Chapter 1:

Introduction

Parts of this work has been used as a contribution to the textbook:

Longevity Genes: A Blueprint for Aging

Exome and Whole Genome Sequencing In Aging and Longevity

Erik B. van den Akker[1,2], Joris Deelen[1,3], P. Eline Slagboom[1,3], Marian Beekman[1,3]

[1.] Molecular Epidemiology, Leiden University Medical Center, Leiden, The Netherlands

[2.] The Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands

[3.] Netherlands Consortium for Healthy Ageing, Leiden, Netherlands

Springer: In press

1. Aging: a Common Suspect in Common Disease

A steadily growing life expectancy of the general western population[1] urges further research into age-associated mechanisms responsible for the gradual decline of health throughout the course of life. Calendar age is the major risk factor for the onset and progression of virtually all common disease affecting the general population of the western world today[2], suggesting that processes of aging are involved in the etiology of many diseases. Indeed, aging is characterized by a progressive and systemic loss of function, which gradually leads to a state of senescence on the cellular, tissular and organismal level, thus affecting the general capacity for maintaining bodily homeostasis[3]. Though seemingly inevitable, aging does not occur at an equal pace across species[4] or even within our own species. Whereas some experience an accelerated rate of aging, as exemplified by patients suffering from progeroid syndromes[5], others seem capable of delaying or evading at least some of the detrimental aspects of aging, as observed in members of long-lived families[6-8]. Hence, by studying the factors affecting the rate of aging, we expect to identify determinants that modulate the capacity for maintaining the bodily homeostasis as the common denominator of age-associated disease.

2. Factors Affecting the Rate of Human Aging

Unlike other traits, aging itself is not driven by any specific molecular mechanism per se, but instead seems to be the integrated result of all corrective and compensatory mechanisms failing to deal with the stochastic damage accumulated over life[9]. Despite its stochastic origin, the accumulation of damage does converge into some consistently observed processes characterizing the aging phenotype. In a landmark paper titled "The Hallmarks of Aging"[10] these processes of aging are comprehensively described and conceptualized around nine main processes co-occurring with aging (Figure 1). Though the causality of some of these nine hallmarks has yet not been irrefutably proven, each of them is likely to occur during aging and is thought to at least aggravate the consequences of aging by further contributing to the loss of the bodily homeostasis.

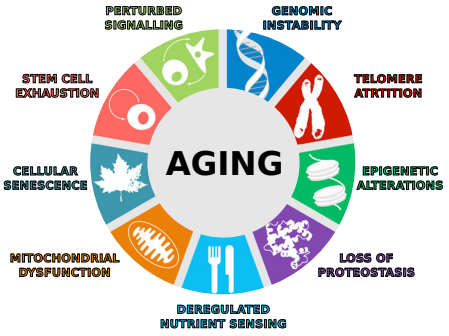

FIGURE 1: Nine recurrently observed processes that occur with aging. Processes that are commonly observed during aging are: genomic instability, telomere attrition, epigenetic alterations, loss of proteostasis, deregulated nutrient sensing, mitochondrial dysfunction, cellular senescence, stem cell exhaustion and perturbed signalling. Figure adapted from López-Otín *et al.*[10].

Thus far, human aging and its relation to health have predominantly been studied in the context of two parallel though complementary lines of research: biomarkers and genetics. Another source

of information comes from systems approaches mainly performed in cell systems and model organisms in which physiological processes related to aging can be perturbed, measured from numerous biological perspectives and integrated to understand the response of the system to the challenge. Here we will focus on biomarker and genetic research into human aging.

2.1 Biomarkers of human aging

Biomarker research is aimed at discovering quantitative parameters that mark biological age. Levels of such biomarkers do not only correspond with the absolute quantity of time that has passed (calendar age), but also mark the mechanisms responsible for the deteriorating health and increasing frailty that occurs with advancing age. Hence, as outlined by Deelen *et al.*[11], ideal biomarkers for biological aging should correlate with calendar age in cross-sectional or in longitudinal studies with repeated measurements, while also displaying correlations with established physiological parameters of health such as systolic blood pressure or insulin resistance. Furthermore, ideal biomarkers of aging are able to discriminate individuals with either accelerated (e.g. progeroid syndromes) or decelerated (e.g. familial longevity) aging phenotypes from those derived of the general population, and are prospective of future clinical endpoints such as morbidity and mortality. Hence the identification of ideal biomarkers of aging would enable us to objectively monitor the rate of biological aging of individuals and potentially allows us to differentiate and understand different mechanisms promoting aging.

