

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/32015> holds various files of this Leiden University dissertation.

Author: Akker, Erik Ben van den

Title: Computational biology in human aging : an omics data integration approach

Issue Date: 2015-02-18

Computational Biology in Human Aging

An Omics Data Integration Approach

Erik Ben van den Akker

Computational Biology in Human Aging

Ir. E.B. van den Akker

The cover displays a view from the north coast of Spain, a region historically known for its adept sailors and daring explorers of the world seas. In this bay, many embarked and sailed off in the unknown, driven by their curiosity, to claim new land for king and country, often with little details on their final destinations. Many parallels exist between these early scientific endeavours and current projects. In this thesis, we set out to link two worlds, epidemiology and bioinformatics, convinced that the synergy between these two fields would allow us to probe deeper for the factors contributing to healthy aging and longevity. Little was known on omics data integration in aging, but still we embarked on what was going to be a very exciting journey. Driven by our curiosity.

Financial support for the printing of this thesis was provided by the Netherlands Consortium of Healthy Ageing (NGI 050-060-810).

PhD thesis with summary in Dutch

ISBN: 978-94-6295-084-9

© 2015 **E.B. van den Akker**

All rights reserved. No part of this thesis may be reproduced, stored in a retrieval system or transmitted in any other form by any means, without the permission of the author, or when appropriate, of the publisher of the represented published articles.

Cover design: Proefschriftmaken.nl || Uitgeverij BOXPress

Printed & Lay Out by: Proefschriftmaken.nl || Uitgeverij BOXPress

Published by: Uitgeverij BOXPress, 's-Hertogenbosch

Computational Biology in Human Aging
An Omics Data Integration Approach

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
ter verdedigen op woensdag 18 februari 2015
klokke 11.15 uur

door

Erik Ben van den Akker

Geboren te 's-Hertogenbosch
in 1981

Promotiecommissie

Promotor: Prof. Dr. P.E. Slagboom

Prof. Dr. Ir. M.J.T. Reinders
Delft University of Technology

Co-promotor: Dr. M. Beekman

Overige leden: Prof. Dr. B.J. Zwaan
Wageningen University Research Centre

Prof. Dr. L. Wessels
Netherlands Cancer Institute
Delft University of Technology

Dr. J.P. De Magalhães
University of Liverpool

Contents

Chapter 1:	Introduction	7
Chapter 2:	Integrating Protein-Protein Interaction Networks with Gene-Gene Co-Expression Networks improves Gene Signatures for Classifying Breast Cancer Metastasis	21
Chapter 3:	Meta-Analysis on Blood Transcriptomic Studies Identifies Consistently Co-Expressed PPI Modules as Robust Markers of Human Aging	41
Chapter 4:	Germ line and Somatic Characteristics of the Long-Lived Genome	61
Chapter 5:	A novel life span regulating locus at chr13q34 influencing serum triiodothyronine level	89
Chapter 6:	An R package for generic access and handling of genomic data	117
Chapter 7:	Discussion	125
Chapter 8:	Nederlandse Samenvatting	141
Appendix	List of Publications	152
	Curriculum Vitae	155
	Dankwoord	157

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 population¹ 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 today², 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 homeostasis³. Though seemingly inevitable, aging does not occur at an equal pace across species⁴ or even within our own species. Whereas some experience an accelerated rate of aging, as exemplified by patients suffering from progeroid syndromes⁵, others seem capable of delaying or evading at least some of the detrimental aspects of aging, as observed in members of long-lived families⁶⁻⁸. 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 life⁹. 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.

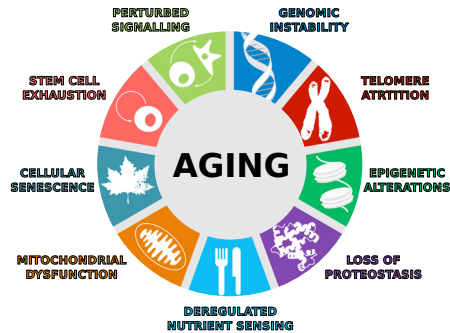


FIGURE 1: Nine recurrently observed processes that occur with aging. Processes that are commonly observed during aging are: genomic instability, telomere attrition, epigenetic alterations, loss of proteostasis, deregulated nutrient sensing, mitochondrial dysfunction, cellular senescence, stem cell exhaustion 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 years¹². 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 cortisol¹³, free triiodothyronine¹⁴ 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 directly²¹ or by targeting hub-genes in regulatory networks²³. 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 species²⁷. 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 5q33.3³⁷.

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 individuals³⁸ or

on a whole genome scale in very limited numbers of individuals³⁹⁻⁴³, 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 performed⁴⁵⁻⁴⁸, however, none of the reported loci display any mutual consistency, nor have they been independently confirmed³⁶. 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 led 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 diseases⁴⁹,

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 rationale 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.

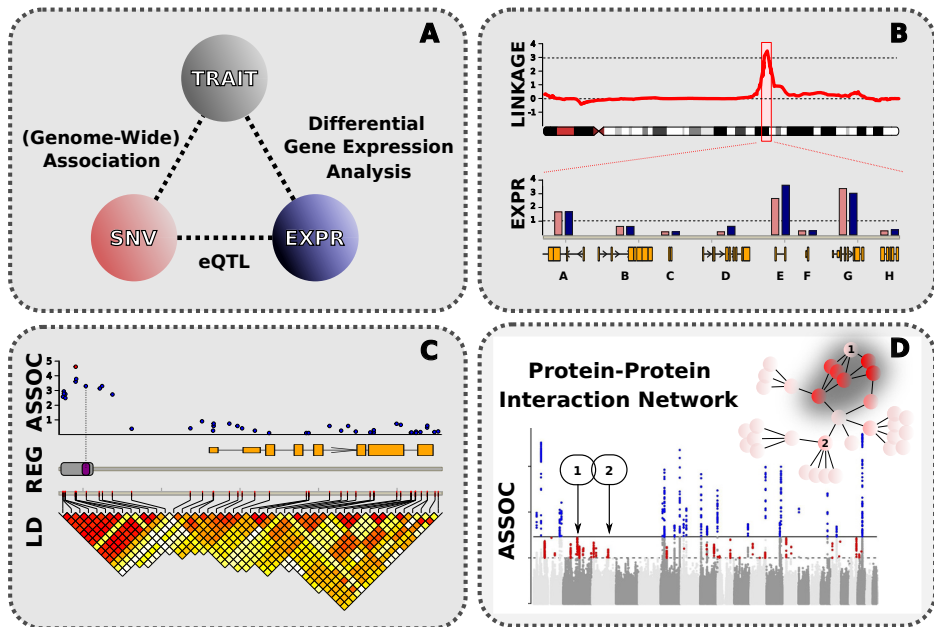


FIGURE 2: Examples of approaches for data integration. **A)** False positive findings are reduced by enforcing significant correlations between trait-SNP (GWAS), SNP-expression (eQTL) and trait-expression. **B)** Genomic Convergence as originally applied by Hauser *et al.*⁵¹. Loci displaying significant 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)** DAPPLE⁵⁸ 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 approach⁵¹ 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 disease⁵¹. 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 were 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, HaploReg⁵³ not only contains results of eQTL studies⁵⁴, but also employs genetic data from the 1000 Genomes Project⁵⁵ for inferring

correlations with nearby genetic markers and epigenetic data from the ENCODE⁵⁶ and Roadmap Epigenomics⁵⁷ projects for inferring overlaps with regulatory domains (Figure 2C). Other well-known examples of algorithms incorporating prior knowledge is DAPPLE⁵⁸, an algorithm developed to test for functional coherence between hits derived from GWAS studies using previously measured networks of Protein-Protein Interaction data⁵⁹ (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. SIFT⁶⁰ or PolyPhen⁶¹) or cross-species conservation (e.g. GERP⁶²), 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 were under strong negative selection in the general population, but recurrently disrupted in tumour samples⁶³. 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 initiative⁶⁵ 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 Project⁶⁶ 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 function⁶⁷. 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 GenomeEUtwin. *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 USA* **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).

Chapter 2:

Integrating Protein-Protein Interaction Networks with Gene-Gene Co-Expression Networks improves Gene Signatures for Classifying Breast Cancer Metastasis

Erik B. van den Akker^{1,2}, Bas Verbruggen², Bas T. Heijmans^{1,3}, Marian Beekman^{1,3}, Joost N. Kok^{1,4,5}, P. Eline Slagboom^{1,3}, Marcel J.T. Reinders^{2,5}

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 of Healthy Ageing

⁴ Algorithms, Leiden Institute of Advanced Computer Science, University Leiden, Leiden, The Netherlands

⁵ Netherlands Bioinformatics Centre

Journal of Integrative Bioinformatics, 8(2):188, 2011

1. Abstract

Multiple studies have illustrated that gene expression profiling of primary breast cancers throughout the final stages of tumor development can provide valuable markers for risk prediction of metastasis and disease sub typing. However, the identification of a biologically interpretable and universally shared set of markers proved to be difficult. Here, we propose a method for *de novo* grouping of genes by dissecting the protein-protein interaction network into disjoint sub networks using pair wise gene expression correlation measures. We show that the obtained sub networks are functionally coherent and are consistently identified when applied on a compendium composed of six different breast cancer studies. Application of the proposed method using different integration approaches underlines the robustness of the identified sub network related to cell cycle and identifies putative new sub network markers for metastasis related to cell-cell adhesion, the proteasome complex and JUN-FOS signaling. Although gene selection with the proposed method does not directly improve upon previously reported cross study classification performances, it shows great promises for applications in data integration and result interpretation.

2. Introduction

A crucial step in breast cancer diagnosis and subsequent therapy is the assessment of the tumor's capacity to metastasize. An erroneous diagnosis can either lead to overtreatment or could potentially allow already spread tumors to develop in distant tissues. Since the first leads to a significant amount of unnecessary burden for the patient, while the latter is the predominant cause of death in breast cancer patients¹, a lot of effort has been invested to improve personalized risk profile predictions by employing gene expression assays. However, as whole genome assays are delivering an increasing list of transcriptomic disease markers, the low mutual overlap between different studies becomes apparent. More importantly, obtained sets of prognostic markers from one study show a significant drop in prediction performance when applied to another study². Current methods for gene set grouping may be less successful when performed on a single gene basis, due to the underlying heterogeneity of the disease as well as the fact that due to secondary effects many genes seem to correlate with the phenotype². Consequently, resulting gene sets purely selected on single gene ranking are often uninformative from a biological point of view.

In response, several types of analyses were developed, which incorporated prior biological knowledge to ensure the biological interpretability of the selected gene set³⁻⁵. Genes can for instance be grouped on similar function, localization or pathway membership. However, as many genes are still not assigned to

relevant groupings and moreover, all relevant groupings themselves might still not be known, the effectiveness of such an approach might be severely compromised⁶.

To deal with the low coverage of predefined functional groupings, several methods have been developed to create groupings *de novo*, by, for instance, exploiting data on physical interactions between proteins. Over the last few years, this type of data has consistently been gathered and integrated with other types of interactions⁷, like lethal-lethal⁸, co-citations, or cellular co-localization interactions to produce large interaction networks. These so called Protein-Protein Interaction (PPI) networks contain modules that can be linked to cellular functions⁹. The use of these networks for the simultaneous task of relevant gene set discovery and prediction optimization was popularized by the work of Chuang *et al.*⁶. In this method, sub networks are seeded once at every node in the network and are iteratively grown by greedily adding the best neighbor, until a certain gene set summary statistic no longer improves. Resulting sub networks have been used as input for classification showing an improvement in cross study classification compared to single gene based signatures as well as providing hypothetical biological mechanisms underlying the studied phenotype⁶.

Although this is clearly an improvement over previously published methods, we fear that the capacity to generalize over studies is compromised by the greedy aspect with which the seeded sub networks are grown. Given the fact that many genes seem to correlate with the studied phenotype, as

they are most probably co-expressed due to downstream effects, a considerable part of the data may be viewed as intrinsic biological replicates independently assessing the state of a select number of ongoing cellular processes. In view of this, we would rather like to use *all* informative genes involved in such a process to robustly characterize the cell's transcriptomic state instead of using the genes from a local greedy search only.

A second drawback of greedy network approaches becomes apparent with the growing amount of protein-protein interactions that becomes available. New data predominantly interconnects genes within existing networks, rather than that it connects previously unlinked genes to existing networks. This contributes to the 'small-world' phenomenon¹⁰, referring to a situation where almost every gene in a network is only a few connections away from any other gene. As a consequence, the informative property of localized network sub selection is lost to global and thus less interpretable sub network solutions. A proper biological interpretation is even further compromised if overlap between identified sub networks is allowed. Under these circumstances, numerous highly similar and equally likely solutions will be produced, biasing the selection towards a select set of predictive network hubs, thereby basically reducing the algorithm to a computationally inefficient global ranking method.

Anticipating the previously described problems in selecting genes, we here propose a non-greedy method for dissecting the interaction network in a set of disjoint sub networks. We expect

that by incorporating both pair wise gene expression correlation measures, as protein-protein interactions functionally more coherent sub networks will be selected. We hypothesize that building such sub networks will not only generalize better across datasets in predicting the risk of metastasis as they exploit the available information maximally, but as well be more informative about the involved biological processes.

3. Experimental Procedures

3.1 Materials

In this study six publicly available microarray data sets of breast cancer samples measured on the HG U133A platform (Affymetrix) were employed to test our hypothesis. Raw expression data was downloaded at the NCBI's ftp server¹¹ under the accessions: GSE7390¹², GSE3494¹³, GSE6532¹⁴, GSE1456¹⁵, GSE2034¹⁶ and GSE11121¹⁷. Data was normalized, log₂ transformed and summarized per probe set using the RMA procedure in the Affy package¹⁸ of R^{19,20} at default settings. Replicate and duplicate samples were removed. See Table 1 for an overview of the employed studies.

A recent annotation was downloaded from the Affymetrix website²¹ to map all "_at", "_s_at" and "_x_at" probe sets to Ensembl Transcript IDs. Mappings to Ensembl gene IDs and protein IDs obtained from the Ensembl site²² and protein-protein interactions obtained from STRING²³ were used to map probe sets to the protein-protein interaction network. Probe sets missing annotations

Study	Accession	#	# Rep/Dup	Missing	"POOR"	"GOOD"
Desmedt	GSE7390	198	174 ¹	24	31	119
Miller	GSE3494	251	232 ²	37	37	158
Loi	GSE6532	327	186 ³	47	32	107
Pawitan	GSE1456	159	156	6	35	115
Wang	GSE2034	286	286	11	95	180
Schmidt	GSE11121	199	199	19	27	153

TABLE 1: OVERVIEW OF STUDIES. Statistics on the six studies employed. Accession, #, # Rep/Dup, Missing, "POOR" and "GOOD" refer to the accession code and the number of samples available at GEO, the number of samples after removal of replicates and duplicate samples, the number of samples with incomplete metadata or prematurely ended censoring, the number of "POOR" prognosis samples and the number of "GOOD" samples respectively. 1) Replicates (Desmedt/Loi) were removed from Desmedt. 2) Replicates (Desmedt/Miller) were removed from Miller. 3) Duplicates (Miller/Loi) were removed from Loi.

to genes, transcripts or proteins, as well as probe sets mapping to multiple genes or probe sets not associated with any interaction data were excluded for further analysis. When multiple probe sets were annotated to the same gene, "_at" probes were preferred over "_s_at" probes and "_s_at" probes over "_x_at" probes. When this did not enforce a decision the probe set with the highest standard deviation was selected. Preprocessing resulted in a mapping of 9,290 probe sets representing 9,290 unique genes to a network of 169,566 undirected interactions.

