbSNP ID	MAF cases	MAF controls	P_{CCA}	OR	Population
1	rs10927851	0.293	0.342	0.0002 (C)	0.80	German LLI
	rs10927851	0.308	0.271	0.0298 (C)	1.20	Danish LLI
11	rs78489201	0.055	0.075	0.0202 (G)	0.72	German centenarian subset
	rs78489201	0.088	0.077	0.0503 (G)	1.16	combined French and Danish LLI

For abbreviations see legend to Table 4-6

Table 4-11: **Frequency distribution between cases and controls**: Case-control frequency distribution in (1) German and Danish LLI and (2) German subgroup and combined French and Danish LLI for two SNVs that showed a significant association signal in the German longevity sample and nominal significance in the replication sample.

4.3 Method 2: Low-frequency variants with functional impact

4.3.1 Selection of variants

For this method, SNVs called from the SOLiD and Illumina sequencing data were combined and variants that showed an intersection between the two platforms (SOLiD and Illumina) were selected for further investigation. As SOLiD and Illumina employ different sequencing techniques, it is expected that the intersection would represent true positive variants among large proportions of putative false-positive calls and sequencing artefacts randomly distributed over the genome.

	(1) German female 108 years	(2) German male 109 years	(3) French female >114 years	(4) Spanish female >110 years	(5) German female 108 years	(6) German male 106 years	Total
SOLiD (whole genome sequencing)	3,264,818	3,923,324	2,695,673	NA	NA	NA	6,465,384
Illumina (whole genome sequencing)	4,013,012	4,071,554	4,022,164	4,081,702	NA	NA	7,130,986
Illumina (exome sequencing)	26,223	26,790	26,767	27,178	18,456	18,481	69,666

Table 4-12: Total number of variants generated by SOLiD and Illumina technology in six centenarians.

The SNVs generated by SOLiD and Illumina sequencing for all six centenarians (two German females, two German males, one French female and one Spanish female) were annotated using snpActs (http://snpacts.ikmb.uni-kiel.de/) and Annovar (Wang et al. 2010). In total, 6,465,384 variants were called from the SOLiD sequenced data (genome) and 7,162,264 variants from the Illumina sequenced data (genome and exome) as presented in Table 4-12. The variants were filtered for exonic regions and SNVs that showed an intersection from both sequencing technologies were chosen for further analysis (see Figure 4-6).

The list was further refined by removing variants that showed a MAF>10%, when compared with the 1000G or the ESP database. This reduced the list to 2,959 variants, including variants with no listed MAF in either 1000G or ESP database. Out of the 2,959 exonic variants, 2,888 coding variants were evaluated with eight different prediction tools to select variants of high functional interest.

	SOLiD sequencing	Illumina sequencing	Overlap	Genotype concordance
Exonic SNVs (missense, nonsense, cancel-start, read-through, splice-sites and unknown)	25,321	23,245	12,528	99.92%

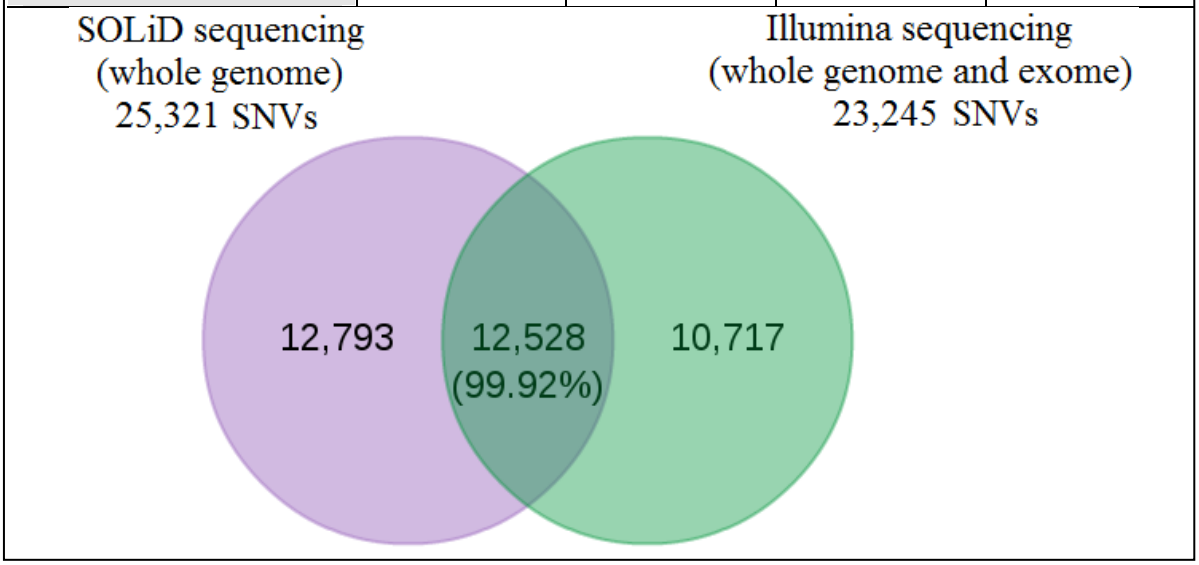

Figure 4-6: **Intersection of exonic variants between SOLiD and Illumina technology**: Variants that showed an intersection between both sequencing technologies (12,528 SNVs) were selected for further analysis.

Prediction tools are advantageous to prioritize variants that may affect the structure or function of proteins. The scores for each variant from all tools were computed by Carolin Knecht from the IMIS, University of Kiel. The input information was protein sequences, amino acid substitution and/or UniProt IDs. Table 4-13 shows the total number of variants that were successfully evaluated for each tool. For each SNV, the binary decisions of each tool were aggregated by summation, whereby -1 implies 'no effect' or neutral, 1 indicates 'effect' or 'damaging' for each tool and 'error' suggests that no prediction could be made by the tools for some variants (most of the tools give errors for predicting the effect of a variant when the amino acid substitution does not match with the given protein sequence). SNPs&GO uses UniProt IDs instead of protein

sequences for input and as many of the IDs were not present in the embedded UniProt database and hence, this tool has given maximum errors. The term 'damaging' used in this context refers to a change in amino acid leading to either loss-of-function or gain-of-function, thereby likely to influence human longevity in either direction. Variants that were predicted to have an effect on the amino acid change in either seven or eight tools were selected for further investigation.

	Grantham score	PhyloP	MutPred	SNAP	Pmut	SIFT	Polyphen2	SNPs&GO
Neutral (Effect: -1)	2306 (79.8%)	1565 (54.2%)	2204 (76.3%)	1888 (65.3%)	1809 (62.6%)	1833 (63.5%)	1502 (52%)	1880 (65.1%)
Damaging (Effect: 1)	582 (20.2%)	1322 (45.78%)	612 (21.2%)	913 (31.6%)	982 (34%)	742 (25.7%)	966 (33.4%)	213 (7.4%)
Error	0	1 (0.02%)	72 (2.5%)	87 (3.01%)	97 (3.4%)	313 (10.8%)	420 (14.6%)	795 (27.5%)

Table 4-13: **Number of variants evaluated with each prediction tool**: Eight different prediction tools were chosen to evaluate 2,959 exonic variants.

In order to prioritize the SNVs of interest, different lists were generated based on various criteria, as shown in Figure 4-7.

Variants selected based on longevity genes and pathways list

Known longevity genes and pathways from various model organisms listed in the NetAge database (Tacutu et al. 2010) were used as a filter mask to overlay with the source list of variants (exonic variants with MAF≤10% in 1000G and ESP). Fifteen variants were located in genes involved in insulin and/or mTOR signaling (Supplementary Table 12-8). These 15 variants were then checked for their functional effect with eight different prediction tools. Out of 15 variants, only one variant (rs146426104) with an extremely low MAF (<1%) was predicted to be 'damaging' by seven of the eight tested tools. The remaining 14 variants were predicted to be neutral in their effect by most of the tools. Two of the variants (rs3208856 and rs17313469) were selected even though they were predicted neutral, because they were found in two individuals, which is much more frequent than expected by chance compared to their MAF of 3 to 5% in the European population of the 1000G.

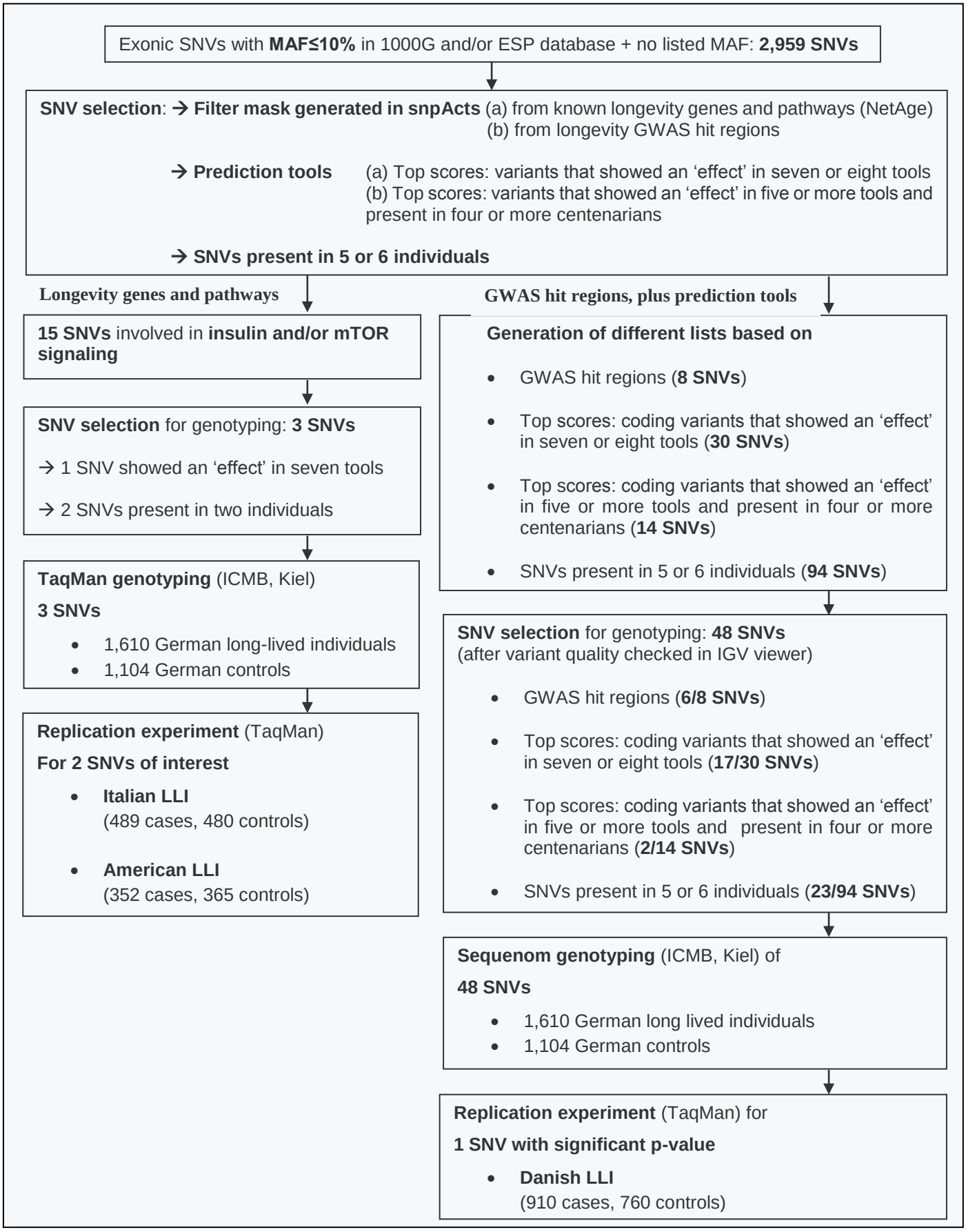

Figure 4-7: **Selection of variants for Method 2**: Selection of variants from six individuals sequenced with SOLiD and Illumina technology to select low-frequency variants based on various criteria.

These three variants as shown in Table 4-14 were then selected for genotyping using the TaqMan technology in the German longevity sample (LLI, *n* = 1,610 and younger controls, *n* = 1,104).

Chr	dbSNP ID	source	Pathway	Gene	MAF*	No. of tools predicted damaging effect [0;8]	No. of tools predicted neutral effect [0;8]
12	rs146426104	(2) German male	Insulin signaling	ACACB	0.006098	7	1
19	rs3208856	(3) French female, (6) German male	Insulin signaling	CBLC	0.036585	3	5
X	rs17313469	(3) French female, (6) German male	Insulin signaling	PHKA2	0.057851	1	7

*MAF: minor allele frequency observed in either 1000G or ESP database

Table 4-14: **Variants selected from genes involved in mTOR or insulin signaling**: Three variants were present in genes involved in mTOR/insulin signaling and were selected because rs146426104 was predicted as "damaging" by seven different tools, rs3208856 and rs17313469 were present in two centenarians.

Variants selected based on longevity GWAS associated hit regions

Our previous longevity GWAS data used for this approach has been described in detail elsewhere (Nebel et al. 2011). It comprises 664,472 autosomal SNPs genotyped in 763 LLI (mean age: 99.7 years) and 1,085 controls (mean age: 60.2 years) from Germany. Associated variants were 'clumped together' using the PLINK clumping algorithm and these 'clumped' regions were used as targets for further SNV selection. There were 325 clumped regions formed with a p-value less than 0.001. The whole variant source list (2,959 SNVs) was then overlaid with the 'clumped' regions. Eight variants were present within the range of the associated hit regions as shown in Table 4-15.

Variants selected based on top scores from prediction tools

Out of the 2,888 variants, 30 SNVs (Supplementary Table 12-9) were chosen based on top scores determined by the binary decisions of each tool. These SNVs were predicted as 'damaging' by seven or eight out of eight tools and each variant was present either in one, two or three individuals.

Variants selected based on prediction tools and their presence in four or more individuals

Fourteen variants were selected (Supplementary Table 12-10) that were predicted as 'damaging' by five or more tools out of eight and were present in four or more individuals, which is much

more frequent than expected by chance, as these SNVs did not even have a MAF listed in the 1000G or ESP databases.

				GWAS associated hit regions	
dbSNP ID	Gene	source	MAF*	p-value‡	Location
rs17123306	KANK4	(2) German male	0.018293	0.00045	Chr1:62732420..62760411 [KANK4]
rs199619070	TTN	(3) French female	0.000363	0.000197	Chr2:179503210..179589768 [TTN]
rs17452588	TTN	(2) German male	0.009214	0.000197	Chr2:179503210..179589768 [TTN]
rs61764030	UGT1A3	(3) French female	0.002442	0.000194	Chr2:234628575..234651799 [UGT1A6,UGT1A8,UGT1A9, UGT1A7,UGT1A10,UGT1A4, UGT1A5,UGT1A3,DNAJB3]
rs200305979	DPYSL5	(3) French female	0.000116	2.45e-05	Chr2:27132821..27299597 [TMEM214,MAPRE3, DPYSL5,AGBL5]
rs3749971	OR12D3	(2) German male, (5) German female	0.095089	0.000696	Chr6:29342774..29461729 [OR2H1,OR12D2,OR12D3, OR11A1,OR10C1,MAS1L]
chr17_44128052	KANSL1	(3) German female	NA	0.000307	Chr17:43801694..44197602 [STH,MAPT, KANSL1, IMP5,CRHR1]
rs35653278	ZNF750	(2) German male	0.0609756	0.000796	Chr17:80776042..80796235 [ZNF750,TBCD]

*MAF: minor allele frequency observed in either 1000G or ESP database
‡pvalue: GWAS pvalue

Table 4-15: **Variants selected based on longevity GWAS hit regions:** Eight variants were selected that were present within the range of longevity GWAS associated hit regions.

Variants selected based on their presence in either five or six individuals

Ninety-four SNVs out of 2,888 were found to be present in five or six individuals. The 94 SNVs were selected independent of the scores from the prediction tools, as their MAF was not even listed in the 1000G or ESP databases and hence they were present more frequently than expected by chance. Many variants with no listed MAF in either the 1000G or the ESP database were selected, because they occurred more frequently than expected by chance among all six individuals. If these variants are very rare, they might have no listed allele frequency in either of the databases. However, on the other hand, if those SNVs have not been called accurately, they could also be false positives. Therefore, to make sure that all selected SNVs are of good quality,

each variant was visualized manually using the Integrative Genomics Viewer (IGV). This comprised a list of 23 SNVs of interest, present in five or six individuals.

Altogether, 48 SNVs were selected for further genotyping using Sequenom technology in our German LLI (listed in Supplementary Table 12-11).

4.3.2 Analysis of selected SNVs

Variants that showed an intersection between both the SOLiD and Illumina platforms were filtered, followed by prioritizing the SNV list based on various criteria (see Figure 4-8). A total of 51 SNVs (see Figure 4-7) were selected for subsequent association testing using the TaqMan and Sequenom technology. German LLI (*n* = 1,610, age range: 95-110 years) were compared to younger controls (*n* = 1,104, age range: 60-75 years) matched for ancestry, gender and geographical origin within Germany.

SNVs selected based on longevity genes and pathways list

Three variants based on target regions (NetAge pathway list) and prediction tools were selected and genotyped using the TaqMan technology. No significant association was observed in the whole German sample (see Table 4-16), but one SNV (rs3208856) showed a significant association signal with an allelic p-value of 0.038 in the centenarian subset (see Table 4-17).

Chr.	dbSNP ID	Gene	MAF cases *n* = 1,610	MAF controls *n* = 1,104	P_{CCA}	OR	95% C.I.
19	rs3208856	CBLC	0.04589 (T)	0.03676	0.1033	1.26	0.9536 - 1.665
X	rs17313469	PHKA2	0.02344 (G)	0.03062	0.1352	0.7597	0.5292-1.091
12	rs146426104	ACACB	0.001593 (T)	0.0009328	0.517	1.709	0.3313-8.818

For abbreviations see legend to Table 4-6

Table 4-16: **Longevity association statistics in German LLI for three low-frequency SNVs:** Association statistics for three SNVs selected based on genes involved in mTOR/insulin signaling in German LLI (n=1,610) and younger controls (n=1,104).

Chr.	dbSNP ID	Gene	MAF cases *n* = 910	MAF controls *n* = 760	P_{CCA}	OR	95% C.I.
19	rs3208856	CBLC	0.05081 (T)	0.03676	**0.03878**	1.403	1.016-1.936
X	rs17313469	PHKA2	0.02023 (G)	0.03062	0.07308	0.6537	0.4094-1.044
12	rs146426104	ACACB	0.003406 (T)	0.0009328	0.09692	**3.66**	0.7092-18.89

For abbreviations see legend to Table 4-6

Table 4-17: **Longevity association statistics in German centenarian subgroup for three low-frequency SNVs:** Association statistics for three SNVs selected based on genes involved in mTOR/insulin signaling in German centenarians subset (n=745) and younger controls (n=1,104).

Two SNVs (rs3208856 and rs146426104) were genotyped for replication in the Italian and American longevity samples. The SNV rs3208856 was selected due to the nominally significant P_{CCA} in the centenarian subset and rs146426104 was selected due to its very high odds ratio in the centenarian subset.

However, as we can see in Table 4-18, the genetic association signal observed in Germans could not be confirmed in the Italian LLI sample and in addition, one of the SNVs (rs146426104) turned out to be monomorphic in this longevity sample.

Chr.	dbSNP ID	Gene	MAF cases *n* = 489	MAF controls *n* = 480	P_{CCA}	OR	95% CI
19	rs3208856	CBLC	0.02206 (T)	0.029	0.4438	0.7444	0.3488-1.58
12	rs146426104	ACACB	0	0	-	-	-

For abbreviations see legend to Table 4-6

Table 4-18: **Longevity association statistics for replication in Italian LLI for two low-frequency SNVs:** Association statistics for replication for two SNVs selected based on genes involved in mTOR/insulin signaling in Italian longevity sample.

In the American longevity sample, one of the two SNVs, rs3208856, confirmed the association signal with an allelic p-value of 0.000189 (see Table 4-19). The allele frequency for case-control distribution in the American longevity sample was similar to the original findings in the Germans.

Chr.	dbSNP ID	Gene	MAF cases *n* = 352	MAF controls *n* = 365	P_{CCA}	OR	95% C.I.
19	rs3208856	CBLC	0.0696 (T)	0.0274	**0.000189**	2.656	1.562-4.51
12	rs146426104	ACACB	0.007246 (T)	0.005495	0.6937	1.321	0.329-5.30

For abbreviations see legend to Table 4-6

Table 4-19: **Longevity association statistics for replication in American LLI:** Association statistics for replication for two SNVs selected based on genes involved in mTOR/insulin signaling in American longevity sample.

SNVs selected based on GWAS hit regions and prediction tools

Furthermore, lists were generated based on various criteria as shown in Figure 4-7, where 48 SNVs that may have a potential influence on the longevity phenotype were chosen for further genotyping with the Sequenom technology.

Out of 48 variants, 31 SNVs were successfully genotyped and the remaining SNVs either failed the assay design or turned out to be monomorphic in our German population (Supplementary Table 12-12). Three of the 31 tested SNVs showed a significant association signal with an allelic p-value of less than 0.05 in the whole sample (see Table 4-20) and the centenarian subset (see Table 4-21).

Chr.	dbSNP ID	Gene	MAF cases *n* = 1,614	MAF controls *n* = 1,104	P_{CCA}	OR	95% C.I.
20	rs35761929	JAG1	0.1246 (C)	0.07678	3.7e-08	1.712	1.411-2.076
17	rs35653278	ZNF750	0.1259 (A)	0.1002	0.004491	1.294	1.083-1.546
11	rs34898047	MICALCL	0.009036 (A)	0.0176	0.006923	0.5089	0.3089-0.8384

For abbreviations see legend to Table 4-6

Table 4-20: **Longevity association statistics in German LLI for low-frequency variants:** Association statistics for three significant SNVs selected based on GWAS hit regions and prediction tools in German LLI (n=1,614) and younger controls (n=1,104).

Chr.	dbSNP ID	Gene	MAF cases *n* = 910	MAF controls *n* = 760	P_{CCA}	OR	95% C.I.
20	rs35761929	JAG1	0.1096 (C)	0.07678	0.0008297	1.481	1.175-1.866
17	rs35653278	ZNF750	0.1302 (A)	0.1002	0.005778	1.344	1.089-1.659
11	rs34898047	MICALCL	0.007891 (A)	0.0176	0.01569	0.4439	0.2257-0.8732

For abbreviations see legend to Table 4-6

Table 4-21: **Longevity association statistics in German centenarian subgroup for low-frequency variants:** Association statistics for three significant SNVs selected based on GWAS hit regions and prediction tools in German centenarian subset (n=748) and younger controls (n=1,104).

Based on the above analysis, it was decided to replicate only the top-ranking SNV (rs35761929) in a larger Danish sample that consisted of 910 LLI (age range: 94-100 years) and 760 younger controls (age range: 60-72 years). The SNV rs35761929 had 89% power to replicate the observed association with an OR of 1.7 as presented in Figure 4-8.

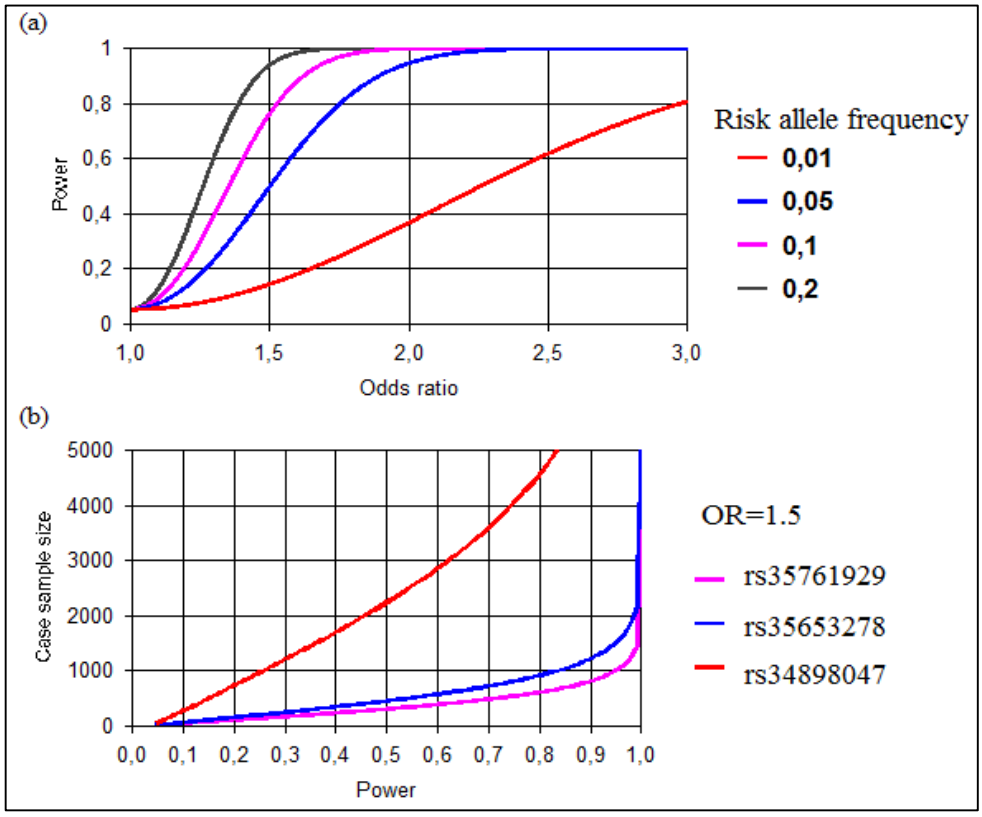

Figure 4-8: **Power and sample size calculation:** (a) Power calculation diagram showing the expected power in relation to sample size, odds ratio and risk allele frequency for the Danish sample of 910 LLI and 760 controls with a significance level of 0.05. (b) Sample size calculation diagram showing the required number of cases (with a case-

control ratio of 1) in relation to the expected power, given a significance level of 0.05, an OR of 1.5 and minor allele frequency similar to that of observed in German LLI.

The other two SNVs (rs35653278 and rs34898047) had a power of less than 40% to replicate in the Danish LLI. Considering a case-control ratio of 1, a sample size of 1,000 (for rs35653278) to 4,500 cases (for rs34898047) would be required to gain 80% power for the replication of SNVs rs35653278 and rs34898047, assuming an OR of 1.5.

Hence, the SNV rs35761929, which showed a significant association with a P_{CCA} of 3.7e-08 in the German LLI, was further investigated for a replication experiment using independent Danish LLI data comprising 910 LLI and 760 younger controls. However, the genetic association observed in Germans could not be confirmed in the Danish population as seen below in Table 4-22.

Chr.	dbSNP ID	Gene	MAF cases *n* = 910	MAF controls *n* = 760	P_{CCA}	OR	95% C.I.
20	rs35761929	JAG1	0.09116 (C)	0.08508	0.5383	1.079	0.8474-1.373

For abbreviations see legend to Table 4-6

Table 4-22: **Longevity association statistics for replication in Danish LLI for low-frequency variant:** Association statistics for replication of one SNV selected based on prediction tools in Danish LLI (n=910) and younger controls (n=760).

SNV list	dbSNP ID	Gene	Effect allele	Direction of effect cases≥85 years	Pvalue cases≥85 years	Direction of effect cases≥90 years	Pvalue cases≥90 years
Method 2: Pathway SNVs	rs3208856	CBLC	T	+	**0.026**	+	0.34
Method 1: SNVs with MAF 1 to 50%	rs17054318	PLEKHG1	T	-	0.24	-	0.37
Method 1: SNVs with MAF 1 to 50%	rs35719359	PCCA	T	-	0.34	+	0.52
Method 1: SNVs with MAF 1 to 50%	rs10927851	FBLIM1	T	+	0.38	+	0.28
Method 2: Prediction tools	rs35761929	JAG1	T	+	0.43	+	0.51
Method 1: SNVs with MAF 1 to 50%	rs3093921	PARP2	G	+	0.83	-	0.69

Table 4-23: **Longevity association statistics for SNVs of interest in meta-analysis discovery sample:** Association statistics for SNVs that showed a significant association signal in the German longevity sample and are present in the discovery-phase meta-analysis of 7,729 cases (≥85 years) and 16,121 controls (≥65 years).

For a further comprehensive investigation, all the SNVs (12 SNVs) that showed a significant association in our German longevity sample from both the approaches were checked for a replication signal in the discovery-phase of the meta-analysis comprising 7,729 LLI of European descent (≥85 years) and 16,121 younger controls, which was recently published (Deelen et al. 2014). Out of 12 SNVs, six SNVs were found to be present in the discovery-phase meta-analysis (Table 4-23), where only one SNV rs3208856 showed a nominally significant signal of 0.026 in cases aged ≥85 years.

All together, 167 SNVs were selected for genotyping based two different approaches. Out of these 167 SNVs, 12 that showed a significant association in our Germany longevity sample were further typed for a replication experiment in different longevity populations (Denmark, France, Italy, USA). Most of the variants failed to confirm the initial association signal apart from one SNV (rs3208856) in the American longevity sample. The SNV rs3208856 (C/T) was selected for genotyping because it was present in two centenarians and the variant was predicted by PhyloP to be conserved and 'damaging' by SIFT and SNAP. It is a missense variant located on the *CBLC* gene and is involved in the insulin pathway. The amino acid substitution is from histidine (His) to tyrosine (Tyr) at position 405 (p.His405Tyr). Replication in the American sample reached a pvalue of 0.000189 with a frequency distribution similar to that in our German sample, where the minor allele 'T' is overrepresented in centenarians as compared to controls. Hence, the *CBLC* variant can be regarded as a very promising candidate that influences longevity but needs further investigation and confirmation in additional larger longevity samples.