2.1.1 Existing biomarkers of human aging: A compelling example of the use of biomarkers associated with health and aging employed in epidemiological research is the Framingham risk score (FRS), which is the estimated individual risk for development of a cardiovascular event within 10 years[12]. The FRS is a composite score taking into account blood pressure, total cholesterol level, HDL cholesterol level and smoking status. Future composite scores for aging should not only incorporate biomarkers indicative for cardiovascular health, but also for many other pathophysiological processes such as aging of the neuromuscular system. By studying aging populations or by comparing members of long-lived families with population controls, potential additional and independent biomarkers of aging are currently being investigated. Biomarkers that distinguish healthy from unhealthy aging groups are for example cortisol[13], free triiodothyronine[14] and fasting glucose serum levels[15,16]. However, risk prediction at the individual level on the basis of these biomarkers is not possible yet. Though each of these traits can be used to objectively assess particular aspects of the aging human system, it is not immediately apparent how they are caused by the hypothesized aging mechanisms listed earlier (Figure 1). Therefore, a challenge remains in translating the molecular events that occur during aging to the age-associated deterioration that only becomes apparent on the whole body

level and which is marked by the existing biomarkers for biological aging.

2.1.2 Omics-derived biomarkers of human aging: In contrast to the existing biochemical and physical parameters, use of genomic, metabolic or proteomic data sources may have the benefit that they directly probe at the molecular level with an unbiased approach. However, the construction of age-associated signatures that are both consistent as interpretable has proven to be challenging with these types of data sources. For instance, limited mutual overlap has been reported thus far for studies probing the aging transcriptome[17,18]. Possible reasons for this could lie in the variable technical circumstances under which these studies have been performed, but also the limited study sizes, low expected signal-to-noise ratios and the high tissue specificity are likely to contribute to the observed inconsistency. Some compelling similarities have been observed on the pathway level across tissues and even across species[19,20] incriminating amongst others electron transport chain and ribogenesis as potential aging promoting mechanisms. Hence, studies into the aging transcriptome have provided some interesting insights into the mechanisms promoting aging. However, significant progress in this field, let alone future translation to the clinic, is severely hampered by the large inconsistencies generally observed between studies.

Another popular omics platform for discovery of biomarkers of biological aging is Illumina's HumanMethylation450k BeadChip array, designed for probing the human methylome. Using this platform, highly robust and tissue independent methylation markers for chronological age have been identified[21,22]. However, whereas gene expression arrays provide interpretable though noisy age-associated signatures, methylation arrays provide highly predictive though poorly understood signatures of aging. Quite unexpectedly, loci coming from large-scale meta-analyses on age-associated changes of methylation levels hardly shed any insights in the age-associated changes in gene regulation, be it either by affecting the expression of nearby genes directly[21] or by targeting hub-genes in regulatory networks[23]. This remarkable absence of any relation with regulatory mechanisms thus questions the importance of DNA methylation changes in the biology of aging. Hence, it has been shown that methylation signatures are highly predictive of calendar age though as of yet are highly uninformative on the processes driving biological aging.

To conclude, many challenges still lie in the field of omics-based biomarkers for aging as the current combination of platforms and methods provide signatures that either lack the robustness or have as of yet a highly disputable role in the etiology of aging. Therefore, additional efforts should go into increasing our capacity to comprehend the results coming from such sources before aging processes observed at the molecular level can be translated to effects for health on the whole body level.