"POOR" and "GOOD" prognosis of samples was assessed using metadata obtained from the NCBI's ftp server¹¹ as well. "POOR" refers to the occurrence of a distant metastatic event or a relapse within five years after surgery. Subjects were selected for the "GOOD" prognosis subgroup when an event free survival of at least five years was reported. Whereas some studies contained information on distant metastatic events, others reported relapses of breast cancer. When both were available, the reports on distant metastatic events were used.

3.2 Methods

3.2.1 Proposed method for dissecting the protein-protein interaction network in disjoint co-regulated sub networks:

Sub networks are created through evidence-based filtering of edges between genes using two types of evidence: physical interaction data and expression correlations between any pair of genes. Let E_{ij} be the gene expression matrix with probe set i and subject j , where $i = 1$ to M and $j = 1$ to N . An $M \times M$ correlation matrix C is computed, where C_{pq} is defined to be the correlation between gene p and gene q over all N samples. Threshold T_{COR} is applied on C to obtain a binary matrix C^T , where $C^T_{pq} = 1$ indicates sufficient and $C^T_{pq} = 0$ indicates insufficient correlation between genes p and q .

Based on a distance matrix equal to $1 - \text{abs}(C)$, the genes are hierarchically clustered (average linkage). The clustering dendrogram is thresholded at $1 - T_{COR}$, creating a grouping matrix G with dimensions $M \times M$, where $G_{pg} = 1$ indicates co-membership of a gene cluster, and G_{pg}

= 0 indicates an assignment to different clusters of gene p and q .

Let matrix \mathbf{P} contain the protein-protein interactions, with \mathbf{P}_{pq} ranging from 1 to 999 indicating the confidence level associated in case an interaction is reported and $\mathbf{P}_{pq} = 0$ if no interactions are known. Threshold T_{ppi} is applied to \mathbf{P} to obtain a binary matrix \mathbf{P}^T , where $\mathbf{P}^T = 1$ indicates a presence and $\mathbf{P}^T = 0$ indicates an absence of known interactions with a sufficient confidence level. The binary correlation matrix \mathbf{C}^T is overlaid with the grouping matrix \mathbf{G} and the binary protein-protein interaction matrix \mathbf{P}^T to yield sub network matrix \mathbf{S} :

$$\mathbf{S}_{pq} = \mathbf{G}_{pq} \mathbf{C}_{pq}^T \mathbf{P}_{pq}^T \forall pq \quad (1)$$

where $\mathbf{S}_{pq} = 1$ indicates an absolute correlation equal to or exceeding T_{cor} between genes p and q , they are assigned to the same cluster and a physical interaction with a confidence level exceeding T_{ppi} between the proteins of these genes has been reported. $\mathbf{S}_{pq} = 0$ indicates that at least one of these conditions is not met.

Correlations between the breast cancer outcome status and gene-expression data per sub network were evaluated using the global test as summary statistic⁵. This test uses ridge regression to model the relation between breast cancer outcome (response variable) and a set of gene expressions (input variables), while correcting for the mutual correlation structure between the input variables. Obtained sub networks were filtered on significance by applying threshold T_S . Genes within significant sub networks rendered the gene sets used to determine cross study prediction

performances and similarities in feature selection.

Since the thresholded gene expression (GE) network (\mathbf{C}^T) is overlaid with the thresholded PPI network (\mathbf{P}^T), both thresholds, T_{ppi} and T_{cor} are crucial in determining the connectivity of the resulting network. To balance the influence of both sources of information, T_{ppi} and T_{cor} are chosen such that roughly equal amounts of interactions are obtained for the thresholded GE and PPI networks. As the overlay network rapidly becomes sparser at PPI quality scores exceeding 500 ('medium confidence score' in STRING), T_{ppi} was set to 500 and consequently T_{cor} was set to 0.6.

3.2.2 Competing methods for gene

selection: Forward filters were trained as described by van Vliet *et al.*²⁴. In short, a double cross fold loop procedure²⁵ was employed splitting the data in a validation and a training set (5 folds). The latter is split in an inner training set and an inner test set (10 folds). The additional cross fold setting within the training set implements a strict separation between data used for optimizing the predictor and its evaluation. The optimal number of genes is determined within the inner set by training and evaluating a classifier for up to 200 top ranking genes. Gene ranking was done using absolute Welch's t-statistic. Once the optimal signature size is determined a classifier is trained on the ranked outer training set, which in turn is evaluated in the left out validation set. This procedure is repeated 20 times, thus producing $20 \times 10 \times 5 = 1000$ unbiased estimates of the optimal signature size. A final predictive gene set

was produced by thresholding the ranked gene list learned on the whole study with the mean over all optimal signature sizes.

Greedy network signatures were obtained by re-implementing the work by Chuang *et al.*⁶ in R²⁰ using identical settings for all parameters, with the exception that sub network performances were evaluated using a Welch's t-statistic instead of the Mutual Information. Gene sets were obtained by enlisting all unique genes within significant sub networks.

3.2.3 Measures of gene set similarity: The Jaccard index²⁶ and odds ratio²⁷ were used to assess the similarity in gene selection between two different studies. The Jaccard index is used to assess the overlap in gene selection and equals the probability for a gene being implicated by both studies, given that it was implicated by at least one study. The odds ratio is used to indicate the consistency in gene selection and is a relative measure of risk representing the increase in likelihood for a gene to be selected, when also selected in another study, compared to a gene being selected, when not selected in another.

3.2.4 Evaluation of Cross Study Prediction Performances: All prediction performances were determined by employing a Nearest Mean Classifier using the cosine-correlation as a distance measure and the Area Under the Curve (AUC) of the Receiver Operator Curve (ROC) as an evaluation measure. Cross study evaluation of the prediction performance was done using two different settings. In the first setting, denoted as “*passing GeneSet*”, a classifier was trained

in a five cross fold setting on the gene set indicated by the first study while employing data of the second study. This procedure was repeated 100 times and the mean classification performance over 100×5 folds was reported as the final performance. In the second setting, denoted as “*passing Classifier*”, a classifier was trained on data of the first study and was evaluated using data of a second study. Prediction performances of integration approaches were determined by using five studies as input while evaluating on the sixth. In the “early” integration approach, data integration occurs at the beginning as five studies are jointly analyzed to select the genes. The “late” integration approach creates a consensus gene set by intersecting the results of selected genes per study.

3.2.5 Sub network visualization: Sub networks were visualized using the RCytoscape²⁸ package in R²⁰ to connect to Cytoscape version 2.8.1²⁹. Nodes were colored according to the sign and magnitude of respectively the calculated Welch's t-test statistic and the accompanying p-value (green: higher expressed in “POOR” outcome compared to “GOOD” and red vice versa).

4. Results

4.1 Data is dissected in functionally coherent sub networks

Using the proposed methodology, disjoint sub networks were created for six well studied publically available breast cancer studies¹²⁻¹⁷ using $T_{cor} = 0.6$, $T_{ppi} = 500$ and

$T_S = 0.05$. Resulting sub networks were visualized using Cytoscape²⁹ (Figure 1). Obtained sub networks varied in sizes ranging from 2 up to 192 genes and were either enriched (e.g. Figure 1: B) or depleted (e.g. Figure 1: A) of predictive markers. Furthermore, genes within resulting sub networks showed a preference to be either jointly down or up-regulated, leading to the observation that hardly any significant sub network (sub networks with a red bounding box in Figure 1) contained oppositely correlating gene expressions with respect to the studied phenotype.

In order to assess whether application of the method led to a biologically meaningful dissection of the data, DAVID³⁰ was used to test for enrichments in functional gene annotations using GO FAT categories. GeneRIF descriptions were inspected for common denominators in case the enrichment analysis returned a-specific or

no functional annotations. Sub networks that showed significant associations with respect to the studied phenotype often also showed significant GO enrichments for hallmark processes of breast cancer. For example, for the Desmedt study in Figure 1: B is enriched for cell cycle phase; I for response to estrogen stimulus; and J for DNA replication. When not related to breast cancer, sub networks could be attributed to processes in lymphocytes or fat tissue. Sub networks enriched for the terms cell cycle phase (GO:0022403), leukocyte activation (GO:0045321) and proteinaceous extracellular matrix (GO:0005578) were seen in all six studies (Figure 1 sub networks A, B and C respectively).

4.2 Eight sub networks are consistently identified

To get a more thorough view whether the observed dissection in functionally

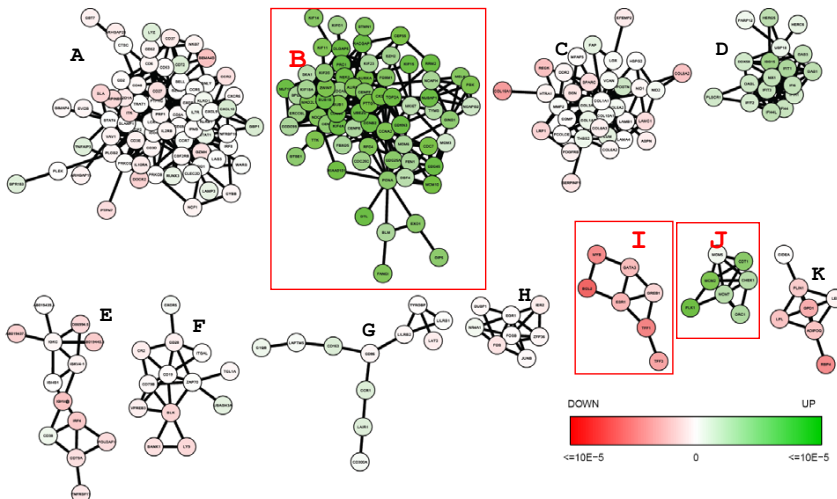


FIGURE 1: AN OVERVIEW OF SUB NETWORKS IDENTIFIED IN THE DESMEDT STUDY. Disjoint sub networks of varying sizes were obtained from the Desmedt study of which the largest are depicted here. Genes are colored according to the p-value of the Welch's-t-test on the expression between "POOR" and "GOOD" outcome subjects (green is higher expressed in "POOR"). A red bounding box around a sub network indicates a significant sub network score obtained with the global test on the gene set indicated by the sub network.

coherent sub networks was consistent between studies, we extended our analyses beyond overlaps in Gene Ontology terms by employing pair wise similarity. For this analysis we calculated Jaccard indices²⁶ between sub networks extracted from the six studies and clustered the obtained similarity matrix. The analysis was limited to sub networks with a minimal size of 7 genes yielding 9 to 16 sub networks per study and a total of 83 sub networks (Figure 3). Cluster analysis shows groupings of six sub networks each derived in a different study implicating a high degree of consistency of detected sub networks between the studies (Figure 2). Besides

the previously consistently identified functionalities: leukocyte activation, proteinaceous extracellular matrix and cell cycle phase (Figure 2, clusters VII, VI and V respectively), five other sub networks with a-specific or no GO enrichments were consistently identified. Common denominators extracted from GeneRIF indicated functionalities related to JUN / FOS signaling for cluster I, interferon induced proteins including ubiquitins for cluster II, Adiponectin / lipid storage for cluster III, Chains of immunoglobulin for cluster IV and immune related genes for cluster VIII.

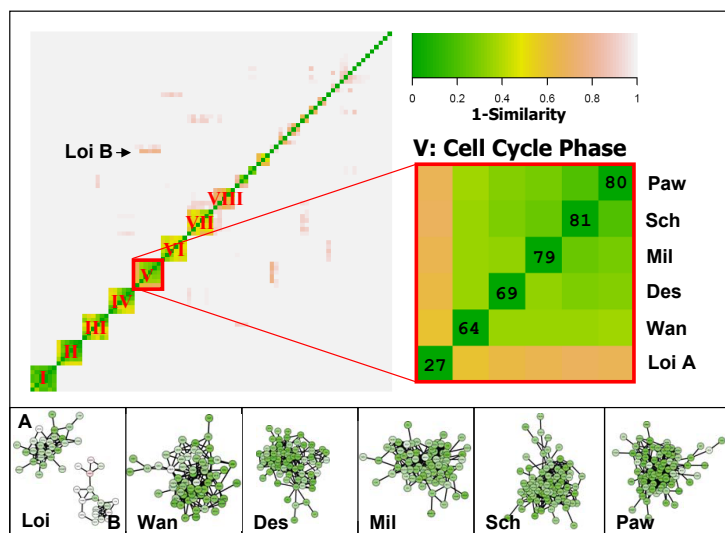


FIGURE 2: OVERLAP BETWEEN BREAST CANCER STUDIES. Pair wise similarities were calculated between sub networks obtained from the six studies using Jaccard indices. The resulting similarity matrix was hierarchically clustered and was depicted as a heat map in the upper left corner. The heatmap is symmetric along the diagonal and each row or column represents a unique sub network identified in one of the studies. The grouping belonging to cluster V (Cell Cycle Phase) is blown up to the right. Numbers on the diagonal indicate the number of genes within the identified sub networks. Extensive similarities are observed between sub networks from the six studies except for comparisons involving Loi, caused by the low number of genes found in the Loi study. Icons of sub networks at the bottom represent the sub networks for the different studies that were clustered together in cluster V, which are all also enriched for Cell Cycle Phase. Note that whereas for the Loi study two small sub networks were identified, others studies only returned a single large sub network.