5 Discussion

In this project, we combined innovative genetic platforms (next-generation sequencing plus high throughput genotyping) with a contemporary study design (case-control association studies) and statistical/bioinformatics methods (such as SNV evaluation by prediction tools) to identify new variants that contribute to exceptional longevity. To reach this goal, we performed whole genome and exome sequencing of six centenarians (108 to114 years) of European origin using SOLiD and Illumina technologies. Variants for further genotyping investigation have been selected based on two different approaches.

The first approach focused on SNVs with MAF 1 to 50% that might have a functional impact, where seven SNVs showed a significant association signal with a p-value less than 0.05 in either the whole German longevity sample or centenarian subset, but none of the initial findings could be confirmed with a positive replication in independent French and Danish longevity samples.

Our second approach focused on selecting low-frequency variants (MAF≤10%) that showed an intersection between SOLiD and Illumina technologies. Several criteria, as listed in figure 4-7 were implemented to select SNVs for further genotyping. Known longevity genes and pathways from the NetAge database (Tacutu et al. 2010) were used as filter masks for the variant selection. Two SNVs (rs3208856 and rs146426104) that showed a significant association signal in the German centenarian subset were typed for a replication experiment in the Italian and American LLI populations. No significant association was observed in the Italian population but the analysis in the American longevity samples, one SNV (rs3208856), confirmed the signal with an allelic p-value of 0.000189.

Furthermore, the longevity GWAS data (Nebel et al. 2011) was used to generate associated hit regions that were used as targets to overlay with the original SNV list. The functional impact of all low-frequency variants were calculated with eight different prediction tools and SNVs that were present in four or more individuals were also selected. This comprised 48 potential functional SNVs that were selected for genotyping, where three SNVs (rs35761929, rs35653278 and rs34898047) showed a significant association signal in the whole German longevity sample. Based on power calculations, only one of the three SNVs (rs35761929) that had a power of 89% to replicate the observed association signal in the Danish LLI sample was selected further for a replication experiment. However, the genetic association observed in Germans for rs35761929 could not be confirmed in the Danish longevity sample.

5.1 Coverage and variant calling performance for SOLiD and Illumina sequencing data

The term coverage generally refers to the average number of reads that align to each base within the sequence. For example, a whole genome sequenced at 30-fold coverage (30x) means that, on average, each base in the genome is covered by 30 sequencing reads (Sims et al. 2014). Coverage is an essential aspect of next-generation sequencing, as a higher coverage allows for a higher confidence for detection of genetic variants. Usually, high coverage regions tend to have higher calling qualities and low coverage regions tend to have lower variant calling qualities. In general, 20x is deemed necessary for reliable sequence variation calling in data from Illumina and SOLiD platforms (Rieber et al. 2013).

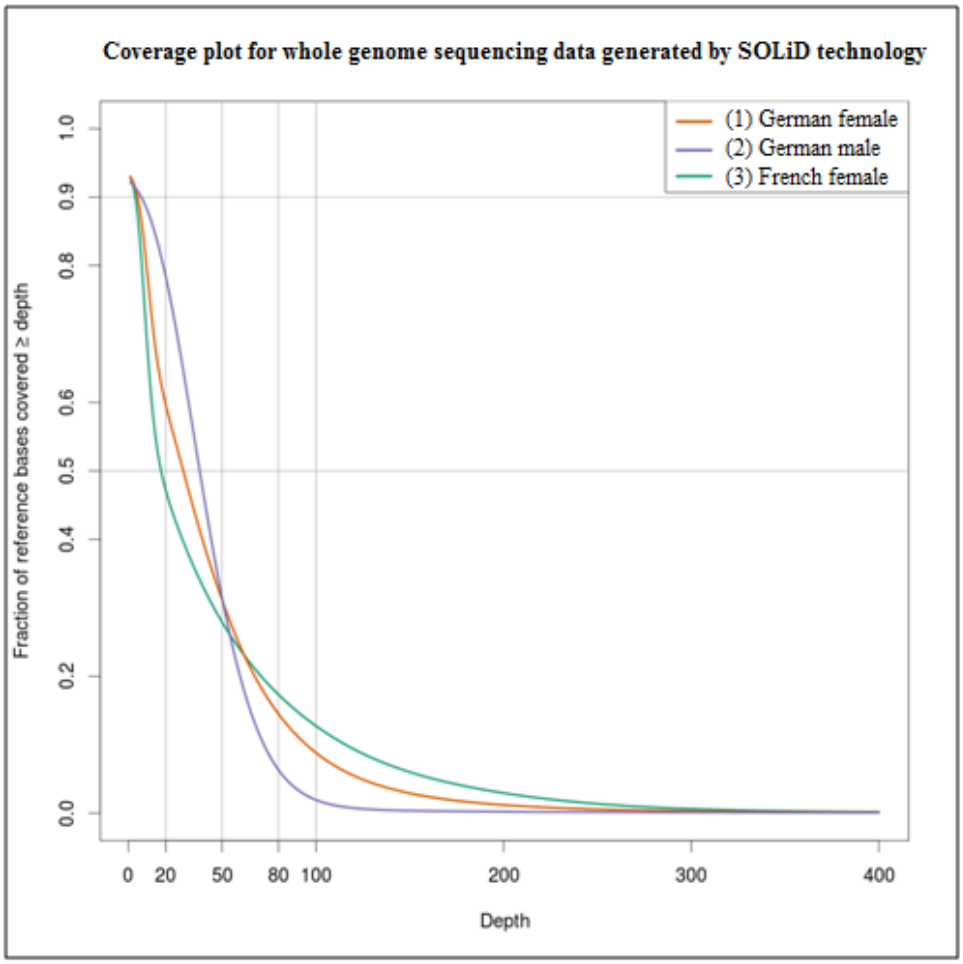

Figure 5-1: Coverage plot for three samples sequenced with SOLiD technology.

Three centenarians ((1) German female and (2) male, plus (3) French female) were sequenced with SOLiD technology using four different libraries per individual. One paired-end library [50 + 35 bp (SOLiD™ Paired-End Library Construction Kit)] and three genomic mate-pair libraries [50

+ 50 bp (SOLiD™ Long Mate-Paired Library Construction Kit)] were generated per sample. The coverage for all three genomes sequenced on SOLiD technology is shown in Figure 5-1. More than 90% of the genome has been covered by at least one read. An average coverage of 43x was generated and 60% of reads were covered at 20x, which is comparable to the coverage attained by other whole genome sequencing studies (Venter et al. 2001; Ratan et al. 2013).

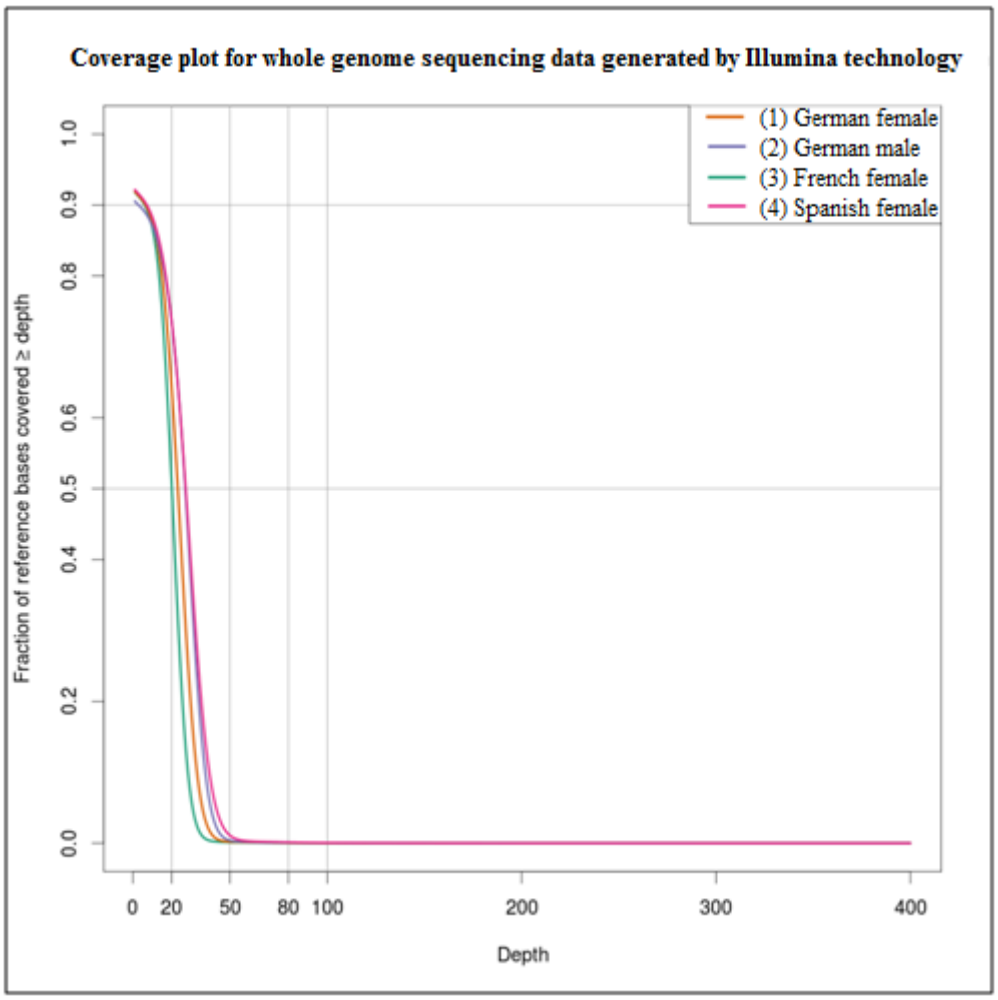

Figure 5-2: Coverage plot for four samples sequenced with Illumina technology.

In addition, four centenarians ((1) German female and (2) male, plus (3) one French female and (4) one Spanish female) were sequenced using the Illumina technology, where the library generation was prepared for paired-end sequencing using the 'PE-102-1001-paired-end sequencing sample prep kit'. The coverage for all four genomes sequenced on Illumina technology is shown in Figure 5-2. More than 90% of the genome has been covered by at least one read. An average coverage of 30x was generated and 80% of reads are covered at 20x, which is again comparable to the coverage attained by other whole genome sequencing studies (Ratan et al. 2013; Rieber et al. 2013).

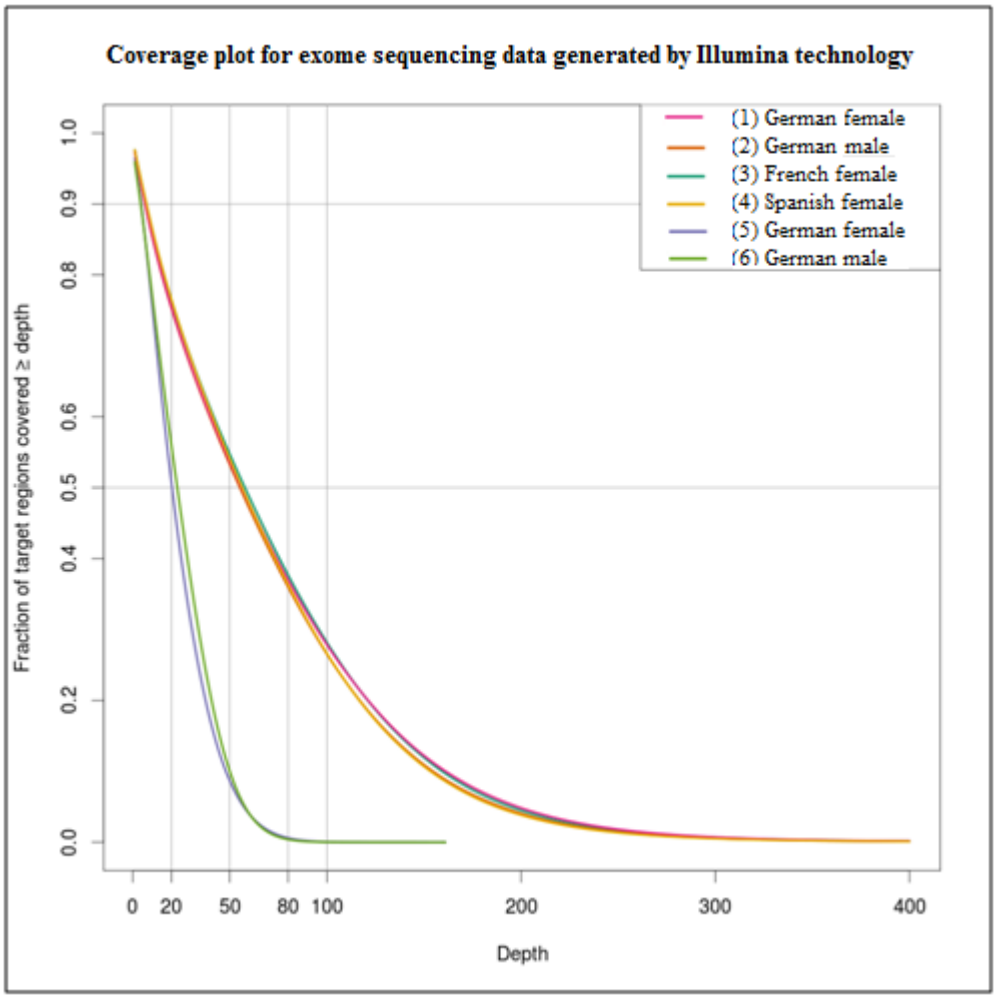

Figure 5-3: Coverage plot for six exomes sequenced with Illumina technology.

Altogether, six centenarians ((2),(6) two German females; (1), (5) two German males; (3) one French female; (4) one Spanish female) were exome sequenced on the Illumina GA machine using Agilent SureSelect and NimbleGen 2.1M target enrichment kit. More than 90% of the genome has been covered by at least one read. An average coverage of 60x was generated and at least 68% of reads were covered at 20x with Agilent SureSelect. Exome sequencing using NimbleGen 2.1M target gave an average coverage of 30x with 60% of the reads covered at 20x, as shown in Figure 5-3. These results are also comparable to other published studies (O'Rawe et al. 2013; Wang et al. 2013).

The coverage exhibited by Illumina technology (Figure 5-2 and 5-3) is observed to be more uniform and displays the least sample-to-sample variation among all individuals when compared to centenarians sequenced with the SOLiD technology (Figure 5-1). For the whole genome sequencing, the Illumina technology covered 80% of the bases at 20x, whereas for the SOLiD technology 60% of the bases were covered. This indicates the shortcomings of the SOLiD

technology in comparison to Illumina technology, by leaving considerable fraction of the genome uncovered. Earlier studies have indicated that less uniform coverage shows a necessity for a higher overall sequencing depth to cover a similar fraction of the genome (Lam et al. 2012; Rieber et al. 2013). Overall, the coverage obtained for both SOLiD and Illumina technologies is comparable to other published studies (Venter et al. 2001; Ratan et al. 2013; Sims et al. 2014). It has also been suggested that integrating sequencing data from different platforms offers the potential to combine the strengths of different technologies and reduce false-positive variants (Rieber et al. 2013).

Identifying a reliable list of SNVs is important when analysing NGS data; hence, various articles have suggested sequencing the same samples, using at least two separate sequencing platforms (Lam et al. 2012; O'Rawe et al. 2013; Ratan et al. 2013). In our dataset, we have three individuals ((1) German female and (2) male, and (3) French female) sequenced for the whole genome and exome with SOLiD and Illumina technology and one individual ((4) Spanish female) sequenced for whole genome and exome with the Illumina technology. False positive variants arising from a single sequencing technology lead to a poor quality of SNVs by the inclusion of non-existing genetic variants, mainly due to issues relating to coverage and SNV quality (Ratan et al. 2013). The best approach for comprehensive variant detection is to sequence genomes with both platforms. Alternatively, supplementing whole genome sequencing with exome sequencing can help in obtaining a more comprehensive set of exonic variants, as shown by O'Rawe (O'Rawe et al. 2013). Whole exome sequencing (WES) yields a higher depth of coverage in most exonic regions, whereas whole genome sequencing (WGS) offers a uniform and comprehensive coverage that may cover regions missed by exon capture in the detection of coding variants. As SOLiD and Illumina employ different techniques, it is expected that the overlap of our sequencing data would represent true-positive SNVs among large proportions of putative false-positive calls and sequencing artefacts randomly distributed over the genome. The combination of exome sequencing and WGS on different platforms, together with multiple variant callers, can provide a powerful means to maximise sensitivity and specificity for any personal genome (O'Rawe et al. 2013; Ratan et al. 2013). Therefore, SNVs generated from both technologies were combined and those variants that show an intersection between SOLiD and Illumina (assuming they represent reliable variant calls) were selected for further investigation.

	SOLiD (whole genome)	Illumina (whole genome)	Illumina (whole exome)	Union
Total number of SNVs	6,465,384	7,130,986	69,666	9,041,484
Total number of novel SNVs	1,875,418 (29%)	983,854 (14%)	5,267 (7.5%)	2,675,839 (29.59%)
synonymous-coding SNVs	18,644	18,874	23,704	34,762
missense SNVs	24,296	19,225	21,275	41,100
cancel-start SNVs	50	28	29	71
read-through SNVs	63	49	24	90
nonsense SNVs	614	206	205	805
SNVs in acceptor	178	47	37	219
SNVs in donor	152	154	81	333
SNVs in 5'UTR	10,497	11,112	1,009	15,278
SNVs in 3'UTR	46,791	45,821	1,281	62,270
SNVs in UTR-splice sites	411,845	456,382	430	579,861
SNVs in introns	1,976,671	2,139,273	21,172	2,731,104
unknown/intergenic SNVs	3,975,578	4,439,810	415	5,575,585
SNVs overlapping with Venter genome	733,068	898,380	17,455	910,591
SNVs overlapping with Watson genome	1,019,716	1,124,813	12,849	1,131,262
SNVs overlapping with Yoruban genome	763,891	880,567	15,163	898,407

Table 5-1: **Summary of SNV distribution in all six centenarians**: Summary statistics of SNV detection and annotation for samples sequenced with SOLiD and Illumina technology.

The variant calling was performed with at least two different SNV callers, SAMtools and GATK (plus diBayes for SOLiD sequenced data) and the results obtained were merged for further analysis. The overlap between DiBayes, SAMtools and GATK variant calling was over 85% for each sample sequenced on SOLiD technology. The overlap between SAMtools and GATK variant calling was over 95% for each sample sequenced on Illumina technology. Both tools use a Bayesian approach to call the variants (Yu and Sun 2013). On one hand, SAMtools is considered good for low coverage data as it uses all reads, while on the other hand, GATK drops reads with low mapping quality and produces variants of only high quality. Usually, it might be wise to

choose an overlap among two or more variant calling programs instead of using just one algorithm (Yu and Sun 2013). However, this could result in a high false negative rate, with many true variants being missed; therefore, the results of a combined SNV calling from both technologies were chosen for further analysis. Table 5-1 gives a detailed overview of the detected variants with both technologies.

A total of 9,041,484 SNVs was detected for samples sequenced on SOLiD and Illumina technology, 29% of the detected SNVs were novel or were not reported on the dbSNP 135 database. The heterozygous/homozygous ratio was 1.4 and the transition/transversion ratio calculated for the entire SNV data set is 1.8 which is according to the expected value published (Levy et al. 2007; DePristo et al. 2011). A comparison with other published genomes showed that 10% of the variants were found to be present in Venter (Levy et al. 2007) and 12.5% of the variants were observed in Watson (Wheeler et al. 2008).

5.2 Challenges of next-generation sequencing

Six centenarians were sequenced using two different technologies (SOLiD and Illumina) of next-generation sequencing to generate a large amount of data to identify new genetic variants that contribute to exceptional longevity. WGS offers the most comprehensive and unbiased approach to study sequence variation, ranging from SNVs and small indels to large structural variation and copy number variations (Snyder et al. 2010). Exome sequencing is a comparatively cost-effective approach that captures only the coding regions of DNA with a high depth of coverage. WES has fewer false positives than WGS, and a greater sensitivity due to the higher coverage achieved, but concerns have previously been raised that it misses important information such as non-coding and structural variations (Wang et al. 2013). WGS is advantageous for studying regulatory sequences and copy-number information, but exome sequencing can identify several thousand single nucleotide variants, both common and rare, associated with the studied phenotype (O'Rawe et al. 2013). The cost differential between the two methods has reduced gradually and some researchers are using the combination of the two to get as much information out of their data as possible (O'Rawe et al. 2013; Ratan et al. 2013). However, although the price of sequencing is decreasing, re-sequencing of the variants/genes using conventional sequencing techniques and further follow-ups, such as genotyping experiments increase the cost of the approach. This step is frequently needed for the proper validation of variants (Majewski et al. 2011). In our project, we have not only combined two technologies of SOLiD and Illumina but also the two methods of WGS and

WES. Furthermore, to substantiate our initial findings, we have added case-control association studies by genotyping LLI from five independent populations (Germany, France, Denmark, Italy and USA).

NGS can help in understanding the genetic mechanisms in biological systems but, at the same time, it can also give rise to new challenges, especially with processing, analysing and interpreting the data. Although a number of software packages are constantly being developed for processing data, second-generation sequencing introduces errors at a fairly high rate compared with the traditional Sanger sequencing (de Magalhães et al. 2010). These errors can arise during the PCR amplification step prior to sequencing or during library preparation. Duplicate reads contribute to false positives derived from PCR-associated errors and are therefore routinely removed during analysis (Koboldt et al. 2010). For example, in our pipeline, Picard's MarkDuplicates was implemented to remove PCR-duplicates from sequencing reads, thereby reducing the sequencing error. Analysis of NGS data can be troublesome particularly due to the given short-read length and the huge volume of data. Gaps present in the human reference genome can lead to ambiguity and misalignments in short read sequences, thereby reducing the possibility of identifying 'true' variants. One way to mitigate this issue is using paired-end reads (de Magalhães et al. 2010; Koboldt et al. 2010), which has been achieved in our project with sequences generated by the Illumina technology. Independent base calling algorithms and software tools have been developed to improve base calling accuracy and reduce systematic errors. Base quality tends to deteriorate towards the ends of reads and hence low quality ends may need to be trimmed to improve the overall data quality. This is automatically done when using the BWA aligner to map the sequenced data to the human genome reference.

Data volume represents a major challenge for data transfer, storage, backup, and analysis. Currently, whole genome sequencing analysis, including read alignment to a reference genome, alignment clean-up and variant calling, with a coverage of 30 to 50x, yields more than 100 Gb of data, and 10 to 12 Gb of data space is required for exome sequencing to achieve at least 20x coverage for 80 to 90% of targeted bases (Meldrum et al. 2011; Puckelwartz et al. 2014). Larger sequencers require Linux servers with multiple cores and large amounts of RAM at a significant capital cost with dedicated human resources to maintain computing clusters. Sophisticated hardware has been set up at ICMB to conquer the computational task of analysing WGS and WES. A high-performance compute cluster with over 900 CPU cores, storage infrastructure encompassing over 1.8PB hard disk and 1.5PB tape archive, as well as a number of stand-alone

high-performance computers, are at our disposal. More recently, cloud-based computing has emerged as a solution for the above limitations pertaining to data volume, which has been described in more detail by Thakur et al. (Thakur et al. 2012).

Identifying 'true' variants among thousands of variants detected by implementing efficient filtering strategies is another bioinformatics challenge. Filtering strategies can remove a large fraction of false variant calls that are due to sequencing and alignment errors, but they also tend to remove true rare variants that are important for genetic studies (Peng et al. 2013). Therefore, it is important to cross-check if the variants correspond to NCBI's dbSNP. A high overlap with dbSNP would suggest the reported variants are real polymorphisms. This can be followed by narrowing down to regions of interest (for example, coding regions) and then evaluating the impact of amino acid change of the variants with prediction tools. This can help to filter and prioritize potential functional variants for further analysis. The development of streamlined, highly automated pipelines for data analysis is critical and one of the possible solutions to address some of these issues (D'Antonio et al. 2013). Therefore, in our project, state-of-the-art tools, including BWA, GATK and ANNOVAR, were integrated into a custom automated pipeline for generating, annotating and analysing sequence variants. In order to filter and prioritize variants, eight different prediction tools were implemented to narrow down the variant list for further follow-up with genotyping experiments.

5.3 Methods implemented for selection of variants

SNVs were selected for further genotyping investigation with two different approaches.

Method 1: SNVs that may have a functional impact

In the first approach four centenarians were sequenced for whole genome and exome analysis and it focused on SNVs that might have a functional impact. The SNV frequencies were compared with variants listed in the 1000G and ESP databases. Binomial testing and p-values were calculated based on allelic frequencies, taking into consideration the sample size. Variants that were present in at least two samples with significant pvalues ($p<0.05$) with respect to the 1000G and ESP databases and were found by PhyloP to be conserved were selected for further investigation. The selected variants constituted 116 exonic SNVs, which were genotyped in our German LLI sample and the seven detected significant association signals were typed for replication in independent French and Danish longevity samples.

Method 2: Low-frequency variants with functional impact

The above approach did not yield new validated longevity associated SNVs, therefore we subsequently implemented a new strategy for a more intensive follow-up for the discovery of new longevity influencing variants by selecting low-frequency SNVs. We used the same four genome and exome sequenced centenarians along with two additional exome sequenced German centenarians. Here, variants were selected that showed an intersection between the two technologies (SOLiD and Illumina), as this should include fewer false-positive variants arising from either of the platforms. As explained in section 5.1 and 5.2, many studies recommend the best approach for an accurate variant detection is to combine two or more sequencing platforms (Ratan et al. 2013; Rieber et al. 2013; Sims et al. 2014). Each platform complement one another as SOLiD and Illumina employs different sequencing techniques, and thus increases the specificity and sensitivity of variant detection (Metzker 2010; O'Rawe et al. 2013).

The list was further reduced by retaining low-frequency variants with MAF≤10% compared with the 1000G and ESP databases for European population. The 1000G and ESP databases as commonly used as a references for comparison with the sequencing data to differentiate between common (MAF>10%) and rare variants (MAF<1%). Most of the MAFs for common variants are similar across the diverse ethnic groups but there are many low-frequency variants that differ in MAFs significantly between various ethnic groups, especially between European and Asian ethnicities (Romualdi et al. 2002; Cross et al. 2010). Many researchers recommend to use ethnically appropriate databases to avoid false negative results for SNVs that may be of a lower frequency in one ethnic group but of higher frequency in other groups (Viennas et al. 2012). Therefore, since the sequenced centenarians are from European ancestry, the allele frequencies were compared to 'CEU' population from the 1000 genomes database and 'European-American' from the ESP database.

Low-frequency variants with MAF≤10% were chosen because relevant variants with a higher frequency should have already been detected by the previous longevity GWAS studies (Newman et al. 2010; Malovini et al. 2011; Nebel et al. 2011; Sebastiani et al. 2012). Also, low-frequency variants may go undetected in a GWAS study as the statistical power to detect such variants with MAF≤10% is much lower. Many studies suggests that variants that influence longevity are likely to be low-frequency variants with large effects sizes and may also contribute to the missing heritability (Vaupel 2010; Newman and Murabito 2013; Garagnani et al. 2014; Zuk et al. 2014) as those low-frequency variants are predicted to have functional consequences (Marth et al. 2011;

Casals et al. 2013). Further, since evolutionary theories predict a trade-off between fertility and longevity, it has been suggested that functional rare variants with large effects under natural conditions may reduce reproduction and thus, increase longevity (Kirkwood and Rose 1991; Westendorp and Kirkwood 1998; Kirkwood 2005; Mitteldorf 2010; Vaupel 2010). So far, only two rare potentially functional missense SNVs, Ala37Thr and Arg407His, located in the IGF-1 and IGF1R genes have been reported to be associated with longevity. The rare variants are overrepresented among centenarians compared to controls (Suh et al. 2008). But so far, the results have not been validated in independent populations.

SNVs selected based on longevity genes and pathways list

Known longevity genes and pathways listed in the NetAge database were used as filter masks for variant selection. The NetAge database contains gene sets and miRNA-regulated protein-protein interaction networks for longevity and aging-associated processes. The database consists of over 450 longevity-associated genes, out of which 120 genes are involved in various pathways (Tacutu et al. 2010). Various studies have shown that insulin and mTOR signaling play a key role in lifespan extension in model organisms (Barzilai et al. 2012). Thus, to prioritize our list, we overlaid the variants with genes that were involved primarily in insulin/mTOR signaling. Three variants located in genes involved in the two pathways were selected for genotyping in our German population. A replication experiment was followed in independent Italian and American longevity samples to validate the two significant association signals observed. Here, the SNV rs3208856 replicated with a p-value of 0. 000189 in the American LLI sample, but no association was confirmed in the Italian LLI.

SNVs selected based on GWAS hit regions and prediction tools

Furthermore, the original variant list was overlaid with the GWAS hit regions and also SNVs that appeared more frequently than expected by chance in four or more individuals were selected. Also, the functional effect of selected SNVs was determined by evaluating the variants with eight prediction tools: Grantham (Grantham, 1974); PMut (Ferrer-Costa, et al., 2005); Screening for Non-acceptable Polymorphisms (SNAP) (Bromberg and Rost, 2007); Sorting Intolerant From Tolerant (SIFT) (Kumar, et al., 2009); SNPs&GO (Calabrese, et al., 2009); MutPred (Li, et al., 2009); Polymorphism Phenotyping (Polyphen-2) (Adzhubei, et al., 2010); and PhyloP (Pollard, et al., 2010). With the use of exome and genome sequencing, prioritizing and interpreting candidate variants within a biological context to the studied phenotype remains a challenge. Longevity, being a complex phenotype, is presumably influenced by a combination of many small-effect

variants located in different genes, thus affecting broader functional networks (Yashin et al. 2010; Brooks-Wilson 2013; Deelen et al. 2013). With the help of different prediction tools that classify variants as functional or neutral, effective low-frequency variants could be selected and prioritized. The relevance of prediction tools to choose the SNVs of interest has been described precisely by Carolin Knecht from the IMIS (Knecht and Krawczak, 2013), who was actively involved in this project. SNVs located in the coding region are often non-synonymous, changing a single amino acid in the encoded protein sequence. For such SNVs, biochemical and physical properties as well as information about functional sites and structure of the protein can be used for the prediction of their potential impact. To study the possibility of a particular genetic variant having a functional consequence, many free online software tools have been developed. But, as described before (Thusberg, et al., 2011), the most favourable choice of a tool depends on the study and there is no 'best' tool available. Different tools implement different algorithms to make predictions regarding functionality of mutated proteins; therefore, the basic idea behind the implementation of all the eight different tools was to take advantage of their possible complementary performance at classifying functionally relevant SNVs.