2.2 Genetics of human aging

Since lifespan regulation has a heritable component of approximately 25%[24,25] in the general population, the second branch of

aging research focuses on the identification of genetic determinants that specifically characterize cases exhibiting either accelerated (e.g. progeroid syndromes) or decelerated (e.g. human longevity) phenotypes of aging. Genetic studies into human longevity are mostly inspired by the findings of lifespan regulating genes using a systems approach in animal studies, such as the insulin-like receptor *daf-16*, initially discovered to modulate life span regulation in *C. elegans*[26]. Interestingly, many more genes in *C. elegans* and *D. melanogaster* that are functionally related to this homologue of human *FOXO3A* have been found to consistently modulate life span across multiple species[27]. Hence, such systems genetics approaches into aging provide valuable starting points for the search of genetic loci modulating the rate of human aging and life span regulation.

Novel loci for human aging and longevity may be identified by comparing the frequencies of common variants between long-lived cases and younger population controls in an association analysis. Such association analyses performed using either a candidate approach, or on a genome-wide scale (GWAS) has yielded thus far three robust and independently confirmed longevity loci: *FOXO3A*[28-31], *APOE*[32-36] and an intergenic locus on chromosome 5q33.3[37].

A relatively unexplored second option for obtaining longevity loci is to sequence the genome of extremely long-lived individuals for rare variants with a large predicted impact. Though very promising, this approach has thus far only been applied on a candidate gene basis in a cohort of long-lived individuals[38] or on a whole genome scale in very limited numbers of individuals[39-43], which makes its use for research into human aging at this point hard to assess.

A third source of potential human longevity loci might come from family-based studies (linkage analysis) with a history of extended survival[6,44]. Thus far, several linkage studies into longevity have been performed[45-48], however, none of the reported loci display any mutual consistency, nor have they been independently confirmed[36]. Hence, genetic studies into aging and life span regulation have known a very limited number of successes judged by the standards set in the genetics field and have been far less successful as compared to other commonly studied multifactorial traits.

2.3 Challenges and opportunities in aging research

Thus far many of the available genetic, transcriptomic, methylome and metabolome data sources on human aging have been analysed in isolation. Whereas this approach has lead to the identification of some biochemical biomarkers for aging, considerably less progress is made with the analysis of genomic data sources on aging. As a result, little is known how aging mechanisms on the molecular and cellular level affect health and aging on the whole body level. Reasons for the lagging insights derived from genomic data sources surpass the relative novelty of these data types, since similar tools for genomic research have been very successfully applied in studying other complex traits. For instance the era of GWAS has brought many novel loci for age-associated traits and diseases[49],

but not in human longevity research.

In effect, the analysis of genomics data on aging is hampered for two main reasons affecting either the discovery of molecular biomarkers or genetic markers. First, the stochastic nature of aging makes biomarker research into this field very distinct from studying other traits, as it is an intrinsically passive mechanism that acts on many processes in parallel and on all systemic levels simultaneously. Hence, much signal is expected to correlate in the analysis for aging biomarkers, though few molecular entities are actually independent or causal for the studied aging phenotype.

Secondly, investigation of the genetic determinants modulating health, the rate of aging and ultimately life span regulation is hampered by the extreme heterogeneity of the studied traits. This poses the possibility that the non-consistent results of genetic screens for life span regulation each constitute actual independent mechanisms for modulating the rate of biological aging.

Interestingly, both the issues of stochasticity and heterogeneity refer to a lack of power that can be solved by analysing data sources on aging jointly instead of analysing each of them in isolation, as is currently the standard. Hence, a huge opportunity lies in the application and development of methods for the integrated analysis of genomic aging data resources.

3. Approaches for Data Integration

Analyses of genomic data sources directed to investigating aging and life span regulation are especially prone to overfitting and therefore deserve special attention from a methodological point of view. An analysis is said to overfit when features are extracted from the data that do not reflect the general characteristics of the studied phenotype, but instead focus on irrelevant features that happen to coincide with the studied phenotype in that particular experimental setting only. Approaches with a reduced chance of fitting noise are indicated as robust and can generally be achieved by applying either two of the following concepts for data integration: the joint analysis of genomic data sources, or the incorporation of prior knowledge.