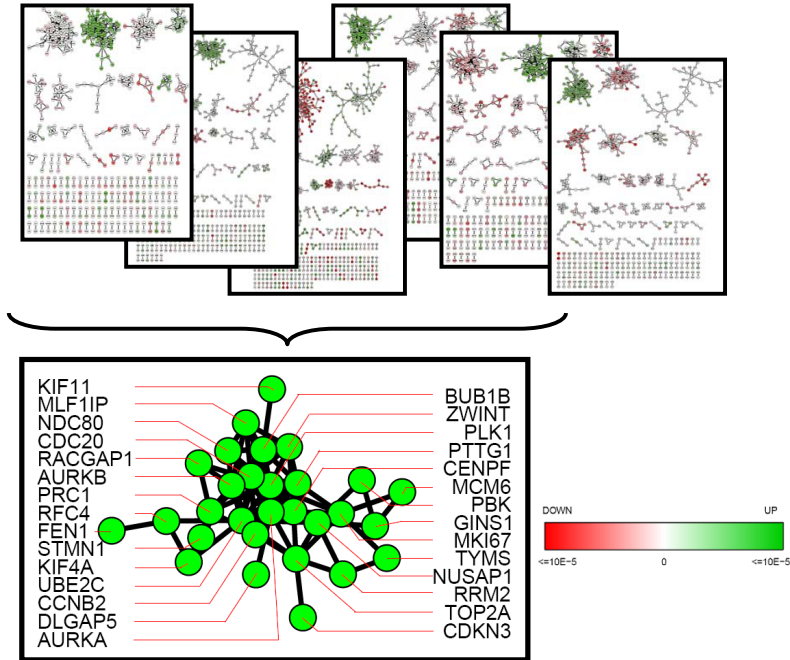


FIGURE 3: SCHEMATIC OVERVIEW OF THE CONSTRUCTION OF A CONSENSUS NETWORK. Detected sub networks are depicted at the top from left to right for the Desmedt, Muller, Loi, Pawitan, Wang and Schmidt study respectively. A consensus sub network was constructed with genes present in significant sub networks ($\alpha = 0.05$) in all six studies and is depicted at the bottom. Edges in the consensus sub network are drawn when confidence values of reported PPI interactions exceed T_{ppi} .

4.3 A “late” integration approach reveals a functionally coherent set of consensus genes putatively involved in metastasis

A consensus gene set of 29 interconnected proteins was retrieved by selecting the genes that were part of a significant sub network throughout *all* six studies (“late” integration, Figure 3). Closer inspection revealed that the majority of these genes have already been implicated as potential therapeutic targets in the treatment of either breast cancer or other types of cancer. This consensus gene set appears to play a pivotal role in the regulation of the cell cycle as not only a considerable enrichment for terms involving the cell cycle ($p = 2.6$

$\times 10^{-16}$), but as well an enrichment for proteins with known activating capacities was found (5 out of 29 are protein kinases, $p = 0.0033$). Interestingly, all genes are on average higher expressed within the “POOR” labeled samples compared to the “GOOD” labeled samples, fitting the cancer’s hallmark of a shortened cell cycle time. Moreover, all these genes are connected to each other by at least one (predicted) physical interaction exceeding $T_{ppi} = 500$, thereby suggesting a plausible molecular mechanism how primary breast tumors acquire or maintain their metastatic capacities.

4.4 An “early” integration approach reveals new sub network markers

We showed that application of the proposed method to six different data sets studying an identical phenotype led to a highly reproducible dissection of the data in at least eight distinct processes. Besides these eight broadly picked up processes, additional smaller clusters are visible along the diagonal in Figure 2, suggesting that there might be more ongoing processes in primary breast tumor tissue that are harder to detect. By applying the proposed method to the data from the six studies concatenated (“early integration”), three new putative sub network markers for metastasis were identified in addition to the eight previously established sub network markers (Figure 4). These three new putative sub network markers for metastasis (Figure 4: A to C) could be related to: unfolded protein binding (GO:0051082), cell-cell adhesion (GO:0016337) and proteasome complex (GO:0000502). All previously established sub network markers now dropped below the set significance threshold $T_3 \leq 0.05$ and showed a significant enrichment for at least a single GO term. The newly established sub networks B (cell-cell adhesion) and C (proteasome complex) and the previously established sub network markers I (JUN & FOS signaling) and V (Cell Cycle Phase) remained significant even after a Bonferroni correction for multiple testing (sub networks with red bounding box in Figure 4). All genes identified by the “late” integration approach were again part of significant sub networks found in the “early” approach, predominantly sub network V (26 out of 29), except for the gene

STMN1. We therefore can view cluster V in Figure 4 as an extension of the consensus sub network in Figure 3, containing 22 more candidate genes.

4.5 A more consistent gene selection is performed compared to other methods

Consistency in gene selection by the proposed method was compared to a classical gene ranking approach known as forward filtering, as described by van Vliet *et al.*²⁴ (Experimental Procedures 3.2.2) and a greedy network approach, as described by Chuang *et al.*⁶. Forward filters were used to find optimal predicting gene sets using either all available probes on the array (Table 2: FWD, $n = 22,283$) or all genes mapped to the protein-protein interaction network (Table 2: FWDNetw, $n = 9,290$). When starting with a reduced set of initial genes (FWDNetw), only a few additional genes were required for obtaining predictors with very similar prediction performances than when started with the set of all genes (FWD). Both network approaches selected considerably more genes as compared to both settings in which the forward filter was employed. This observation was most extreme for the greedy network approach of Chuang *et al.* (Table 2: ChuangNetw) for which from 11.6% to 23.0% of the genes mapped to the PPI network ($n = 9,290$) were selected in hundreds of overlapping sub networks. Application of the proposed method (Table 2: CoRegNetw) resulted in the identification of comprehensible numbers of disjoint co-regulated sub networks and implicating only 1.4% to 5.5% of the genes mapped to the PPI network.

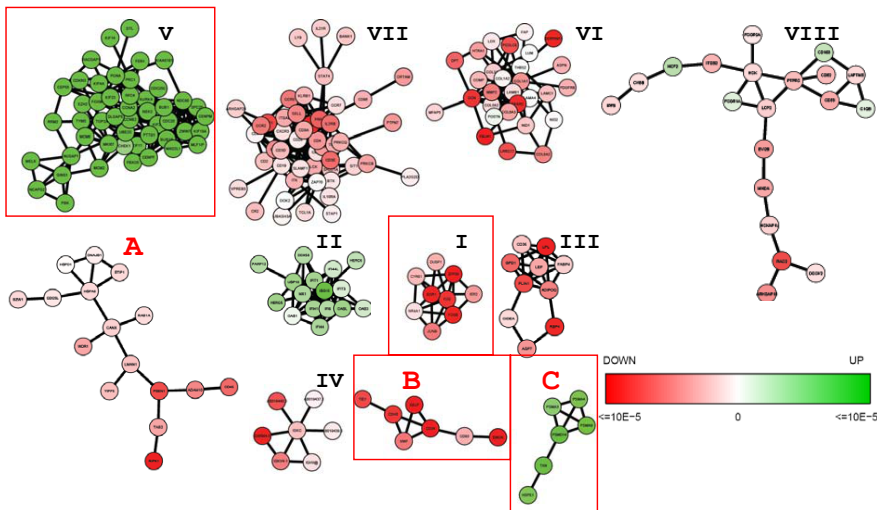


FIGURE 4: SUB NETWORK MARKERS IDENTIFIED WITH AN EARLY INTEGRATION APPROACH. Data of the six studies was concatenated prior to applying the procedure for sub network identification. Resulting sub networks marked with black roman numerals correspond to the reported eight consistently identified sub networks, also indicated in Figure 2. Sub networks A, B and C were newly identified and were enriched for the GO terms: unfolded protein binding (GO:0051082), cell-cell adhesion (GO:0016337) and proteasome complex (GO:0000502) respectively. Significant sub networks ($T_s \leq 0.05$) showing a functional enrichment for at least one GO category were reported for this analysis only. Sub networks marked by red bounding boxes remained significant after correction for multiple testing.

	FWD	FWDNetw	ChuangNetw	CoRegNetw		
	# genes [%]	# genes [%]	# genes [%]	# netw. [μ]	# genes [%]	# netw. [μ]
Des	49 (0.22)	51 (0.55)	1437 (15.5)	356 (14.4)	130 (1.4)	25 (5.2)
Mil	21 (0.09)	28 (0.30)	2137 (23.0)	662 (14.2)	240 (2.6)	35 (6.9)
Loi	59 (0.26)	75 (0.80)	1098 (11.8)	317 (13.0)	515 (5.5)	80 (6.4)
Paw	48 (0.22)	44 (0.47)	1237 (13.3)	293 (13.7)	290 (3.1)	52 (5.8)
Wan	55 (0.25)	60 (0.65)	1004 (10.8)	423 (11.6)	184 (2.0)	38 (4.8)
Sch	65 (0.29)	56 (0.60)	1696 (18.3)	331 (14.8)	172 (1.9)	22 (7.8)

TABLE 2: RESULTS OF SELECTING PREDICTIVE GENES USING DIFFERENT METHODS ON SIX BREAST CANCER STUDIES. Forward filters (following van Vliet *et al.*²⁴) were used to extract the optimal number of predictive genes (columns # genes (%) refer to the number and percentage of selected genes) when initially starting with all genes on the array (FWD) or all genes mapped to the PPI network (FWDNetw). The method proposed in this article (CoRegNetw) was also compared to the network approach of Chuang *et al.*⁶ (ChuangNetw) and for methods the number of sub networks (# netw.) and average sub network sizes (μ) were reported also.

Consistency of selected genes across different studies using the four previously introduced methods was assessed by calculating (1) Jaccard indices indicating gene set similarities and (2) odds ratios indicating the increase in risk for genes of being selected as a result of a previous selection in another study. The proposed network approach (CoRegNetw) considerably outperformed both ranking settings (FWD and FWDNetw) for all pair wise comparisons between studies for both criteria (Table 3). Whereas the mean Jaccard index was 2.5% and 2.7% for the ranking approaches, respectively, our method showed a mean Jaccard index of 25.9%. Chuang's greedy network approach was outperformed for all odds ratios (Table 3, panel C and D below diagonal), but not for all Jaccard indices (Table 3, panel C and D above diagonal). Although pair wise comparisons involving the Loi study showed lower similarities for our method compared to those observed when employing the method proposed by Chuang *et al.*, the mean Jaccard index of our method still substantially outperformed the means calculated on all other methods (21.9% for CoRegNetw versus 2.7%, 2.5%, and 16.7% for respectively FWD, FWDNetw and ChuangNetw).

4.6 Network approaches do not outperform classical ranking approaches in a cross study prediction evaluation

We next were interested whether our method for a highly reproducible dissection in functionally coherent sub networks would improve the robustness

of cross study prediction performances. We evaluated the prediction performances in two settings. In both settings a gene set is derived from a first study. In the first setting, denoted "*passing GeneSet*", this gene set is then passed to a second study, where the actual prediction rule is build and evaluated using a proper cross validation. In the second setting, this gene set is used to train a prediction rule with the first study and is evaluated only on the second study. This setting is denoted as "*passing Classifier*" (Figure 5 and 6).

In the "*passing GeneSet*" setting (Figure 5), network approaches either outperform or show comparable classification performances as compared to classical rankings. Notably, when evaluating on the Loi study Chuang's approach, it shows a considerable improvement compared to the other methods and when evaluating on the Schmidt study our method considerably outperforms other methods. Prediction performances of the two integration approaches "early" and "late" were evaluated as well. Whereas the "early" integration approach (dark blue diamonds) improves or at least not significantly worsens the prediction performances upon the mean single study approaches (yellow diamonds), the "late" integration approach shows an adverse effect. Especially for the Loi study, the "late" integration approach seems to fail.

In the clinically more relevant "*passing Classifier*" setting (Figure 6), variations in prediction performances have increased, as expected, compared to the "*passing GeneSet*". Now, classical ranking approaches consistently outperform

A: FWD

OR\JI	Des	Mil	Loi	Paw	Wan	Sch
Des		<i>0.01</i>	<i>0.00</i>	<i>0.08</i>	<i>0.09</i>	<i>0.05</i>
Mil	9.6		<i>0.00</i>	<i>0.03</i>	<i>0.00</i>	<i>0.02</i>
Loi	1.0	1.0		<i>0.00</i>	<i>0.01</i>	<i>0.00</i>
Paw	37.3	21.1	1.0		<i>0.04</i>	<i>0.05</i>
Wan	44.8	1.0	2.9	16.4		<i>0.02</i>
Sch	17.4	15.4	1.0	17.8	5.5	

B: FWDNetw

OR\JI	Des	Mil	Loi	Paw	Wan	Sch
Des		<i>0.04</i>	<i>0.00</i>	<i>0.09</i>	<i>0.09</i>	<i>0.02</i>
Mil	23.0		<i>0.00</i>	<i>0.01</i>	<i>0.00</i>	<i>0.04</i>
Loi	1.0	1.0		<i>0.01</i>	<i>0.01</i>	<i>0.00</i>
Paw	47.4	7.9	2.9		<i>0.03</i>	<i>0.02</i>
Wan	38.5	1.0	2.1	11.8		<i>0.02</i>
Sch	6.9	20.8	1.0	8.1	5.9	

C: ChuangNetw

OR\JI	Des	Mil	Loi	Paw	Wan	Sch
Des		<i>0.21</i>	<i>0.16</i>	<i>0.20</i>	<i>0.15</i>	<i>0.19</i>
Mil	3.3		<i>0.16</i>	<i>0.19</i>	<i>0.17</i>	<i>0.22</i>
Loi	3.0	2.6		<i>0.14</i>	<i>0.12</i>	<i>0.15</i>
Paw	3.9	3.2	2.6		<i>0.13</i>	<i>0.18</i>
Wan	3.0	3.1	2.5	2.5		<i>0.14</i>
Sch	3.0	2.9	2.5	3.0	2.4	

D: CoRegNetw

OR\JI	Des	Mil	Loi	Paw	Wan	Sch
Des		<i>0.25</i>	<i>0.07</i>	<i>0.25</i>	<i>0.32</i>	<i>0.33</i>
Mil	71.3		<i>0.07</i>	<i>0.42</i>	<i>0.20</i>	<i>0.38</i>
Loi	8.8	4.7		<i>0.08</i>	<i>0.08</i>	<i>0.05</i>
Paw	76.2	123.1	4.7		<i>0.22</i>	<i>0.35</i>
Wan	117.6	32.2	6.9	39.3		<i>0.22</i>
Sch	126.7	139.1	5.6	120.6	44.3	

TABLE 3: GENE SET SIMILARITIES. Gene set similarities calculated between gene sets obtained from significant gene lists (FWD and FWDNetw) or significant sub networks (the method of Chuang et al. ChuangNetw and the method proposed in this paper CoRegNetw) within each single study. Shown similarity measures are the Jaccard index (above diagonal, italic) or odds ratio (below diagonal, not italic) grouped per method (Panels A to D). Pair wise comparisons depicted in bold are outperforming all competing methods, the comparisons depicted not in bold are outperformed by at least one other method.