The tools that predicted an effect for almost all 2,888 coding SNVs were Grantham score and PhyloP. The Grantham matrix predicts the effect of the variant based on differences in physio-chemical properties between amino acids, thus establishing a clear relationship between the severity of amino acid replacement and the likelihood of clinical observation (Grantham 1974). PhyloP scores measure the evolutionary nucleotide conservation at individual alignment sites. Scores for each coding variant were extracted from the 'phyloP46wayAll' table from the UCSC Table Browser (Pollard et al. 2010; Karolchik et al. 2014). Approximately, 96 to 97% of the selected variants could be predicted with MutPred, SNAP and PMut. MutPred is a Random Forest-based classification method that estimates effect of the amino acid substitution based on properties relating to protein structure, function and evolution. SNAP and PMut are based on a neural network that uses *in-silico* derived protein information (e.g. secondary structure, conservation) in order to make predictions for a missense variant (Ferrer-Costa et al. 2005; Bromberg and Rost 2007). The remaining tools, SIFT, PolyPhen2 and SNPs&GO predicted 70 to 90% of the selected variants. SIFT follows the principle of predicting the potential effect of a non-synonymous variant based on sequence similarity using mathematical operations. Predictions rely on the assumption that mutations in evolutionary conserved regions are more likely to affect protein function (Kumar et al. 2009; Knecht and Krawczak 2013). PolyPhen2 implements the naive Bayesian classifier and uses a blend of sequence and structure based attributes to predict the

effect of mutation (Adzhubei et al. 2010). SNPs&GO is based on the Support Vector Machine (SVM) method that uses various types of structural and functional annotation, such as protein sequence, evolutionary information and functions encoded in gene ontology terms, to predict whether a given mutation can be classified as deleterious or neutral (Calabrese et al. 2009). The combination of all the eight different tools should allow the best possible identification of functionally relevant SNVs from an extensive whole genome data set. Apart from prediction tools, variants that appeared more frequently than expected by chance in all six centenarians were also selected. Applying these criteria, 48 SNVs were chosen for direct genotyping in our German longevity sample. Based on power calculations, the top-ranking SNV (rs35761929) that was significantly associated in the German LLI sample (P_{CCA}=3.7e-08, OR=1.712) was further investigated in a replication experiment in an independent Danish longevity population, but could not confirm the previous observed association signal.

5.4 Potential influencing factors for association studies

The most common potential influencing factors for association studies in longevity research are population stratification, population-specificity, phenotype heterogeneity and sample size and statistical power, which might be the reason for lack of replication of significant associations in most longevity studies so far.

5.4.1 Population stratification

Population stratification, the most cited reason for non-replication of genetic association results, is the mixture of individuals from heterogeneous genetic backgrounds (Cardon and Palmer 2003). When cases and controls have different allele frequencies due to events attributable to gene flow between two different populations; or if their frequencies differ due to a demographic expansion into a scarcely populated environment, leading to a partial admixture with indigenous populations, genetic drift or differential selection, a study is said to have population stratification (Cavalli-Sforza and Piazza 1993). Unrecognized population stratification can lead to both false-positive and false-negative findings and can obscure the true association signals if not appropriately corrected (Li et al. 2010). Concerns about population stratification can be addressed by matching the control individuals as close as possible to the LLI in terms of their age (born only one or two generations apart), ethnicity, geographic origin and environmental factors, to avoid false-positive findings (Bloss et al. 2011). In addition, a major problem in selecting the control individuals is that there is a possibility for some control individuals to become LLI themselves. It

is also recommended not to choose a very young control group, because instead of an effect on longevity, the frequency difference may reflect a change in population structure over time (Nebel and Schreiber 2005). Further, inconsistent replication can partly be attributed to genetic stratification among LLI due to cohort differences in survival probability (Nygaard et al. 2014). A recently published report by Nygaard *et al.* compared the cohort specificity of variants in the *APOE* and *FOXO3A* gene at age 95+ and 100+ in 2,712 individuals from the genetically homogeneous Danish birth cohorts (1895–96, 1905, 1910–11, and 1915) and showed that there is a decrease in the allele frequencies of the investigated variants in more recent birth cohorts. The results of this study suggest that birth year and population-dependent differences in selection pressure may also be a part of the explanation for the general lack of replication. As the genetic variations related to longevity are currently expected to be rare and/or have small effects, even modest cohort effects could, when unaccounted for, confound results and leave true associations undiscovered. The possibility of population stratification in our German longevity sample is unlikely, as previous association findings for genetic longevity research have been identified and validated (Nebel et al. 2005; Flachsbart et al. 2009; Nebel et al. 2009) in our German longevity sample. Further, in Germany, the genetic differentiation in population structure is considered to be very low (Steffens et al. 2006). The younger controls recruited were chosen to match the LLI as closely as possible in terms of ancestry, gender, and geographical origin within the country, thus minimizing any systematic genetic differences between the samples that might arise because of very low levels of undetected population structure (Nebel et al. 2005). According to the Human Mortality Database, the chances of a 60 year old female becoming a LLI is 1.5% and for a 75 year old, it is 1.8%. Hence, we can estimate, out of the 1,104 unrelated younger controls; approximately 18 individuals may become LLI themselves, which is a statistically negligible proportion. The replication samples from France, Denmark, Italy and USA were also matched for gender and geographical origin with healthy unrelated younger controls and tested for population stratification (Blanché et al. 2001; Geesaman et al. 2003, Soerensen et al. 2010, Anselmi et al. 2009, Boyden and Kunkel 2010).

In the American longevity sample, one of the two pathway SNVs, rs3208856, confirmed the association signal with an allelic p-value of 0.000189 (OR=2.656). However, since the U.S. Caucasians comprise immigrants from various European countries, the American LLIs represent an admixed population. Hence, the positive replication might be a false-positive result caused by population stratification as it was shown previously for the *MTTP* gene (Puca et al. 2001; Nebel et al. 2005). However, it has been reported that the American longevity sample has been corrected

for population structure (Geesaman et al. 2003; Boyden and Kunkel 2010). Furthermore, the control group in this population is very young (0 to 35 years), which can often lead to false-positive findings (i.e. instead of an effect on longevity, the frequency difference may reflect a change in population structure due to recent immigration) (Nebel et al. 2005). Therefore, even though our initial finding for one SNV, rs3208856, was confirmed with a positive replication in the American population, it maybe due to chance or other influencing variables.

5.4.2 Population-specificity

Another reason for the failed replication might be attributed to population-specific effects, as longevity in different populations is likely to be influenced by varying sets of interacting genetic and environmental factors (Caliebe et al. 2010). Gene variants found to be associated with human longevity in one population rarely replicate in other populations. For example, in 2010 a study investigated the polymorphism rs1333049 associated with coronary artery disease in Northern Italians and showed the frequency of the C allele of rs1333049 was significantly lower in centenarians compared to young controls (Emanuele et al. 2010). Similar results was observed in another Southern European (Mediterranean) cohort in Spain in 2014, however the findings were not replicated in the Japanese, a population of different ethnic and geographic origin (Pinós et al. 2014).

The difficulty in replicating the observed association signals with human longevity can be noticed in our study. Although the German and Danish populations are quite close in sample size, age range and ethnicity, no association was confirmed for variants selected from both approaches. The importance of potentially functional low-frequency variants has emerged recently (Cirulli and Goldstein 2010) and they are mostly population-specific. In another study, low-frequency variants between 14 populations from the 1000 Genomes Project Phase 1 data were compared using statistical methods. The results showed significant differences in low frequency variants across these 14 populations and additionally, populations that were closely related also showed evident differences (Moore et al. 2013). This indicates that it may be even more difficult to replicate low-frequency variant signals in different populations. As our second approach focused on low-frequency variants, the lack of replication in the Danish and Italian longevity sample might be due to the low MAF of the variant (0.6% to 3%) and due to the environmental and geographical differences (Lescai et al. 2009).

5.4.3 Phenotype heterogeneity

Further, the lack of consistent findings may also be explained by phenotype heterogeneity among LLI. As life expectancy has improved over the past two centuries, the probability to survive to extreme ages in developed countries has increased by 50-100% per decade (Oeppen and Vaupel 2002; Vaupel 2010). Some centenarians might be considered 'phenocopies', i.e. individuals who display the same phenotype, but have attained extreme survival by taking advantage of different environment contributions thus, diluting the genetic component of survival to ages above 85 years (De Benedictis and Franceschi 2006; Deelen et al. 2014).

A study by Perls and co-workers on three different groups of centenarians (100–104 years), semisupercentenarians (105–109 years) and supercentenaians (110–119 years) showed a progressive delay in the onset of age-related diseases (e.g. cancer, cardiovascular diseases, dementia and stroke) with increasing age. It was observed that the frequency of survivors (LLI diagnosed with age-associated disease before the age of 80) decreases and the frequency of escapers (LLI who celebrated their 100th birthday without the diagnosis of the age-associated diseases investigated) increases with age. So 8% of supercentenarians were survivors and 69% escapers, respectively, compared with 12% and 56% in semisupercentenarians, and 17% and 30% in centenarians (Andersen et al. 2012). The results show that the phenotype of centenarians can still be very heterogeneous in contrast to supercentenarians, who exhibit more compression of morbidity and disability. It was further reported that for all different groups investigated, males were observed to be healthier than females in terms of cognitive and physical functional status. Therefore, choosing centenarians and supercentenarians as old as possible with accurate geriatric assessment as ‘cases’ maybe more useful for discovering potential genetic variants that influence exceptional longevity.

5.4.4 Sample sizes and statistical power

Discrepancy among all the association results might be due to the different sample sizes used and an overall lack of power for the investigated phenotype. Small sample sizes and over-interpretation of marginal results lead to failure to replicate the initial association signals. In the recently published longevity GWAS meta-analysis, the discovery-phase that consisted of 7,729 cases (above 85 years) and 16,121 controls (below 65 years) showed a genome-wide significant association with human longevity only at the well-known *TOMM40*/*APOE*/*APOC1* locus. However, an additional genome-wide significant locus (rs2149954 on chromosome 5q33.3) was observed subsequently in the joint analysis of the discovery and replication phase comprising

12,736 cases (above 90 years) and 76,268 controls (below 65 years) (Deelen et al. 2014). Hence, even the investment of approximately 90,000 samples did not yield many new insights in the genetic basis of human longevity.

In our first approach, none of the seven SNVs that showed a significant association in the German longevity sample could be replicated in the French longevity sample even though the sample had a power of 80% to replicate the observed association of the top-ranking SNV, rs10927851 (LLI: P_{CCA}=0.002, OR=0.80). According to power calculations, a sample size of 2,000 cases (with a case-control ratio of 1 and assuming OR=1.2) would have been required to replicate the observed association of the top-ranking SNV, rs10927851 (LLI: P_{CCA}=0.002, OR=0.80) with a power of 80%. Hence, to increase statistical efficiency, a meta-analysis was performed by combining the French and Danish longevity samples to give 2,179 cases and 2,594 younger controls. The statistical power of this meta-analysis to replicate the detected association signals was 85%, but yet the genetic association signals observed in Germans could not be confirmed. In the second approach as well, the top-ranking SNV rs35761929 (LLI: P_{CCA}=3.7e-08, OR=1.7) that had a power of 89% to replicate the observed association in the Danish longevity sample of 910 cases and 760 controls did not yield a positive replication result. Furthermore, two SNVs (rs3208856 and rs146426104) located in genes involved in insulin/mTOR signaling were followed-up by a replication experiment in the Italian and American longevity populations. The Italian longevity sample had a power of only 18% to replicate the observed findings with 489 cases and 480 controls and hence, no significant association was observed in the Italian population. At least, a sample size of 3,500 individuals (with a case-control ratio of 1) would have been required to replicate the observed association of the top-ranking SNV, rs3208856 with a power of 80%. To further clarify the association signals observed for all 12 SNVs in our German longevity samples from both the approaches, the discovery-phase of the recent longevity GWAS meta-analysis data comprising over 20,000 individuals was used as a replication sample (Deelen et al. 2014). Here, only one SNV (rs3208856) showed a nominal significant p-value of 0.026 as shown in Table 4-23. Though the allele frequency difference between cases and controls was negligible, the distribution followed the same direction as observed in the German and American longevity sample, with a small increase of the minor allele in the LLI compared with the younger controls. The discovery-phase comprises of data originating from seven European populations (Deelen et al. 2014). However, it has been reported that true association signals may be concealed when combining data from populations with different lifestyles and genetic backgrounds, even if well-matched for ethnicity (Brooks-Wilson 2013).

Tan and co-workers used the Danish life tables and simulations to assess the power for different sample sizes of centenarians; their results show that small samples of centenarians or even supercentenarians (several hundred) provide power to detect only common alleles with large effects and, to detect variants with small effects, large samples of centenarians (more than 1,000) would be needed (Tan et al. 2008). Therefore, selecting only centenarians instead of LLI for 'cases' as shown by Tan *et al*., might be more likely to increase power for detection of genetic variants (Tan et al. 2008; Bloss et al. 2011).

5.5 Summary of findings

5.5.1 Study findings

In the presented study, whole genome and exome sequences of six centenarians by integrating two different technologies (SOLiD and Illumina) were generated and in total 167 SNVs were selected implementing two different approaches for further investigation. Relevant association signals were followed-up in independent longevity samples from France, Italy, USA and Denmark. No significant replication signal was observed for most of our initial results, but the analysis in the American longevity population supported one of our findings (rs3208856: P_{CCA}=0.000189, OR=2.656). It was further supported for association in the discovery-phase meta-analysis data of cases aged ≥85 years (Deelen et al. 2014) with a p-value of 0.026. The SNV rs3208856 (C/T) is a missense variant (p.His405Tyr) located on the *CBLC* gene (Casitas B-lineage Lymphoma Proto-Oncogene, E3 Ubiquitin Protein Ligase C) that is involved in the insulin pathway. *CBLC* plays an important role in the regulation of growth, development, metabolism, and survival. Studies have reported that c-CBL may promote the ubiquitylation of both insulin and IGF1 receptors (Sehat et al. 2008). Molero *et al.* studied mice that lack the *CBLC* gene and observed that this led to reduced adiposity, presumably through increased energy expenditure, thus improving peripheral insulin sensitivity (Molero et al. 2004). Yu *et al*. showed that Drosophila Cbl (dCbl) regulates longevity and carbohydrate metabolism through down regulating the production of Drosophila insulin-like peptides (dILPs) in the brain (Yu et al. 2012). *CBLC* can be a promising candidate but needs further investigation and confirmation in additional larger longevity samples.

Method 1: SNVs that may have a functional impact

Out of 116 SNVs selected for genotyping, the seven SNVs that showed a significant association in our German longevity sample were located in seven different genes: Filamin Binding LIM Protein 1 (*FBLIM1*); Poly (ADP-Ribose) Polymerase 2 (*PARP2*); NLR Family, Apoptosis

Inhibitory Protein (*NAIP*); Propionyl CoA Carboxylase Alpha Polypeptide (*PCCA*); Pleckstrin Homology Domain Containing-Family G (*PLEKHG1*); Proteoglycan-3 (*PRG3*); and Tankyrase-1 Binding Protein-1 (*TNKS1BP1*). *FBLIM1* (rs10927851) is an important component of the cell–matrix adhesions implicated in cell motility, growth and survival, and is a mortality risk gene involved in Alzheimer's disease and skin atrophy (Rebhan et al. 1997; Tacutu et al. 2010). *PARP2* (rs3093921) is an active player in base excision repair, and interacts with *PARP1* and *XRCC1* to synthesize ADP-ribose polymers. *PARP-1* plays an important role in various aging-related processes such as DNA repair, apoptosis, and inflammation (Schreiber et al. 2002). It has been shown in a study that maximal oligonucleotide-stimulated poly-(ADP-ribosyl)-ation is significantly higher in permeabilized lymphoblastoid cell lines from centenarians compared with younger controls, but follow-up studies have failed to show any association of *PARP* and human longevity (De Benedictis et al. 1998; Cottet et al. 2000). *NAIP* (rs61757629) is involved in apoptosis signaling pathways and also plays a role in neurodegenerative diseases such as spinal muscular atrophy (SMA) (Rebhan et al. 1997; Tacutu et al. 2010). Studies have shown that absence of *NAIP*, as an apoptotic suppressor, may modulate cell death or survival (Akutsu et al. 2002). *PCCA* (rs35719359) is primarily involved in metabolism of lipids and lipoproteins; mutations in this gene lead to propionic acidemia, an autosomal recessive disorder (Ugarte et al. 1999). Genetic variations in the *PLEKHG1* (rs17054318) protein signal transduction are associated with panic disorders (Rebhan et al. 1997). *PRG3* (rs34108746) localizes in the cytoplasm and induces apoptosis; it is also associated with Crohn's disease (Ohiro et al. 2002). *TNKS1BP1 (*rs78489201) is involved in nucleotide metabolism and is associated with prostate cancer and prostatitis (Rebhan et al. 1997).

Method 2: Low-frequency variants with functional impact

SNVs selected based on longevity genes and pathways

Various studies have shown that the insulin/IGF-1 pathway is highly conserved and regulates lifespan in organisms ranging from invertebrates to mammals (van Heemst et al. 2005; Barzilai et al. 2012). Two SNVs that showed a significant association signal in the German centenarian subgroup were located in the *CBLC* (rs3208856) and *ACACB* gene (rs146426104). The association signal observed for *CBLC* (rs3208856) was also confirmed in the American longevity sample. Acetyl Coenzyme A carboxylase β (*ACACB*) is involved in the regulation of metabolism and genetic variation in the *ACACB* gene is associated with obesity and diabetes (Riancho et al. 2011). Other studies have also shown that continuous fatty acid oxidation in *ACACB* knock-out

mice increases insulin sensitivity, thereby concluding that common variants within the *ACACB* locus appear to regulate adipose gene expression in humans (Ma et al. 2011).

SNVs selected based on GWAS hit regions and prediction tools

Here, out of 48 SNVs selected for genotyping, three SNVs showed a significant association in both the whole German sample and the centenarian subset. The three SNVs were located in *JAG1* (rs35761929), *ZNF750* (rs35653278) and *MICALCL* (rs34898047). The Jagged-1 (*JAG1)* gene encodes a cell surface protein and belongs to the Delta/Serrate domain (DSL) family (Rebhan et al. 1997). Recent studies have reported the involvement of the *JAG1* gene in bone formation and that activation of the *JAG1* gene is associated with increased bone mineral deposition (Kung et al. 2010). The Okinawa centenarian study has shown that the long-lived Okinawans have about 20% fewer hip fractures than the mainland Japanese population. The mainland Japanese begin to lose significantly more calcium from their bones than the Okinawans, suggesting that the Okinawans preserve their bone density at healthy levels for longer periods of time than other Japanese (Suzuki et al. 1995). Zinc Finger Protein-750 (*ZNF750)* encodes a protein with a nuclear localization site and a C2H2 zinc finger domain and is involved in cell differentiation. *ZNF750* has previously been reported to be associated with autosomal dominant forms of psoriasis or psoriasiform dermatitis and may serve an important function in keratinocyte differentiation or immune response in the skin. It has been shown that insufficient levels of *ZNF750* could lead to a downstream effect that fails to repress a stimulated immune response in psoriasis (Birnbaum et al. 2011). Molecule Interacting with CasL C-terminal like *(MICALCL)* is a protein-coding gene involved in the intracellular signal transduction pathway. It participates in the control of cytoskeleton dynamics and may establish a direct link between cell oxido-reduction metabolism and cytoskeleton rearrangements (Terman et al. 2002).

Although, most of the variants, apart from rs3208856, could not be replicated in independent longevity samples (France, Denmark, Italy and USA), they may be interesting enough to warrant further investigation in large-scale meta-analysis for a more stringent phenotype (e.g. 100 years and older).

5.5.2 Genetic profiles of centenarians

Variants known to influence longevity

In addition to the above analysis, the sequencing data of four centenarians were used to observe their genetic profiles by investigating those genes that have a significant impact on exceptional

survival. Most of the variants investigated below are intronic variants and therefore, only centenarians that have been genome-sequenced were used. Apart from *APOE* (Schächter et al. 1994; Blanché et al. 2001; Deelen et al. 2011; Nebel et al. 2011) and *FOXO3A* (Willcox et al. 2008; Anselmi et al. 2009; Flachsbart et al. 2009; Li et al. 2009c; Pawlikowska et al. 2009; Soerensen et al. 2010) that have been validated repeatedly as longevity influencing genes, we also included the recent GWAS-identified longevity locus on chromosome 5q33.3 that may influence survival in the general European population (Deelen et al. 2014). Using the LongevityMap database (Budovsky et al. 2013), five intronic variants in APOE and FOXO3A that are associated significantly with human longevity were selected (Table 5-2).

Chromosome/ Gene	dbSNP ID	Alleles	Function	×Minor allele	¥MAF in 1000G	*1	*2	*3	*4
chr 19 *TOMM40*/*APOE* (Deelen et al. 2011; Nebel et al. 2011)	rs2075650	A/G	Intron	G ↓	0.16	-	AG	-	-
chr 6 *FOXO3A* (Willcox et al. 2008; Anselmi et al. 2009; Flachsbart et al. 2009; Soerensen et al. 2010)	rs2802288	A/G	Intron	A ↑	0.33	-	GG	AG	AG
	rs7762395	A/G	Intron	A ↑	0.16	GA	-	-	GA
	rs9400239	C/T	Intron	T ↑	0.24	-	CC	CC	CT
	rs3800231	A/G	Intron	A ↑	0.24	-	AG	GG	GG
5q33.3 (closest gene *EBF1*) (Deelen et al. 2014)	rs2149954	C/T	Intron	T ↑	0.34	CT	-	-	CT

×Minor allele associated with longevity that is overrepresented (↑) or underrepresented (↓) in long-lived individuals
¥Minor allele frequencies reported in the 1000 Genomes European population
* (1) German female, (2) German male, (3) French female, (4) Spanish female
- No variants detected in the sequencing data

Table 5-2: List of variants from literature that were significantly linked to exceptional human longevity

Many studies have shown the *APOE* ε4 allele predisposes to both Alzheimer's and cardiovascular diseases and is associated with increased mortality (Corder et al. 1993; Schächter et al. 1994). Among LLI, the probability of carrying the *APOE* ε4 allele is lower and the probability of carrying *APOE* ε2 allele is higher (Schächter et al. 1994; Christensen et al. 2006; Bennet et al. 2007). Further, since the variant, rs2075650, is in linkage disequilibrium with the *APOE*-defining alleles (rs429358 and rs7412), studies have shown the likelihood of carrying the 'G' allele in rs2075650 of *TOMM40* is lower in LLI (Deelen et al. 2011; Nebel et al. 2011). However, out of the four centenarians, it could be evaluated in only one, where the German male was heterozygous for rs2075650.

FOXO3A is associated with insulin signaling pathway and is known to play an important role in apoptosis, stress resistance and metabolism (Carter and Brunet 2007). Many studies have not only shown that variations in *FOXO3A* are associated with longevity, but the association is significantly stronger in centenarians than nonagenarians (Willcox et al. 2008; Anselmi et al. 2009; Flachsbart et al. 2009; Li et al. 2009c; Pawlikowska et al. 2009; Soerensen et al. 2010). All four centenarians carry a cluster of variants for *FOXO3A*, but we investigated only those variants that are reported in literature to be significantly associated with the longevity phenotype in more than one study. The variant rs2802288 is present in three centenarians and only two of them (French and Spanish females) carried the effective minor allele 'A'. For rs7762395, both German and Spanish females carry the variant with the effective allele 'A'. For the other two variants, rs9400239 and rs3800231, only one out of three centenarians carry the effective minor allele 'T' and 'A'.

The recently identified longevity GWAS locus (rs2149954) on chromosome 5q33.3 that may influence extreme survival was also investigated in our centenarians. The GWAS-study by Deelen *et al.* reported a higher frequency of the minor allele 'T' in LLI and that rs2149954 influences longevity by a decrease in the risk of mortality due to stroke. From Table 5-2, it is observed that only the German and Spanish female carry the intronic variant with the effective allele.

The above analysis, although limited, suggests that to achieve extreme survival all genetic variants associated with longevity reported to-date might not necessarily be present in all LLIs. A similar observation was reported by Sebastiani *et al.* (2012), where the female and male supercentenarian genome sequenced did not carry the effective minor alleles for most of the

variants that are reported in the literature and are suggested to influence longevity. Hence, rare variants or common variants with rather small effects that are yet to be discovered may play a vital role in the genetic variation of human longevity.

Disease-associated variants in centenarians

As mentioned earlier, various studies have shown that LLI carried a similar number of disease variants compared to the general population (Beekman et al. 2010; Sebastiani et al. 2011). Therefore, we wanted to assess the number of disease-associated variants present in our sequencing data of all six centenarians and if it was comparable to younger controls. For this analysis, we chose four random in-house control sequences (kindly provided by Prof. Dr. Andre Franke) that were exome-sequenced at ICMB or at the BGI Institute in China. The exonic variants of the four younger controls were compared with our exome-sequenced centenarians using snpActs (http://snpacts.ikmb.uni-kiel.de/). Since the sample size presented in this study is too small to implement statistical analysis, we followed an analysis similar to that performed by Sebastiani *et al.* (2012), where only the number of disease-associated variants present in the centenarians and controls were reported.

snpActs is linked with the Human Gene Mutation Database (HGMD) and thus, it was possible to filter for known disease-associated variants. All variants that were annotated by a HGMD tag were considered. Further, we compared the number of disease-associated variants for five common age-associated diseases: Alzheimer's disease, cancer, cardiovascular diseases, diabetes and stroke. The keywords used for the search were 'alzheimer', ‘dementia’, 'cancer', 'diabetes', 'cardiovascular diseases', 'heart failure', 'coronary heart disease' and 'stroke'.

Our analysis presented in this section is not intended to be comprehensive and only an initial attempt to have a general view of the disease-associated variants in centenarians. The overall differences observed in Table 5-3 and 5-4 might also be influenced by different sequencing platforms, mapping algorithms and annotation methods implemented for variant calling, rather than changes linked to longevity (Sebastiani et al. 2011). To allow a proportionate comparison, the number of risk alleles detected was normalized to the total number of centenarians and controls used.

	*1	*2	*3	*4	*5	*6	×C-1	×C-2	×C-3	×C-4	Total risk allele count cases	Total risk allele count controls	Disease risk alleles per individual cases-controls
Alzheimer's	7	9	9	9	7	10	8	9	7	7	57	39	10-10
¥het/hom	4/3	5/4	5/4	8/1	4/3	7/3	6/2	8/1	5/2	4/3			
Cancer	65	69	73	74	59	62	57	72	69	58	529	352	88-88
het/hom	43/22	45/24	50/23	56/18	39/20	42/20	37/20	50/22	39/30	34/24			
Cardiovascular diseases	13	15	9	20	11	13	12	12	11	11	105	56	18-14
het/hom	8/5	8/7	7/2	19/1	4/7	11/2	8/4	8/4	10/1	10/1			
Diabetes	17	20	23	19	11	21	18	27	20	20	152	112	25-28
het/hom	12/5	15/5	17/6	9/10	5/6	12/9	12/6	16/11	17/3	13/7			
Stroke	2	1	1	2	1	2	2	1	1	2	11	6	2-2
het/hom	2/0	1/0	1/0	2/0	0/1	1/1	2/0	1/0	1/0	2/0			
Number of exonic variants	26,223	26,790	26,767	27,178	18,456	18,481	28,274	47,033	26,819	23,668	3,823	2,319	637-580
×Number of disease variants	488 (1.9%)	489 (1.8%)	498 (1.9%)	505 (1.9%)	424 (2.3%)	430 (2.3%)	421 (1.5%)	511 (1.1%)	427 (1.6%)	371 (1.6%)			
het/hom	325/163	318/171	322/176	361/144	251/173	268/162	289/132	334/177	280/147	238/133			

* cases: (1) German female, (2) German male, (3) French female, (4) Spanish female, (5) German female, (6) German male
× controls: (C-1) Control 1 male, BGI, (C-2) Control 2 male, ICMB, (C-3) Control 3 male, BGI, (C-4) Control 4 female, BGI
¥ number of heterozygous variants to number of homozygous variants
×other diseases including Alzheimer's disease, asthma, arthritis, bipolar disorder, cancer, cardiovascular diseases, cataract, diabetes, hypertension, muscular dystrophy, Parkinson's, stroke, etc.