A very commonly used example of a joint analysis of genomic data sources is a so-called eQTL analysis, which is sometimes performed in addition to a normal GWAS or a whole genome expression analysis (Figure 2A). The aim of such an analysis is to determine whether SNPs influence the expression of (nearby) genes, hence the term expression Quantitative Trait Loci or eQTLs. Besides inferring clues for the mechanistic causality of an observed trait association, the rational for this approach was exemplified by a study of Nicolae *et al.*[50] showing that eQTLs, as a subset of all SNPs, are enriched for trait-associations. Incorporation of eQTL analyses is thus likely to reduce false positive findings, next to providing additional mechanistic insights.

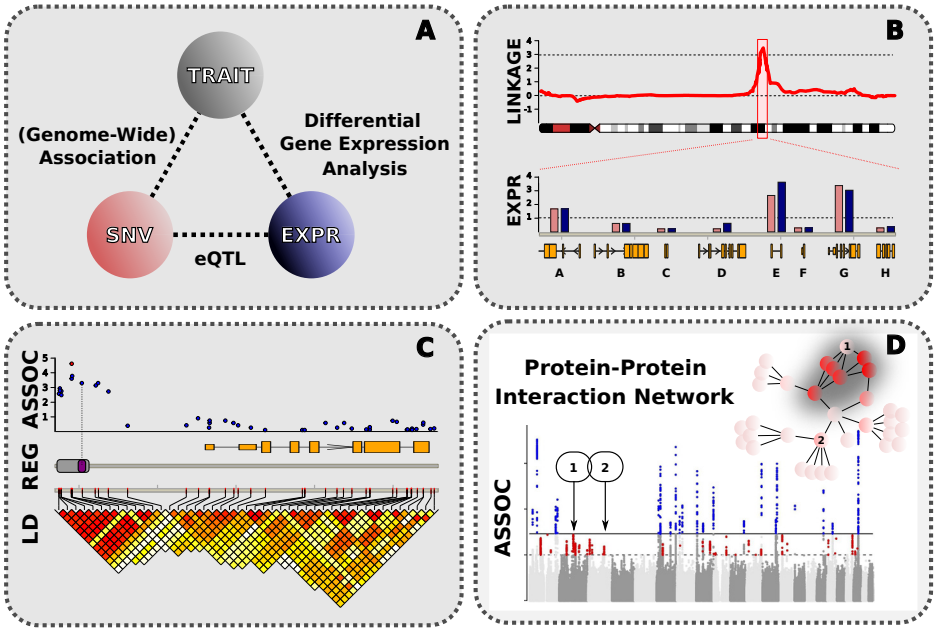

FIGURE 2: Examples of approaches for data integration. A) False positive findings are reduced by enforcing significant correlations between trait-SNP (GWAS), SNP-expression (eQTL) and trait-expression. **B**) Genomic Convergence as originally applied by Hauser *et al.*[51]. Loci displaying significant linkage are scrutinized using a differential gene expression analysis. Note that in this example only three out of seven genes are expressed in the studied tissue (A, E and G) of which gene E seems to exhibit the largest differences between cases (red) and controls (blue). **C**) Hits coming from GWASs (ASSOC) are interpreted in their genomic context in HaploReg, which integrates various publically available resources. The top associated SNV (red) is in strong Linkage Disequilibrium (LD) with other nearby SNVs (red, yellow and white indicate respectively high, mediocre and low R^2). Histone marks for enhancers have been found (REG: grey) in the studied tissue and an eQTL for the upstream region has been reported (REG: purple). Together, these results link the top SNV (red) to the upstream gene. **D**) DAPPLE[58] maps GWAS hits to Protein-Protein Interaction to identify functionally coherent clusters of genes involved in the studied trait. These modules serve several purposes, for instance, candidate loci 1 and 2 (ASSOC) can be prioritized using their proximity to other significantly associated genes (red) in the Protein-Protein Interaction Network. In this example, gene 1 seems to be a more plausible candidate as compared to gene 2, due to their network neighbourhood.