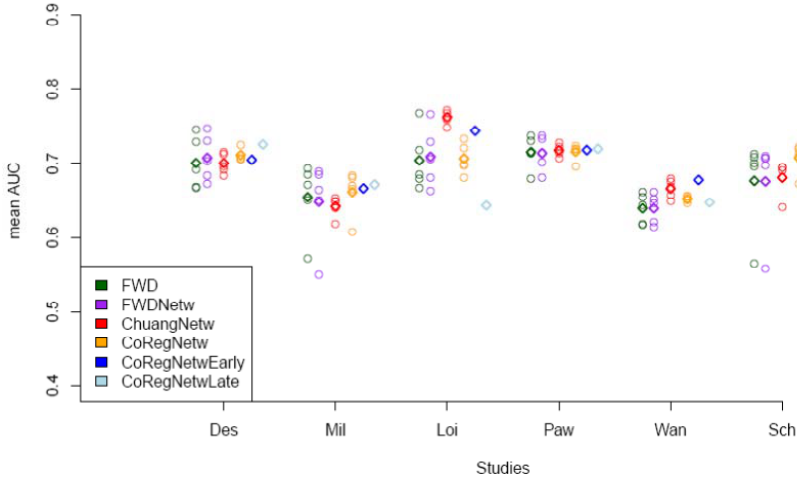


FIGURE 5: CROSS STUDY PREDICTION PERFORMANCES OF SEVERAL METHODS GROUPED PER EVALUATION STUDY IN THE "PASSING GENESet" SETTING. Circles indicate results of cross study prediction performances involving a single study for training, diamonds show results involving five studies for training. The latter can either be a summarization statistic (mean) or be the result of an integration approach (CoRegNetEarly and CoRegNetLate).

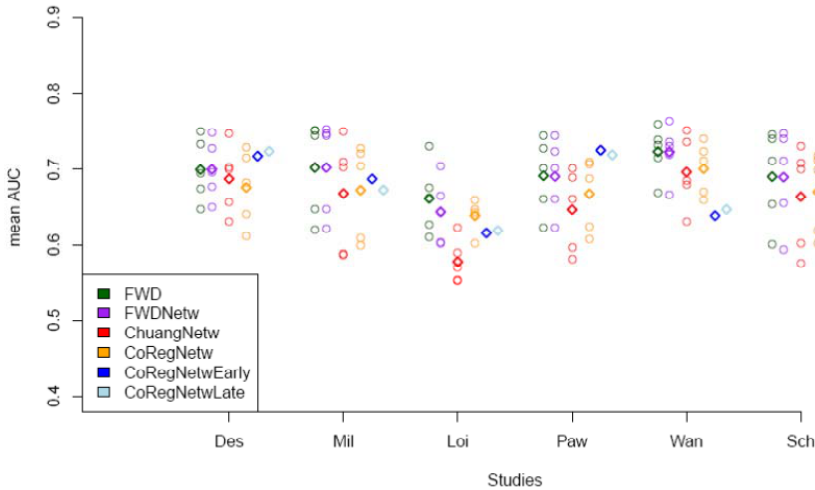


FIGURE 6: CROSS STUDY PREDICTION PERFORMANCES OF SEVERAL METHODS GROUPED PER EVALUATION STUDY IN THE "PASSING CLASSIFIER" SETTING. Circles indicate results of cross study prediction performances involving a single study for training, diamonds show results involving five studies for training. The latter can either be a summarization statistic (mean) or be the result of an integration approach (CoRegNetEarly and CoRegNetLate).

network approaches. Notably, Chuang's method applied to the Loi study now shows the worst overall performance. The "early" and "late" integration approaches now show a correlated behavior across data sets, improving upon mean single study performances (yellow diamond) in four out of six times and improving upon both ranking approaches in three out of six evaluations. The integration approaches especially seem to deteriorate prediction performances for the Loi and Wang Study.

5. Discussion

We proposed a method for *de novo* grouping of genes by dissecting the protein-protein interaction network into disjoint sub networks using pair wise gene expression correlation measures. By selecting sub networks significantly correlated with phenotypic outcome, we expected that this would result in a functionally more coherent gene selection as compared to competing risk profile predictors. We verified this by applying the proposed method and two competing methods to a breast cancer compendium composed of six different studies. Furthermore, we investigated whether the expected consistency in gene selection would have benefits for risk prediction of metastasis.

Experiments on the breast cancer compendium have shown that the proposed methodology leads to a *functionally coherent* dissection of genes into sub networks. Furthermore, similarity analyses showed that a considerable amount of these sub networks are picked up *consistently* across studies, suggesting that previously

reported low overlaps in predictive gene sets can not be attributed to differences in ongoing basal processes picked up by the different studies. The observation that sub networks were consistently identified underlines the weaknesses of previous methods that purely rely on pre-defined functional groupings for their analyses and interpretation.

Quite contrary to classical gene ranking approaches, extensive overlap between predictive gene sets derived from different studies is observed when employing the proposed method. A consensus gene set that consisted of genes that were part of significant sub networks in *all* six studies was predominantly composed of genes previously implicated in a wide variety of cancers, and was heavily enriched for both the GO term "cell cycle phase" as for the presence of proteins with known regulatory capacities (kinases). This so called "late" integration approach improves robustness in gene selection but at the cost of power to detect potential candidate genes. This was clearly illustrated by the fact that the Loi study alone was most decisive for the gene composition of the consensus gene set, due to its relatively small significant sub network representing cell cycle phase.

A consistent overlap between studies also cleared the way for an "early" integration approach where the data of all studies is concatenated before detecting sub networks. This approach confirmed and extended the consensus sub network found by the late integration approach and identified potential new sub network markers involved in JUN & FOS signaling, cell-cell adhesion and the proteasome complex.

When comparing consistency in gene set selection across studies over different methods, the proposed method always significantly outperforms classical ranking approaches. Chuang's greedy network approach⁶ is outperformed as well except for comparisons involving the Loi study. However, on average Chuang's method is outperformed using this metric. Moreover when odds ratios for the risk of reselection over the risk of no reselection were compared, our method substantially outperforms Chuang's method for all pair wise comparisons. This suggests that once a gene is implicated by our method in one study, the chance that it will be implicated again in another study is much higher.

Despite the observed consistency in selection of gene sets, no improvements in classification performance were observed when compared to competing methods in the clinically most relevant evaluation setting ("*passing Classifier*"). Moreover, when no integration approach was employed to exploit the presence of multiple studies, all network approaches were outperformed by the classical gene ranking approaches, suggesting that the higher interpretability comes at the expense of predictive power. In the work of Chuang *et al.*⁶ an evaluation setting similar to the one denoted as "*passing GeneSet*" was used. Indeed we confirmed that in such a setting, network approaches either outperform or show comparable classification performances as compared to classical rankings. However, we would like to issue a word of caution when interpreting the classification results while employing the "*passing GeneSet*" setting. The results with the overall

highest prediction performance in the "*passing GeneSet*" setting were created by applying Chuang's feature selection on the Loi study. Meanwhile, these results also show the largest discrepancy with the setting denoted as "*passing Classifier*", where it shows the overall lowest prediction performance. We hypothesize that other studies might be particularly uninformative about the Loi study, as this study is the only one in which the majority is treated with tamoxifen, thereby negating or possibly reversing previously observed relations between gene expressions and outcome.

When considering the "*passing Classifier*" setting, integration approaches seem to deteriorate prediction performances especially for two studies: Loi en Wang. In case of the Loi study, integration approaches are expected to be even more sensitive for the previously described disruptive effects of tamoxifen on relations between gene expressions and outcome. Due to the larger amounts of training data, more specific predictors are obtained, which are less capable to generalize when underlying processes are differing. The drop in prediction performance can be explained by the fact that the Wang study is the only one with a balanced number of "POOR" and "GOOD" outcomes. Other studies have a much lower incidence of "POOR" outcome class and therefore training on these studies will focus the classifier mainly on recognizing the more heterogeneous subset of "GOOD" outcome subjects.

Whereas integration approaches showed some adverse effects in the "*passing GeneSet*" evaluation setting,

correlated prediction performances were observed in the “*passing Classifier*” setting. When ignoring the Loi and Wang study, “early” integration approaches seem to only slightly outperform “late” integration approaches. This observation is especially relevant when considering integration of data measured on different platforms in which an “early” integration approach is not feasible.

Employing several analytic strategies, we consistently found a gene sub network involved in an established hallmark of cancer, cell cycle phase, which is persistent over-expressed in all six breast cancer studies in the “POOR” labeled samples compared to the “GOOD” labeled samples. Moreover, application of the proposed method in an “early” integration approach revealed new putative sub network markers, implicating molecular mechanisms involved in cell-cell adhesion, proteasome complex and JUN & FOS signaling to be involved in metastasis. Although not directly improving previously reported cross study classification performances, knowledge-based decomposition of measured gene expression data into co-regulated modules seems to result in a consistent and biologically relevant feature selection and might therefore have a general applicability beyond the field of breast cancer.

6. Acknowledgements

This work was supported by a grant from the Medical Delta (<http://www.medicaldelta.nl>).

7. References

1. Weigelt, B., Peterse, J.L. & van 't Veer, L.J. Breast cancer metastasis: markers and models. *Nat Rev Cancer* **5**, 591-602 (2005).
2. Ein-Dor, L., Kela, I., Getz, G., Givol, D. & Domany, E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* **21**, 171-8 (2005).
3. Tian, L. *et al.* Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A* **102**, 13544-9 (2005).
4. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
5. Goeman, J.J., van de Geer, S.A., de Kort, F. & van Houwelingen, H.C. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **20**, 93-9 (2004).
6. Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D. & Ideker, T. Network-based classification of breast cancer metastasis. *Mol Syst Biol* **3**, 140 (2007).
7. Snel, B., Lehmann, G., Bork, P. & Huynen, M.A. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res* **28**, 3442-4 (2000).
8. Michaut, M. *et al.* Protein complexes are central in the yeast genetic landscape. *PLoS Comput Biol* **7**, e1001092 (2011).
9. Zhang, C., Liu, S. & Zhou, Y. Fast and accurate method for identifying high-quality protein-interaction modules by clique merging and its application to yeast. *J Proteome Res* **5**, 801-7 (2006).
10. Barabasi, A.L. & Oltvai, Z.N. Network biology: understanding the cell's

- functional organization. *Nat Rev Genet* **5**, 101-13 (2004).
11. <ftp://ftp.ncbi.nih.gov/pub/geo>. Gene Expression Omnibus (GEO) is a database repository of high throughput gene expression data and hybridization arrays, chips, microarrays.
 12. Desmedt, C. *et al.* Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res* **13**, 3207-14 (2007).
 13. Miller, L.D. *et al.* An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci U S A* **102**, 13550-5 (2005).
 14. Loi, S. *et al.* Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics* **9**, 239 (2008).
 15. Pawitan, Y. *et al.* Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res* **7**, R953-64 (2005).
 16. Wang, Y. *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671-9 (2005).
 17. Schmidt, M. *et al.* The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res* **68**, 5405-13 (2008).
 18. Gautier, L., Cope, L., Bolstad, B.M. & Irizarry, R.A. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307-15 (2004).
 19. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.
 20. R-Core-Team. R: A Language and Environment for Statistical Computing. (2013).
 21. <http://www.affymetrix.com/support>. Affymetrix Support.
 22. <http://www.biomart.org/biomart/martview>. Biomart: A repository for genomic annotations.
 23. von Mering, C. *et al.* STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* **31**, 258-61 (2003).
 24. van Vliet, M.H. *et al.* Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability. *BMC Genomics* **9**, 375 (2008).
 25. Wessels, L.F. *et al.* A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics* **21**, 3755-62 (2005).
 26. Jaccard, P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise de Sciences Naturelles*, 547-579 (1901).
 27. Edwards, A.W.F. The measure of association in a 2x2 table. *JSTOR* **126**, 1-28 (1968).
 28. Shannon, P.T., Grimes, M., Kutlu, B., Bot, J.J. & Galas, D.J. RCytoscape: tools for exploratory network analysis. *BMC Bioinformatics* **14**, 217 (2013).
 29. Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431-2 (2011).
 30. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).

Chapter 3:

Meta-Analysis on Blood Transcriptomic Studies Identifies Consistently Co-Expressed PPI Modules as Robust Markers of Human Aging

Erik B. van den Akker^{1,2,§}, Willemijn M. Passtoors^{1,§}, Rick Jansen³, Erik W. van Zwet⁴, Jelle J. Goeman⁴, Marc Hulsman², Valur Emilsson⁵, Marcus Perola⁶, A.H.M. Gonneke Willemsen⁷, Brenda W.J.H. Penninx³, Bas T. Heijmans¹, Andrea B. Maier⁸, Dorret I. Boomsma^{3,7}, Joost N. Kok^{1,9}, P. Eline Slagboom^{1,10}, Marcel J.T. Reinders², Marian Beekman^{1,10}

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

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

3. Department of Psychiatry, VU University Medical Center, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands

4. EMGO Institute for Health and Care Research, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands

5. Medical Statistics, Leiden University Medical Center, Leiden, Netherlands

6. Icelandic Heart Association, Kópavogur, Iceland

7. National Institute for Health and Welfare, Helsinki, Finland

8. Department of Biological Psychology, VU University, Amsterdam, The Netherlands

9. Gerontology and Geriatrics, VU University Medical Center, Amsterdam, The Netherlands

10. Algorithms, Leiden Institute of Advanced Computer Science, University of Leiden, Leiden, Netherlands

11. Netherlands Consortium for Healthy Ageing, Leiden, Netherlands

§ Joint First Authors.

1. Abstract

The bodily decline that occurs with advancing age strongly impacts on the prospects for future health and life expectancy. Despite the profound role of age in disease etiology, knowledge about the molecular mechanisms driving the process of aging in humans is limited. Here, we used an integrative network-based approach for combining multiple large-scale expression studies in blood (2,539 individuals) with protein-protein Interaction (PPI) data for the detection of consistently co-expressed PPI modules that may reflect key processes that change throughout the course of normative aging. Module detection followed by a meta-analysis on chronological age identified fifteen consistently co-expressed PPI modules associated with chronological age, including a highly significant module ($p = 3.5 \times 10^{-38}$) enriched for 'T-cell activation' marking age-associated shifts in lymphocyte blood cell counts ($R^2 = 0.603$; $p = 1.9 \times 10^{-10}$). Adjusting the analysis in the compendium for the 'T-cell activation' module showed five consistently co-expressed PPI modules that robustly associated with chronological age and included modules enriched for 'Translational elongation', 'Cytolysis' and 'DNA metabolic process'. In an independent study of 3,535 individuals, four of five modules consistently associated with chronological age, underpinning the robustness of the approach. We found three of five modules to be significantly enriched with aging-related genes, as defined by the GenAge database, and association with prospective survival at high ages for one of the modules including *ASF1A*. The hereby-detected age-associated and consistently co-expressed PPI modules therefore may provide a molecular basis for future research into mechanisms underlying human aging.

2. Introduction

A steadily growing life expectancy of the general western population throughout the past two centuries¹ has imposed the urgency for understanding the adverse effects of aging for public health and its relation to the observed large variation in healthy lifespan². Age-dependent detrimental processes strongly attenuate prospects for future health, with chronological age being the major risk factor for mortality and virtually all common diseases in the western world³. Aging is a systemic ailment marked by a gradual metabolic decline eventually leading to a state of senescence on both the cellular and organismal level that seems to be caused by the accumulation of damage over time⁴. Despite their profound role for disease etiology, the existing knowledge concerning the molecular mechanisms driving biological aging processes in humans is limited.