Table 5-3: **Number of disease-associated variants** that are involved in major age-related diseases and total number of disease-associated variants that are present in centenarians and controls

	*1	*2	*3	*4	*5	*6	×C-1	×C-2	×C-3	×C-4	Total protective allele count cases	Total protective allele count controls	Protective alleles per individual cases-controls
Alzheimer's	0	0	0	0	0	0	0	2	0	0	0	2	0-2
¥het/hom	0/0	0/0	0/0	0/0	0/0	0/0	0/0	2/0	0/0	0/0			
Cancer	7	5	5	4	8	5	5	6	3	3	44	26	7-7
het/hom	6/1	4/1	3/2	4/0	3/5	4/1	2/3	3/3	2/1	1/2			
Cardiovascular diseases	5	4	6	4	5	5	2	2	2	5	38	13	6-3
het/hom	2/3	3/1	5/1	3/1	4/1	3/2	2/0	1/1	1/1	5/0			
Diabetes	2	4	3	3	2	2	3	3	2	2	26	12	4-3
het/hom	1/1	2/2	1/2	1/2	1/1	0/2	3/0	2/1	2/0	1/1			
Stroke	1	0	1	1	0	0	0	0	0	0	5	0	1-0
het/hom	1/0	0/0	0/1	1/0	0/0	0/0	0/0	0/0	0/0	0/0			
Number of protective variants	34	41	36	34	31	26	24	38	23	25	266	145	44-36
het/hom	24/10	29/12	25/11	26/8	20/11	14/12	17/7	24/14	16/7	18/7			

* cases: (1) German female, (2) German male, (3) French female, (4) Spanish female, (5) German female, (6) German male
× controls: (C-1) Control 1 male, BGI, (C-2) Control 2 male, ICMB, (C-3) Control 3 male, BGI, (C-4) Control 4 female, BGI
¥ number of heterozygous variants to number of homozygous variants

Table 5-4: **Number of protective alleles** that are involved in major age-related diseases and total number of protective variants that are present in centenarians and controls

Out of the total exonic variants detected, the six centenarians have 2% annotated for the disease-associated variants in the coding region and the controls only 1.5%. In both the groups, more than 60% of the variants were heterozygous (Table 5-3). The centenarians and controls carried a similar number of disease-risk alleles associated with Alzheimer's, cancer (such as leukemia, breast cancer, colon cancer, prostate cancer, lung cancer, etc), cardiovascular diseases, diabetes and stroke. However, when considering the total number of disease variants associated, the centenarians carry a higher number of disease risk alleles than controls. Interestingly, as shown in Table 5-4, centenarians carry slightly more protective variants than controls that may compensate for the damaging effects caused by the disease-associated variants.

The 'longevity-enabling' variants may act to buffer the deleterious effects of genes that cause age-related diseases, which may explain the higher frequency in disease-associated variants observed among centenarians in Table 5-3. The frequencies of the deleterious variants is higher among LLI because the protective variants may allow these disease-related genes to accumulate with extreme lifespan (Bergman et al. 2007). Our above analysis is also supported by other studies, where it was reported that the distribution of risk alleles is similar in LLI and younger controls (Beekman et al. 2010). Further, it was previously hypothesized that supercentenarians may have a lower number of disease-associated variants when compared to centenarians (Andersen et al. 2012) but, in our study, both the supercentenarians (French and Spanish females) carry a similar number of disease associated variants as compared to centenarians and controls. A similar observation was reported by Sebastiani *et al.* (2012), where the two supercentenarians investigated carried a similar number of disease-associated variants compared to younger subjects. The protective variants present in the centenarians are thought to delay the onset or reduce the severity of age-related diseases (Sebastiani and Perls 2012) or as suggested above, rare variants (i.e. variants without a link to a specific disease) may play an important role in influencing extreme survival by targeting the longevity-assurance mechanism (Schächter et al. 1993).

5.6 Conclusion and outlook

Researchers have hypothesized that the LLI are enriched for longevity-associated variants, which not only compensate for the damaging effects of disease variants but also offer protection (Bergman et al. 2007; Sebastiani and Perls 2012). Detection of these longevity-associated variants may help explain why LLI delay or escape age-related diseases. To detect such small effective variants, extremely large sample sizes are required. It is also possible to perform meta-analyses

for identification of candidate genes with modest influence. Some researchers have argued that lifelong exposure to different environmental factors may be one of the main determinants of healthy longevity at older ages (Harris et al. 1992). So far, the number of genetic findings that influence human longevity is limited and the loci that could explain the familial clustering of longevity have not yet been identified (Deelen et al. 2013). Until now, candidate gene studies and GWAS have yielded *APOE* (Schächter et al. 1994; Deelen et al. 2011; Nebel et al. 2011) and *FOXO3A* (Willcox et al. 2008; Anselmi et al. 2009; Flachsbart et al. 2009; Li et al. 2009c; Soerensen et al. 2010) as the only two genes influencing human lifespan that have been confirmed and replicated in various populations. Recently, a GWAS meta-analysis was performed with 7,729 LLI of European descent and 16,121 younger controls, where besides *TOMM40*/*APOE*/*APOC1,* one additional locus, rs2149954 on chromosome 5q33.3 showed genome wide significance (Deelen et al. 2014). This is the first longevity GWAS, which had sufficient power to detect lifespan-regulating loci with relatively small effects (OR = 1.10, P = 1.74×10^{-8}). However, GWAS of complex late-onset diseases, such as Alzheimer's disease, with sample sizes comparable to the described longevity meta-analysis, have identified 11 new susceptibility loci (Lambert et al. 2013). Hence, even larger GWAS (50,000 LLI) may be required to identify additional longevity loci, preferably in the most stringent phenotype, i.e. the oldest old (Deelen et al. 2014). On the other hand, as mentioned before, the association between genetic variants and human longevity can also be population-specific. Hence, meta-analysis studies, where different populations well matched for ethnicity but with different genetic backgrounds are combined, may still obscure true signals (Brooks-Wilson 2013). Also, GWAS lack the sensitivity to identify causal variants (MAF≤0.05) and can explain only a small proportion of the heritability (less than 10%) for complex traits (Schork et al. 2009). One of the new approaches to overcome the shortcomings of GWAS may be sequencing and therefore, today, with the use of recent technological advances, next-generation sequencing (NGS) should act as a powerful tool to identify associations between genetic variants and human longevity (de Magalhães, et al., 2010). Inadequate results pertaining to the discovery of new genetic variants was observed by Sebastiani *et al.* who reported the whole genome sequencing of a male and a female supercentenarian (Sebastiani et al. 2011). However, in our study, by implementing state-of-the-art technologies for sequencing, combined with conventional case-control association studies, sequencing of whole genome and exome of six centenarians, one potentially functional missense variant (rs3208856) was confirmed in a replication experiment in the American longevity sample (rs3208856: P_{CCA}=0.000189, OR=2.656). The distribution of the allele frequency of the variant in both German and American longevity sample points in a similar direction, where the minor allele 'T'

is overrepresented in centenarians as compared to controls. It was further supported for association in the discovery-phase meta-analysis data of cases aged ≥85 years with a nominal p-value of 0.026. The variant on the *CBLC* gene can be a promising candidate that could potentially provide important insights into the genetic and molecular basis of human longevity but needs further investigation and confirmation in additional larger longevity samples. A limitation of our study is the small sample size: six individuals might not be sufficient to identify new genetic variants associated with the complex longevity phenotype. To identify such low-frequency and rare variants that influence human longevity, analysis of the genomes on many more centenarians must be performed (Deelen et al. 2013).

For future work, it may be interesting to identify rare variants by sequencing a large number of supercentenarians as they are more phenotypically homogeneous compared to centenarians, therefore have a higher genetic contribution to exceptional longevity (Hitt et al. 1999; Robine and Vaupel 2001) and hence, should also be associated with the increased ability to identify genetic variants that influence longevity (Schoenhofen et al. 2006; Andersen et al. 2012). This may be possible in the near future due to the continous rapid decrease in the sequencing costs (Sims et al. 2014). It has been shown in our study (Table 5-3) and also previously that centenarians carry a similar number of disease-associated variants compared to younger controls (Beekman et al. 2010; Sebastiani et al. 2011). Therefore, the ideal subjects for the investigation of genetic variants would be ‘escapers’ or individuals who attained their 100th year of life without the diagnosis of common age-associated illnesses such as heart disease, stroke, diabetes, cancer or Alzheimer's disease (Evert et al. 2003). Studies have shown clustering of longevity within families (Schoenmaker et al. 2006; Newman et al. 2011; Sebastiani et al. 2013), therefore, LLI from families with a history of longevity may prove to be the most informative subjects to sequence using NGS as they are considered to be more enriched for familial and genetic effects on longevity (Deelen et al. 2013). Another option may be to focus on male centenarians, as they are less heterogeneous, and tend to have significantly better cognition and physical function compared to their female counterparts (Franceschi et al. 2000; Franceschi and Bonafè 2003; Schoenhofen et al. 2006). Furthermore, the role of genes that may influence human lifespan requires thorough testing through functional studies before they are considered ‘longevity-enabling’ genes. An integration of alternate approaches such as the investigation of structural variation (Kuningas et al. 2011), gene–gene interactions (Tan et al. 2002), transcriptome studies (Passtoors et al. 2012) or epigenetic mechanisms such as miRNA studies (ElSharawy et al. 2012) and methylation patterns (Bell et al. 2012) are likely to improve our understanding of the

interplay of the various genetic and environmental factors that influence human longevity (Christensen et al. 2006; de Magalhães et al. 2009; Sebastiani et al. 2011). This might be achieved in the future by using a systems biology approach that combines and quantifies genetics and omics-based fields and can handle the increasing amount of data generated by new high-throughput technologies, thereby ideally able to provide insight into the complex mechanisms underlying the longevity phenotype (Cevenini et al. 2010).

6 Summary

The genetic contribution to adult human lifespan is ~25-30% and is assumed to be determined by rare variants or common variants with rather small effects. The current hypothesis is that long-lived individuals (LLI) are enriched with longevity-associated variants that may compensate for the damaging effects of disease-associated variants and are thought to be of rather low frequency. In this project, we combined next-generation sequencing with case-control association studies to identify new exonic longevity influencing variants as those are more likely to be functionally relevant due to amino acid substitution. To reach this goal, we performed whole genome and exome sequencing of six centenarians (108-114 years) of European origin using two different technologies (SOLiD and Illumina). A fraction of the detected single nucleotide variants (SNVs) were selected for a follow-up by genotyping based on two different approaches. The first approach focused on SNVs with minor allele frequencies (MAF) 1 to 50%, which resulted in 116 SNVs that were genotyped in our German sample of 1,610 LLI and 1,104 controls. Seven significant association signals were obtained and further investigated for a replication experiment in independent French (1,269 LLI and 1,834 younger controls) and Danish populations (910 LLI and 760 controls), but none of the associations could be confirmed. The second approach was an intensive follow-up by focusing on low-frequency variants (MAF≤10%). Using eight different bioinformatic prediction tools to evaluate the functional impact of SNVs and overlaying the initial SNV list with locations associated with genome-wide association (GWAS) hit regions resulted in 48 variants that were selected for genotyping, where three SNVs showed a significant association signal in the German longevity sample. The top-ranking SNV (P_{CCA}=3.7e-08, OR=1.7) was selected for a replication experiment in the Danish population but could not be confirmed. In addition, longevity genes and pathways from known model organisms were used as filter masks for the variant selection. Two SNVs showed a significant association signal in the German centenarian subset and were investigated in a subsequent replication experiment in an Italian (489 LLI and 480 controls) and an American (352 LLI and 365 controls) study population. No significant association was observed in the Italian population. However, the case-control analysis in the American longevity sample confirmed the association signal for one SNV (P_{CCA}=0.000189, OR=2.65). The SNV is a missense variant located in the *CBLC* gene, which is involved in the insulin pathway and plays an important role in the regulation of growth, development, metabolism, and survival. Hence, this *CBLC* missense SNV can be regarded as another promising longevity candidate but needs further investigation and confirmation in additional larger samples.

7 Zusammenfassung

Der genetische Einfluss auf die Lebensspanne liegt beim erwachsenen Menschen bei etwa 25-30%, und es wird angenommen, dass dieser Beitrag durch seltene oder häufige Varianten mit eher geringen Effekten bestimmt wird. Die aktuelle Hypothese ist, dass bei langlebigen Individuen (LLI) eine Anreicherung langlebigkeitsassoziierter Varianten vorliegt, welche die schädlichen Effekte von krankheitsassoziierten Varianten kompensieren, und dass solche Varianten eine eher niedrige Frequenz aufweisen. In diesem Projekt kombinierten wir die *next generation sequencing* Methode mit Fall-Kontroll-Assoziationsstudien, um neue exonische Langlebigkeits-Varianten zu identifizieren, welche durch Änderungen der Aminosäuresequenz potentiell funktionell relevant sein könnten. Um dieses Ziel zu erreichen, führten wir mit zwei verschiedenen Technologien (SOLiD und Illumina) eine Gesamtgenom- und Exom-Sequenzierung von sechs Hundertjährigen (108–114 Jahre) europäischen Ursprungs durch. Einige der detektierten Einzelbasenvarianten (*single nucleotide variants* = SNVs) wurden anhand von zwei verschiedenen Ansätzen für eine nachfolgende Genotypisierung ausgewählt. Im ersten Ansatz lag der Fokus auf exonischen SNVs mit einer Frequenz des seltenen Alles (*minor allele frequency* = MAF) von 1-50%. Daraus resultierten 116 SNVs, die in unserer deutschen Stichprobe von 1.610 LLI und 1.104 Kontrollen typisiert worden sind. Hier zeigten sich sieben signifikante Assoziationssignale, die in einem folgenden Replikationsexperiment in einer unabhängigen französischen (1.269 LLI und 1.834 jüngere Kontrollen) und dänischen Stichprobe (910 LLI und 760 Kontrollen) untersucht wurden. Allerdings konnte hierbei keine dieser Assoziationen bestätigt werden. Der zweite Ansatz war ein intensives Follow-up zur Entdeckung neuer Langlebigkeits-Varianten mit Fokussierung auf niederfrequenten Varianten (MAF≤10%). Die Verwendung von acht verschiedenen bioinformatischen Vorhersage-Tools zur Bewertung der funktionellen Bedeutung der exonischen SNVs sowie ein Abgleich der SNV-Ausgangsliste mit assoziierten GWAS-Hit-Regionen (Genomweite Assoziationsstudie) resultierte in 48 Varianten, wobei drei SNVs ein signifikantes Assoziationssignal in der deutschen Langlebigkeitsstichprobe zeigten. Der am stärksten assoziierte SNV (P_{CCA}=3,7e-08; OR=1,7) wurde für ein Replikationsexperiment in der dänischen Sammlung ausgewählt, konnte dort aber nicht bestätigt werden. Des Weiteren wurden aus Modellorganismen bekannte Langlebigkeits-Gene und -Signalwege als Filtermaske für die Auswahl der Varianten verwendet. Zwei dieser SNVs zeigten ein signifikantes Assoziationssignal in der deutschen Hundertjährigen-Subpopulation und wurden im Folgenden in einem Replikationsexperiment in einer italienischen (489 LLI und 480 Kontrollen) und einer amerikanischen (352 LLI und 365 Kontrollen) Analysepopulation

untersucht. In der italienischen Population wurde keine signifikante Assoziation beobachtet. Dafür bestätigte aber die Fall-Kontroll-Assoziationsanalyse in der amerikanischen Langlebigkeitsstichprobe das Assoziationssignal für einen SNV (P_{CCA}=0,000189; OR=2,65). Dieser SNV ist eine nicht-synonyme Variante und liegt im *CBLC*-Gen, welches in den Insulin-Signalweg involviert ist und eine wichtige Rolle bei der Regulation von Wachstum, Entwicklung, Metabolismus und für Überlebensvorgänge spielt. Somit kann dieser nicht-synonyme SNV im *CBLC*-Gen als neue vielversprechende Langlebigkeits-Variante betrachtet werden, die jedoch noch eine weitere Untersuchung und Bestätigung in zusätzlichen, größeren Stichproben benötigt.

8 References

Abbott MH, Murphy EA, Bolling DR, Abbey H. 1974. The familial component in longevity. A study of offspring of nonagenarians. II. Preliminary analysis of the completed study. *Johns Hopkins Med J* **134**(1): 1-16.

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA, Consortium GP. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422): 56-65.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* **7**(4): 248-249.

Akutsu T, Nishio H, Sumino K, Takeshima Y, Tsuneishi S, Wada H, Takada S, Matsuo M, Nakamura H. 2002. Molecular genetics of spinal muscular atrophy: contribution of the NAIP gene to clinical severity. *Kobe J Med Sci* **48**(1-2): 25-31.

Albani D, Mazzuco S, Polito L, Batelli S, Biella G, Ongaro F, Gustafson DR, Antuono P, Gajo G, Durante E et al. 2011. Insulin-like growth factor 1 receptor polymorphism rs2229765 and circulating interleukin-6 level affect male longevity in a population-based prospective study (Treviso Longeva--TRELONG). *Aging Male* **14**(4): 257-264.

Andersen SL, Sebastiani P, Dworkis DA, Feldman L, Perls TT. 2012. Health span approximates life span among many supercentenarians: compression of morbidity at the approximate limit of life span. *J Gerontol A Biol Sci Med Sci* **67**(4): 395-405.

Andrews S. 2010. FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

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

Applied Biosystems SOLiD™ 4 System. 2010. Library Preparation Guide. http://tools.lifetechnologies.com/content/sfs/manuals/SOLiD4_Library_Preparation_man.pdf

Atzmon G, Rincon M, Rabizadeh P, Barzilai N. 2005. Biological evidence for inheritance of exceptional longevity. *Mech Ageing Dev* **126**(2): 341-345.

Ayyadevara S, Alla R, Thaden JJ, Shmookler Reis RJ. 2008. Remarkable longevity and stress resistance of nematode PI3K-null mutants. *Aging Cell* **7**(1): 13-22.

Bartke A. 2005. Minireview: role of the growth hormone/insulin-like growth factor system in mammalian aging. *Endocrinology* **146**(9): 3718-3723.

Barzilai N, Atzmon G, Schechter C, Schaefer EJ, Cupples AL, Lipton R, Cheng S, Shuldiner AR. 2003. Unique lipoprotein phenotype and genotype associated with exceptional longevity. *JAMA* **290**(15): 2030-2040.

Barzilai N, Huffman DM, Muzumdar RH, Bartke A. 2012. The critical role of metabolic pathways in aging. *Diabetes* **61**(6): 1315-1322.

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

Beekman M, Nederstigt C, Suchiman HE, Kremer D, van der Breggen R, Lakenberg N,

Alemayehu WG, de Craen AJ, Westendorp RG, Boomsma DI et al. 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.

Bell JT, Tsai PC, Yang TP, Pidsley R, Nisbet J, Glass D, Mangino M, Zhai G, Zhang F, Valdes A et al. 2012. Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet* **8**(4): e1002629.

Bellizzi D, Rose G, Cavalcante P, Covello G, Dato S, De Rango F, Greco V, Maggiolini M, Feraco E, Mari V et al. 2005. A novel VNTR enhancer within the SIRT3 gene, a human homologue of SIR2, is associated with survival at oldest ages. *Genomics* **85**(2): 258-263.

Bennet AM, Di Angelantonio E, Ye Z, Wensley F, Dahlin A, Ahlbom A, Keavney B, Collins R, Wiman B, de Faire U et al. 2007. Association of apolipoprotein E genotypes with lipid levels and coronary risk. *JAMA* **298**(11): 1300-1311.

Bergman A, Atzmon G, Ye K, MacCarthy T, Barzilai N. 2007. Buffering mechanisms in aging: a systems approach toward uncovering the genetic component of aging. *PLoS Comput Biol* **3**(8): e170.

Bethesda M. 2010. Database of Single Nucleotide Polymorphisms (dbSNP) : National Center for Biotechnology Information, National Library of Medicine. (dbSNP Build ID: 135).

Birnbaum RY, Hayashi G, Cohen I, Poon A, Chen H, Lam ET, Kwok PY, Birk OS, Liao W. 2011. Association analysis identifies ZNF750 regulatory variants in psoriasis. *BMC Med Genet* **12**: 167.

Blanché H, Cabanne L, Sahbatou M, Thomas G. 2001. A study of French centenarians: are ACE and APOE associated with longevity? *C R Acad Sci III* **324**(2): 129-135.

Bloss CS, Pawlikowska L, Schork NJ. 2011. Contemporary human genetic strategies in aging research. *Ageing Res Rev* **10**(2): 191-200.

Bonafè M, Barbieri M, Marchegiani F, Olivieri F, Ragno E, Giampieri C, Mugianesi E, Centurelli M, Franceschi C, Paolisso G. 2003. Polymorphic variants of insulin-like growth factor I (IGF-I) receptor and phosphoinositide 3-kinase genes affect IGF-I plasma levels and human longevity: cues for an evolutionarily conserved mechanism of life span control. *J Clin Endocrinol Metab* **88**(7): 3299-3304.

Boyden SE, Kunkel LM. 2010. High-density genomewide linkage analysis of exceptional human longevity identifies multiple novel loci. *PLoS One* **5**(8): e12432.

Bromberg Y, Rost B. 2007. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res* **35**(11): 3823-3835.

Brooks-Wilson AR. 2013. Genetics of healthy aging and longevity. *Hum Genet* **132**(12): 1323-1338.

Budovsky A, Craig T, Wang J, Tacutu R, Csordas A, Lourenço J, Fraifeld VE, de Magalhães JP. 2013. LongevityMap: a database of human genetic variants associated with longevity. *Trends Genet* **29**(10): 559-560.

Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. 2009. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum Mutat* **30**(8): 1237-1244.

Caliebe A, Kleindorp R, Blanché H, Christiansen L, Puca AA, Rea IM, Slagboom E, Flachsbart F, Christensen K, Rimbach G et al. 2010. No or only population-specific

effect of PON1 on human longevity: a comprehensive meta-analysis. *Ageing Res Rev* **9**(3): 238-244.

Cardon LR, Palmer LJ. 2003. Population stratification and spurious allelic association. *Lancet* **361**(9357): 598-604.

Carter ME, Brunet A. 2007. FOXO transcription factors. *Curr Biol* **17**(4): R113-114.

Casals F, Hodgkinson A, Hussin J, Idaghdour Y, Bruat V, de Maillard T, Grenier JC, Gbeha E, Hamdan FF, Girard S et al. 2013. Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLoS Genet* **9**(9): e1003815.

Cavalli-Sforza LL, Piazza A. 1993. Human genomic diversity in Europe: a summary of recent research and prospects for the future. *Eur J Hum Genet* **1**(1): 3-18.

Cevenini E, Bellavista E, Tieri P, Castellani G, Lescai F, Francesconi M, Mishto M, Santoro A, Valensin S, Salvioli S et al. 2010. Systems biology and longevity: an emerging approach to identify innovative anti-aging targets and strategies. *Curr Pharm Des* **16**(7): 802-813.

Chan Y, Lim ET, Sandholm N, Wang SR, McKnight AJ, Ripke S, Daly MJ, Neale BM, Salem RM, Hirschhorn JN et al. 2014. An excess of risk-increasing low-frequency variants can be a signal of polygenic inheritance in complex diseases. *Am J Hum Genet* **94**(3): 437-452.

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

Christensen K, Thinggaard M, Oksuzyan A, Steenstrup T, Andersen-Ranberg K, Jeune B, McGue M, Vaupel JW. 2013. Physical and cognitive functioning of people older than 90 years: a comparison of two Danish cohorts born 10 years apart. *Lancet* **382**(9903): 1507-1513.

Cirulli ET, Goldstein DB. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* **11**(6): 415-425.

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

Consortium WTCC. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**(7145): 661-678.

Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA. 1993. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**(5123): 921-923.

Cottet F, Blanché H, Verasdonck P, Le Gall I, Schächter F, Bürkle A, Muiras ML. 2000. New polymorphisms in the human poly(ADP-ribose) polymerase-1 coding sequence: lack of association with longevity or with increased cellular poly(ADP-ribosyl)ation capacity. *J Mol Med (Berl)* **78**(8): 431-440.

Crimmins EM, Finch CE. 2006. Infection, inflammation, height, and longevity. *Proc Natl Acad Sci U S A* **103**(2): 498-503.

Cross DS, Ivacic LC, Stefanski EL, McCarty CA. 2010. Population based allele frequencies of disease associated polymorphisms in the Personalized Medicine Research Project. *BMC Genet* **11**: 51.

D'Antonio M, D'Onorio De Meo P, Paoletti D, Elmi B, Pallocca M, Sanna N, Picardi E, Pesole

G, Castrignanò T. 2013. WEP: a high-performance analysis pipeline for whole-exome data. *BMC Bioinformatics* **14 Suppl 7**: S11.

De Benedictis G, Carotenuto L, Carrieri G, De Luca M, Falcone E, Rose G, Cavalcanti S, Corsonello F, Feraco E, Baggio G et al. 1998. Gene/longevity association studies at four autosomal loci (REN, THO, PARP, SOD2). *Eur J Hum Genet* **6**(6): 534-541.

De Benedictis G, Franceschi C. 2006. The unusual genetics of human longevity. *Sci Aging Knowledge Environ* **2006**(10): pe20.

de Magalhães JP, Curado J, Church GM. 2009. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics* **25**(7): 875-881.

de Magalhães JP, Finch CE, Janssens G. 2010. Next-generation sequencing in aging research: emerging applications, problems, pitfalls and possible solutions. *Ageing Res Rev* **9**(3): 315-323.

de Magalhães JP, Toussaint O. 2004. GenAge: a genomic and proteomic network map of human ageing. *FEBS Lett* **571**(1-3): 243-247.

Deelen J, Beekman M, Capri M, Franceschi C, Slagboom PE. 2013. Identifying the genomic determinants of aging and longevity in human population studies: progress and challenges. *BioEssays : news and reviews in molecular, cellular and developmental biology* **35**(4): 386-396.

Deelen J, Beekman M, Uh HW, Broer L, Ayers KL, Tan Q, Kamatani Y, Bennet AM, Tamm R, Trompet S et al. 2014. Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. *Hum Mol Genet.(in press)*

Deelen J, Beekman M, Uh HW, Helmer Q, Kuningas M, Christiansen L, Kremer D, van der Breggen R, Suchiman HE, Lakenberg N et al. 2011. Genome-wide association study identifies a single major locus contributing to survival into old age; the APOE locus revisited. *Aging Cell* **10**(4): 686-698.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**(5): 491-498.

Di Cianni F, Campa D, Tallaro F, Rizzato C, De Rango F, Barale R, Passarino G, Canzian F, Gemignani F, Montesanto A et al. 2013. MAP3K7 and GSTZ1 are associated with human longevity: a two-stage case-control study using a multilocus genotyping. *Age (Dordr)* **35**(4): 1357-1366.

Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, Abraham C, Regueiro M, Griffiths A et al. 2006. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314**(5804): 1461-1463.

Dupont WD, Plummer WD. 1990. Power and sample size calculations. A review and computer program. *Control Clin Trials* **11**(2): 116-128.

Ellinghaus D, Zhang H, Zeissig S, Lipinski S, Till A, Jiang T, Stade B, Bromberg Y, Ellinghaus E, Keller A et al. 2013. Association between variants of PRDM1 and NDP52 and Crohn's disease, based on exome sequencing and functional studies. *Gastroenterology* **145**(2): 339-347.

ElSharawy A, Keller A, Flachsbart F, Wendschlag A, Jacobs G, Kefer N, Brefort T, Leidinger P, Backes C, Meese E et al. 2012. Genome-wide miRNA signatures of human longevity. *Aging Cell* **11**(4): 607-616.

Emanuele E, Fontana JM, Minoretti P, Geroldi D. 2010. Preliminary evidence of a genetic association between chromosome 9p21.3 and human longevity. *Rejuvenation Res* **13**(1): 23-26.

Engberg H, Oksuzyan A, Jeune B, Vaupel JW, Christensen K. 2009. Centenarians--a useful model for healthy aging? A 29-year follow-up of hospitalizations among 40,000 Danes born in 1905. *Aging Cell* **8**(3): 270-276.

Evert J, Lawler E, Bogan H, Perls T. 2003. Morbidity profiles of centenarians: survivors, delayers, and escapers. *J Gerontol A Biol Sci Med Sci* **58**(3): 232-237.

Ferrer-Costa C, Gelpí JL, Zamakola L, Parraga I, de la Cruz X, Orozco M. 2005. PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics* **21**(14): 3176-3178.

Flachsbart F, Caliebe A, Kleindorp R, Blanche H, von Eller-Eberstein H, Nikolaus S, Schreiber S, Nebel A. 2009. Association of FOXO3A variation with human longevity confirmed in German centenarians. *Proceedings of the National Academy of Sciences of the United States of America* **106**(8): 2700-2705.

Franceschi C, Bonafè M. 2003. Centenarians as a model for healthy aging. *Biochem Soc Trans* **31**(2): 457-461.

Franceschi C, Motta L, Valensin S, Rapisarda R, Franzone A, Berardelli M, Motta M, Monti D, Bonafè M, Ferrucci L 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.

Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R et al. 2010. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* **42**(12): 1118-1125.

Friedman DB, Johnson TE. 1988. A mutation in the age-1 gene in Caenorhabditis elegans lengthens life and reduces hermaphrodite fertility. *Genetics* **118**(1): 75-86.

Fries JF. 1980. Aging, natural death, and the compression of morbidity. *N Engl J Med* **303**(3): 130-135.

Garagnani P, Pirazzini C, Giuliani C, Candela M, Brigidi P, Sevini F, Luiselli D, Bacalini MG, Salvioli S, Capri M et al. 2014. The three genetics (nuclear DNA, mitochondrial DNA, and gut microbiome) of longevity in humans considered as metaorganisms. *Biomed Res Int* **2014**: 560340. *(in press)*

Gavrilova NS, Gavrilov LA, Evdokushkina GN, Semyonova VG, Gavrilova AL, Evdokushkina NN, Kushnareva YE, Kroutko VN, Andreyev AYu. 1998. Evolution, mutations, and human longevity: European royal and noble families. *Hum Biol* **70**(4): 799-804.