Another example of a joint analysis on omics data sources is a so-called genomic convergence approach[51] in which the relatively low-resolution results coming from genome wide linkage analysis are fine mapped using a differential gene expression analysis (Figure 2B). The original paper coining this term successfully used Serial Analysis of Gene Expression (SAGE) data, to prioritize the thousands of candidate genes that were identified with a linkage analysis on families suffering from Parkinson's disease[51]. Later the concept of genomic convergence was extended to the sequential use and intersection of significant results of any combination of omics data sources that also included whole genome gene expression profiling, as exemplified by Wheeler *et al.*[52]. Here, loci influencing kidney aging where found by a sequential use of a differential gene expression profiling on age, followed by

an eQTL analysis for the significantly age-associated genes, which delivered the final 101 prioritized loci to be tested with the actual phenotype of interest. The aim of such an approach is again to control the statistical power, while gaining additional mechanistic insights. Many more examples exist for the joint interpretation of multiple genomic data sources, but in general all these approaches are aimed at improving the power by including data on additional measurements, and not additional samples per se. Hence, whenever the number of available samples is limited, as is often the case when studying human aging and life span regulation, additional power can be gained by applying approaches for the joint analysis of data sources.

Multiple genomic data sources assayed on an overlapping group of individuals are not always available, but fortunately, much can also be gained from results created in previously performed independent experiments. The number and types of such annotations stored by online databases is rapidly expanding, as are the number of algorithms employing this information that can be readily applied for improving one's own analysis. The incorporation of such prior knowledge is in general performed to aid in the interpretation or prioritization of results or for introducing additional constraints in the analysis of genomics data (regularization) to prevent overfitting. A very straightforward example is that of databases integrating results of genomic approaches to aid in the interpretation of GWAS results. For instance, HaploReg[53] not only contains results of eQTL studies[54], but also employs genetic data from the 1000 Genomes Project[55] for inferring correlations with nearby genetic markers and epigenetic data from the ENCODE[56] and Roadmap Epigenomics[57] projects for inferring overlaps with regulatory domains (Figure 2C). Other well-known examples of algorithms incorporating prior knowledge is DAPPLE[58], an algorithm developed to test for functional coherence between hits derived from GWAS studies using previously measured networks of Protein-Protein Interaction data[59] (Figure 2D). Using this algorithm, it was shown that genes in loci associated to height and lipid levels assemble into significantly interconnected modules. Hence, both these examples for GWAS result interpretation imply that false positive rates can be reduced using measures derived from prior knowledge.

Many algorithms exist for prioritising variants obtained from sequencing experiments using prior information. Besides predicting the putative impact of coding variants using established gene models (e.g. SIFT[60] or PolyPhen[61]) or cross-species conservation (e.g. GERP[62]), more recent algorithms are also able to prioritise variants residing in non-coding regions by exploiting public genetic data resources for inferring the relative sensitivity of genomic regions to perturbations[63,64]. The latter concept was elegantly exploited for prioritizing candidate cancer driver mutations by revisiting previously assayed sequencing data and assessing which motifs where under strong negative selection in the general population, but recurrently disrupted in tumour samples[63]. To conclude, a positive side effect of many of the methods for data integration is that often also additional biological insights are

gained, by revealing some of the molecular interactions. Therefore, approaches for data integration are not only useful in aging research for the purpose of dealing with statistical issues related to power, but for probing the essence of molecular aging mechanisms as well.

4. Aim and Outline of this Thesis

The aim of this thesis was to develop state-of-the-art integrative algorithms for the comprehensive and robust analysis of omics data sets, and to apply them to elucidate molecular pathways driving the rate of human aging.

To develop methodology for a comprehensive and robust analysis of gene expression data in **Chapter 2**, we explored employing Protein-Protein Interaction (PPI) data for grouping gene-expression data into comprehensive modules of functionally related genes (Figure 2D). We investigated whether the expression of such gene modules jointly could serve as robust biomarkers. In this chapter we revisited six expression data sets previously assayed for investigating indicators of prospective outcome of patients undergoing breast cancer surgery. Like the aging phenotype, breast cancer outcome is a very heterogeneous and complex phenotype that demands advanced methodology for the robust analysis and comprehensive interpretation of assayed omics data. Novel methodology for calling co-expressed PPI modules from gene expression data was introduced and cross-study reproducibility, cross-study prediction accuracy and comprehensibility of the thus obtained biomarkers was investigated.