Construction of consistent age-associated signatures has proven to be challenging as a multitude of gene expression studies have identified age-associated genes so far, though with limited mutual overlap^{5,6}. This inconsistency is most likely due to variable technical circumstances, small study sizes, and low signal-to-noise ratios, typically observed when analyzing the aging transcriptome. More similarity was observed at the pathway level, across tissues and even species^{7,8} suggesting that the analysis of the aging transcriptome by functionally grouped gene sets is a promising alternative for the classical individual-gene analyses.

Rather than employing literature-based sets of genes sharing similar

biological functions, so-called network approaches are increasingly used, which infer functional clusters of genes from the expression data itself by exploiting gene co-expression patterns hidden within the data⁹. Alternatively, changes in these gene co-expression patterns that occur with age might be used for inferring a functional grouping from the data¹⁰. However, co-expression patterns may contain spurious gene-gene correlations¹¹, which makes the use of multiple data sources simultaneously or the integration with other additional information sources on functional relationships between genes desirable.

Established modulators of aging processes in model organisms were reported to spatially cluster within networks constructed of protein-protein interaction (PPI) data^{12,13}. Hence, PPI networks can be exploited for prioritizing new aging-associated genes^{14,15} or for refining modules of co-expressed genes that are correlated during the course of aging¹⁶. We previously demonstrated that the inference of these so-called co-expressed PPI modules has a high reproducibility across multiple expression datasets in breast cancer¹⁷, and here we extend this algorithm to combine multiple gene expression datasets on aging.

Though many algorithms for network inference exist¹⁸, relatively little attention has gone to the problem of network inference and subsequent associations with a phenotype using multiple heterogeneous expression data sources simultaneously. Merging the expression data into a single set and using this for network inference clearly surpasses the differences in correlation structures present within each

Study	Tissue	Ethnicity	# start total ^{††}	# end total [†]	# males (%) [†]	mean age [†]	min age [†]	max age [†]
SAFHS*	Lymphocytes	Mexican Americans (USA)	1240	1240	506 (40.8%)	39.3	15	94
IFB_A	Peripheral blood	Caucasian (Icelandic)	904 ^A	411	198 (48.2%)	48.8	19	84
IFB_B	Peripheral blood	Caucasian (Icelandic)	904 ^A	434	180 (41.5%)	46.2	20	76
DILGOM	Peripheral blood	Caucasian (Finnish)	518 ^B	454	195 (43.0%)	51.6	30	70

TABLE 1: DESCRIPTIVES OF THE DATASETS COMPOSING THE COMPENDIUM.

(*) Expression and phenotypic data were obtained from ArrayExpress under accessions: E-TABM-305

(††) Number of individuals with matching phenotypic data per study when obtained.

(A) Data of IFB was measured in two batches. This figure indicates the total number of individuals before preprocessing or removal of duplicates across batches.

(B) A small batch was detected and all samples belonging to it were removed.

(†) Statistics computed after preprocessing.

dataset. Irrespective of the type of network inference chosen, we propose to handle such heterogeneity by integrating the gene-gene similarity measures obtained across expression datasets using a suitable meta-analysis setting. Thus, in the approach described in this paper, we employ a meta-analysis for inferring a consistent gene-gene network that serves as a basis for identifying consistently co-expressed PPI modules, which are subsequently analyzed with respect to chronological age across datasets using again a meta-analysis.

To robustly characterize the changes of the blood transcriptome associated with chronological age, we have build a compendium using three large-scale transcriptomic studies¹⁹⁻²¹ generated in blood comprising 2,539 individuals on which we applied our integrative network

approach. For comparison, two types of individual-gene meta-analyses were performed as well, which in combination with an enrichment analysis yielded only broad terms for age-associated cellular processes. Application of our integrative network-based approach, yielded five consistently co-expressed PPI modules showing robust age associations and functional enrichments for ‘Translational elongation’, ‘Cytolysis’ and ‘DNA metabolic process’, which seem to reflect downstream mTOR signaling events or cell-cycle checkpoints. Finally, we show that four of five modules replicate in an independent cohort, and that they are enriched for known longevity- and aging-related genes and that the expression of one module associates with prospective survival at old age.

3. Results

3.1 The largest transcriptome compendium for normative aging

To robustly characterize the changes of the blood transcriptome throughout the course of normative aging in the range of 15-94 years, we built a gene expression compendium using three large-scale transcriptomic studies performed in blood: the San Antonio Family Heart Study (SAFHS)¹⁹, the Icelandic Family Blood (IFB) cohort²⁰ and the Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome (DILGOM) study²¹. Data of IFB were measured in two roughly equally sized batches, from this point on referred to as IFB_A and IFB_B, and was treated as two separate datasets in the downstream analysis. Data quality was critically reassessed and re-annotated yielding a compendium of 9,047 unique genes expressed in 2,539 individuals divided over four datasets (SAFHS: 1,240, IFB_A: 411, IFB_B: 435, DILGOM: 454; Table 1 & Experimental Procedures).

3.2 Limited overlap of age-associated genes between studies within the compendium

The most straightforward method for an integrative analysis across datasets is to first compute the age-association genes per dataset and subsequently inspect the overlap of significant results. A linear model adjusted for gender yielded between 111 (1.2%) and 1,103 (12.2%) significantly age-associated genes per dataset (Bonferroni correction, $\alpha \leq 0.05$), of which 26 genes were significantly associated with age in all four datasets (Figure 1 and Table S1, Supplemental Materials). These results

confirmed the high discrepancy between lists of age-associated genes previously reported in literature, even though now observed in equal or similar tissues^{5,6}.

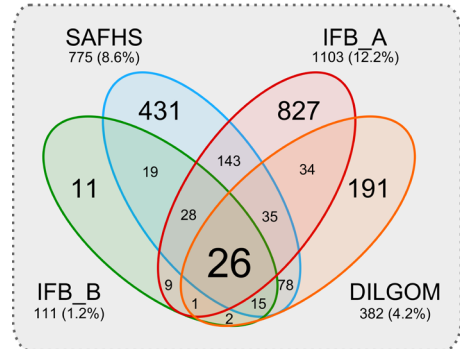


FIGURE 1: Significantly age-associated genes in studies of the blood compendium. A Venn analysis was performed for inspecting the overlap of the significantly age-associated genes found within different studies. The majority of the consistently detected age-associated genes (24 of 26) show a decreased expression with advancing age and include the following: *ARH*, *BACH2*, *CCR7*, *ECRG4*, *EDAR*, *EPHA1*, *EPHX2*, *FAM102A*, *FAM134B*, *FBLN2*, *FCGBP*, *FLNB*, *IL24*, *LRRN3*, *NELL2*, *NMT2*, *NRCAM*, *OXNAD1*, *PDE9A*, *PHGDH*, *PIK3IP1*, *SIRPB2*, *SUSD3*, and *TSGA14*. The remaining 2 consistently age-associated genes showing increased expressions are *ARP10* and *SYT11*. See Table S1 (Supplemental Materials) for more details.

3.3 Rank-based integration of age-associated genes improves consistency between studies

As repeatedly applied cutoffs across multiple heterogeneous datasets may lead to high false exclusion rates of age-associated genes, we investigated whether age-association rankings were consistently high across datasets by applying a rank integration approach^{6,22}. From the 9,047 genes present in the compendium, 247 consistently showed highly ranked differential expressions with age across the four datasets, of which 195 remained

significant after permutation tests (both at $FDR \leq 0.05$) (Experimental Procedures). Of these 195 genes, 128 (65.6%) showed decreased and 67 (34.4%) showed increased expression levels with age. The top 25 genes with increased and decreased expression are displayed in Tables 2 and 3, respectively, and include many of the age-associated genes previously identified, like *LRNN3*, *LEF1*, and *SYT11*²³⁻²⁵. Results for all 9,047 genes in the compendium are provided in Table S2 (Supplemental Materials).

3.4 Functional enrichments of individual-gene analysis are not informative for normative aging

We next identified enriched functional groupings among genes significantly associated with normative aging using DAVID focusing on GO_FAT terms. Whereas the 26 genes from the overlap did not yield any significantly enriched terms, the 195 significant genes obtained with the rank integration approach yielded 11 significant enriched groupings when run at default settings (Tables S3 and S4, Supplemental Materials respectively). Interestingly, enriched terms include 'Glycosylation site: N-linked' ($p = 6.1 \times 10^{-5}$, Benjamini corrected), previously linked to the inflamm-aging theory²⁶. However, as most of the 11 identified terms are rather broadly defined, like 'disulfide bond' or 'signal peptide', little detailed knowledge is gained on potential molecular mechanisms underlying normative aging following the individual-gene analysis approach.

3.5 A novel integrative network approach for detecting consistent co-expressed PPI modules

To improve robustness against noise and increase power, we used a novel integrative network-based approach to explore functional age-associated groupings of genes. The proposed approach detects consistently co-expressed PPI modules across multiple datasets (for details see Experimental Procedures and Data S1, Supplemental Materials). Using the four transcriptomic datasets mapped onto the PPI network, we detected a total of 162 consistently co-expressed PPI modules ranging in size from 2 to 37 genes (see Figure S1, Supplemental Materials for a complete overview). The following steps in our analysis were limited to the subset of 27 co-expressed PPI modules counting at least five genes. Application of DAVID yielded significant functional enrichments for 19 of the 27 identified co-expressed PPI modules (Table S5, Supplemental Materials), suggesting that the applied approach grouped genes according to plausible biological functions.

3.6 Age-associated co-expressed PPI modules point toward T-cell activation

To test whether transcriptional changes of the 27 identified modules associate with chronological age, an expression profile for each module was constructed by determining the mean expression of the genes within a detected co-expressed PPI module per individual. As with the individual-gene analysis, we proceeded by computing the associations of the module expressions with age while adjusting for gender for each dataset separately. Only

Symbol	GeneID	p-value ¹	q-value ¹	p-value ²	q-value ²
GPR56	9289	5.3×10^{-09}	4.8×10^{-05}	1.0×10^{-06}	0.0018
HF1	3075	2.3×10^{-08}	8.1×10^{-05}	1.0×10^{-06}	0.0018
SYT11	23208	2.7×10^{-08}	8.1×10^{-05}	$\leq 5.0 \times 10^{-7}$	0.0018
ARP10	164668	7.3×10^{-08}	1.7×10^{-04}	1.0×10^{-06}	0.0018
B3GAT1 (CD57)	27087	1.1×10^{-07}	2.0×10^{-04}	3.0×10^{-06}	0.0021
SLC1A7	6512	1.8×10^{-07}	2.6×10^{-04}	3.2×10^{-05}	0.0110
IFNG	3458	5.0×10^{-07}	6.4×10^{-04}	1.1×10^{-05}	0.0065
DSCR1L1	10231	6.1×10^{-07}	6.8×10^{-04}	2.0×10^{-06}	0.0021
ARK5	9891	7.9×10^{-07}	7.9×10^{-04}	3.0×10^{-06}	0.0021
PIG13	81563	9.3×10^{-07}	8.8×10^{-04}	1.0×10^{-06}	0.0018
SPUVE	11098	1.1×10^{-06}	8.8×10^{-04}	1.2×10^{-05}	0.0067
PDGFRB	5159	1.2×10^{-06}	8.8×10^{-04}	1.5×10^{-06}	0.0021
EDG8	53637	1.4×10^{-06}	9.4×10^{-04}	7.8×10^{-05}	0.015
MARLIN1	152789	1.5×10^{-06}	9.4×10^{-04}	5.0×10^{-06}	0.0032
TGFBR3	7049	2.0×10^{-06}	0.0012	2.8×10^{-05}	0.011
GZMB	3002	2.4×10^{-06}	0.0013	5.0×10^{-04}	0.050
CX3CR1	1524	2.9×10^{-06}	0.0014	2.9×10^{-05}	0.011
STYK1	55359	3.3×10^{-06}	0.0015	4.8×10^{-05}	0.013
ADRB2	154	3.7×10^{-06}	0.0016	3.0×10^{-06}	0.0021
GAF1	26056	7.1×10^{-06}	0.0029	7.2×10^{-05}	0.015
CTSL	1514	7.7×10^{-06}	0.0030	3.2×10^{-04}	0.040
GFI1	2672	1.1×10^{-05}	0.0040	3.0×10^{-06}	0.0021
TTC38	55020	1.1×10^{-05}	0.0040	7.6×10^{-05}	0.015
AGPAT4	56895	1.2×10^{-05}	0.0041	2.5×10^{-06}	0.0021
GZMA	3001	1.4×10^{-05}	0.0045	3.3×10^{-04}	0.040

TABLE 2: TOP 25 GENES ACCORDING TO THE GENE STATISTIC (U_i) HAVING INCREASED EXPRESSION WITH AGE.

(1) p - and q -values determined using the gamma-distribution of the gene statistic, U_i

(2) p - and q -values determined using permutation of the gene statistic, U_i

one module (Figure 2A), enriched for ‘T-cell activation’, was significantly associated with age in each of the four datasets of the compendium. This module A contains genes commonly employed as markers for assessing the differentiation status of T-cell lineages, such as *CCR7*, *CD28*, and *TNFRSF7* (*CD27*). A fixed-effect meta-analysis on the expression of the different modules across

the datasets showed again that the ‘T-cell activation’ module was most significantly associated with age (Bonferroni corrected $p = 3.5 \times 10^{-38}$) (see also Experimental Procedures). The consistent age association of the ‘T-cell activation’ module, however, raises the concern that the identified modules reflect age-related changes in the proportions of cell populations in blood, as

Symbol	GeneID	p-value ¹	q-value ¹	p-value ²	q-value ²
LRRN3	54674	1.3×10^{-12}	1.2×10^{-8}	$\leq 5.0 \times 10^{-7}$	3.2×10^{-4}
FCGBP	8857	3.2×10^{-10}	1.5×10^{-6}	$\leq 5.0 \times 10^{-7}$	3.2×10^{-4}
CCR7	1236	1.1×10^{-9}	3.2×10^{-6}	$\leq 5.0 \times 10^{-7}$	3.2×10^{-4}
NELL2	4753	2.0×10^{-8}	4.5×10^{-5}	1.0×10^{-6}	3.8×10^{-4}
NRCAM	4897	3.1×10^{-8}	5.6×10^{-5}	$\leq 5.0 \times 10^{-7}$	3.2×10^{-4}
IGJ	3512	1.5×10^{-7}	2.3×10^{-4}	2.6×10^{-4}	0.019
LEF1	51176	1.9×10^{-7}	2.5×10^{-4}	$\leq 5.0 \times 10^{-7}$	3.2×10^{-4}
FAM134B	54463	2.2×10^{-7}	2.5×10^{-4}	$\leq 5.0 \times 10^{-7}$	3.2×10^{-4}
PACAP	51237	2.5×10^{-7}	2.5×10^{-4}	1.5×10^{-6}	4.8×10^{-4}
ITM2C	81618	2.8×10^{-7}	2.5×10^{-4}	3.5×10^{-6}	8.1×10^{-4}
PIK3IP1	113791	3.0×10^{-7}	2.5×10^{-4}	1.0×10^{-6}	3.8×10^{-4}
PDE9A	5152	5.1×10^{-7}	3.8×10^{-4}	1.0×10^{-6}	3.8×10^{-4}
BACH2	60468	6.9×10^{-7}	4.8×10^{-4}	1.0×10^{-6}	3.8×10^{-4}
FLJ12895	65982	9.5×10^{-7}	6.0×10^{-4}	1.5×10^{-6}	4.8×10^{-4}
FAM102A	399665	1.1×10^{-6}	6.0×10^{-4}	$\leq 5.0 \times 10^{-7}$	3.2×10^{-4}
FBLN2	2199	1.1×10^{-6}	6.0×10^{-4}	$\leq 5.0 \times 10^{-7}$	3.2×10^{-4}
FLNB	2317	1.2×10^{-6}	6.0×10^{-4}	$\leq 5.0 \times 10^{-7}$	3.2×10^{-4}
APEG1	10290	1.2×10^{-6}	6.0×10^{-4}	1.0×10^{-6}	3.8×10^{-4}
EPHX2	2053	1.3×10^{-6}	6.0×10^{-4}	1.5×10^{-6}	4.8×10^{-4}
TNFRSF17	608	1.3×10^{-6}	6.1×10^{-4}	1.2×10^{-4}	0.011
MYC	4609	1.6×10^{-6}	6.6×10^{-4}	3.5×10^{-6}	8.1×10^{-4}
NT5E	4907	1.7×10^{-6}	6.6×10^{-4}	1.0×10^{-6}	3.8×10^{-4}
TOSO	9214	1.7×10^{-6}	6.6×10^{-4}	1.0×10^{-6}	3.8×10^{-4}
ARH	26119	3.2×10^{-6}	0.0012	2.0×10^{-6}	6.2×10^{-4}
OXNAD1	92106	3.3×10^{-6}	0.0012	$\leq 5.0 \times 10^{-7}$	3.2×10^{-4}

TABLE 3: TOP 25 GENES ACCORDING TO THE GENE STATISTIC (U_i) HAVING DECREASED EXPRESSION WITH AGE.