Geesaman BJ, Benson E, Brewster SJ, Kunkel LM, Blanché H, Thomas G, Perls TT, Daly MJ, 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.

Gems D, McElwee JJ. 2003. Ageing: Microarraying mortality. *Nature* **424**(6946): 259-261.

GenomeWeb. 2013. X Prize Foundation Shuts Down Genomics Competition. http://www.genomeweb.com/sequencing/x-prize-foundation-shuts-down-genomics-competition

Gentschew L, Flachsbart F, Kleindorp R, Badarinarayan N, Schreiber S, Nebel A. 2013. Polymorphisms in the superoxidase dismutase genes reveal no association with human

longevity in Germans: a case-control association study. *Biogerontology. (in press)*

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

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science* **185**(4154): 862-864.

Gudmundsson H, Gudbjartsson DF, Frigge M, Gulcher JR, Stefánsson K. 2000. Inheritance of human longevity in Iceland. *Eur J Hum Genet* **8**(10): 743-749.

Harris JR, Pedersen NL, McClearn GE, Plomin R, Nesselroade JR. 1992. Age differences in genetic and environmental influences for health from the Swedish Adoption/Twin Study of Aging. *J Gerontol* **47**(3): P213-220.

Hatem A, Bozdağ D, Toland AE, Çatalyürek Ü. 2013. Benchmarking short sequence mapping tools. *BMC Bioinformatics* **14**: 184.

Henderson ST, Johnson TE. 2001. daf-16 integrates developmental and environmental inputs to mediate aging in the nematode Caenorhabditis elegans. *Curr Biol* **11**(24): 1975-1980.

Hercberg S, Galan P, Preziosi P, Roussel AM, Arnaud J, Richard MJ, Malvy D, Paul-Dauphin A, Briançon S, Favier A. 1998. Background and rationale behind the SU.VI.MAX Study, a prevention trial using nutritional doses of a combination of antioxidant vitamins and minerals to reduce cardiovascular diseases and cancers. SUpplementation en VItamines et Minéraux AntioXydants Study. *Int J Vitam Nutr Res* **68**(1): 3-20.

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

Hitt R, Young-Xu Y, Silver M, Perls T. 1999. Centenarians: the older you get, the healthier you have been. *Lancet* **354**(9179): 652.

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

Hwangbo DS, Gershman B, Gersham B, Tu MP, Palmer M, Tatar M. 2004. Drosophila dFOXO controls lifespan and regulates insulin signalling in brain and fat body. *Nature* **429**(6991): 562-566.

Illumina Inc. 2011. Paired End Sample Preparation Guide. http://supportres.illumina.com/documents/myillumina/e5af4eb5-6742-40c8-bcb1-d8b350bcb964/paired-end_sampleprep_guide_1005063_e.pdf

Illumina PPIS. 2014. HiSeq X™ Ten: $1000 human genome and extreme throughput for population-scale sequencing. http://res.illumina.com/documents/datasheet-hiseq-x-ten.pdf

Jurinke C, van den Boom D, Cantor CR, Köster H. 2002. The use of MassARRAY technology for high throughput genotyping. *Adv Biochem Eng Biotechnol* **77**: 57-74.

Kaeberlein M, Powers RW, Steffen KK, Westman EA, Hu D, Dang N, Kerr EO, Kirkland KT, Fields S, Kennedy BK. 2005. Regulation of yeast replicative life span by TOR and Sch9 in response to nutrients. *Science* **310**(5751): 1193-1196.

Kapahi P, Zid BM, Harper T, Koslover D, Sapin V, Benzer S. 2004. Regulation of lifespan in Drosophila by modulation of genes in the TOR signaling pathway. *Curr Biol* **14**(10): 885-890.

Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M et al. 2014. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* **42**(1): D764-770.

Kedes L, Campany G. 2011. The new date, new format, new goals and new sponsor of the Archon Genomics X PRIZE competition. *Nat Genet* **43**(11): 1055-1058.

Kenny EE, Bustamante CD. 2011. SnapShot: Human biomedical genomics. *Cell* **147**(1): 248-248.e241.

Kerber RA, O'Brien E, Smith KR, Cawthon RM. 2001. Familial excess longevity in Utah genealogies. *J Gerontol A Biol Sci Med Sci* **56**(3): B130-139.

Kervinen K, Savolainen MJ, Salokannel J, Hynninen A, Heikkinen J, Ehnholm C, Koistinen MJ, Kesäniemi YA. 1994. Apolipoprotein E and B polymorphisms--longevity factors assessed in nonagenarians. *Atherosclerosis* **105**(1): 89-95.

Kimura KD, Tissenbaum HA, Liu Y, Ruvkun G. 1997. daf-2, an insulin receptor-like gene that regulates longevity and diapause in Caenorhabditis elegans. *Science* **277**(5328): 942-946.

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

Kirkwood TB. 2005. Time of our lives. What controls the length of life? *EMBO Rep* **6 Spec No**: S4-8.

Kirkwood TB. 2008. A systematic look at an old problem. *Nature* **451**(7179): 644-647.

Kirkwood TB, Rose MR. 1991. Evolution of senescence: late survival sacrificed for reproduction. *Philos Trans R Soc Lond B Biol Sci* **332**(1262): 15-24.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* **308**(5720): 385-389.

Knecht C, Krawczak M. 2013. Molecular genetic epidemiology of human diseases: from patterns to predictions. *Hum Genet. (in press)*

Koboldt DC, Ding L, Mardis ER, Wilson RK. 2010. Challenges of sequencing human genomes. *Brief Bioinform* **11**(5): 484-498.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**(7): 1073-1081.

Kung AW, Xiao SM, Cherny S, Li GH, Gao Y, Tso G, Lau KS, Luk KD, Liu JM, Cui B et al. 2010. Association of JAG1 with bone mineral density and osteoporotic fractures: a genome-wide association study and follow-up replication studies. *Am J Hum Genet* **86**(2): 229-239.

Kuningas M, Estrada K, Hsu YH, Nandakumar K, Uitterlinden AG, Lunetta KL, van Duijn CM, Karasik D, Hofman A, Murabito J et al. 2011. Large common deletions associate with mortality at old age. *Hum Mol Genet* **20**(21): 4290-4296.

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

Lam HY, Clark MJ, Chen R, Natsoulis G, O'Huallachain M, Dewey FE, Habegger L, Ashley EA, Gerstein MB, Butte AJ et al. 2012. Performance comparison of whole-genome sequencing platforms. *Nat Biotechnol* **30**(1): 78-82.

Lambert JC Ibrahim-Verbaas CA Harold D Naj AC Sims R Bellenguez C DeStafano AL Bis JC Beecham GW Grenier-Boley B et al. 2013. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* **45**(12): 1452-

1458.

Lamming DW, Sabatini DM. 2011. A radical role for TOR in longevity. *Cell Metab* **13**(6): 617-618.

Lee JH, Flaquer A, Costa R, Andrews H, Cross P, Lantigua R, Schupf N, Tang MX, Mayeux R. 2004. Genetic influences on life span and survival among elderly African-Americans, Caribbean Hispanics, and Caucasians. *Am J Med Genet A* **128A**(2): 159-164.

Lescai F, Blanché H, Nebel A, Beekman M, Sahbatou M, Flachsbart F, Slagboom E, Schreiber S, Sorbi S, Passarino G et al. 2009. Human longevity and 11p15.5: a study in 1321 centenarians. *Eur J Hum Genet* **17**(11): 1515-1519.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G et al. 2007. The diploid genome sequence of an individual human. *PLoS Biol* **5**(10): e254.

Lewis SJ, 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.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009a. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. 2009b. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Li M, Reilly MP, Rader DJ, Wang LS. 2010. Correcting population stratification in genetic association studies using a phylogenetic approach. *Bioinformatics* **26**(6): 798-806.

Li Y, Wang WJ, Cao H, Lu J, Wu C, Hu FY, Guo J, Zhao L, Yang F, Zhang YX et al. 2009c. Genetic association of FOXO1A and FOXO3A with longevity trait in Han Chinese populations. *Human molecular genetics* **18**(24): 4897-4904.

Lin K, Dorman JB, Rodan A, Kenyon C. 1997. daf-16: An HNF-3/forkhead family member that can function to double the life-span of Caenorhabditis elegans. *Science* **278**(5341): 1319-1322.

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. 2012. Comparison of next-generation sequencing systems. *J Biomed Biotechnol* **2012**: 251364.

Livak KJ. 2003. SNP genotyping by the 5'-nuclease reaction. *Methods Mol Biol* **212**: 129-147.

Ma L, Mondal AK, Murea M, Sharma NK, Tönjes A, Langberg KA, Das SK, Franks PW, Kovacs P, Antinozzi PA et al. 2011. The effect of ACACB cis-variants on gene expression and metabolic traits. *PLoS One* **6**(8): e23860.

Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. 2011. What can exome sequencing do for you? *J Med Genet* **48**(9): 580-589.

Malovini A, Illario M, Iaccarino G, Villa F, Ferrario A, Roncarati R, Anselmi CV, Novelli V, Cipolletta E, Leggiero E et al. 2011. Association study on long-living individuals from Southern Italy identifies rs10491334 in the CAMKIV gene that regulates survival proteins. *Rejuvenation Res* **14**(3): 283-291.

Manolio TA, Brooks LD, Collins FS. 2008. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* **118**(5): 1590-1605.

Manton KG, Gu X, Lamb VL. 2006. Change in chronic disability from 1982 to 2004/2005 as measured by long-term changes in function and health in the U.S. elderly population. *Proc Natl Acad Sci U S A* **103**(48): 18374-18379.

Marth GT, Yu F, Indap AR, Garimella K, Gravel S, Leong WF, Tyler-Smith C, Bainbridge M, Blackwell T, Zheng-Bradley X et al. 2011. The functional spectrum of low-frequency coding variation. *Genome Biol* **12**(9): R84.

McGue M, Vaupel JW, Holm N, Harvald B. 1993. Longevity is moderately heritable in a sample of Danish twins born 1870-1880. *Journal of gerontology* **48**(6): B237-244.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**(9): 1297-1303.

Meldrum C, Doyle MA, Tothill RW. 2011. Next-generation sequencing for cancer diagnostics: a practical perspective. *Clin Biochem Rev* **32**(4): 177-195.

Metzker ML. 2010. Sequencing technologies - the next generation. *Nat Rev Genet* **11**(1): 31-46.

Mitchell BD, Hsueh WC, King TM, Pollin TI, Sorkin J, Agarwala R, Schäffer AA, Shuldiner AR. 2001. Heritability of life span in the Old Order Amish. *Am J Med Genet* **102**(4): 346-352.

Mitteldorf J. 2010. Female fertility and longevity. *Age (Dordr)* **32**(1): 79-84.

Molero JC, Jensen TE, Withers PC, Couzens M, Herzog H, Thien CB, Langdon WY, Walder K, Murphy MA, Bowtell DD et al. 2004. c-Cbl-deficient mice have reduced adiposity, higher energy expenditure, and improved peripheral insulin action. *J Clin Invest* **114**(9): 1326-1333.

Moore CB, Wallace JR, Wolfe DJ, Frase AT, Pendergrass SA, Weiss KM, Ritchie MD. 2013. Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data. *PLoS Genet* **9**(12): e1003959.

Morris JZ, Tissenbaum HA, Ruvkun G. 1996. A phosphatidylinositol-3-OH kinase family member regulating longevity and diapause in Caenorhabditis elegans. *Nature* **382**(6591): 536-539.

Mukhopadhyay A, Tissenbaum HA. 2007. Reproduction and longevity: secrets revealed by C. elegans. *Trends Cell Biol* **17**(2): 65-71.

Murabito JM, Yuan R, Lunetta KL. 2012. The search for longevity and healthy aging genes: insights from epidemiological studies and samples of long-lived individuals. *J Gerontol A Biol Sci Med Sci* **67**(5): 470-479.

Nebel A, Croucher PJ, Stiegeler R, Nikolaus S, Krawczak M, Schreiber S. 2005. No association between microsomal triglyceride transfer protein (MTP) haplotype and longevity in humans. *Proc Natl Acad Sci U S A* **102**(22): 7906-7909.

Nebel A, Flachsbart F, Till A, Caliebe A, Blanché H, Arlt A, Häsler R, Jacobs G, Kleindorp R, Franke A et al. 2009. A functional EXO1 promoter variant is associated with prolonged life expectancy in centenarians. *Mech Ageing Dev* **130**(10): 691-699.

Nebel A, Kleindorp R, Caliebe A, Nothnagel M, Blanche H, Junge O, Wittig M, Ellinghaus D, Flachsbart F, Wichmann HE et al. 2011. A genome-wide association study confirms APOE as the major gene influencing survival in long-lived individuals. *Mechanisms of*

ageing and development **132**(6-7): 324-330.

Nebel A, Schreiber S. 2005. Allelic variation and human longevity. *Sci Aging Knowledge Environ* **2005**(29): pe23.

Newman AB, Cauley JA, Murabito J, Lunetta K. 2012. Genetics of Human Longevity and Healthy Aging. In *The Epidemiology of Aging*, pp. 215-235. Springer Netherlands.

Newman AB, Glynn NW, Taylor CA, Sebastiani P, Perls TT, Mayeux R, Christensen K, Zmuda JM, Barral S, Lee JH et al. 2011. Health and function of participants in the Long Life Family Study: A comparison with other cohorts. *Aging (Albany NY)* **3**(1): 63-76.

Newman AB, Murabito JM. 2013. The Epidemiology of Longevity and Exceptional Survival. *Epidemiol Rev. (in press)*

Newman AB, Walter S, Lunetta KL, Garcia ME, Slagboom PE, Christensen K, Arnold AM, Aspelund T, Aulchenko YS, Benjamin EJ 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.

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

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

Nygaard M, Lindahl-Jacobsen R, Soerensen M, Mengel-From J, Andersen-Ranberg K, Jeune B, Vaupel JW, Tan Q, Christiansen L, Christensen K. 2014. Birth cohort differences in the prevalence of longevity-associated variants in APOE and FOXO3A in Danish long-lived individuals. *Exp Gerontol. (in press)*

O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, Bodily P, Tian L, Hakonarson H, Johnson WE et al. 2013. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med* **5**(3): 28.

Oeppen J, Vaupel JW. 2002. Demography. Broken limits to life expectancy. *Science* **296**(5570): 1029-1031.

Ogg S, Paradis S, Gottlieb S, Patterson GI, Lee L, Tissenbaum HA, Ruvkun G. 1997. The Fork head transcription factor DAF-16 transduces insulin-like metabolic and longevity signals in C. elegans. *Nature* **389**(6654): 994-999.

Ohiro Y, Garkavtsev I, Kobayashi S, Sreekumar KR, Nantz R, Higashikubo BT, Duffy SL, Higashikubo R, Usheva A, Gius D et al. 2002. A novel p53-inducible apoptogenic gene, PRG3, encodes a homologue of the apoptosis-inducing factor (AIF). *FEBS Lett* **524**(1-3): 163-171.

Passtoors WM, Beekman M, Deelen J, van der Breggen R, Maier AB, Guigas B, Derhovanessian E, van Heemst D, de Craen AJ, Gunn DA et al. 2013. Gene expression analysis of mTOR pathway: association with human longevity. *Aging Cell* **12**(1): 24-31.

Passtoors WM, Boer JM, Goeman JJ, Akker EB, Deelen J, Zwaan BJ, Scarborough A, Breggen R, Vossen RH, Houwing-Duistermaat JJ et al. 2012. Transcriptional profiling of human familial longevity indicates a role for ASF1A and IL7R. *PLoS One* **7**(1): e27759.

Pavlopoulos GA, Oulas A, Iacucci E, Sifrim A, Moreau Y, Schneider R, Aerts J, Iliopoulos I. 2013. Unraveling genomic variation from next generation sequencing data. *BioData Min* **6**(1): 13.

Pawlikowska L, Hu D, Huntsman S, Sung A, Chu C, Chen J, Joyner AH, Schork NJ, Hsueh WC, Reiner AP et al. 2009. Association of common genetic variation in the insulin/IGF1 signaling pathway with human longevity. *Aging Cell* **8**(4): 460-472.

Peng G, Fan Y, Palculict TB, Shen P, Ruteshouser EC, Chi AK, Davis RW, Huff V, Scharfe C, Wang W. 2013. Rare variant detection using family-based sequencing analysis. *Proc Natl Acad Sci U S A* **110**(10): 3985-3990.

Perls T, Shea-Drinkwater M, Bowen-Flynn J, Ridge SB, Kang S, Joyce E, Daly M, Brewster SJ, Kunkel L, Puca AA. 2000. Exceptional familial clustering for extreme longevity in humans. *J Am Geriatr Soc* **48**(11): 1483-1485.

Perls T, Terry D. 2003. Genetics of exceptional longevity. *Exp Gerontol* **38**(7): 725-730.

Perls TT, Wilmoth J, Levenson R, Drinkwater M, Cohen M, Bogan H, Joyce E, Brewster S, Kunkel L, Puca A. 2002. Life-long sustained mortality advantage of siblings of centenarians. *Proc Natl Acad Sci U S A* **99**(12): 8442-8447.

Petersen B-S. 2014. Discovery of novel rare Crohn's disease variants by next-generation sequencing In *Mathematisch-Naturwissenschaftlichen Fakultät*, Vol PhD. University of Kiel, Kiel. *(PhD thesis)*

Petersen BS, Spehlmann ME, Raedler A, Stade B, Thomsen I, Rabionet R, Rosenstiel P, Schreiber S, Franke A. 2014. Whole genome and exome sequencing of monozygotic twins discordant for Crohn's disease. *BMC Genomics* **15**(1): 564. *(in press)*

Picard. 2009. http://picard.sourceforge.net.

Pinós T, Fuku N, Cámara Y, Arai Y, Abe Y, Rodríguez-Romo G, Garatachea N, Santos-Lozano A, Miro-Casas E, Ruiz-Meana M et al. 2014. The rs1333049 polymorphism on locus 9p21.3 and extreme longevity in Spanish and Japanese cohorts. *Age (Dordr)* **36**(2): 933-943.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**(1): 110-121.

Puca AA, Daly MJ, Brewster SJ, Matise TC, Barrett J, Shea-Drinkwater M, Kang S, Joyce E, Nicoli J, Benson E et al. 2001. A genome-wide scan for linkage to human exceptional longevity identifies a locus on chromosome 4. *Proc Natl Acad Sci U S A* **98**(18): 10505-10508.

Puckelwartz MJ, Pesce LL, Nelakuditi V, Dellefave-Castillo L, Golbus JR, Day SM, Cappola TP, Dorn GW, Foster IT, McNally EM. 2014. Supercomputing for the parallelization of whole genome analysis. *Bioinformatics*. *(in press)*

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**(3): 559-575.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**(6): 841-842.

Ragoussis J. 2006. Genotyping technologies for all. *Drug Discovery Today: Technologies* **3**(2): 115-122.

Ratan A, Miller W, Guillory J, Stinson J, Seshagiri S, Schuster SC. 2013. Comparison of sequencing platforms for single nucleotide variant calls in a human sample. *PLoS One*

8(2): e55089.

Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. 1997. GeneCards: integrating information about genes, proteins and diseases. *Trends Genet* **13**(4): 163.

Reed T, Dick DM, Uniacke SK, Foroud T, Nichols WC. 2004. Genome-wide scan for a healthy aging phenotype provides support for a locus near D4S1564 promoting healthy aging. *J Gerontol A Biol Sci Med Sci* **59**(3): 227-232.

Riancho JA, Vázquez L, García-Pérez MA, Sainz J, Olmos JM, Hernández JL, Pérez-López J, Amado JA, Zarrabeitia MT, Cano A et al. 2011. Association of ACACB polymorphisms with obesity and diabetes. *Mol Genet Metab* **104**(4): 670-676.

Rieber N, Zapatka M, Lasitschka B, Jones D, Northcott P, Hutter B, Jäger N, Kool M, Taylor M, Lichter P et al. 2013. Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. *PLoS One* **8**(6): e66621.

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

Robine J, Vaupel JW. 2001. Supercentenarians: slower ageing individuals or senile elderly? *Exp Gerontol* **36**(4-6): 915-930.

Robine JM, Allard M. 1998. The oldest human. *Science* **279**(5358): 1834-1835.

Robine JM, Cheung SL, Saito Y, Jeune B, Parker MG, Herrmann FR. 2010. Centenarians Today: New Insights on Selection from the 5-COOP Study. *Curr Gerontol Geriatr Res* **2010**: 120354.

Roche NimbleGen Inc. 2009. NimbleGen Sequence Capture 2.1M Human Exome Array. http://www.atlas-biolabs.de/nimblegen

Romualdi C, Balding D, Nasidze IS, Risch G, Robichaux M, Sherry ST, Stoneking M, Batzer MA, Barbujani G. 2002. Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. *Genome Res* **12**(4): 602-612.

Rose G, Dato S, Altomare K, Bellizzi D, Garasto S, Greco V, Passarino G, Feraco E, Mari V, Barbi C et al. 2003. Variability of the SIRT3 gene, human silent information regulator Sir2 homologue, and survivorship in the elderly. *Exp Gerontol* **38**(10): 1065-1070.

Rowe JW, Kahn RL. 1997. Successful aging. *Gerontologist* **37**(4): 433-440.

Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL, Jørgensen JE, Weigel D, Andersen SU. 2009. SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat Methods* **6**(8): 550-551.

Schoenhofen EA, Wyszynski DF, Andersen S, Pennington J, Young R, Terry DF, Perls TT. 2006. Characteristics of 32 supercentenarians. *J Am Geriatr Soc* **54**(8): 1237-1240.

Schoenmaker M, de Craen AJ, de Meijer PH, Beekman M, Blauw GJ, Slagboom PE, Westendorp RG. 2006. Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. *Eur J Hum Genet* **14**(1): 79-84.

Schork NJ, Murray SS, Frazer KA, Topol EJ. 2009. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* **19**(3): 212-219.

Schreiber V, Amé JC, Dollé P, Schultz I, Rinaldi B, Fraulob V, Ménissier-de Murcia J, de Murcia G. 2002. Poly(ADP-ribose) polymerase-2 (PARP-2) is required for efficient base excision DNA repair in association with PARP-1 and XRCC1. *J Biol Chem* **277**(25): 23028-23036.

Schächter F, Cohen D, Kirkwood T. 1993. Prospects for the genetics of human longevity. *Hum Genet* **91**(6): 519-526.

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

Sebastiani P, Perls TT. 2012. The genetics of extreme longevity: lessons from the new England centenarian study. *Front Genet* **3**: 277.

Sebastiani P, Riva A, Montano M, Pham P, Torkamani A, Scherba E, Benson G, Milton JN, Baldwin CT, Andersen S et al. 2011. Whole genome sequences of a male and female supercentenarian, ages greater than 114 years. *Frontiers in genetics* **2**: 90.

Sebastiani P, Solovieff N, Dewan AT, Walsh KM, Puca A, Hartley SW, Melista E, Andersen S, Dworkis DA, Wilk JB et al. 2012. Genetic signatures of exceptional longevity in humans. *PLoS One* **7**(1): e29848.

Sebastiani P, Sun FX, Andersen SL, Lee JH, Wojczynski MK, Sanders JL, Yashin A, Newman AB, Perls TT. 2013. Families Enriched for Exceptional Longevity also have Increased Health-Span: Findings from the Long Life Family Study. *Front Public Health* **1**: 38.

Sehat B, Andersson S, Girnita L, Larsson O. 2008. Identification of c-Cbl as a new ligase for insulin-like growth factor-I receptor with distinct roles from Mdm2 in receptor ubiquitination and endocytosis. *Cancer Res* **68**(14): 5669-5677.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**(1): 308-311.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. 2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* **15**(2): 121-132.

Snyder M, Du J, Gerstein M. 2010. Personal genome sequencing: current approaches and challenges. *Genes Dev* **24**(5): 423-431.

Soerensen M. 2012. Genetic variation and human longevity. *Dan Med J* **59**(5): B4454.

Soerensen M, Christensen K, Stevnsner T, Christiansen L. 2009. The Mn-superoxide dismutase single nucleotide polymorphism rs4880 and the glutathione peroxidase 1 single nucleotide polymorphism rs1050450 are associated with aging and longevity in the oldest old. *Mech Ageing Dev* **130**(5): 308-314.

Soerensen M, Dato S, Christensen K, McGue M, Stevnsner T, Bohr VA, Christiansen L. 2010. Replication of an association of variation in the FOXO3A gene with human longevity using both case-control and longitudinal data. *Aging Cell* **9**(6): 1010-1017.

Steffens M, Lamina C, Illig T, Bettecken T, Vogler R, Entz P, Suk EK, Toliat MR, Klopp N, Caliebe A et al. 2006. SNP-based analysis of genetic substructure in the German population. *Hum Hered* **62**(1): 20-29.

Suh Y, Atzmon G, Cho MO, Hwang D, Liu B, Leahy DJ, Barzilai N, Cohen P. 2008. Functionally significant insulin-like growth factor I receptor mutations in centenarians. *Proc Natl Acad Sci U S A* **105**(9): 3438-3442.

SureSelect Target Enrichment Kit. 2010. Agilent Technologies: SureSelect Target Enrichment for Illumina Paired-End Sequencing Library. http://dnatech.genomecenter.ucdavis.edu/uploads/SureSelect_IlluminaPaired.pdf

Suzuki M, Akisaka M, Ashitomi I, Higa K, Nozaki H. 1995. [Chronological study concerning ADL among Okinawan centenarians]. *Nihon Ronen Igakkai Zasshi* **32**(6): 416-423.

Tabatabaie V, Atzmon G, Rajpathak SN, Freeman R, Barzilai N, Crandall J. 2011. Exceptional longevity is associated with decreased reproduction. *Aging (Albany NY)* **3**(12): 1202-1205.

Tacutu R, Budovsky A, Fraifeld VE. 2010. The NetAge database: a compendium of networks for longevity, age-related diseases and associated processes. *Biogerontology* **11**(4): 513-522.

Tacutu R, Craig T, Budovsky A, Wuttke D, Lehmann G, Taranukha D, Costa J, Fraifeld VE, de Magalhães JP. 2013. Human Ageing Genomic Resources: integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Res* **41**(Database issue): D1027-1033.

Tan Q, De Benedictis G, Ukraintseva SV, Franceschi C, Vaupel JW, Yashin AI. 2002. A centenarian-only approach for assessing gene-gene interaction in human longevity. *Eur J Hum Genet* **10**(2): 119-124.

Tan Q, Zhao JH, Zhang D, Kruse TA, Christensen K. 2008. Power for genetic association study of human longevity using the case-control design. *Am J Epidemiol* **168**(8): 890-896.

Tazearslan C, Cho M, Suh Y. 2012. Discovery of functional gene variants associated with human longevity: opportunities and challenges. *J Gerontol A Biol Sci Med Sci* **67**(4): 376-383.

Terman JR, Mao T, Pasterkamp RJ, Yu HH, Kolodkin AL. 2002. MICALs, a family of conserved flavoprotein oxidoreductases, function in plexin-mediated axonal repulsion. *Cell* **109**(7): 887-900.

Thakur RS, Bandopadhyay R, Chaudhary B, Chatterjee S. 2012. Now and next-generation sequencing techniques: future of sequence analysis using cloud computing. *Front Genet* **3**: 280.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**(2): 178-192.

Ugarte M, Pérez-Cerdá C, Rodríguez-Pombo P, Desviat LR, Pérez B, Richard E, Muro S, Campeau E, Ohura T, Gravel RA. 1999. Overview of mutations in the PCCA and PCCB genes causing propionic acidemia. *Hum Mutat* **14**(4): 275-282.

User Guide: Applied Biosystems. 2010. Applied Biosystems SOLiD™ System BioScope™ Software for Scientists Guide https://tools.lifetechnologies.com/content/sfs/manuals/cms_082377.pdf

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

Vaupel JW. 1995. *Exceptional longevity: from prehistory to the present*. Springer, Berlin.

Vaupel JW. 2004. *The Biodemography of Aging. In: Waite, L. J. (Ed.), Aging, health and public policy. Demographic and economic perspectives*. Population Council, New York.

Vaupel JW. 2010. Biodemography of human ageing. *Nature* **464**(7288): 536-542.

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

Vellai T, Takacs-Vellai K, Zhang Y, Kovacs AL, Orosz L, Müller F. 2003. Genetics: influence of TOR kinase on lifespan in C. elegans. *Nature* **426**(6967): 620.

Venter JC Adams MD Myers EW Li PW Mural RJ Sutton GG Smith HO Yandell M Evans CA Holt RA et al. 2001. The sequence of the human genome. *Science* **291**(5507): 1304-

1351.

Viennas E, Gkantouna V, Ioannou M, Georgitsi M, Rigou M, Poulas K, Patrinos GP, Tzimas G. 2012. Population-ethnic group specific genome variation allele frequency data: a querying and visualization journey. *Genomics* **100**(2): 93-101.

Waldron I. 1983. Sex differences in illness incidence, prognosis and mortality: issues and evidence. *Soc Sci Med* **17**(16): 1107-1123.

Waldron I. 1993. Recent trends in sex mortality ratios for adults in developed countries. *Soc Sci Med* **36**(4): 451-462.

Waldron I. 1995. Contributions of biological and behavioural factors to changing sex differences in ischaemic heart disease mortality. In *In: Adult mortality in developed countries: from description to explanation*, (ed. GC Alan D. Lopez, Tapani Valkonen), pp. 161-178. Clarendon Press, Oxford, England.

Wallace JE. 1996. Gender differences in beliefs of why women live longer than men. *Psychol Rep* **79**(2): 587-591.

Walter S, Atzmon G, Demerath EW, Garcia ME, Kaplan RC, Kumari M, Lunetta KL, Milaneschi Y, Tanaka T, Tranah GJ et al. 2011. A genome-wide association study of aging. *Neurobiol Aging* **32**(11): 2109.e2115-2128.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**(16): e164.