In **Chapter 3** the methodology for calling co-expressed PPI modules was further developed adopting a meta-analysis framework for both the module inference and following associations with phenotypes of interest. Aim was to investigate the benefits for studying the aging transcriptome with aid of the newly developed methodology for a module based meta-analysis as opposed to the traditional individual gene meta-analysis. For this purpose, we revisited four transcriptomic datasets previously measured in blood (~2.500 samples) and employed an additional independent dataset (~3.500 samples) for replicating the obtained associations with chronological age. The potential application of the thus obtained age-associated co-expressed PPI modules as biomarkers for healthy aging was further studied in a small independent set of nonagenarians (~50 samples) derived from the Leiden Longevity Study (LLS).

To dive deeper into the genetics underlying the rate of aging and longevity, the whole genome sequence of 218 long-lived cases of the Leiden Longevity Study (LLS) was compared with that of 98 population controls provided by the BBMRI-NL biobanking initiative[65] in **Chapter 4**. The analysis of whole genome sequencing data in the current study, but also in general for other studies, is heavily underdetermined and the objective was to investigate strategies for including prior knowledge to appropriately deal with this statistical issue. In this chapter prediction tools, similarly as discussed in Figure 2C, were employed that incorporate

prior knowledge to limit the initial analysis to those variants with the highest prior probability of disrupting a gene's functioning. Moreover, variant frequencies from a large-scale sequencing project, the Exome Sequencing Project[66] were incorporated to assess the significance of the joint presence, or burden, of these disruptive variants in long-lived cases.

Long-lived families are characterized by an attenuated thyroid function[14,67], suggesting a shared genetic basis for attenuation of the thyroid function and the longevity phenotype. In **Chapter 5** we set out to elucidate this pleiotropic genetic mechanism by investigating the 239 nonagenarian sibships from the LLS displaying the most profound family history of excess survival (FH(+)), a trait previously associated with attenuation of the thyroid function[67]. For the analysis, we pursued a variation on the two-step genomic convergence approach (Figure 2B). First, genome-wide linkage analyses for familial longevity in the whole LLS (415 sibships) identified suggestive linkage at chr13q34, that was highly specific to the FH(+) subset and almost absent in the remaining 176 sibships without such a marked family history (FH(-)). For the second fine-mapping step of the variants under the linkage peak, we investigated which of the thyroid parameters was most characteristic to the FH(+) subset. The FH(+) subset exhibited a significantly lower serum free triiodothyronine level, the active thyroid hormone itself (fT3), as compared to the FH(-)) subset. Therefore we hypothesized that variants at chr13q34 might explain the observed pleiotropic interaction between longevity and an attenuated thyroid signalling, by lowering serum fT3 levels. Hence, the second fine-mapping step was performed by Quantitative Trait Loci (QTL) analyses, correlating free triiodothyronine (fT3) serum levels to NGS variants, to probe for causal variants underlying both the attenuated fT3 signalling as human longevity in this locus.

Finally, during this thesis we have encountered several bioinformatics tasks that are routinely performed during projects for genomic data integration. To generalize and standardize the execution of such highly similar though demanding tasks over different types of omics data sets, we implemented the R package SATORi (Standardized Access To Omics in R). In **Chapter 6** we exemplify its use with publically available omics data sets and comment on some of the considerations made in the design of this package.

5. References

1. Oeppen, J. & Vaupel, J.W. Demography. Broken limits to life expectancy. *Science* **296**, 1029-31 (2002).
2. Hitt, R., Young-Xu, Y., Silver, M. & Perls, T. Centenarians: the older you get, the healthier you have been. *Lancet* **354**, 652 (1999).
3. Kirkwood, T.B. & Austad, S.N. Why do we age? *Nature* **408**, 233-8 (2000).
4. Jones, O.R. *et al.* Diversity of ageing across the tree of life. *Nature* **505**, 169-73 (2014).
5. Navarro, C.L., Cau, P. & Levy, N. Molecular bases of progeroid syndromes. *Hum Mol Genet* **15 Spec No 2**, R151-61 (2006).
6. Perls, T.T. *et al.* Life-long sustained mortality advantage of siblings of centenarians. *Proc Natl Acad Sci U S A* **99**, 8442-7 (2002).