(1) p - and q -values determined using the gamma-distribution of the gene statistic, U_i

(2) p - and q -values determined using permutation of the gene statistic, U_i

previously reported²⁷, rather than changes in gene expression.

3.7 T-cell activation module expression marks blood lymphocyte counts

To investigate the relation between the expression of the ‘T-cell activation’ module and the proportions of blood

cell populations, for which we have no data in the compendium, we revisited a transcriptomic dataset on peripheral blood measured in the Leiden Longevity Study (LLS)²⁵ (Data S1, Supplemental Materials). Using the expression data of 50 middle-aged and 50, 90-year-old individuals, we first confirmed the association with age

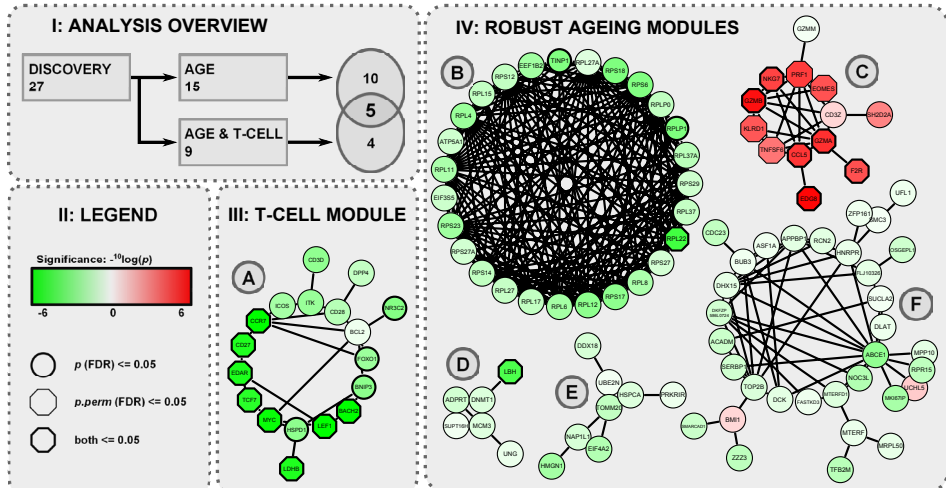


FIGURE 2: Overview main results of the integrative network-based approach. Panel 1: Overlap of the PPI network and cluster analysis of the transcriptomic data reveals 27 modules, 15 are significantly associated with age, 9 are significantly associated with age when corrected for the ‘T-cell activation’ module expression, and the 5 most robust findings are found in the overlap. Panel 2: Legend: Genes are represented by nodes, whose shape and color reflect the results of the individual-gene statistic (U_i). The red and green colors denote a correlating or anti-correlating relationship of gene expression with age, respectively. The intensity of the coloring indicates the significance of the gamma-distributed transformed rank product statistics. Nodes marked by a thick bordering or a hexagon shaped bordering represent genes with FDR adjusted p -values ≤ 0.05 for respectively the analytical and permutation-based approach. Panel 3: The co-expressed PPI module that is enriched for ‘T-cell activation’. Panel 4: B-F: 5 co-expressed PPI modules with expressions robustly associated with age. B, C and D: modules enriched for ‘Translational elongation’, ‘Cytolysis’, and ‘DNA metabolic process’, respectively. Node’s shape and color reflect the results of the individual-gene statistic (U_i).

of the expression of the ‘T-cell activation’ module ($p = 3.7 \times 10^{-5}$), and subsequently observed a significant correlation between the expression of the ‘T-cell activation’ module and lymphocyte counts ($R^2 = 0.603$, $p = 1.9 \times 10^{-10}$). These findings suggest that the previously observed age associations in the blood compendium are most probably confounded by the age-associated decline in lymphocyte counts. We also conclude that the expression of the ‘T-cell activation’ module could serve as a proxy for the age-associated decline in lymphocyte counts in the compendium.

3.8 Five co-expressed PPI modules associate with age independent of T-cell activation

Based on these findings, we adapted the fixed-effect meta-analysis to reanalyze the 27 modules in the compendium while adjusting for gender as well as the expression of the ‘T-cell activation’ module. This revealed nine modules significantly associated with chronological age, of which five also showed a significant association without adjusting for ‘T-cell activation’ (Figure 2B-F). These five modules thus exhibit the most robust expression changes with age and include (i) a large consistently down-regulated ribosomal module ($p = 9.4 \times 10^{-19}$), enriched

for ‘Translational elongation’ ($p = 4.5 \times 10^{-46}$); (ii) an up-regulated module containing among others several granzymes and the perforin gene ($p = 2.9 \times 10^{-24}$), enriched for ‘Cytolysis’ ($p = 9.4 \times 10^{-05}$); and (iii) a down-regulated module containing the *PARP1* (*ADPRT*) gene ($p = 3.1 \times 10^{-39}$) enriched for ‘DNA metabolic process’ ($p = 0.0036$). The two remaining modules were both down-regulated with advancing age and lacked any significant functional enrichments (Figure 2E,F; $p = 3.9 \times 10^{-11}$ and $p = 2.5 \times 10^{-18}$, respectively).

3.9 Replication of co-expressed PPI modules as robust markers for aging

We conducted an independent replication study of the identified network modules as robust markers for chronological age using gene expression data from the Netherlands Twin Register and Netherlands Study of Depression and Anxiety (NTR & NESDA) consortium ($N = 3535$)²⁸ assayed on individuals within age range 17-79 years (Data S1, Supplemental Materials). An association analysis between the mean expression of a module and chronological age, adjusted for sex and the mean expression of the ‘T-cell activation’ module, yielded significant results for four of the five identified modules, all with directions corresponding to those found in the compendium (Table S6, Supplemental Materials). These results emphasize the robustness of the findings produced by our approach and confirm that the mean module expression in whole blood of module B, C, E, and F may be considered as robust markers of chronological age.

3.10 Co-expressed PPI modules are enriched for GenAge longevity and aging genes

As a validation of the identified modules, we computed whether aging-related genes stored by GenAge¹², a database providing a comprehensive overview of aging-related genes in humans and model systems, were enriched within modules A–F (Figure 2) (Data S1, Supplemental Materials). Whereas module A was supported by human derived annotations only (OR = 12.1, 95% CI 2.88–39.2, $p = 6.95 \times 10^{-4}$), module B was solely based on knowledge derived from model organisms (OR = 16.9, 95% CI 7.26–39.1, $p = 2.52 \times 10^{-10}$) (Table S7, Supplemental Materials). Modules D, E, and F had annotations balanced over both sources, and therefore, the significance of the joint enrichment was assessed by using a resampling approach (Data S1, Supplemental Materials), which yielded significant enrichments for modules E ($p = 0.016$) and F ($p = 0.0029$). These findings provide additional evidence that the joint expression of these modules may play a relevant role in human aging.

3.11 Module F associates with prospective survival at old age

To investigate whether the identified modules could potentially serve as biomarkers, we studied the microarray data assayed on 50 nonagenarian individuals from the Leiden Longevity Study²⁵. A left truncated Cox proportional hazard model adjusted for sex and cell counts indicates that the mean expression of module F associates with prospective survival beyond the age of 90 years ($N = 50$, $N_{\text{death}} = 45$, HR

= 0.265, 95% CI 0.12-0.57, $p = 0.001$). By showing that module F associates with prospective survival at old age, we illustrate its potential biological relevance.

Interestingly, the *ASF1A* gene is part of module F and has previously been identified by our group as one of the genes that was differentially expressed in blood of members of long-lived families as compared to similarly aged controls at middle age²⁵. To confirm that the expression of the *ASF1A* gene in module F also associates with prospective survival at old age, we analyzed the gene expression of *ASF1A* measured with RT-qPCR in 74 nonagenarians from the Leiden Longevity Study (of which 24 overlapped with the micro-array experiment) for association with prospective survival. Because we observe a similar association ($N_{\text{death}} = 64$, HR = 0.54, 95% CI 0.34-0.85, $p = 0.008$) (Figure 3), these results indicate that modules, of which the expression in blood is consistently associated with chronological age across various datasets, may associate with variation in lifespan, and therefore provide valid gene targets for studying relevant biological endpoints in human aging.

4. Discussion

Age-associated changes in gene expression may provide meaningful leads to pathways affected by and involved in aging, though are generally difficult to detect consistently⁶. Therefore, we constructed a large compendium of human whole blood expression studies¹⁹⁻²¹ comprising 2,539 individuals on which we performed a novel integrative network-based analysis. This

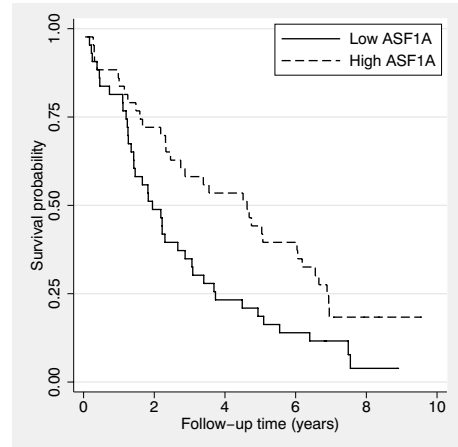


FIGURE 3: Expression of *ASF1A* associates with prospective survival in nonagenarians. High expression of *ASF1A* confers a prospective survival benefit at old age.

yielded fifteen consistently age-associated co-expressed PPI modules. Because the most significant age-associated module appeared to correlate with lymphocyte cell counts in an independent gene expression dataset, the expression of this module, enriched for ‘T-cell activation’, was subsequently used as a proxy for possible confounding shifts in the distribution of lymphocyte subsets. This enabled the identification of five age-associated modules (Figure 2 Panel I and IV), including three modules enriched for ‘Translational elongation’, ‘Cytolysis’ and ‘DNA metabolic process’ (Figure 2B–D). Replication in an independent cohort confirmed these findings for four of five modules (Figure 2B, C, E and F), underpinning the robustness of the proposed approach. The enrichments against a database for aging-related genes (Figure 2B, E and F) emphasize the relevance of these biological findings for aging research, which is even further substantiated by the fact that the

mean expression of module F associates with prospective survival at old age.

4.1 Mitochondrion-related aging

Two of the identified modules are down-regulated with age and seem to be related to the mitochondrion, though lacking any significant functional enrichment (Figure 2E and F). Despite the absence of functional enrichments, both modules were significantly enriched for aging-related genes, as defined by GenAge, implying that known age-related single genes can be put into a novel biological perspective by our network approach.

Module F (Figure 2F) contains several mitochondrial factors and enzymes, like, for instance, the mitochondrial transcription termination factor *MTERF*, the *ACADM* enzyme used for fatty acid metabolism, or the mitochondrial tRNA synthetase *IARS2*, whose homolog was shown to increase lifespan upon disruption in worms²⁹. This module also includes several genes previously associated with age or age-associated diseases such as the mitotic checkpoint protein *BUB3*, previously associated with accelerated aging in mice³⁰, and the cell-cycle checkpoint protein *APBBP1* found in increased quantities in the brain affected by Alzheimer's disease³¹. This broad range of gene characteristics composing the module could be explained by the fact that the functionality of mitochondria is not confined to cellular energy metabolism alone, but also seems to make up an integral part of multiple cell signaling cascades including cell-cycle control and cell death³².

Interestingly, module F also includes the *ASF1A* histone chaperone of which we

previously have shown that its expression associates with familial longevity in the Leiden Longevity Study²⁵. We revisited the RT-qPCR data assayed on 74 nonagenarians and now show that the expression of *ASF1A* also associates with prospective survival. This result illustrates that modules, of which the expression in blood is consistently associated with chronological age across various datasets, may associate with variation in lifespan, and therefore provide valid gene targets for studying relevant biological endpoints in human aging.

The other mitochondrion-related module (Figure 2E) contains the heat shock protein *HSPCA* (*HSP90*) and the mitochondrial receptor *TOMM20*, which jointly play a central role in translocating pre-proteins into the mitochondria³³. They seem to be consistently co-expressed in blood with *EIF4A2*, a eukaryotic translation initiation factor and *DDX18*, an ATP-dependent RNA helicase, of which the worm homologs were shown to extend lifespan upon disruption^{29,34}. To summarize, this module seems to relate to aging by influencing protein translation and mitochondrial translocation efficiency.