Wang Z, Liu X, Yang BZ, Gelernter J. 2013. The role and challenges of exome sequencing in studies of human diseases. *Front Genet* **4**: 160.

Westendorp RG, Kirkwood TB. 1998. Human longevity at the cost of reproductive success. *Nature* **396**(6713): 743-746.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**(7189): 872-876.

Willcox BJ, Donlon TA, He Q, Chen R, Grove JS, Yano K, Masaki KH, Willcox DC, Rodriguez B, Curb JD. 2008. FOXO3A genotype is strongly associated with human longevity. *Proceedings of the National Academy of Sciences of the United States of America* **105**(37): 13987-13992.

Willcox BJ, Willcox DC, He Q, Curb JD, Suzuki M. 2006a. Siblings of Okinawan centenarians share lifelong mortality advantages. *J Gerontol A Biol Sci Med Sci* **61**(4): 345-354.

Willcox DC, Willcox BJ, Hsueh WC, Suzuki M. 2006b. Genetic determinants of exceptional human longevity: insights from the Okinawa Centenarian Study. *Age (Dordr)* **28**(4): 313-332.

Yashin AI, Wu D, Arbeev KG, Ukraintseva SV. 2010. Joint influence of small-effect genetic variants on human longevity. *Aging (Albany NY)* **2**(9): 612-620.

Ye K, Beekman M, Lameijer EW, Zhang Y, Moed MH, van den Akker EB, Deelen J, Houwing-Duistermaat JJ, Kremer D, Anvar SY et al. 2013. Aging as accelerated accumulation of somatic variants: whole-genome sequencing of centenarian and middle-aged monozygotic twin pairs. *Twin Res Hum Genet* **16**(6): 1026-1032.

Young RD, Coles SL. 2014. Validated Worldwide Supercentenarians, Living and Recently Deceased. *Rejuvenation Res. (in press)*

Yu X, Sun S. 2013. Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC Bioinformatics* **14**: 274.

Yu Y, Sun Y, He S, Yan C, Rui L, Li W, Liu Y. 2012. Neuronal Cbl controls biosynthesis of insulin-like peptides in Drosophila melanogaster. *Mol Cell Biol* **32**(18): 3610-3623.

Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, Daly MJ, Neale BM, Sunyaev SR, Lander ES. 2014. Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A* **111**(4): E455-464.

9 Declaration

Herewith, I confirm that the submitted thesis is completely the result of my own work. Apart from the advice of my supervisors, all sources and cooperation partners are listed within the thesis. This thesis has not been submitted elsewhere. It has been carried out in strict accordance with the rules of good scientific practice of the *Deutsche Forschungsgesellschaft*.

_____ _____

Signature Date

10 Curriculum vitae

First and Last Name: Nandini Badarinarayan

Address: Gurlittstrasse 3
24106 Kiel

Date of birth: 1st September 1984

Nationality: Indian

School education:

06/1987 – 02/2000 St Mary's High School, Pune, India
(Indian Certificate of Secondary Education)

Higher education:

06/2000 – 03/2002 Fergusson College, Pune, India
(Higher Secondary Certificate)
Major subjects: Biology, Chemistry, Physics, Maths

09/2004 – 06/2008 Visvesvaraya Technological University, Bangalore, India
(Bachelor in Engineering: Biotechnology)

01/2009 – 01/2010 University of Leicester, United Kingdom
(Master of Science: Bioinformatics)

01/2011 – 07/2014 PhD student at Institute of Clinical Molecular Biology
Christian-Albrechts-University, Kiel

11 Acknowledgements

I am fortunate to have received sound advice and motivating words of encouragement from all the people that I was associated with during this project. I take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project:

Prof. Dr. Stefan Schreiber, Prof. Dr. Philip Rosenstiel and Prof. Dr. Andre Franke for giving me the opportunity to accomplish my Ph.D. at the Institute of Clinical Molecular Biology and for providing excellent working conditions.

Prof. Dr. Tal Dagan for kindly agreeing to review my thesis.

Prof. Dr. Almut Nebel for her guidance, helpful advice and being a great supervisor. Her comments were always encouraging, extremely perceptive and appropriate, which helped me at all times during research and writing of my thesis.

PD Dr. Friederike Flachsbart for being the most supportive and understanding supervisor. Not only was Friederike easily approachable when I had doubts, but she guided me out of the woods every time I felt I lost my way. She also helped in proof-reading my thesis and provided many considerable suggestions. Her patience and positive energy helped me overcome many criticial situations that emerged unexpectedly during the course of my Ph.D.

Furthermore, I would also like to thank our collaborators at IMIS and CRG, specially Prof. Dr. Michael Krawczak, Dr. Amke Caliebe, Carolin Knecht and Daniel Trujillano for their inputs and friendly advice. I express my sincere gratitude to the whole institute of ICMB and all my colleagues for supporting me every way possible. My special thanks goes to Geetha Venkatesh, Liljana Gentschew, Britt Petersen, Michael Forster, Ingo Thomsen and Björn Stade, because without their generous support the completion of this work would have been difficult. I am also grateful to the Next-gen sequencing and genotyping platforms at ICMB for their continuous high-quality work in terms of processing, sequencing and genotyping of samples and thereby making this work possible.

It is a pleasure to thank my friends, Richa and Vasudev, who made my stay in Kiel more fun and a memorable experience. Most importantly, none of this would have been possible without the love and patience of my family- my parents, brother and sister-in-law. They have been a constant source of support and strength all these years.

12 Supplementary material

Table 12-1: Selected examples of genes identified influencing lifespan in model organisms (Kuningas et al. 2008)

Organism	Gene name/description	Function	Reference
Canenorhabditis elegans			
age-1	Phosphatidylinositol kinase	Insulin signaling	(Morris et al. 1996)
daf-2	Insulin receptor like gene	Insulin signaling	(Kimura et al. 1997)
daf-16	Forkhead transcription factor	Regulation of metabolic and development pathways	(Ogg et al. 1997)
Drosophila melanogaster			
Chico	Insulin receptor substrate	Insulin signaling	(Clancy et al. 2001)
Mus musculus			
Gh	Growth hormone	Insulin signaling, tissue proliferation	(Bartke 2005)

Table 12-2: Genome-wide association studies with discovery and replication samples in humans (Murabito et al. 2012)

Reference	Discovery sample	Replication sample	Gene	SNV	P value	Odds Ratio
(Newman et al. 2010) ‖	CHARGE cohorts (AGES, CHS, FHS, and RS) 1,836 individuals age >90 y; 1,955 individuals who died between ages 55–80 y	Leiden Longevity Study: 940 long-lived (mean age 94); 744 partners of offspring (mean age 60); Danish 1905 cohort: 1,644 long-lived (mean age 93); 2,007 younger Danish twins (mean age 57)	MINPPI	rs9664222	6.8×10^{-7}	0.82
(Walter et al. 2011) ‖	CHARGE cohorts (AGES, ARIC, BLSA, CHS, FHS, HABC, InCHIANTI, RS, and SHIP), 25,007 participants age ≥55 y at baseline (55% women), European origin, 8,444 deaths (mean age 81.1); average follow-up 10.6 y	Four independent samples of European origin; N = 10,411, deaths = 1,295	OTOL1	rs1425609	1.6×10^{-6}	-
(Malovini et al. 2011)	410 long-lived individuals from Southern Italy (90–109 y); 553 younger controls (18–48 y)	116 long-lived individuals (90–109 y); 160 younger controls (18–44 y)	CAMKIV	rs10491334	1.7×10^{-6}	0.55
(Deelen et al. 2011)	Leiden Longevity Study: 403 Long-lived (mean age 94); 1,760 younger controls (mean age 58)	Rotterdam Study: 960 long-lived (mean age 94); 1,825 younger controls (mean age 62) Leiden 85+ Study: 1,208 long-lived (mean age 92); 2,090 younger controls (mean age 35) Danish 1905 cohort: 1,598 long-lived (mean age 93); 1,997 younger controls (mean age 57)	TOMM40 †	rs2075650	3.4×10^{-17}	0.71

(Nebel et al. 2011)	763 long-lived German individuals (mean age 99.7) 1,085 young German individuals (mean age 60.2 y)	754 long-lived German individuals (mean age 96.9) 860 young German individuals (mean age 67.3 y)	APOC1*	rs4420638	1.8×10^{-10}	0.53
(Sebastiani et al. 2012)	801 European ancestry long-living individuals (mean age 104) 914 European ancestry controls (age range 0-75 years)	292 European ancestry long-living individuals, 21 long-living individuals, (mean age 108) 867 controls (age range 0-75 years)	-	-	-	-

Notes: AGES = Age, Gene/Environment Susceptibility-ReyKjavik Study; ARIC = Atherosclerosis Risk in Communities Study; BLSA = Baltimore Longitudinal Study of Ageing; CHARGE = Cohorts for Heart and Aging Research in Genomic Epidemiology; CHS = Cardiovascular Health Study; FHS = Framingham Heart Study; HAAS = Honolulu Asia Aging Study; HABC = Health, Aging and Body Composition Study; HHP = Honolulu Heart Program; InCHIANTI = Invescchiare nel Chianti; RS = Rotterdam Study; SHIP = Study of Health in Pomerania; SNV = single nucleotide variant.

* Explained by linkage equilibrium with the ApoE E4 allele (r2 = .72).

† Explained by moderate linkage disequilibrium with ApoE E4 (r2 = .55, rs429358).

§ Homozygous minor (GG) versus homozygous major (TT) alleles between cases and controls.

|| None of the associations achieved genome-wide significance; only the most significant association in the discovery plus replication stage is provided in the table.

Figure 12-1: **SOLiD™ technology sequencing schema**: (1) Complementary dinucleotide hybridizes to the already primer-bound template sequence and is ligated. (2) After the fluorescence is measured, (3) unextended strands are capped (4) and the dye is cleaved off leaving a free 5' phosphate group. (5) This process is repeated for several cycles until the required length is achieved. (6) The synthesized strand is removed, a new primer with a one-base offset is hybridized and (7) the ligation cycles are repeated. (8) This primer reset process is repeated for five rounds providing dual measurement of each base (Figure from http://www.appliedbiosystems.com).

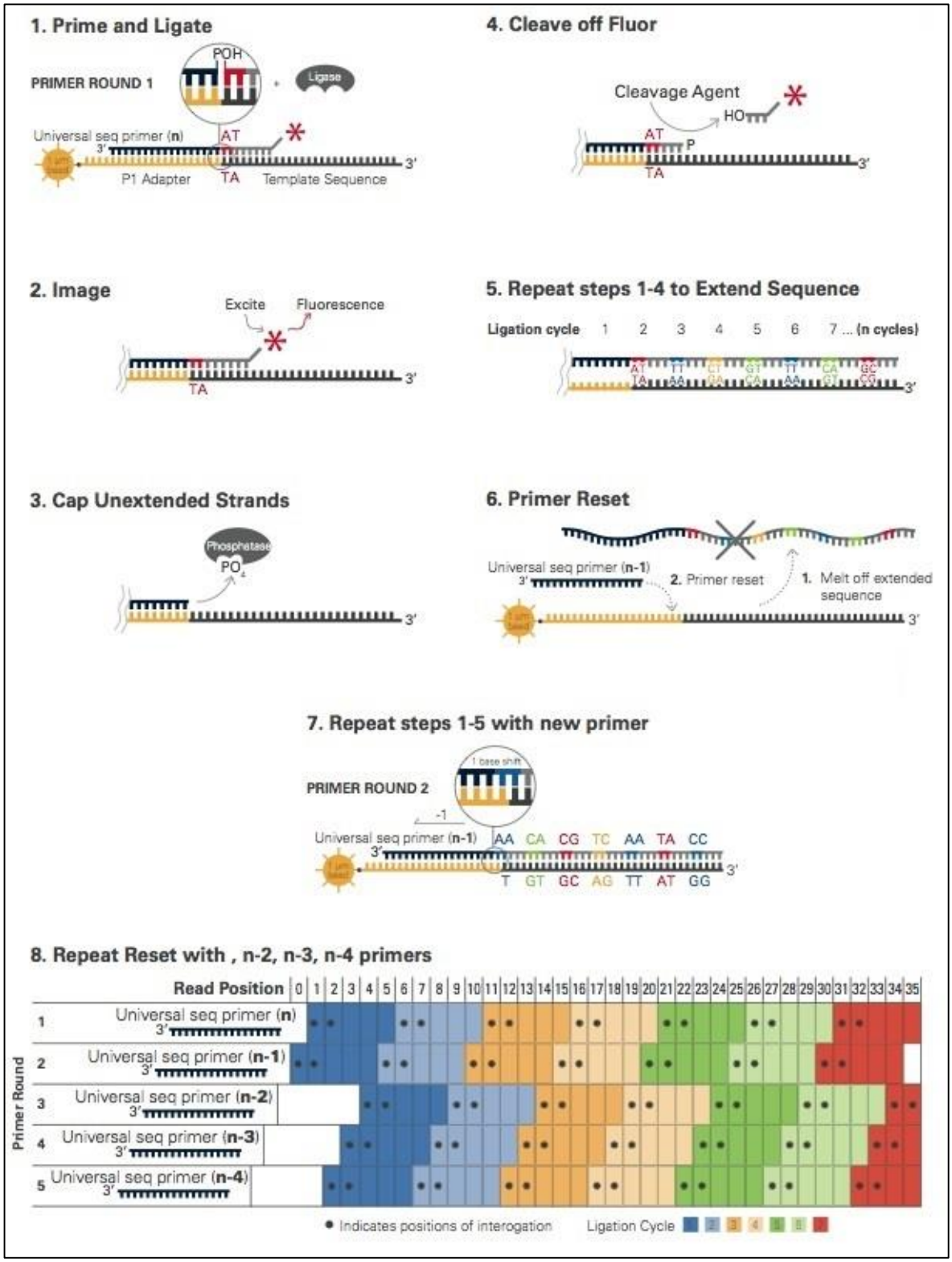

Figure 12-2: **Illumina technology sequencing schema**: Fluorescently labeled nucleotides are incorporated into the complementary strand after the sequencing primer is hybridized. The remaining nucleotides are washed away and the fluorescence signal identifying the base is recorded. The fluorescent label and terminator group are removed and a new cycle of sequencing is started (Figure from Metzker 2010).

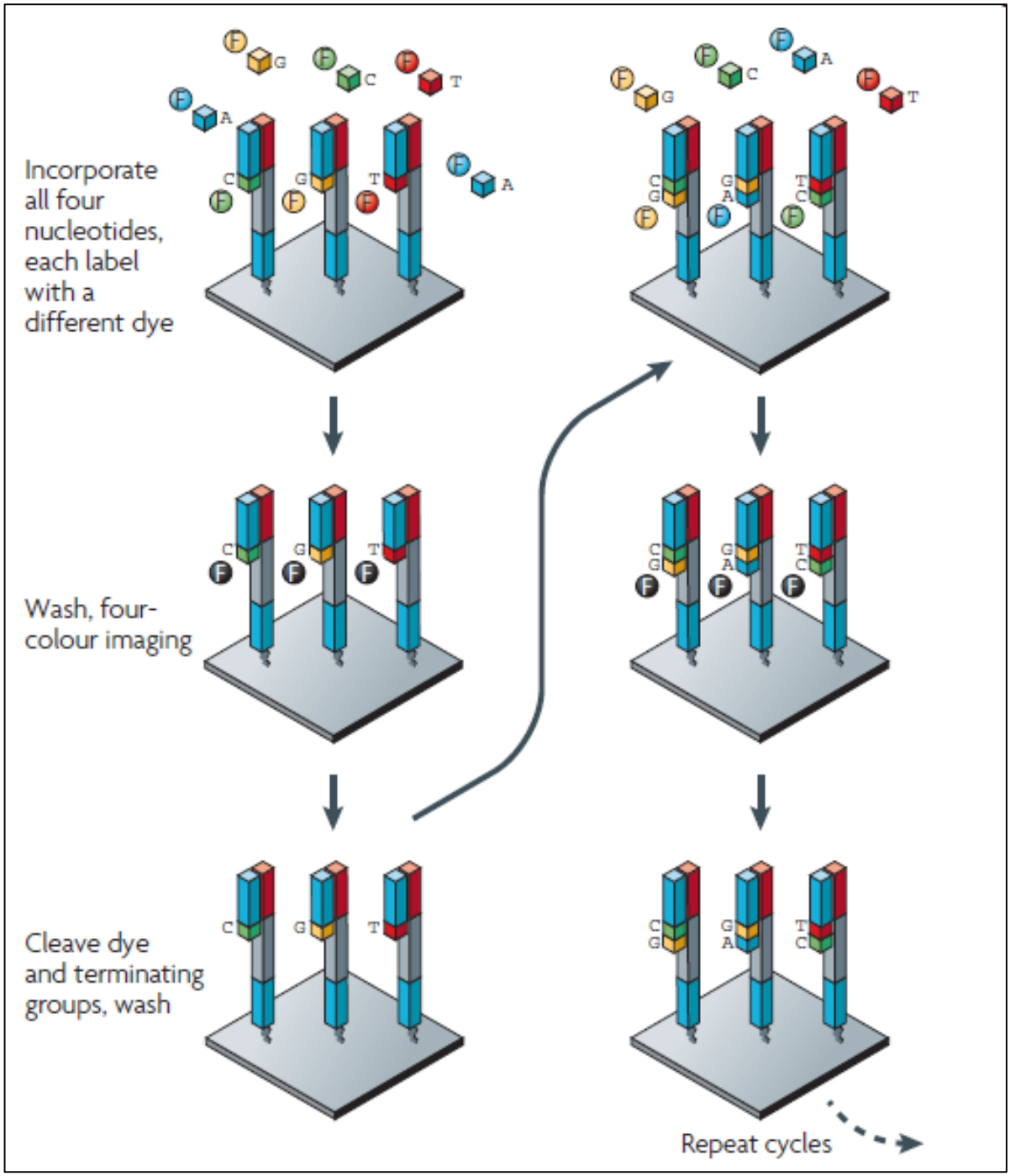

Figure 12-3: **Exome sequencing schema:** Exome sequencing for six individuals were performed using (b) SureSelect and (c) NimbleGen target enrichment kit (Figure from Roche NimbleGen Inc 2009 and SureSelect Target Enrichment Kit 2010)

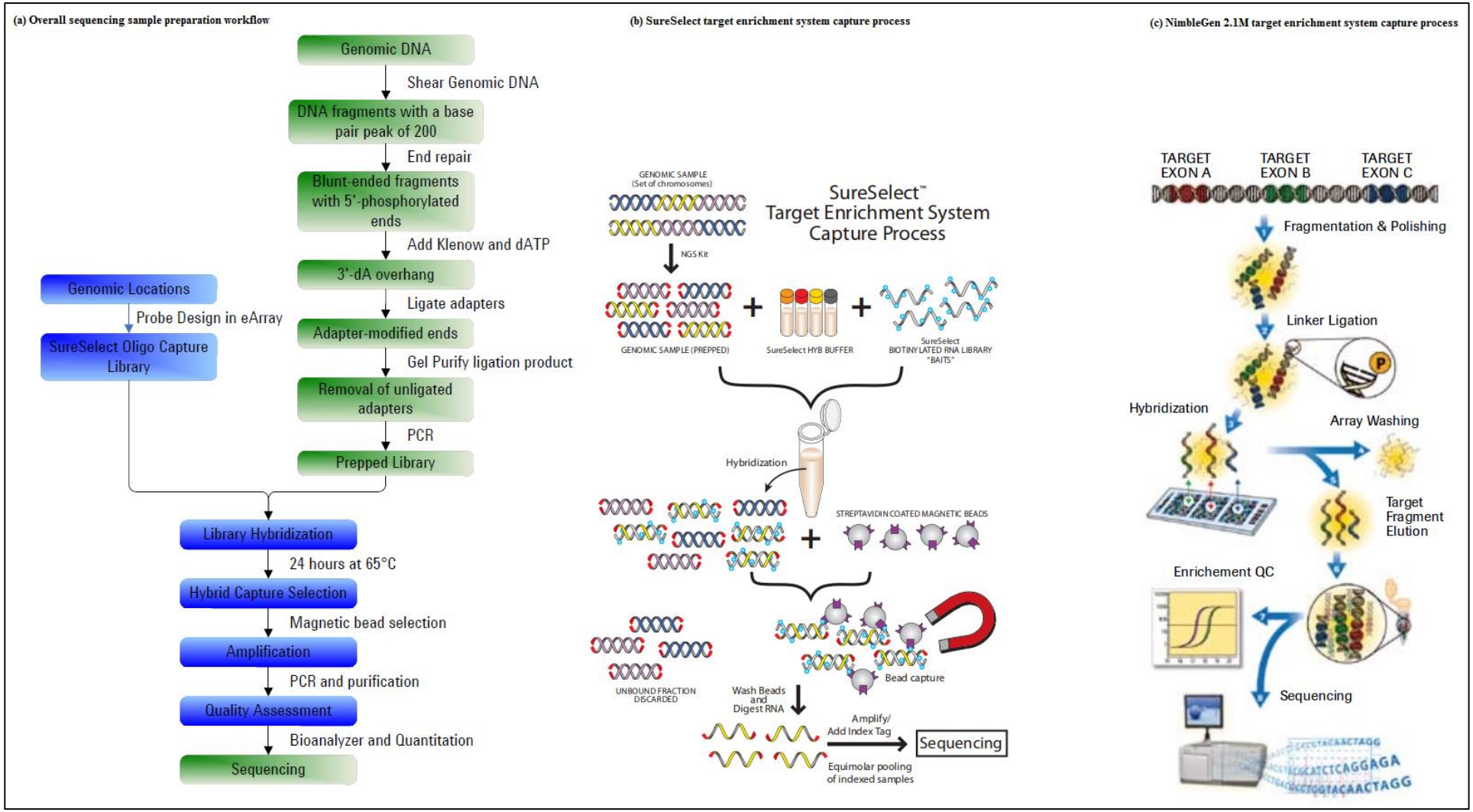

Figure 12-4: **Workflow for whole genome and exome variant calling**: for (a) SOLiD and (b) Illumina sequencing data

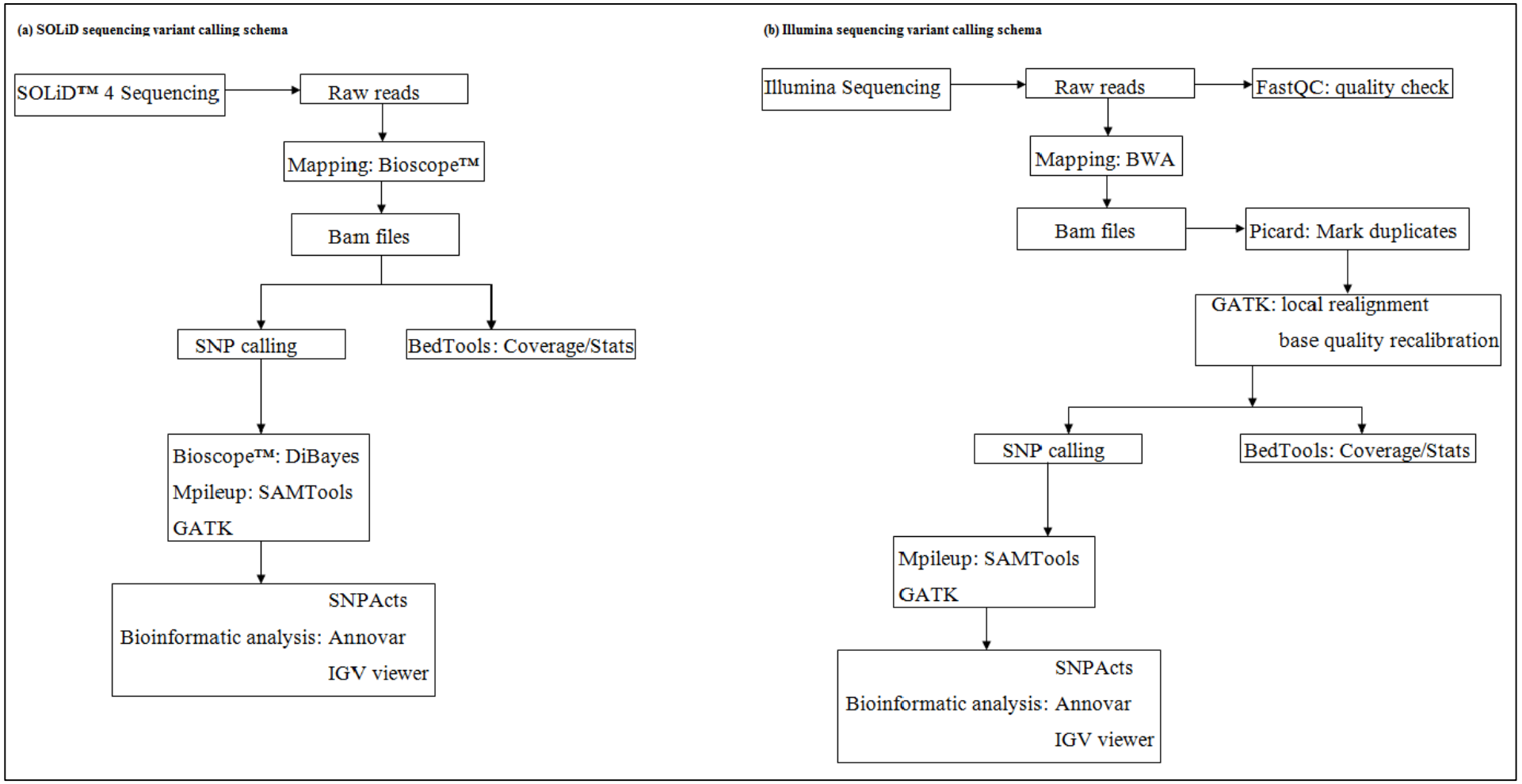

Figure 12-5: **Exome pipeline for SNV calling implemented by CRG, Spain**: The samples were mapped with BWA using hg19 human genome reference. This was followed by local realignment around indels and quality score recalibration done using GATK. The resulting SNV calling was performed with three different variant tools, namely GATK, SAMtools mpileup and SHORE. The three independent SNV predictions were subsequently quality filtered using GATK VariantFiltration and intersected using GATK CombineVariants. Functional annotation was performed using Annovar. (Figure from CRG,Spain)

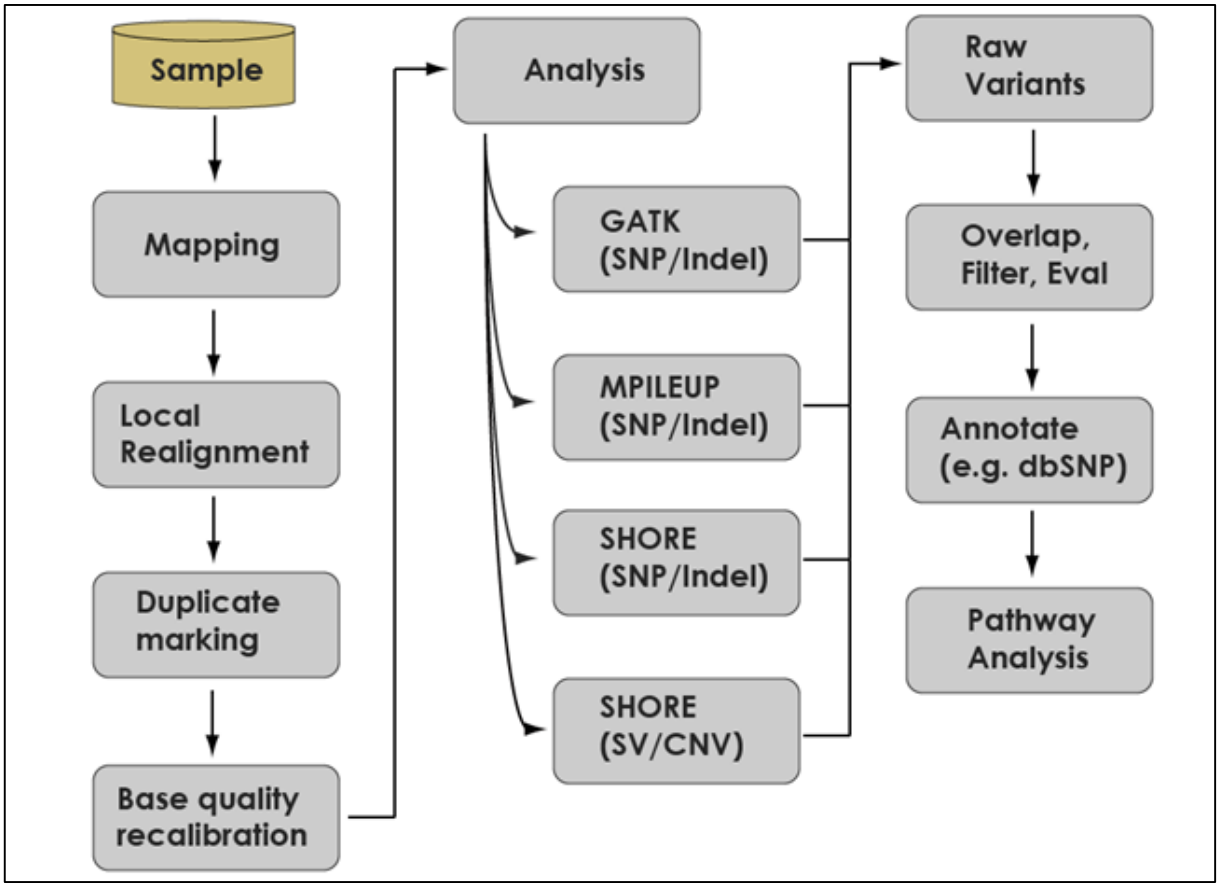

Table 12-3: Functional annotation of variants generated with SOLiD technology using snpActs