7. Westendorp, R.G. *et al.* Nonagenarian siblings and their offspring display lower risk of mortality and morbidity than sporadic nonagenarians: The Leiden Longevity Study. *J Am Geriatr Soc* **57**, 1634-7 (2009).
8. Terry, D.F. *et al.* Lower all-cause, cardiovascular, and cancer mortality in centenarians' offspring. *J Am Geriatr Soc* **52**, 2074-6 (2004).
9. Kirkwood, T.B. Evolution of ageing. *Nature* **270**, 301-4 (1977).
10. Lopez-Otin, C., Blasco, M.A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194-217 (2013).
11. Deelen, J., Beekman, M., Capri, M., Franceschi, C. & Slagboom, P.E. Identifying the genomic determinants of aging and longevity in human population studies: progress and challenges. *Bioessays* **35**, 386-96 (2013).
12. Hankins, T.C. & Wilson, G.F. A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviat Space Environ Med* **69**, 360-7 (1998).
13. Noordam, R. *et al.* Cortisol serum levels in familial longevity and perceived age: the Leiden longevity study. *Psychoneuroendocrinology* **37**, 1669-75 (2012).
14. Rozing, M.P. *et al.* Low serum free triiodothyronine levels mark familial longevity: the Leiden Longevity Study. *J Gerontol A Biol Sci Med Sci* **65**, 365-8 (2010).
15. Rozing, M.P. *et al.* Favorable glucose tolerance and lower prevalence of metabolic syndrome in offspring without diabetes mellitus of nonagenarian siblings: the Leiden longevity study. *J Am Geriatr Soc* **58**, 564-9 (2010).
16. Newman, A.B. *et al.* Health and function of participants in the Long Life Family Study: A comparison with other cohorts. *Aging (Albany NY)* **3**, 63-76 (2011).
17. Passtoors, W.M. *et al.* Genomic studies in ageing research: the need to integrate genetic and gene expression approaches. *J Intern Med* **263**, 153-66 (2008).
18. de Magalhaes, J.P., Curado, J. & Church, G.M. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics* **25**, 875-81 (2009).
19. Partridge, L. & Gems, D. Mechanisms of ageing: public or private? *Nat Rev Genet* **3**, 165-75 (2002).
20. Zahn, J.M. *et al.* Transcriptional profiling of aging in human muscle reveals a common aging signature. *PLoS Genet* **2**, e115 (2006).
21. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol* **14**, R115 (2013).
22. Weidner, C.I. *et al.* Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biol* **15**, R24 (2014).
23. West, J., Widschwendter, M. & Teschendorff, A.E. Distinctive topology of age-associated epigenetic drift in the human interactome. *Proc Natl Acad Sci U S A* **110**, 14138-43 (2013).
24. Skytthe, A. *et al.* Longevity studies in GenomEUtwin. *Twin Res* **6**, 448-54 (2003).
25. Herskind, A.M. *et al.* The heritability of human longevity: a population-based study of 2872 Danish twin pairs born 1870-1900. *Hum Genet* **97**, 319-23 (1996).
26. Kenyon, C., Chang, J., Gensch, E., Rudner, A. & Tabtiang, R. A C. elegans mutant that lives twice as long as wild type. *Nature* **366**, 461-4 (1993).
27. Kenyon, C.J. The genetics of ageing. *Nature* **464**, 504-12 (2010).
28. Willcox, B.J. *et al.* FOXO3A genotype is strongly associated with human longevity. *Proc Natl Acad Sci U S A* **105**, 13987-92 (2008).
29. Flachsbart, F. *et al.* Association of FOXO3A variation with human longevity confirmed in German centenarians. *Proc Natl Acad Sci U S A* **106**, 2700-5 (2009).

30. Pawlikowska, L. *et al.* Association of common genetic variation in the insulin/IGF1 signaling pathway with human longevity. *Aging Cell* **8**, 460-72 (2009).

31. Soerensen, M. *et al.* Replication of an association of variation in the FOXO3A gene with human longevity using both case-control and longitudinal data. *Aging Cell* **9**, 1010-7 (2010).