4.2 Age-associated limitation of protein synthesis

One of the identified modules predominantly consisted of ribosomal proteins and translation elongation factors comprising part of the ribosomal complex (Figure 2B). The module was significantly enriched for 'Translational elongation' and for previous findings in model organisms with respect to aging and longevity. In addition, the module was down-regulated with advancing age fitting previous observations of the aging

blood transcriptome²³⁻²⁵, which could be interpreted as an attempt of the cell to limit global protein synthesis in response to stress arising from damage accumulating throughout lifespan³⁵. Whether caused by response to stress or other factors, the change in protein translation may be ascribed to the mTORC1 complex³⁶. This complex modulates cellular growth and metabolisms by determining the balance between protein synthesis and degradation in response to nutrient availability. Inhibition of mTOR signaling through the mTORC1 complex not only inhibits protein synthesis, but also has been shown to positively affect the lifespan in various invertebrates and mammals³⁷. Moreover, human blood transcriptome studies showed that the gene expression of mTOR pathway is down-regulated with chronological age^{24,38} and is even associated with human familial longevity²⁵. Hence, a consistently down-regulated ribosomal module with advancing age corresponds with the age-associated demise of mTOR signaling. Although it is well established that mTOR signaling links to both lifespan regulation and ‘Translational elongation’, it remains to be determined whether down-regulation of ‘Translational elongation’ is causal for human aging.

4.3 WRN-related cell-cycle checkpoint on DNA integrity

A module down-regulated with age and enriched for ‘DNA metabolic process’ identified in the compendium could not be replicated in the NTR&NESDA cohort (Figure 2D). Interestingly, this module contains the *PARP1* (*ADPRT*) gene, which directly binds to *WRN* to induce apoptosis upon oxidative

stress induced DNA damage and is as such a prime suspect for Werner syndrome³⁹, a premature aging disease. Furthermore, the activity of the Parp1 protein in mononuclear cells has previously been shown to positively correlate with the species-specific lifespan across 13 mammalian species⁴⁰. Taken together, findings in the compendium suggest that the lowered transcription rate of *PARP1* negatively affects DNA integrity and thus lifespan, though more experiments are required to investigate this hypothesis.

4.4 Age-associated shifts in T-cell composition

Another identified module is up-regulated with age and enriched for ‘Cytolysis’ (Figure 2C). It contains several genes used to dispatch virus-infected cells and may reflect the decreased competence for fighting infections in an early stage, caused by an age-related deterioration of the immune system, known as immuno-senescence⁴¹. We can, however, not rule out that the age-associated expression of *GZMA*, *GZMB*, and *PRF1* that are part of this module point to an age-associated shift in T-cytotoxic cells²⁷.

Though identified co-expressed PPI modules may show extensive correlation with confounding factors, we should be careful to dismiss modules as such only. For instance, the ‘T-cell activation’ module (Figure. 2A), which is down-regulated with age, also contained *BNIP3*, an inhibitor of the mTORC1 complex shown to modulate lifespan in worms, flies, and mice³⁷; and *FOXO1*, also displaying an intricate interplay with both complexes of mTOR³⁶, and shown to extend lifespan in various invertebrates⁴². Additionally, human mTOR signaling may

play a central role in orchestrating T-cell maturation and T-cell fate decisions⁴³, and could thereby also explain the age-associated decline in lymphocytes as marked by the ‘T-cell activation’ module. Taken together, these examples illustrate that what is confounding the analysis of the blood transcriptome for molecular mechanisms associated with aging is subjective to debate and might even not be possible to determine given the complex interplay between the different biological levels on which aging acts.

4.5 The proposed network approach into perspective

Network analyses have clear advantages over individual-gene analyses, as they enable the incorporation of useful prior knowledge, which can be exploited for improving the robustness of the analysis and the subsequent interpretation of the results. The improved robustness of the network approach over the individual-gene analyses was reflected by the low mutual overlap between the individual-gene results (Figure 1) as opposed to the high concordance between the results obtained in the compendium and replication cohort. The advantages for the interpretation were clearly illustrated by the modest insights gained from the two different strategies for individual-gene analysis (‘Glycosylation site: N-linked’), as opposed to the detailed gene modules produced by our approach that can serve as a novel basis for further investigation into the molecular mechanisms underlying normative aging. Moreover, our approach is capable of inferring biological coherence from the data, without the explicit need of predefined functional groupings, as was

shown by the enrichments of the identified modules found for genes within the GenAge database.

Though the analysis benefits from incorporating protein-protein interaction data, the type, and source clearly affect the results. To be as inclusive as possible for types and sources of PPI data, we have chosen to employ data obtained from the STRING database, which systematically collects and integrates interaction data derived from various sources for predicting functional relations between gene pairs. This choice results in a vast and comprehensive source of data. However, STRING data are not confined to physical interactions, as is the case with for instance IntAct (<http://www.ebi.ac.uk/intact/>) and unlike KEGG (<http://www.genome.jp/kegg/>), STRING data are not manually curated. For network inference, a trade-off exists between the sparsity and the quality of the employed gene-gene interactions. We made use of a threshold on the quality of reported interactions that are created by STRING by benchmarking the different interaction data sources to KEGG. Varying this threshold would affect the size and nature of the obtained co-expressed PPI modules. As the threshold determines the scale of the analysis, an interesting observation is that the results can be confounded to parts of the global network that do not necessarily overlap with the predefined known biological pathways. The latter is illustrated by the fact that some of our modules are not enriched for biological pathways and could basically be valued as a strong point of our data-driven approach.

4.6 Conclusion

By applying a network approach to multiple blood transcriptomics datasets, we have identified five co-expression PPI modules that associate with chronological age in humans. The confirmation of most of our findings in an independent dataset underpins the robustness of our approach. The modules are significantly enriched for aging-related genes as curated by the GenAge database. This implies that these age-related single genes, in the absence of a clear understanding of their joint functioning belong to a network that finds its basis in protein–protein interactions and will serve as novel input for aging research. We reinforced the biological relevance of one of the modules by showing that it associates with prospective survival beyond 90 years in humans as was observed also for a single known age-related gene in this module (*ASF1A*). These findings collectively warrant further investigations into the biological function of module F and its potential as a biomarker for healthy aging and human longevity.

5. Experimental Procedures

5.1 Creating the blood expression compendium

Analyses were based on gene expression data derived from individuals enrolled in three large cohort studies for which details on sample inclusion and employed expression protocols are provided in depth in the original publications¹⁹⁻²¹. Gene expression and accompanying phenotypic data was obtained from either the original authors or from the public data repository ArrayExpress. Data quality was stringently reexamined per dataset for the presence of outlier samples

or outlier measurements and annotated to a common annotation standard (EntrezGeneID). A detailed description of the data processing and an overview on the resulting sample statistics is given in the Data S1 (Supplemental Materials) and Table 1, respectively.

5.2 Rank integration approach

A rank integration approach^{6,22} was used to identify genes consistently up- or down-regulated with age across multiple heterogeneous datasets. This type of meta-analysis integrates individual-gene statistics across datasets, by ranking the statistics per dataset and assessing the significance of the observed combined ranking using a Gamma distribution⁴⁴ or through permutation. Gender adjusted linear fits between expression and age were used as gene statistics that were obtained by fitting the following multivariate linear regression model:

$$E_{ijk} = \beta_{0ik} + \beta_{1ik}G_{jk} + \beta_{2ik}A_{jk} + \varepsilon_{ijk} \quad (1)$$

where E_{ijk} is the gene expression of gene i for individual j in the k^{th} dataset, with $1 \leq i \leq M$, $1 \leq j \leq N$ and $1 \leq k \leq K$, where G_{jk} and A_{jk} are the gender and age of individual j in the k^{th} dataset, respectively, and where ε_{ijk} is the residual error of gene i for individual j in the k^{th} dataset. Genes were ranked on the regression coefficients between age and expression, β_{2ik} . The rank position of gene i in dataset k is denoted by R_{ik} . Ranks across the datasets were integrated per gene by computing rank product statistics as previously defined by Koziol⁴⁴:

$$RP_i = \sum_{k=1}^K \log(R_{ik}) \quad (2)$$

The significance of the observed rank products was assessed in two ways. Following Koziol, rank products RP_i were transformed using:

$$U_i = -RP_i + K \times \log(M + 1) \quad (3)$$

The significance of the U-statistics could be assessed by employing the gamma distribution⁴⁴

or through permutation as described in the Data S1.

5.3 Extracting co-expressed PPI modules

Genes were mapped to the protein-protein interaction network (STRING v9.0, <http://string-db.org/>), which yielded a compendium of about 81.3% of the initial set of genes ($N = 7,353$) in the compendium. Ranked co-expression matrices were computed for each dataset separately by computing a correlation matrix composed of first-order partial correlations between all pairs of genes adjusted for sex and subsequently assigned a rank to each of them. A higher positive correlation resulted in a higher ranking. The ranked co-expression matrices were integrated by computing rank products as in the section on individual-gene analysis. The resulting gene-gene rank product matrix together with the PPI network matrix was subsequently used as input for the method that identifies co-expressed PPI sub networks as described in Van den Akker *et al.*¹⁷, see also Data S1 (Supplemental Materials). In short, a cluster analysis on the gene-gene rank product matrix yielded co-expressed modules of genes. High confidence co-expressed genes were obtained by applying a threshold on the gene-gene rank product matrix. We obtained co-expressed PPI modules by intersecting the co-expressed gene modules with the PPI network matrix. Co-expressed PPI modules were subsequently visualized using Cytoscape (Data S1, Supplemental Materials).

5.4 Fixed-effect meta-analysis on module expressions across the blood compendium

Gene expression data were summarized per co-expressed PPI module for each dataset separately by taking the mean expression per individual over all genes in the module, resulting in a module expression for each dataset. Associations with age were tested for each co-expressed PPI module, by performing a fixed-effect meta-analysis across the four datasets using a first-order partial correlation between age and the module expression,

computed with the controlling variable gender to adjust for sex differences. Per dataset k , we thus computed:

$$\rho_{a_k m_k \cdot g_k} = \frac{\rho_{a_k m_k} - \rho_{a_k g_k} \rho_{m_k g_k}}{\sqrt{1 - (\rho_{a_k g_k})^2} \sqrt{1 - (\rho_{m_k g_k})^2}} \quad (4)$$

where $\rho_{a_k m_k}$ is the correlation between age and the expression of the n^{th} module across individuals of the k^{th} dataset; $\rho_{a_k g_k}$ is the correlation between age and gender across individuals of the k^{th} dataset and $\rho_{m_k g_k}$ is the correlation between expression of the m^{th} module and gender of individuals in the k^{th} dataset. To correct for multiple controlling variables, higher order partial correlations were computed by repeatedly computing first order partial correlations as described above. The function *metacor* of R package *meta* was used for integrating and testing the meta correlation statistic between age and module expression across the four datasets using default settings. Modules with significant correlations (bonferroni corrected p -value ≤ 0.05) were considered age dependent.

6. Acknowledgements

The research leading to these results has received funding from the Medical Delta (COMO) and the European Union's Seventh Framework Programme (FP7/2007-2011) IDEAL-ageing under grant agreement n° 259679. This study was supported by a grant from the Innovation-Oriented Research Program on Genomics (SenterNovem IGE05007), the Centre for Medical Systems Biology, the Netherlands Consortium for Healthy Ageing (Grant 050-060-810), all in the framework of the Netherlands Genomics Initiative, Netherlands Organization for Scientific Research (NWO) and by Unilever Colworth. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

The authors thank the participants of the SAFHS, IFB, DILGOM, LLS, NTR and NESDA studies for their contributions. Furthermore, we

would like to thank Michael Inouye for helpful discussions.

7. 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. Wilson, P.W. *et al.* Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837-47 (1998).
4. Kirkwood, T.B. Evolution of ageing. *Nature* **270**, 301-4 (1977).
5. 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).
6. 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).
7. Partridge, L. & Gems, D. Mechanisms of ageing: public or private? *Nat Rev Genet* **3**, 165-75 (2002).
8. Zahn, J.M. *et al.* Transcriptional profiling of aging in human muscle reveals a common aging signature. *PLoS Genet* **2**, e115 (2006).
9. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* **4**, Article17 (2005).
10. Southworth, L.K., Owen, A.B. & Kim, S.K. Aging mice show a decreasing correlation of gene expression within genetic modules. *PLoS Genet* **5**, e1000776 (2009).
11. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249-55 (2003).
12. de Magalhaes, J.P. & Toussaint, O. GenAge: a genomic and proteomic network map of human ageing. *FEBS Lett* **571**, 243-7 (2004).
13. Bell, R. *et al.* A human protein interaction network shows conservation of aging processes between human and invertebrate species. *PLoS Genet* **5**, e1000414 (2009).
14. Witten, T.M. & Bonchev, D. Predicting aging/longevity-related genes in the nematode *Caenorhabditis elegans*. *Chem Biodivers* **4**, 2639-55 (2007).
15. Tacutu, R. *et al.* Prediction of *C. elegans* longevity genes by human and worm longevity networks. *PLoS One* **7**, e48282 (2012).
16. Xue, H. *et al.* A modular network model of aging. *Mol Syst Biol* **3**, 147 (2007).
17. van den Akker, E.B. *et al.* Integrating protein-protein interaction networks with gene-gene co-expression networks improves gene signatures for classifying breast cancer metastasis. *J Integr Bioinform* **8**, 188 (2011).
18. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat Methods* **9**, 796-804 (2012).
19. Goring, H.H. *et al.* Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet* **39**, 1208-16 (2007).
20. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423-8 (2008).
21. Inouye, M. *et al.* An immune response network associated with blood lipid levels. *PLoS Genet* **6**, e1001113 (2010).
22. Breitling, R., Armengaud, P., Amtmann, A. & Herzyk, P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett* **573**, 83-92 (2004).
23. Hong, M.G., Myers, A.J., Magnusson, P.K. & Prince, J.A. Transcriptome-wide assessment of human brain and lymphocyte senescence. *PLoS One* **3**, e3024 (2008).
24. Harries, L.W. *et al.* Human aging is characterized by focused changes in gene