	(1) German female	(2) German male	(3) French female
Total number of SNVs	3,264,816	3923,324	2,695,673
Total number of novel SNVs	688,484	770,546	454,305
all synonymous-coding SNVs	9,372	11,192	7,324
all missense SNVs	11,523	12,950	8,393
all cancel-start SNVs	22	27	15
all read-through SNVs	35	40	24
all nonsense SNVs	260	247	164
Novel missense SNVs	4,743	4,625	2,761
Novel cancel-start SNVs	13	14	4
Novel read-through SNVs	12	14	5
Novel nonsense SNVs	206	181	116
SNVs in acceptor	77	70	52
SNVs in donor	62	63	39
SNVs in 5'UTR	5,053	6,237	3,728
SNVS in 3'UTR	22,977	27,281	18,788
SNVs in UTR-Splice sites	206,931	248,370	171,185
SNVs in introns	997,987	1,191,940	824,278
unknown/intergenic SNVs	2,010,513	2,424,907	1,661,682
SNVs overlapping with Venter genome	431,220	1,138,396	82,528
SNVs overlapping with Watson genome	710,950	626,227	42,124
SNVs overlapping with Yh1 genome	480,999	1,115,681	73,537
SNVs overlapping with Yoruban3 genome	479,328	963,730	61,009

Table 12-4: Functional annotation of variants generated with Illumina technology using snpActs

	(1) German female	(2) German male	(3) French female	(4) Spanish female
Total number of SNVs	4,013,012	4,071,554	4,022,164	4,081,702
Total number of novel SNVs	339,537	373,213	343,478	348,019
all synonymous-coding SNVs	9,866	10,268	10,225	11,037
all missense SNVs	9,590	10,276	9,871	10,390
all cancel-start SNVs	10	12	15	16
all read-through SNVs	30	33	29	40
all nonsense SNVs	92	108	87	86
Novel missense SNVs	832	1,056	868	715
Novel cancel-start SNVs	1	0	1	2
Novel read-through SNVs	2	3	2	5
Novel nonsense SNVs	22	28	20	6
SNVs in acceptor	24	30	24	21
SNVs in donor	64	82	66	44
SNVs in 5'UTR	5,617	5,831	5,775	6,334
SNVS in 3'UTR	24,880	25,431	25,270	26,322
SNVs in UTR-Splice sites	253,876	257,817	255,583	262,471
SNVs in introns	1,206,955	1,221,284	1,209,656	1,216,178
unknown/intergenic SNVs	2,502,004	2,540,381	2,505,562	2,548,761
SNVs overlapping with Venter genome	586,254	589,452	584,387	591,360
SNVs overlapping with Watson genome	859,265	860,536	859,697	864,114
SNVs overlapping with Yh1 genome	618,180	620,578	623,079	625,436
SNVs overlapping with Yoruban3 genome	589,687	591,384	589,130	599,648

Table 12-5: Functional annotation of variants generated with exome sequencing using snpActs

	(1) German female	(2) German male	(3) French female	(4) Spanish female	(5) German female	(6) German male
Total number of SNVs	26,223	26,790	26,767	27,178	18,456	18,481
Total number of novel SNVs	1,111	1,078	1,120	1,093	925	885
all synonymous-coding SNVs	9,014	9,145	9,094	9,245	7,107	7,099
all missense SNVs	7,725	7,911	7,916	8,040	5,859	5,915
all cancel-start SNVs	8	9	12	11	10	4
all read-through SNVs	8	11	10	10	8	5
all nonsense SNVs	65	65	59	53	51	59
Novel missense SNVs	455	434	452	417	451	430
Novel cancel-start SNVs	2	0	1	1	1	0
Novel read-through SNVs	0	1	1	0	1	0
Novel nonsense SNVs	17	12	12	4	19	24
SNVs in acceptor	14	14	13	12	5	4
SNVs in donor	12	17	16	20	28	26
SNVs in 5'UTR	348	371	382	397	205	231
SNVS in 3'UTR	480	502	489	532	304	285
SNVs in UTR-Splice sites	166	179	157	190	76	67
SNVs in introns	8,225	8,407	8,440	8,482	4,739	4,718
unknown/intergenic SNVs	153	159	178	185	65	66
SNVs overlapping with Venter genome	9,022	9,300	9,362	9,392	6,562	6,722
SNVs overlapping with Watson genome	5,581	5,907	5,897	5,933	4,130	4,184
SNVs overlapping with Yh1 genome	9,252	9,359	9,309	9,466	6,749	6,814
SNVs overlapping with Yoruban3 genome	7,743	7,658	7,646	7,855	5,666	5,720

Table 12-6: Exonic SNVs with functional impact selected for genotyping for Method 1

Chr-Start-Obs	dbSNP132	Gene	MAF_1000G*	p 1000G‡	MAF ESP*	p ESP‡	No. of Samples	PhyloP_pred¤	No of tools predicted „damaging“ [0;4]±
1-9305445-C	rs34603401	H6PD	0.05	0.005788218	0.109593	0.048267295	3	C	2
1-11884555-G	rs198400	CLCN6	1	0	0.000744	0.005936524	4	C	0
1-16096934-T	rs10927851	FBLIM1	0.59	0.024673938	0.389292	0.026496846	4	C	1
1-38329999-G	rs41311191	INPP5B	0.08	0.021100486	0.109838	0.048544407	2	C	4
1-161132777-A	rs17356051	USP21	0.03	0.022340758	0.042387	0.042425092	2	C	1
1-169519049-C	rs6025	F5	0.99	0.002690078	0.021937	0.012339645	4	C	0
1-186275564-T	rs12128607	PRG4	0.03	0.022340758	0.036903	0.032877126	2	N	1
1-196928188-G	rs41310132	CFHR2	0.009	3.95E-005	0.01284	0.000112953	2	C	2
1-201754444-C	rs16849342	NAV1	0.03	0.001349863	0.04973	0.005700882	3	C	0
1-210412843-T	rs61740848	SERTAD4	0.003	0.000248993	0.006507	0.001155066	2	C	2
1-220161969-C	rs116081500	EPRS	0.002	0.000111107	0.009853	0.002613111	2	N	2
1-220197625-T	rs2230301	EPRS	0.89	0.007106765	0.140732	0.017098252	4	C	0
1-220331205-G	rs2289189	RAB3GAP2	0.04	0.003079679	0.068693	0.013961462	3	N	2
1-222801661-T	rs142088763	MIA3	0.002	0.000111107	0.009724	0.002546451	2	N	1
2-17963123-G	rs77424145	GEN1	0.01	0.002690078	0.018315	0.008727391	2	N	0
2-26667130-G	rs3795958	CCDC164	0.13	0.012929701	0.12335	0.010727311	3	C	0
2-98928494-A	rs7587534	VWA3B	0.98	0.010336892	0.031516	0.024505585	3	N	0
2-118732831-A	rs17512204	CCDC93	0.03	0.001349863	0.058282	0.008878647	3	C	0
2-120129841-G	rs8192506	DBI	0.01	0.002690078	0.02454	0.015280953	2	C	2

2-171400449-C	rs56181206	MYO3B	0.009	0.002187717	0.02232	0.012754626	2	NA	NA
2-172650165-T	rs35565687	SLC25A12	0.02	0.010336892	0.032627	0.026146451	2	C	1
2-211456637-G	rs1047883	CPS1	0.57	0.012311736	0.403049	0.025073282	4	N	1
2-234627536-A	rs6755571	UGT1A4	0.02	0.010336892	0.043688	0.044833189	2	N	0
2-239155053-T	rs934945	PER2	0.17	0.032786296	0.139617	0.016628267	4	C	0
2-239237388-A	rs61742338	TRAF3IP1	0.02	0.010336892	0.029838	0.022114555	2	N	0
3-3887508-G	rs35362954	LRRN1	0.01	0.002690078	0.031047	0.023826643	2	N	2
3-4354697-A	rs6801634	SETMAR	0.12	0.009721613	0.127812	0.012174089	3	N	0
3-45869972-T	rs1129183	LZTFL1	0.04	0.003079679	0.066369	0.01270619	3	C	1
3-48628014-A	rs2228561	COL7A1	0.07	0.01469865	0.101506	0.039602028	2	N	0
3-49138810-C	rs11539148	QARS	0.02	0.010336892	0.042015	0.041746359	2	C	2
3-75714337-G	rs73840323	FRG2C	0.85	0.02135247	0.136364	0.01530678	4	N	0
3-129281980-T	rs2713625	PLXND1	1	0	0.000186	3.60E-010	2	N	0
4-42003671-G	rs1047626	SLC30A9	0.61	0.026402805	0.426287	0.012827489	4	N	0
4-57797467-T	rs114282228	REST	0.007	0.001334084	0.014501	0.005555439	2	N	0
4-71390616-A	rs151041998	AMTN	0.001	2.79E-005	0.008459	0.001936805	2	C	1
4-87770252-A	rs17694522	SLC10A6	0.02	0.010336892	0.040063	0.038257756	2	C	0
4-96106322-G	rs2289043	UNC5C	0.49	0.003323293	0.475181	0.008354808	4	C	2
4-119219909-A	rs28661939	PRSS12	0.12	0.009721613	0.139617	0.016628267	3	N	0
4-178256913-G	rs7689099	NEIL3	0.06	0.009622942	0.098847	0.036957511	2	C	2
5-53815560-C	rs13162502	SNX18	0.11	0.007106765	0.146496	0.019669938	2	N	1
5-70308262-T	rs61757629	NAIP	0.008	0.001735509	0.019814	0.010153107	2	N	2
5-177422876-C	rs7445271	PROP1	1	0	0.000372	0.002972128	4	N	0

6-13470113-T	rs766773	C6orf114	0.9	0.03809179	0.085319	0.02506696	3	N	0
6-30672353-G	rs9262151	MDC1	0.006	1.18E-005	0.017568	0.000284188	2	C	0
6-31556928-A	rs3179003	NCR3	0.03	0.022340758	0.040156	0.038421155	2	N	1
6-31778272-A	rs2227956	HSPA1L	0.89	0.007106765	0.12744	0.012048679	3	N	0
6-46135884-G	rs34109856	ENPP5	0.02	0.010336892	0.037746	0.03427957	2	C	4
6-119327632-T	rs17827619	FAM184A	0.07	0.01469865	0.103971	0.042144098	2	C	1
6-151161086-C	rs17080410	PLEKHG1	0.03	0.022340758	0.035323	0.030314701	2	N	0
6-151161116-A	rs61742396	PLEKHG1	0.03	0.022340758	0.035137	0.030018793	2	C	0
6-151161836-C	rs17054318	PLEKHG1	0.03	0.022340758	0.03523	0.030166594	2	C	0
6-167343141-A	rs11159	RNASET2	0.08	0.021100486	0.088585	0.027699408	2	N	1
7-88965021-A	rs10487075	ZNF804B	0.06	0.009622942	0.079476	0.020731059	3	C	0
7-99032517-A	rs34943973	ATP5J2-PTCD1	0.03	0.022340758	0.044618	0.046586898	2	C	1
7-107427322-C	rs34407351	SLC26A3	0.02	0.010336892	0.032906	0.02656564	2	C	2
7-140301731-T	rs269243	DENND2A	0.96	0.038147228	0.045996	0.049234036	4	N	0
8-17396380-A	rs13259948	SLC7A2	0.05	0.000371751	0.143511	0.01830814	3	N	0
8-17419539-A	rs62622371	SLC7A2	0.03	0.001349863	0.087842	0.027087291	2	C	0
8-124206324-A	rs7813708	FAM83A	0.17	0.032786296	0.155512	0.024187415	4	N	0
9-111678508-T	rs1140064	IKBKAP	0.02	0.010336892	0.021658	0.012041265	2	C	1
9-131483749-A	rs2900268	ZDHHC12	1	0	0.00439	2.56E-008	2	N	0
10-16979714-C	rs41289305	CUBN	0.07	0.001335867	0.14473	0.018856408	3	N	0
10-69934258-G	rs3814182	MYPN	0.47	0.002381129	0.477505	0.008257514	4	N	2
10-71018660-C	rs1111335	HKDC1	0.99	0.002690078	0.006972	0.001323582	3	N	0

11-4673788-A	rs17224476	OR51E1	0.05	0.005788218	0.087842	0.027087291	2	C	0
11-11906050-C	rs34511735	USP47	0.003	0.000248993	0.013219	0.004640387	2	C	3
11-18319180-T	rs7128017	HPS5	0.14	0.016788716	0.130879	0.013242003	4	N	3
11-27720937-T	rs66866077	BDNF	0.01	0.002690078	0.03526	0.030214337	2	N	0
11-45245778-T	rs35090414	PRDM11	0.03	0.001349863	0.064081	0.011538992	2	N	1
11-56000403-G	rs10791893	OR5T2	0.9	0.00502435	0.153374	0.023060091	3	N	0
11-57076820-G	rs78489201	TNKS1BP1	0.04	0.003079679	0.055215	0.007639522	3	N	1
11-57146225-G	rs34108746	PRG3	0.03	0.001349863	0.055308	0.00767543	3	N	0
11-60785263-G	rs61755080	CD6	0.03	0.022340758	0.044618	0.046586898	2	N	0
11-67430762-C	rs1551886	ALDH3B2	0.86	0.016788716	0.149563	0.02113762	3	N	0
11-120187971-A	rs2282537	POU2F3	0.09	0.028886792	0.110801	0.049641922	2	C	0
11-134226244-A	rs61740182	GLB1L2	0.01	0.002690078	0.023239	0.013775555	2	C	3
12-999638-T	rs17755373	WNK1	0.01	0.002690078	0.010225	0.002809957	2	C	1
12-10571091-G	rs2682494	KLRC3	0.99	6.78E-007	0.015123	3.49E-006	2	N	0
12-48144925-C	rs2016123	RAPGEF3	1	0	0.000188	3.72E-010	2	N	0
12-96312686-A	rs12368787	CCDC38	0.06	0.009622942	0.100874	0.038964291	2	C	0
12-108102757-G	rs74918182	PWP1	0.02	0.010336892	0.017382	0.007890379	2	N	0
13-24436475-T	rs11551114	MIPEP	0.07	0.01469865	0.107403	0.04582832	2	N	0
13-39263714-C	rs2496423	FREM2	1	0	0.000186	9.68E-007	3	N	0
13-100962156-G	rs35719359	PCCA	0.01	0.002690078	0.037739	0.034267824	2	C	1
13-113530199-A	rs11616795	ATP11A	0.06	0.009622942	0.077988	0.019702693	2	N	1
14-20822308-G	rs3093921	PARP2	0.009	0.002187717	0.016005	0.006726845	2	C	2
14-33291583-A	rs34711402	AKAP6	0.01	0.002690078	0.026957	0.018260814	2	C	1

14-64447776-C	rs9944035	SYNE2	0.1	0.03809179	0.081213	0.021970285	2	C	1
14-73717720-A	rs741842	PAPLN	0.06	0.009622942	0.100112	0.038202986	2	C	0
14-75574087-T	rs10146482	NEK9	0.53	0.008607098	0.429448	0.012386407	4	N	0
15-42439444-T	rs111633028	PLA2G4F	0.02	0.010336892	0.046384	0.049989717	2	N	1
15-42602621-T	rs35285091	GANC	0.003	0.000248993	0.010039	0.002710677	2	C	4
15-48443699-C	rs2470103	MYEF2	1	0	0.000558	0.004455292	4	C	0
15-86123019-T	rs2061824	AKAP13	0.59	0.024673938	0.374884	0.028911657	4	N	1
15-86123833-C	rs4075256	AKAP13	0.6	0.02531584	0.375627	0.028763442	4	N	1
15-86123988-A	rs4075254	AKAP13	0.6	0.02531584	0.374977	0.028892962	4	C	1
15-86124483-G	rs4843074	AKAP13	0.59	0.024673938	0.375534	0.028781851	4	N	1
15-86124555-A	rs4843075	AKAP13	0.59	0.024673938	0.374977	0.028892962	4	N	1
15-86124946-C	rs7162168	AKAP13	0.59	0.024673938	0.374326	0.029024688	4	N	0
16-2821573-T	rs8017	TCEB2	0.4	0.00065536	0.448782	0.01016834	4	N	1
16-29708350-G	rs9932770	QPRT	1	0	0.001209	9.85E-008	2	N	0
16-58314433-C	rs2241414	PRSS54	0.07	0.01469865	0.095464	0.033739314	3	N	0
16-58314598-T	rs1052276	PRSS54	0.08	0.021100486	0.095464	0.033739314	3	N	0
16-81058354-G	rs11641523	CENPN	0.007	0.001334084	0.017917	0.008365571	2	N	1
17-63683-A	rs117190076	RPH3AL	0.03	0.001349863	0.050009	0.005791143	2	N	1
17-33689926-C	rs4796077	SLFN11	0.97	0.022340758	0.039413	0.037123676	4	N	0
17-48265495-C	rs1800215	COL1A1	0.96	0.000157383	0.029467	4.80E-005	2	C	0
18-30846895-T	rs9965081	C18orf34	0.92	0.021100486	0.095562	0.033830237	3	N	0
19-4442999-G	rs243383	CHAF1A	0.97	5.15E-005	0.017781	6.61E-006	2	N	0
19-11891003-A	rs799193	ZNF441	0.9	0.00502435	0.15514	0.023988719	3	N	0

19-52000624-G	rs3752135	SIGLEC12	0.83	0.032786296	0.112103	0.007610379	4	N	0
20-31383238-A	rs150682895	DNMT3B	0.008	0.001735509	0.009667	0.00251726	2	C	0
22-18901004-T	rs450046	PRODH	0.9	0.03809179	0.095961	0.03420184	3	C	0
22-41574383-C	rs1046088	EP300	0.02	0.010336892	0.026678	0.017904874	2	C	1
22-46931077-C	rs4823850	CELSR1	0.89	0.007106765	0.135922	0.015132866	3	C	2
22-50599466-A	rs2272843	MOV10L1	0.12	0.009721613	0.115449	0.008461633	4	C	0

*MAF_1000G,MAF_ESP: minor allele frequency based on 1000Genomes or NHLBI Exome Sequencing Project database

ǂp_1000G,p_ESP: p-value calculated by binomial testing comparing allelic frequencies of four sample with 1000Genomes or NHLBI Exome Sequencing Project database

¤PhyloP_pred: Predictions based on Phylop; C-conserved, N- Neutral

± No of tools predicted „damaging“: impact of amino acid substitution leading to either loss-of-function or gain-of-function based on Annovar (SIFT,Polyphen-2,LRT and MutationTaster)

Table 12-7: **Association statistics for 116 common SNVs** genotyped using the Sequenom and TaqMan technology in German LLI (n=1,610), centenarian subset (n=748) and younger controls (n=1,104)

Association analysis: German population LLI							
Chr	dbSNP ID	Gene	MAF cases n=1,610	MAF controls n=1,104	P_{CCA}	OR	95% CI
1	rs10927851	FBLIM1	0.293	0.3429	0.0001824	0.7941	0.8923-1.412
14	rs3093921	PARP2	0.03635	0.02148	0.001965	1.719	0.5816-1.77
5	rs61757629	NAIP	0.03153	0.01896	0.004989	1.684	0.837-1.041
6	rs17080410	PLEKHG1	0.03544	0.04762	0.02736	0.7349	0.7232-1.328
13	rs35719359	PCCA	0.06429	0.05	0.02779	1.306	0.7506-1.027
6	rs17054318	PLEKHG1	0.03555	0.04745	0.03067	0.7401	0.6262-1.425
22	rs2272843	MOV10L1	0.1447	0.1651	0.04137	0.8549	0.7686-1.101
6	rs61742396	PLEKHG1	0.03583	0.04678	0.04415	0.7572	0.9165-1.226
15	rs4843075	AKAP13	0.3661	0.3419	0.0714	1.112	0.2299-3.196
13	rs11551114	MIPEP	0.1292	0.1464	0.07176	0.8654	0.8044-1.052
14	rs9944035	SYNE2	0.06596	0.07832	0.0843	0.831	0.6625-1.45
11	rs34108746	PRG3	0.06286	0.07477	0.09097	0.8301	0.04289-1.098
15	rs4075254	AKAP13	0.365	0.3426	0.09102	1.103	0.7037-0.8961
1	rs61740848	SERTAD4	0.005618	0.009554	0.09321	0.5857	0.8404-1.12
15	rs2061824	AKAP13	0.3651	0.343	0.0965	1.101	0.8874-1.159
3	rs6801634	SETMAR	0.1331	0.1489	0.1029	0.8779	0.8072-1.101
15	rs7162168	AKAP13	0.3662	0.3444	0.1036	1.01	0.6945-1.297
3	rs2228561	COL7A1	0.1193	0.1339	0.1107	0.8756	0.9112-1.291
1	rs16849342	NAV1	0.04934	0.0592	0.1135	0.8249	0.7422-1.141
4	rs17694522	SLC10A6	0.05836	0.06824	0.1402	0.8463	1.215-2.432
2	rs6755571	UGT1A4	0.04054	0.04876	0.1493	0.8243	0.7394-1.013
11	rs35090414	PRDM11	0.1	0.1122	0.1525	0.8794	0.8245-1.09
14	rs741842	PAPLN	0.1462	0.1601	0.1623	0.8984	0.7195-2.21
4	rs28661939	PRSS12	0.1436	0.1302	0.1631	1.12	0.7476-1.171
15	rs4843074	AKAP13	0.3604	0.3428	0.1877	1.08	0.6223-1.689
14	rs34711402	AKAP6	0.03513	0.04212	0.1878	0.828	0.675-1.353
11	rs78489201	TNKS1BP1	0.06628	0.0757	0.1883	0.8668	0.9096-1.224
1	rs2289189	RAB3GAP2	0.09551	0.1064	0.191	0.8871	0.6249-1.097
3	rs11539148	QARS	0.05395	0.06221	0.2006	0.8596	0.6896-1.274
10	rs3814182	MYPN	0.4264	0.4439	0.2026	0.9313	0.8367-1.161

1	rs142088763	MIA3	0.01505	0.01099	0.2038	1.375	0.893-1.155
11	rs1551886	ALDH3B2	0.08349	0.09344	0.2042	0.8838	0.7829-1.221
16	rs8017	TCEB2	0.4604	0.4776	0.2164	0.9335	0.797-1.93
19	rs799193	ZNF441	0.2001	0.2138	0.2223	0.9199	0.7373-1.049
16	rs1052276	PRSS54	0.1002	0.1106	0.2238	0.8964	0.7588-1.134
1	rs17356051	USP21	0.06219	0.05454	0.2431	1.15	0.8092-1.32
16	rs11641523	CENPN	0.02884	0.02381	0.262	1.218	0.7426-1.167
16	rs2241414	PRSS54	0.1016	0.1109	0.2827	0.9071	0.6405-1.646
11	rs10791893	OR5T2	0.1191	0.129	0.2837	0.913	0.7437-1.031
17	rs117190076	RPH3AL	0.06452	0.05789	0.3235	1.122	0.6818-1.084
15	rs35285091	GANC	0.01754	0.01419	0.3391	1.24	0.7353-0.994
11	rs7128017	HPS5	0.1155	0.1075	0.3608	1.085	0.8323-1.126
11	rs17224476	OR51E1	0.09846	0.1061	0.364	0.9201	0.7731-1.078
1	rs6025	F5	0.03059	0.02639	0.3653	1.164	0.7004-1.073
2	rs17512204	CCDC93	0.07497	0.08151	0.3805	0.9132	0.6686-1.03
11	rs2282537	POU2F3	0.1376	0.1296	0.396	1.072	0.6159-1.271
6	rs9262151	MDC1	0.01101	0.008748	0.4169	1.261	0.7592-1.084
1	rs34603401	H6PD	0.1754	0.1671	0.4331	1.06	0.7515-1.069
6	rs17827619	FAM184A	0.01161	0.009407	0.4459	1.236	0.7548-1.186
22	rs450046	PRODH	0.06527	0.07052	0.4499	0.9204	0.6733-1.026
2	rs934945	PER2	0.2017	0.1934	0.4507	1.054	0.7304-1.07
2	rs3795958	CCDC164	0.1814	0.1895	0.4512	0.9478	0.8347-1.039
8	rs62622371	SLC7A2	0.1392	0.1464	0.4576	0.9429	1.166-2.432
3	rs1129183	LZTFL1	0.07822	0.0838	0.4651	0.9278	0.7468-1.712
6	rs2227956	HSPA1L	0.1695	0.1621	0.478	1.055	0.773-1.044
1	rs2230301	EPRS	0.1948	0.1874	0.4995	1.049	0.9138-1.139
13	rs11616795	ATP11A	0.09929	0.1049	0.5038	0.9404	0.8627-1.718
4	rs114282228	REST	0.02197	0.02477	0.5074	0.8847	0.9829-1.234
2	rs7587534	VWA3B	0.03701	0.03364	0.5171	1.104	0.9844-1.236
6	rs11159	RNASET2	0.06426	0.06862	0.5276	0.9322	0.963-1.212
22	rs4823850	CELSR1	0.06039	0.06459	0.5342	0.9308	0.9908-1.248
18	rs9965081	C18orf34	0.06054	0.06444	0.5603	0.9355	0.9807-1.234
4	rs151041998	AMTN	0.01901	0.01685	0.5615	1.131	0.6778-1.057
4	rs7689099	NEIL3	0.1193	0.1142	0.5666	1.051	0.848-1.207
11	rs61755080	CD6	0.06122	0.06449	0.631	0.9461	0.8789-1.115
7	rs34407351	SLC26A3	0.04483	0.04758	0.6364	0.9395	0.8603-1.227

8	rs7813708	FAM83A	0.1986	0.2037	0.647	0.9685	0.8184-1.489
1	rs41310132	CFHR2	0.01311	0.01457	0.6496	0.8982	0.7607-1.296
2	rs8192506	DBI	0.04185	0.03955	0.6745	1.061	1.029-1.657
19	rs3752135	SIGLEC12	0.1504	0.1546	0.6752	0.9682	0.7254-1.217
17	rs4796077	SLFN11	0.03151	0.03355	0.6783	0.9372	0.7084-1.403
10	rs41289305	CUBN	0.1715	0.1758	0.6789	0.9701	0.6981-1.319
14	rs10146482	NEK9	0.4853	0.4804	0.7231	1.02	0.7852-1.126
2	rs77424145	GEN1	0.03113	0.03274	0.7436	0.9492	0.7455-1.119
12	rs12368787	CCDC38	0.1093	0.1066	0.7646	1.028	0.9552-1.312
19	rs243383	CHAF1A	0.01767	0.01869	0.7856	0.9446	0.7157-2.136
6	rs766773	C6orf114	0.000311	9 0.0004545	0.7888	0.6861	0.8054-1.397
6	rs34109856	ENPP5	0.05442	0.05275	0.7926	1.033	0.913-1.258
6	rs3179003	NCR3	0.02467	0.02578	0.7983	0.9557	0.8445-1.111
9	rs1140064	IKBKAP	0.02945	0.03065	0.7996	0.9596	0.7335-1.345
4	rs1047626	SLC30A9	0.2412	0.2383	0.8118	1.016	0.7973-1.235
1	rs198400	CLCN6	0.001567	0.001828	0.8182	0.8571	0.5584-0.9671
2	rs61742338	TRAF3IP1	0.03258	0.03156	0.8355	1.033	0.5771-0.9935
8	rs13259948	SLC7A2	0.2173	0.215	0.8365	1.014	0.5628-0.9732
15	rs111633028	PLA2G4F	0.06305	0.06438	0.8433	0.9778	0.9096-1.453
1	rs41311191	INPP5B	0.1288	0.1305	0.8611	0.9855	0.7496-1.159
4	rs2289043	UNC5C	0.3066	0.3087	0.8703	0.9901	0.8375-1.618
2	rs56181206	MYO3B	0.03027	0.03091	0.8941	0.9788	0.7144-1.341
3	rs35362954	LRRN1	0.03348	0.03414	0.896	0.9799	0.7952-1.279
7	rs10487075	ZNF804B	0.1081	0.107	0.8968	1.012	0.887-1.245
17	rs1800215	COL1A1	0.01387	0.01351	0.9124	1.027	0.7665-1.304
11	rs34511735	USP47	0.01963	0.02002	0.9197	0.9801	0.5653-1.427
20	rs150682895	DNMT3B	0.01245	0.01215	0.9221	1.025	0.6497-1.047
2	rs35565687	SLC25A12	0.0558	0.05535	0.944	1.009	0.3113-1.102
7	rs269243	DENND2A	0.06728	0.06776	0.9463	0.9925	0.6316-1.63
1	rs116081500	EPRS	0.01347	0.01328	0.9518	1.015	0.9133-1.204
7	rs34943973	ATP5J2-PTCD1	0.04387	0.04417	0.9581	0.9929	0.7413-1.062
12	rs17755373	WNK1	0.009694	0.009554	0.9589	1.015	0.8398-2.251
11	rs61740182	GLB1L2	0.03321	0.03342	0.9653	0.9933	0.6337-1.072
12	rs74918182	PWP1	0.02602	0.0261	0.9856	0.9968	0.9189-1.21
1	rs12128607	PRG4	0.0444	0.04441	0.9983	0.9997	0.7584-1.408