32. Deelen, J. *et al.* Genome-wide association study identifies a single major locus contributing to survival into old age; the APOE locus revisited. *Aging Cell* **10**, 686-98 (2011).

33. Nebel, A. *et al.* A genome-wide association study confirms APOE as the major gene influencing survival in long-lived individuals. *Mech Ageing Dev* **132**, 324-30 (2011).

34. Sebastiani, P. *et al.* Genetic signatures of exceptional longevity in humans. *PLoS One* **7**, e29848 (2012).

35. Schachter, F. *et al.* Genetic associations with human longevity at the APOE and ACE loci. *Nat Genet* **6**, 29-32 (1994).

36. Christensen, K., Johnson, T.E. & Vaupel, J.W. The quest for genetic determinants of human longevity: challenges and insights. *Nat Rev Genet* **7**, 436-48 (2006).

37. Deelen, J. *et al.* Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. *Human molecular genetics*, ddu139 (2014).

38. Han, J. *et al.* Discovery of novel non-synonymous SNP variants in 988 candidate genes from 6 centenarians by target capture and next-generation sequencing. *Mech Ageing Dev* **134**, 478-85 (2013).

39. Ye, K. *et al.* Aging as accelerated accumulation of somatic variants: whole-genome sequencing of centenarian and middle-aged monozygotic twin pairs. *Twin Research and Human Genetics* **16**, 1026-1032 (2013).

40. Gierman, H.J. *et al.* Whole-Genome Sequencing of the World's Oldest People. *PLoS One* **9**, e112430 (2014).

41. Sebastiani, P. *et al.* Whole genome sequences of a male and female supercentenarian, ages greater than 114 years. *Front Genet* **2**, 90 (2011).

42. Holstege, H. *et al.* Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis. *Genome Res* **24**, 733-42 (2014).

43. Cash, T.P. *et al.* Exome sequencing of three cases of familial exceptional longevity. *Aging Cell* **13**, 1087-90 (2014).

44. Schoenmaker, M. *et al.* Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. *Eur J Hum Genet* **14**, 79-84 (2006).

45. Puca, A.A. *et al.* A genome-wide scan for linkage to human exceptional longevity identifies a locus on chromosome 4. *Proc Natl Acad Sci U S A* **98**, 10505-8 (2001).

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

47. Edwards, D.R. *et al.* Successful aging shows linkage to chromosomes 6, 7, and 14 in the Amish. *Ann Hum Genet* **75**, 516-28 (2011).

48. Beekman, M. *et al.* Genome-wide linkage analysis for human longevity: Genetics of Healthy Aging Study. *Aging Cell* **12**, 184-93 (2013).

49. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001-6 (2014).

50. Nicolae, D.L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* **6**, e1000888 (2010).

51. Hauser, M.A. *et al.* Genomic convergence: identifying candidate genes for Parkinson's disease by combining serial analysis of gene expression and genetic linkage. *Hum Mol Genet* **12**, 671-7 (2003).

52. Wheeler, H.E. *et al.* Sequential use of transcriptional profiling, expression quantitative trait mapping, and gene

association implicates MMP20 in human kidney aging. *PLoS Genet* **5**, e1000685 (2009).

53. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**, D930-4 (2012).
54. Fan, C., Sun, Y., Yang, J., Ye, J. & Wang, S. Maternal and neonatal outcomes in dichorionic twin pregnancies following IVF treatment: a hospital-based comparative study. *Int J Clin Exp Pathol* **6**, 2199-207 (2013).
55. Abecasis, G.R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
56. Bernstein, B.E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
57. Bernstein, B.E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**, 1045-8 (2010).
58. Rossin, E.J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet* **7**, e1001273 (2011).
59. Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* **25**, 309-16 (2007).
60. Ng, P.C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-4 (2003).
61. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-9 (2010).
62. Davydov, E.V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025 (2010).
63. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).
64. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
65. Boomsma, D.I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet* **22**, 221-7 (2014).
66. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-9 (2012).
67. Rozing, M.P. *et al.* Familial longevity is associated with decreased thyroid function. *J Clin Endocrinol Metab* **95**, 4979-84 (2010).