Association analysis: German population centenarian subset							
Chr	dbSNP ID	Gene	MAF cases n=745	MAF controls n=1,104	P_{CCA}	OR	95% CI
1	rs10927851	FBLIM1	0.2896	0.3429	0.0009571	0.7811	0.6745-0.9046
11	rs34108746	PRG3	0.053	0.07477	0.01022	0.6926	0.5226-0.9179
5	rs61757629	NAIP	0.03151	0.01896	0.01559	1.683	1.099-2.578
11	rs78489201	TNKS1BP1	0.05579	0.0757	0.02021	0.7214	0.5471-0.9513
11	rs7128017	HPS5	0.1288	0.1075	0.05181	1.227	0.9981-1.509
4	rs17694522	SLC10A6	0.05331	0.06824	0.06581	0.7688	0.5806-1.018
14	rs9944035	SYNE2	0.06286	0.07832	0.08077	0.7893	0.6049-1.03
15	rs4843075	AKAP13	0.3703	0.3419	0.08191	1.132	0.9844-1.301
15	rs4075254	AKAP13	0.3696	0.3426	0.09289	1.125	0.9806-1.291
14	rs34711402	AKAP6	0.03134	0.04212	0.09363	0.7356	0.5131-1.055
15	rs2061824	AKAP13	0.3698	0.343	0.09625	1.124	0.9794-1.289
13	rs35719359	PCCA	0.06275	0.05	0.09637	1.272	0.9573-1.69
22	rs4823850	CELSR1	0.05139	0.06459	0.1006	0.7845	0.5869-1.049
15	rs7162168	AKAP13	0.3701	0.3444	0.1147	1.119	0.9731-1.286
6	rs17054318	PLEKHG1	0.03653	0.04745	0.117	0.7613	0.5408-1.072
1	rs142088763	MIA3	0.01701	0.01099	0.1209	1.557	0.8858-2.737
6	rs17080410	PLEKHG1	0.03696	0.04762	0.125	0.7676	0.547-1.077
15	rs4843074	AKAP13	0.3672	0.3428	0.1312	1.112	0.9686-1.278
16	rs11641523	CENPN	0.03197	0.02381	0.1361	1.354	0.9076-2.02
6	rs61742396	PLEKHG1	0.03711	0.04678	0.1557	0.7854	0.5623-1.097
6	rs9262151	MDC1	0.01355	0.008748	0.1663	1.557	0.8278-2.927
14	rs3093921	PARP2	0.02859	0.02148	0.1758	1.341	0.8756-2.054
10	rs41289305	CUBN	0.1599	0.1758	0.2072	0.892	0.7467-1.065
1	rs17356051	USP21	0.06421	0.05454	0.2223	1.189	0.8999-1.572
3	rs1129183	LZTFL1	0.07263	0.0838	0.2259	0.8562	0.6658-1.101
1	rs61740848	SERTAD4	0.006073	0.009554	0.2494	0.6334	0.2893-1.387
1	rs34603401	H6PD	0.1819	0.1671	0.252	1.108	0.9294-1.322
3	rs6801634	SETMAR	0.1354	0.1489	0.2601	0.895	0.7378-1.086
17	rs4796077	SLFN11	0.02721	0.03355	0.2788	0.8058	0.5448-1.192
14	rs10146482	NEK9	0.4623	0.4804	0.2893	0.9301	0.8133-1.064
1	rs12128607	PRG4	0.03736	0.04441	0.3031	0.8349	0.592-1.177
11	rs35090414	PRDM11	0.102	0.1122	0.3331	0.8994	0.7254-1.115
18	rs9965081	C18orf34	0.05669	0.06444	0.3389	0.8725	0.6597-1.154
1	rs6025	F5	0.03171	0.02639	0.3412	1.208	0.8178-1.786
1	rs2289189	RAB3GAP2	0.09717	0.1064	0.3672	0.9042	0.7265-1.125
4	rs28661939	PRSS12	0.1404	0.1302	0.3776	1.09	0.8997-1.321
1	rs41310132	CFHR2	0.01822	0.01457	0.3879	1.255	0.7487-2.103
2	rs8192506	DBI	0.04521	0.03955	0.3999	1.15	0.8304-1.593
22	rs2272843	MOV10L1	0.1547	0.1651	0.4079	0.9254	0.7701-1.112

2	rs6755571	UGT1A4	0.04292	0.04876	0.411	0.8748	0.6358-1.204
1	rs2230301	EPRS	0.1982	0.1874	0.4149	1.072	0.9072-1.266
16	rs8017	TCEB2	0.4638	0.4776	0.4158	0.9464	0.8289-1.081
19	rs243383	CHAF1A	0.02235	0.01869	0.4462	1.02	0.7502-1.919
1	rs41311191	INPP5B	0.122	0.1305	0.4568	0.9262	0.7567-1.134
2	rs61742338	TRAF3IP1	0.03605	0.03156	0.4586	1.148	0.7972-1.652
14	rs741842	PAPLN	0.1511	0.1601	0.4612	0.9338	0.7783-01. Dez
4	rs7689099	NEIL3	0.1221	0.1142	0.4632	1.079	0.8804-1.323
6	rs34109856	ENPP5	0.04749	0.05275	0.4815	0.8952	0.6575-1.219
12	rs74918182	PWP1	0.02245	0.0261	0.4852	0.8569	0.5553-1.323
16	rs1052276	PRSS54	0.1034	0.1106	0.4911	0.9276	0.7489-1.149
3	rs2228561	COL7A1	0.1267	0.1339	0.525	0.9381	0.7703-1.142
6	rs2227956	HSPA1L	0.1701	0.1621	0.527	1.06	0.8857-1.268
11	rs61755080	CD6	0.05927	0.06449	0.5278	0.9141	0.6916-1.208
8	rs13259948	SLC7A2	0.2238	0.215	0.5282	1.053	0.8963-1.238
13	rs11616795	ATP11A	0.1116	0.1049	0.5299	1.071	0.8639-1.329
2	rs17512204	CCDC93	0.07582	0.08151	0.5383	0.9245	0.7199-1.187
4	rs114282228	REST	0.02162	0.02477	0.5425	0.8701	0.5557-1.362
2	rs56181206	MYO3B	0.03441	0.03091	0.5555	1.117	0.7725-1.616
11	rs34511735	USP47	0.02289	0.02002	0.5595	1.147	0.7237-1.817
3	rs11539148	QARS	0.0579	0.06221	0.5919	0.9264	0.7006-1.225
10	rs3814182	MYPN	0.4352	0.4439	0.6038	0.9654	0.8453-1.103
12	rs12368787	CCDC38	0.1119	0.1066	0.6218	1.055	0.8519-1.308
2	rs7587534	VWA3B	0.03068	0.03364	0.6244	0.9092	0.621-1.331
20	rs150682895	DNMT3B	0.01399	0.01215	0.6335	1.153	0.6413-2.074
7	rs10487075	ZNF804B	0.102	0.107	0.634	0.948	0.7607-1.181
8	rs62622371	SLC7A2	0.1408	0.1464	0.6383	0.9557	0.7913-1.154
6	rs17827619	FAM184A	0.01093	0.009407	0.6529	1.164	0.6009-2.253
19	rs3752135	SIGLEC12	0.1492	0.1546	0.6539	0.9587	0.7971-1.153
15	rs111633028	PLA2G4F	0.06073	0.06438	0.6545	0.9396	0.7151-1.235
16	rs2241414	PRSS54	0.1061	0.1109	0.6583	0.9525	0.7677-1.182
11	rs1551886	ALDH3B2	0.08919	0.09344	0.6618	0.9501	0.7553-1.195
11	rs61740182	GLB1L2	0.03605	0.03342	0.6693	1.082	0.7546-1.55
13	rs11551114	MIPEP	0.1415	0.1464	0.6801	0.9611	0.7959-1.161
1	rs16849342	NAV1	0.05601	0.0592	0.6843	0.9429	0.71-1.252
4	rs151041998	AMTN	0.01865	0.01685	0.6885	1.109	0.6685-1.84
17	rs117190076	RPH3AL	0.06098	0.05789	0.702	1.057	0.7962-1.403
12	rs17755373	WNK1	0.01083	0.009554	0.705	1.135	0.59-2.181
7	rs34943973	ATP5J2-PTCD1	0.04155	0.04417	0.706	0.938	0.6727-1.308
22	rs450046	PRODH	0.07355	0.07052	0.7268	1.046	0.8114-1.349
1	rs198400	CLCN6	0.001361	0.001828	0.7319	0.7439	0.1361-4.066
6	rs3179003	NCR3	0.02751	0.02578	0.7508	1.069	0.7085-1.613

11	rs10791893	OR5T2	0.1255	0.129	0.762	0.9694	0.7929-1.185
8	rs7813708	FAM83A	0.1997	0.2037	0.7695	0.9754	0.8254-1.152
6	rs766773	C6orf114	0.0006748	0.0004545	0.7785	1.485	0.0928-23.76
1	rs116081500	EPRS	0.01429	0.01328	0.7972	1.077	0.6118-1.896
6	rs11159	RNASET2	0.06667	0.06862	0.818	0.9695	0.7449-1.262
2	rs77424145	GEN1	0.03138	0.03274	0.8214	0.9571	0.6541-1.04
4	rs2289043	UNC5C	0.3052	0.3087	0.8226	0.9835	0.8506-1.137
2	rs35565687	SLC25A12	0.05374	0.05535	0.8337	0.9692	0.7241-1.297
2	rs3795958	CCDC164	0.1872	0.1895	0.8562	0.9845	0.8318-1.165
17	rs1800215	COL1A1	0.01298	0.01351	0.8903	0.9599	0.5362-1.718
4	rs1047626	SLC30A9	0.2402	0.2383	0.8986	1.01	0.8636-1.182
3	rs35362954	LRRN1	0.03487	0.03414	0.9075	1.022	0.7086-1.474
11	rs17224476	OR51E1	0.1073	0.1061	0.9099	1.013	0.815-1.258
2	rs934945	PER2	0.1946	0.1934	0.9289	1.008	0.8506-1.194
11	rs2282537	POU2F3	0.1286	0.1296	0.9291	0.9911	0.8136-1.207
15	rs35285091	GANC	0.01429	0.01419	0.9817	1.007	0.5761-1.758
7	rs269243	DENND2A	0.06764	0.06776	0.9894	0.9982	0.7649-1.303
9	rs1140064	IKBKAP	0.03061	0.03065	0.9949	0.9987	0.6805-1.466
7	rs34407351	SLC26A3	0.04762	0.04758	0.9952	1.001	0.7338-1.365
19	rs799193	ZNF441	0.2139	0.2138	0.996	1	0.8514-1.176

Min AF: minor allele frequency**;**
P_{CCA}, p-value obtained from an allele-based case–control comparison, using a $\chi 2$-test with 1 degree of freedom (df)**;**
OR: Odds ratio for attaining old age with the minor allele in controls as reference allele;
95% CI, 95% confidence interval for OR

Table 12-8: Variants overlaid with genes involved in insulin pathway/mTOR pathway

Individual	Pathway	Chr	Start	dbSNP135	Gene	MAF*
(3) French female	Insulin / mTOR	1	9777599	rs61755420	PIK3CD	0.03
(2) German female	Insulin	5	176308303	rs145827614	HK3	0.0158
(1) German female	Insulin	9	134501369		RAPGEF1	0.000363
(3) French female	Insulin	10	70987024	rs145939161	HKDC1	0.0225
(2) German male	Insulin	10	71008316		HKDC1	0.000349
(2) German male	Insulin	10	97154424	rs147078270	SORBS1	0.0026
(2) German male	Insulin	11	67200812	rs55987642	RPS6KB2	0.04
(2) German male	Insulin	12	109604776	rs146426104	ACACB	0.0013
(1) German female	Insulin	12	21695439	rs61733199	GYS2	0.02
(1) German female	mTOR	12	132399687	rs12827141	ULK1	0.01
(1) German female	Insulin	14	55510166		SOCS4	-
(1) German female	Insulin	16	47684830	rs34667348	PHKB	0.01
(3) French female	mTOR	17	19741877	rs34670978	ULK2	0.03
(3) French female, (6)German male	Insulin	19	45296806	rs3208856	CBLC	0.02
(3) French female, (6) German male	Insulin	X	18972497	rs17313469	PHKA2	0.03

*MAF: minor allele frequency based on 1000Genomes or NHLBI Exome Sequencing Project database

Table 12-9: Top scores of coding variants “effective” in seven or eight prediction tools

Chr-Start	dbSNP135	Gene	Substitution	MAF*	No. of samples	No. of Tools worked [0;8]	No of tools predicted damaging effect [0;8]	Tool that did not work
1_172411189		PIGC	R192C	-	1	8	8	-
1_207867854	rs41303261	CR1L	C207Y	0.054878	1	8	8	-
1_42628591	rs142157365	GUCA2A	C112S	0.001395	1	8	8	-
3_183960695		ALG3	R354C	-	1	8	8	-
3_186338564	rs35457250	AHSG	R317C	0.00609756	1	8	8	-
3_75786243	rs139633377	ZNF717	C844Y	-	2	7	7	SNPs&GO
3_75787240	rs138742243	ZNF717	C512R	-	2	7	7	SNPs&GO
3_75787405	rs141106119	ZNF717	G457R	-	3	7	7	SNPs&GO
3_75787416	rs10442977	ZNF717	C453Y	-	2	7	7	SNPs&GO
3_75787996	rs142456725	ZNF717	G260R	-	3	7	7	SNPs&GO
4_190878563	rs137858630	FRG1	A148D	-	3	8	8	-
5_149360630	rs78676079	SLC26A2	R492W	0.0426829	1	8	8	-
6_129571272	rs36044314	LAMA2	G600R	0.0121951	1	8	8	-
6_132206079	rs28933977	ENPP1	R774C	0.0487805	1	8	8	-
6_49663567	rs36069724	CRISP2	C196R	0.00609756	1	8	8	-
7_142460339		PRSS1	C171Y	-	2	8	8	-
8_17166805		MTMR7	G378R	-	1	8	8	-
9_117165068		DFNB31	D897A	-	1	8	8	-
10_118351309		PNLIPRP1	G26R	-	1	8	8	-
10_50943387		OGDHL	R974W	-	1	8	8	-
11_12379949	rs34898047	MICALCL	R671C	0.00609756	1	7	7	SNPs&GO
11_55563336	rs76383258	OR5D14	Q102L	0.103659	1	8	8	-
12_123345509	rs34149579	HIP1R	C938F	0.0426829	2	8	8	-
12_53217726	rs116963732	KRT79	E364V	0.00609756	1	8	8	-
15_80191338	rs139874813	ST20	C59R	0.00609756	1	7	7	SNPs&GO
17_15522455	rs188826833	CDRT1	W124C	0.000582	1	7	7	SNPs&GO
19_18502861	rs34666550	LRRC25	C285Y	0.0243902	1	8	8	-
19_43430060		PSG7	G370R	0.001744	1	7	7	SIFT
19_48519241	rs3745751	ELSPBP1	C100W	0.0487805	1	8	8	-
20_10622501	rs35761929	JAG1	P871R	0.0731707	2	8	8	-

*MAF: minor allele frequency based on 1000Genomes or NHLBI Exome Sequencing Project database

Table 12-10: Top scores of variants effective in five or more tools and present in four or more individuals

dbSNP135	Substitution	Gene	MAF*	No. of samples	No. of Tools worked [0;8]	No. of tools predicted damaging effect [0;8]	No. of tools predicted neutral effect [0;8]	Tool that did not work
rs76417519	R671W	IGSF3	-	6	6	5	1	SNPs&GO, Polyphen-2
rs61955126	C302R	SETD8	-	6	6	5	1	SNPs&GO, Polyphen-2
rs61786577	R476C	IGSF3	-	6	6	6	-	SNPs&GO, Polyphen-2
rs73979896	A185V	KCNJ12, KCNJ18	-	5	8	6	2	-
rs1782241	Y120C	OR2T27	-	4	8	6	2	-
rs76780359	E191K	NCOR1	-	4	8	6	2	-
rs75029097	G145S	KCNJ12, KCNJ18	-	4	8	6	2	-
rs74496366	M475R	CDC27	-	4	6	5	1	SNPs&GO, Polyphen-2
rs143791478	C565S	ZNF717	-	4	7	6	1	SNPs&GO
rs74776730	F348V	ZNF717	-	4	7	6	1	SNPs&GO
rs9885916	Y345C	PRIM2	-	4	7	6	1	SIFT
rs9885751	R350C	PRIM2	-	4	7	6	1	SIFT
rs76265595	E139K	KCNJ12, KCNJ18	-	4	8	7	1	-
rs2310687	P592A	OTOP1	-	4	8	7	1	-

*MAF: minor allele frequency based on 1000Genomes or NHLBI Exome Sequencing Project database

Table 12-11: Low-frequency variants selected on various criteria for genotyping

Chr	Start	dbSNP135	Gene	MAF*	Selection criteria
1	888659	rs3748597	NOC2L	0.0731707	variants present in 6 or 5 individuals
1	3807593	rs4274008	C1orf174	0.0121951	variants present in 6 or 5 individuals
1	15812432	rs6429745	CELA2B	0.000233	variants present in 6 or 5 individuals
1	33065947	rs704886	ZBTB8A	0.000116	variants present in 6 or 5 individuals
1	33160878	rs360042	SYNC	0.0304878	variants present in 6 or 5 individuals
1	42628591	rs142157365	GUCA2A	0.001395	Top hit of SNVs from whole list
1	48697733	rs212991	SLC5A9	0.001395	variants present in 6 or 5 individuals
1	62734089	rs17123306	KANK4	0.0182927	GWAS associated hit region
1	117142641	rs76417519	IGSF3	-	Top scores of SNVs found in 4/5/6/ individuals
1	172411189		PIGC	-	Top hit of SNVs from whole list
1	207867854	rs41303261	CR1L	0.054878	Top hit of SNVs from whole list
2	27167536	rs200305979	DPYSL5	0.000116	GWAS associated hit region

2	179581897		TTN	0.000363	GWAS associated hit region
2	234638245	rs61764030	UGT1A3	0.00609756	GWAS associated hit region
3	11643465	rs2276749	VGLL4	0.0304878	variants present in 6 or 5 individuals
3	33055721	rs4302331	GLB1	0.001806	variants present in 6 or 5 individuals
3	183960695		ALG3	-	Top hit of SNVs from whole list
3	186338564	rs35457250	AHSG	0.00609756	Top hit of SNVs from whole list
4	95578588	rs13107595	PDLIM5	0.006279	variants present in 6 or 5 individuals
5	36269551	rs1035480	RANBP3L	0.0182927	variants present in 6 or 5 individuals
5	43509348	rs6872851	C5orf34	-	variants present in 6 or 5 individuals
5	149360630	rs78676079	SLC26A2	0.0426829	Top hit of SNVs from whole list
6	49663567	rs36069724	CRISP2	0.00609756	Top hit of SNVs from whole list
6	79708000	rs7747479	PHIP	0.0426829	variants present in 6 or 5 individuals
6	129571272	rs36044314	LAMA2	0.0121951	Top hit of SNVs from whole list
6	132206079	rs28933977	ENPP1	0.0487805	Top hit of SNVs from whole list
7	25267963	rs886354	NPVF	0.0731707	variants present in 6 or 5 individuals
7	50611735	rs6264	DDC	-	variants present in 6 or 5 individuals
7	56136260	rs4245575	SUMF2	0.00609756	variants present in 6 or 5 individuals
7	64291991	rs4236203	ZNF138	0.000233	variants present in 6 or 5 individuals
8	22886020	rs13265018	TNFRSF10B	0.0853659	variants present in 6 or 5 individuals
10	99240758	rs2275586	MMS19	0.0182927	variants present in 6 or 5 individuals
11	7059960	rs12801277	NLRP14	0.000466	variants present in 6 or 5 individuals
11	12379949	rs34898047	MICALCL	0.00609756	Top hit of SNVs from whole list
11	20529886	rs6483700	PRMT3	-	variants present in 6 or 5 individuals
11	34152939	rs2957516	NAT10	-	variants present in 6 or 5 individuals
12	53217726	rs116963732	KRT79	0.00609756	Top hit of SNVs from whole list
12	123345509	rs34149579	HIP1R	0.0426829	Top hit of SNVs from whole list
12	123892095	rs61955126	SETD8	-	Top scores of SNVs found in 4/5/6/ individuals
12	129299446	rs33990080	SLC15A4	0.103659	variants present in 6 or 5 individuals
15	80191338	rs139874813	ST20	0.00609756	Top hit of SNVs from whole list
16	11002927	rs7197779	CIITA	0,09	variants present in 6 or 5 individuals
17	15522455	rs188826833	CDRT1	0.000582	Top hit of SNVs from whole list
17	44128052		KANSL1	NA	GWAS associated hit region
17	80789468	rs35653278	ZNF750	0.0609756	GWAS associated hit region
19	18502861	rs34666550	LRRC25	0.0243902	Top hit of SNVs from whole list
19	48519241	rs3745751	ELSPBP1	0.0487805	Top hit of SNVs from whole list
20	10622501	rs35761929	JAG1	0.0731707	Top hit of SNVs from whole list

*MAF: minor allele frequency based on 1000Genomes or NHLBI Exome Sequencing Project database

Table 12-12: **Association statistics for 48 SNVs** selected based on GWAS hit regions and prediction tools genotyped using the Sequenom technology in German LLI (n=1,610), centenarian subset (n=748) and younger controls (n=1,104)

Association analysis: German population LLI							
Chr	dbSNP ID	Gene	MAF cases n=1,610	MAF controls n=1,104	P_{CCA}	OR	95% CI
20	rs35761929	JAG1	0.1246	0.07678	3.7e-08	1.712	1.411-2.076
17	rs35653278	ZNF750	0.1259	0.1002	0.004491	1.294	1.083-1.546
11	rs34898047	MICALCL	0.009036	0.0176	0.006923	0.5089	0.3089-0.8384
6	rs36069724	CRISP2	0.01839	0.02652	0.05019	0.688	0.4722-1.002
7	rs886354	NPVF	0.08481	0.09963	0.06713	0.8375	0.6926-1.013
1	rs3748597	NOC2L	0.06546	0.05417	0.09298	1.223	0.9667-1.548
5	rs78676079	SLC26A2	0.026	0.03343	0.1169	0.772	0.5582-1.068
3	rs2276749	VGLL4	0.05064	0.06008	0.1438	0.8345	0.6546-1.064
2	rs179581897	TTN	0.001638	0.0004625	0.2173	3.546	0.414-30.37
3	rs183960695	ALG3	0.0006614	0		0.2326	-
7	rs6264	DDC	0	0.0004625	0.2347	0	0
6	rs7747479	PHIP	0.05182	0.05735	0.3845	0.8981	0.7049-1.144
7	rs4245575	SUMF2	0.0009785	0.00185	0.396	0.5284	0.1181-2.363
2	rs200305979	DPYSL5	0.0003347	0	0.4	-	-
12	rs61955126	SETD8	0.0003218	0	0.4084	-	-
19	rs34666550	LRRC25	0.01152	0.009174	0.4333	1.259	0.7068-2.242
2	rs61764030	UGT1A3	0.4894	0.4995	0.4914	0.9601	0.8551-1.078
5	rs1035480	RANBP3L	0.02146	0.0189	0.5243	1.138	0.7638-1.697
19	rs3745751	ELSPBP1	0.05892	0.06238	0.6114	0.9411	0.7446-1.189
10	rs2275586	MMS19	0.02106	0.02309	0.623	0.9103	0.6256-1.324
12	rs34149579	HIP1R	0.06054	0.06333	0.6829	0.9531	0.7568-1.2
1	rs360042	SYNC	0.0388	0.0408	0.7192	0.949	0.7136-1.262
1	rs142157365	GUCA2A	0.003896	0.003305	0.7286	Jan 18	0.4636-3.001
6	rs28933977	ENPP1	0.03268	0.03104	0.7446	1.055	0.7653-1.454
8	rs13265018	TNFRSF10B	0.09302	0.09074	0.7817	1.028	0.8472-1.247
3	rs4302331	GLB1	0.001641	0.001388	0.8177	1.183	0.2825-4.957
12	rs33990080	SLC15A4	0.09045	0.09198	0.8534	0.9817	0.8072-1.194
1	rs41303261	CR1L	0.05431	0.05324	0.8662	1.021	0.8004-1.303
4	rs13107595	PDLIM5	0.003032	0.002838	0.8999	1.069	0.3798-3.007
3	rs35457250	AHSG	0.008731	0.008988	0.9231	0.9712	0.5361-1.759

6	rs36044314	LAMA2	0.01437	0.01464	0.9378	0.9816	0.6165-1.563
15	rs139874813	ST20	0.002262	0.002367	0.9382	0.9556	0.3029-3.015
1	rs4274008	C1orf174	-	-	-	-	failed
1	rs6429745	CELA2B	0	0	-	-	monomorphic
1	rs704886	ZBTB8A	0	0	-	-	monomorphic
1	rs212991	SLC5A9	0	0	-	-	monomorphic
1	rs17123306	KANK4	-	-	-	-	failed
1	rs76417519	IGSF3	-	-	-	-	failed
1	1_172411189	PIGC	0	0	-	-	monomorphic
5	rs6872851	C5orf34	0	0	-	-	monomorphic
7	rs4236203	ZNF138	-	-	-	-	failed
11	rs12801277	NLRP14	0	0	-	-	monomorphic
11	rs6483700	PRMT3	0	0	-	-	monomorphic
11	rs2957516	NAT10	0	0	-	-	monomorphic
12	rs116963732	KRT79	0	0	-	-	monomorphic
16	rs7197779	CIITA	0	0	-	-	monomorphic
17	rs115420242	CDRT1	0	0	-	-	monomorphic
17	17_44128052	KANSL1	0	0	-	-	monomorphic
Association analysis: German population centenarian subset							
Chr	dbSNP ID	Gene	MAF cases n=745	MAF controls n=1,104	P_{CCA}	OR	95% CI
20	rs35761929	JAG1	0.1096	0.07678	0.0008297	1.481	1.175-1.866
17	rs35653278	ZNF750	0.1302	0.1002	0.005778	1.344	1.089-1.659
11	rs34898047	MICALCL	0.007891	0.0176	0.01569	0.4439	0.2257-0.8732
3	rs183960695	ALG3	0.001435	0	0.07867	-	-
7	rs886354	NPVF	0.08285	0.09963	0.09329	0.8164	0.644-1.035
6	rs36069724	CRISP2	0.0194	0.02652	0.1751	0.7262	0.4565-1.155
1	rs3748597	NOC2L	0.06456	0.05417	0.1964	1.205	0.9077-1.6
5	rs78676079	SLC26A2	0.02586	0.03343	0.2006	0.7677	0.5116-1.152
10	rs2275586	MMS19	0.01697	0.02309	0.2101	0.7305	0.4462-1.196
2	rs200305979	DPYSL5	0.0007163	0	0.2182	-	-
12	rs61955126	SETD8	0.0007112	0	0.219	-	-
12	rs34149579	HIP1R	0.05429	0.06333	0.2681	0.849	0.6354-1.135
3	rs2276749	VGLL4	0.05165	0.06008	0.2911	0.8521	0.6329-1.147
7	rs4245575	SUMF2	0.0007123	0.00185	0.375	0.3845	0.04294-3.444
7	rs6264	DDC	0	0.0004625	0.4203	0	-

19	rs34666550	LRRC25	0.01194	0.009174	0.4394	1.305	0.6632-2.569
19	rs3745751	ELSPBP1	0.05692	0.06238	0.5073	0.9071	0.68-1.21
8	rs13265018	TNFRSF10B	0.09726	0.09074	0.5161	1.08	0.8567-1.361
12	rs33990080	SLC15A4	0.08613	0.09198	0.5579	0.9305	0.7311-1.184
6	rs7747479	PHIP	0.06178	0.05735	0.5848	1.082	0.815-1.437
2	rs61764030	UGT1A3	0.4908	0.4995	0.6107	0.9658	0.8446-1.104
6	rs36044314	LAMA2	0.01648	0.01464	0.6645	1.128	0.6548-1.942
4	rs13107595	PDLIM5	0.003608	0.002838	0.6909	1.272	0.3875-4.176
1	rs41303261	CR1L	0.05587	0.05324	0.7344	1.052	0.7838-1.413
2	rs179581897	TTN	0.0007143	0.0004625	0.7567	1.545	0.09654-24.72
1	rs360042	SYNC	0.04239	0.0408	0.8175	1.041	0.7418-1.46
5	rs1035480	RANBP3L	0.01793	0.0189	0.8348	0.9478	0.5724-1.569
15	rs139874813	ST20	0.002122	0.002367	0.8805	0.896	0.2138-3.755
1	rs142157365	GUCA2A	0.003556	0.003305	0.9003	1.076	0.3409-3.398
3	rs35457250	AHSG	0.008596	0.008988	0.9034	0.956	0.4626-1.976
6	rs28933977	ENPP1	0.03175	0.03104	0.9068	1.024	0.693-1.512
3	rs4302331	GLB1	0.001433	0.001388	0.9721	1.033	0.1723-6.187
1	rs4274008	C1orf174	-	-	-	-	failed
1	rs6429745	CELA2B	0	0	-	-	monomorphic
1	rs704886	ZBTB8A	0	0	-	-	monomorphic
1	rs212991	SLC5A9	0	0	-	-	monomorphic
1	rs17123306	KANK4	-	-	-	-	failed
1	rs76417519	IGSF3	-	-	-	-	failed
1	1_172411189	PIGC	0	0	-	-	monomorphic
5	rs6872851	C5orf34	0	0	-	-	monomorphic
7	rs4236203	ZNF138	-	-	-	-	failed
11	rs12801277	NLRP14	0	0	-	-	monomorphic
11	rs6483700	PRMT3	0	0	-	-	monomorphic
11	rs2957516	NAT10	0	0	-	-	monomorphic
12	rs116963732	KRT79	0	0	-	-	monomorphic
16	rs7197779	CIITA	0	0	-	-	monomorphic
17	rs115420242	CDRT1	0	0	-	-	monomorphic
17	17_44128052	KANSL1	0	0	-	-	monomorphic

For abbreviations refer to Supplementary Table 11-7