- expression and deregulation of alternative splicing. *Aging Cell* **10**, 868-78 (2011).
25. Passtoors, W.M. *et al.* Transcriptional profiling of human familial longevity indicates a role for ASF1A and IL7R. *PLoS One* **7**, e27759 (2012).
 26. Dall'olio, F. *et al.* N-glycomic biomarkers of biological aging and longevity: A link with inflammaging. *Ageing Res Rev* **12**, 685-98 (2013).
 27. Derhovanessian, E. *et al.* Hallmark features of immunosenescence are absent in familial longevity. *J Immunol* **185**, 4618-24 (2010).
 28. Boomsma, D.I. *et al.* Genome-wide association of major depression: description of samples for the GAIN Major Depressive Disorder Study: NTR and NESDA biobank projects. *Eur J Hum Genet* **16**, 335-42 (2008).
 29. Smith, E.D. *et al.* Quantitative evidence for conserved longevity pathways between divergent eukaryotic species. *Genome Res* **18**, 564-70 (2008).
 30. Baker, D.J. *et al.* Early aging-associated phenotypes in Bub3/Rae1 haploinsufficient mice. *J Cell Biol* **172**, 529-40 (2006).
 31. Chen, Y., Liu, W., McPhie, D.L., Hassinger, L. & Neve, R.L. APP-BP1 mediates APP-induced apoptosis and DNA synthesis and is increased in Alzheimer's disease brain. *J Cell Biol* **163**, 27-33 (2003).
 32. McBride, H.M., Neuspiel, M. & Wasiak, S. Mitochondria: more than just a powerhouse. *Curr Biol* **16**, R551-60 (2006).
 33. Fan, A.C. *et al.* Interaction between the human mitochondrial import receptors Tom20 and Tom70 in vitro suggests a chaperone displacement mechanism. *J Biol Chem* **286**, 32208-19 (2011).
 34. Curran, S.P. & Ruvkun, G. Lifespan regulation by evolutionarily conserved genes essential for viability. *PLoS Genet* **3**, e56 (2007).
 35. Clemens, M.J. Initiation factor eIF2 alpha phosphorylation in stress responses and apoptosis. *Prog Mol Subcell Biol* **27**, 57-89 (2001).
 36. Laplante, M. & Sabatini, D.M. mTOR signaling at a glance. *J Cell Sci* **122**, 3589-94 (2009).
 37. Johnson, S.C., Rabinovitch, P.S. & Kaeblerlein, M. mTOR is a key modulator of ageing and age-related disease. *Nature* **493**, 338-45 (2013).
 38. Passtoors, W.M. *et al.* Gene expression analysis of mTOR pathway: association with human longevity. *Aging Cell* **12**, 24-31 (2013).
 39. von Kobbe, C. *et al.* Central role for the Werner syndrome protein/poly(ADP-ribose) polymerase 1 complex in the poly(ADP-ribosyl)ation pathway after DNA damage. *Mol Cell Biol* **23**, 8601-13 (2003).
 40. Grube, K. & Burkle, A. Poly(ADP-ribose) polymerase activity in mononuclear leukocytes of 13 mammalian species correlates with species-specific life span. *Proc Natl Acad Sci U S A* **89**, 11759-63 (1992).
 41. Pawelec, G. & Solana, R. Immunosenescence. *Immunol Today* **18**, 514-6 (1997).
 42. Calnan, D.R. & Brunet, A. The FoxO code. *Oncogene* **27**, 2276-88 (2008).
 43. Chi, H. Regulation and function of mTOR signalling in T cell fate decisions. *Nat Rev Immunol* **12**, 325-38 (2012).
 44. Koziol, J.A. Comments on the rank product method for analyzing replicated experiments. *FEBS Lett* **584**, 941-4 (2010).

Supplemental Materials

Supplemental Materials are accessible online: <http://onlinelibrary.wiley.com/doi/10.1111/accel.12160/supinfo>

Overview

Data S1

Supplemental Methods

Table S1

Results of 26 significantly age-associated genes in all four datasets

Table S2

Results of the individual-gene rank product test

Table S3

Enrichment analyses on the 26 significantly age-associated genes in all four datasets

Table S4

Enrichment analyses on 195 significant genes obtained with an individual-gene rank product test

Figure S1

Overview of detected consistently co-expressed PPI modules.

Table S5

Gene enrichment analyses using DAVID on the 27 consistently co-expressed PPI modules counting at least 5 genes

Table S6

Replication of the identified modules as robust markers of chronological age in the NTR & NESDA cohort.

Table S7

GenAge enrichment analyses of identified co-expressed PPI modules.

Chapter 4:

Germ line and Somatic Characteristics of the Long-Lived Genome

Erik B. van den Akker^{1,2}, Steven J. Pitts³, Joris Deelen^{1,4}, Matthijs H. Moed¹, Shobha Potluri³, H. Eka D. Suchiman¹, Nico Lakenberg¹, Wesley J. de Dijcker¹, Anton J.M. de Craen⁵, Jeanine J. Houwing-Duistermaat⁶, Genome of the Netherlands Consortium⁷, David R. Cox^{3†}, Marian Beekman^{1,4}, Marcel J.T. Reinders², P. Eline Slagboom^{1,4}

¹ Molecular Epidemiology, Leiden University Medical Center, Leiden, The Netherlands

² The Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands

³ Rinat-Pfizer Inc, South San Francisco, United States of America

⁴ Netherlands Consortium of Healthy Ageing

⁵ Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands

⁶ Medical Statistics, Leiden University Medical Center, Leiden, The Netherlands

⁷ Genome of the Netherlands Consortium members are listed in the Supplemental Materials.

† In memoriam

1. Abstract

Human longevity has an estimated heritability of approximately 25% in the population at large, which remains largely unexplained by known common genetic variation. The missing heritability in the longevity phenotype might be explained by rare disruptive variants that can be readily measured by the current sequencing techniques. Here we report the results of a whole genome sequencing study into familial longevity comparing the genomes of 218 independent nonagenarians originating from families with a multi-generational history of extended survival into old age and 98 ethnicity-matched random population controls. An exome-wide comparison did not reveal any robust differences in the overall prevalence of rare disruptive variants between the genomes of long-lived cases and random population controls. In contrast, recurrent rare disruptive variants were identified in two key epigenetic genes, e.g. *TET2* and *DNMT3A*, in long-lived cases exclusively, which suggests that a reduced functionality in these genes relates to longevity. Read depth evidence and Sanger re-sequencing data, however, indicated that the variants identified in *TET2* and *DNMT3A* were in general of somatic origin, and should therefore be discarded as potential heritable factors underlying familial longevity. Somatic variation in these genes is generally regarded as an indicator of age-associated outgrowth of myeloid progenitor cells, a pre-malignant phase, that marks the aging hematopoietic stem cell compartment and an increased susceptibility to leukemia. Although nonagenarian carriers of somatic disruptive variants in *TET2* and *DNMT3A* may exhibit signs of a shift in blood cell composition, they did not display a significantly compromised survival during a 10-year follow up. To conclude we found no robust evidence for the long-lived genome to carry either an overall excess or depletion of germ line rare disruptive variants. We do observe an increased prevalence of somatic variation in specific loci likely to stimulate clonal outgrowth.

2. Introduction

In western societies, life expectancy has been steadily growing over the past two centuries¹, yet striking variations in life span are observed among the population at large². Human life span regulation is an extraordinary complex outcome and is largely determined by chance and factors from the environment, though a modest contribution of heritable components (~25%) is also expected in the general population³. The propensity to become long-lived nevertheless clearly runs in families⁴⁻⁶ and seems to relate to the capacity to delay or evade age-associated disease. Offspring of nonagenarians, centenarians and super centenarians display a lower prevalence of cardiovascular disease, type II diabetes and cancer⁴⁻⁶, as compared to the general population, thus suggesting that human longevity is caused by genetic factors modifying risk of age-associated disease. However, compared to the general population, the genomes of nonagenarians do not show a depletion of common disease susceptibility alleles identified by genome-wide association studies (GWASs)⁷, nor did GWASs for longevity revealed sufficient loci to explain the heritability of longevity⁸. Since GWASs predominantly focus on analysing common variants (Minor Allele Frequency $\geq 1\%$), we hypothesize that the missing heritability of the longevity phenotype might be explained by rare coding variants with disruptive impact on the gene's functioning.

Rare disruptive variants can modify disease risk, like common variants, by affecting the expression or structure of

translated proteins, which may contribute to longevity in two ways. First, the genome is reported to contain on average about 100 rare disruptive variants per individual that severely limit or totally negate the functionality of the associated proteins⁹. Hence a genome-wide depletion of such rare disruptive variants might implicate a more complete or better functioning proteome, improving the capacity to maintain the bodily homeostasis. Moreover, such a genome-wide depletion of variants might also point to an improved fidelity of the DNA repair system as compared to the general population^{10,11}. Secondly, a targeted knockdown of a single gene in model organisms can already give rise to a long-lived species¹². Hence, a local enrichment of rare disruptive variants in the genomes of long-lived individuals might implicate that a similar loss of function of the gene originating from that particular locus promotes longevity in humans. Though both genetic mechanisms are plausible, little evidence exists to date whether the genetic propensity for human longevity relates more closely to a fitter proteome or the targeted disruption of particular gene functions.

The first NGS efforts to study rare variants in longevity involve study designs with few extreme cases. The genomes of super-centenarians and centenarians were sequenced in order to describe genetic features of exceptional longevity¹³⁻¹⁷. Obviously, these analyses have a very limited statistical power for revealing evidence in favour of any of the two proposed genetic mechanisms for longevity mentioned above. However, also these very extreme cases do not show a depletion of

common disease susceptibility alleles as identified by genome-wide association studies (GWASs), in line with work of Beekman *et al.*⁷. Using a more targeted approach, 988 candidate longevity genes were sequenced in 6 centenarians to identify novel non-synonymous SNVs¹⁸, which were subsequently tested in larger case control studies and suggested *PMS2* and *GABRR3* as novel candidate longevity genes. These initial studies provide some first insights into genetic backgrounds that are conducive to exceptional longevity.

To investigate potential genetic mechanisms for human longevity involving rare disruptive variants, whole-genome sequencing was performed by Complete Genomics on DNA derived of 218 nonagenarian participants of the Leiden Longevity Study (LLS). The Leiden Longevity Study consists of sib pairs of which female members reached at least 91 years of age and male members 89 years of age. First-degree family members of these nonagenarian siblings show a 30% survival advantage as compared to their birth cohort¹⁹. Moreover, offspring of these nonagenarians exhibit a propensity for healthy aging already at middle age, as indicated by their significantly lowered incidence of hypertension, type II diabetes and use of cardiovascular medication, as compared to population controls⁴. We therefore hypothesize that LLS families show healthy aging and longevity by their genetic predisposition. To further identify genetic variation that predisposes to familial longevity, we compared the genomes of these 218 unrelated long-lived cases with those of 98 younger population controls of the Biobanking

and Biomolecular Resources Research Infrastructure of the Netherlands (BBMRI-NL) consortium^{20,21}.

3. Results

3.1 Study design and variant detection

We explored the human genome for rare variants contributing to human longevity using whole genome sequencing data of 218 independent long-lived cases from the LLS (median age 93.7, $N_{\text{male}} = 82$ (37.6%)) and 98 population controls of the BBMRI biobanking initiative (median age 57, $N_{\text{male}} = 39$ (39.6%)) (Experimental Procedures 5.1). DNA sequencing and subsequent variant calling was performed by Complete Genomics (Complete Genomics Inc., Mountain View California) (median read depth >30x) on genetic material isolated from peripheral blood. Sequencing data were subjected to a stringent quality control prior to performing the analyses. For the following analysis we considered Single Nucleotide Variants (SNVs), small deletions (DELS) and insertions (INSS) called at high quality and with a minimal call rate of 95% in both long-lived cases and population controls. For a more detailed description of variant detection and quality control see Experimental Procedures 5.2.

3.2 Depletion of coding variation in longevity genomes

The genome-wide burden of disruptive genetic variants in long-lived cases compared to the population controls was investigated for all variants in the coding sequence (CDS) jointly and for variants

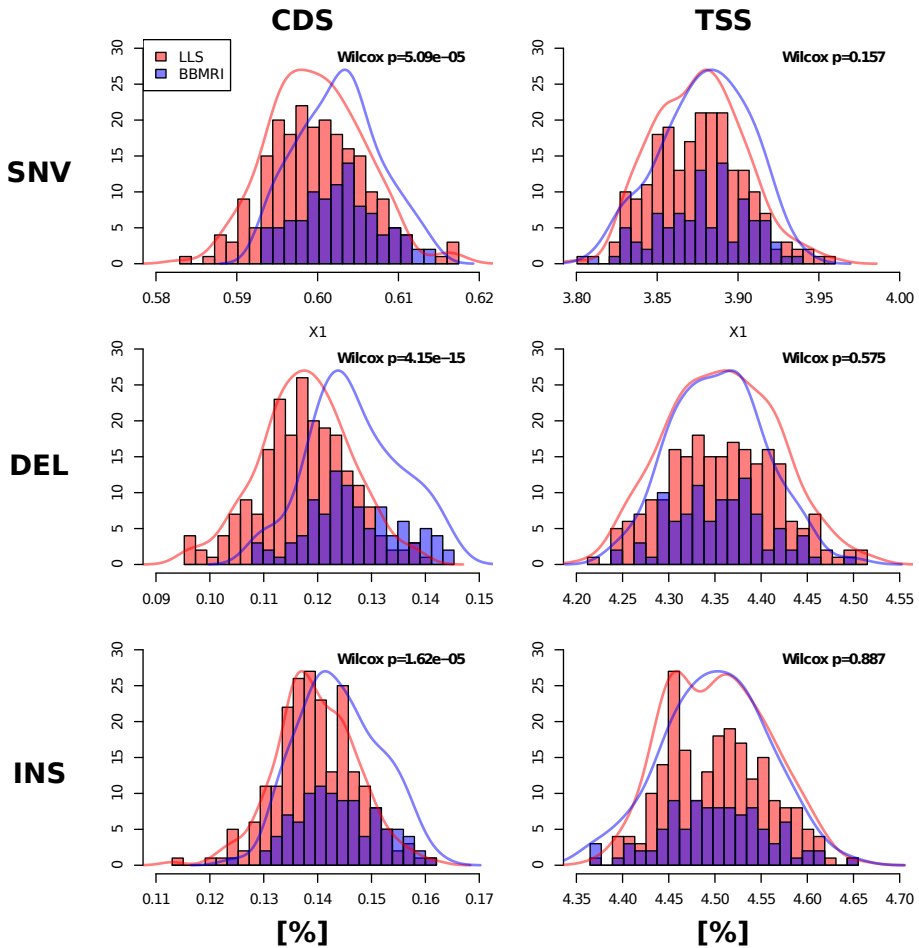


FIGURE 1: DEPLETION OF CODING VARIANTS IN GENOMES OF LONG-LIVED INDIVIDUALS. Distributions of proportions of variants annotated to the CDS (coding sequence) or sequence upstream of the Transcription Start Site (TSS, 0-7.5kb) for each of the three small variant types (SNV, DEL, INS) are displayed for long-lived cases (LLS; red) and random population controls (BBMRI; blue) respectively. Test results for differences in these distributions are reported in the upper right corner (Wilcoxon Rank-Sum test). Whereas a significant depletion of coding variants was observed for all small variant types in long-lived cases (LLS) compared to population controls (BBMRI), no such association was observed for the proportion of variants annotated to TSS.

categorized per impact (e.g. missense or nonsense) and type (single nucleotide variant: SNV, small deletions: DEL or insertions: INS). Counts per thus formed categories were normalized per individual on the totals of variants observed for each variant type to negate biases from overall differences in variant calling between the cohorts. Using this approach, we detect a

lowered proportion of variants annotated to the CDS in nonagenarians cases compared to the population controls for all types of variants (Wilcoxon Rank-Sum test: SNV: $p=5.09 \times 10^{-5}$, DEL: $p=4.15 \times 10^{-15}$ and INS: $p=1.62 \times 10^{-5}$; Figure 1, left column). As a negative control, we tested for differences in proportions of variants annotated up to 7.5 kb upstream of the Transcription

Start Site (TSS) and did not observe any

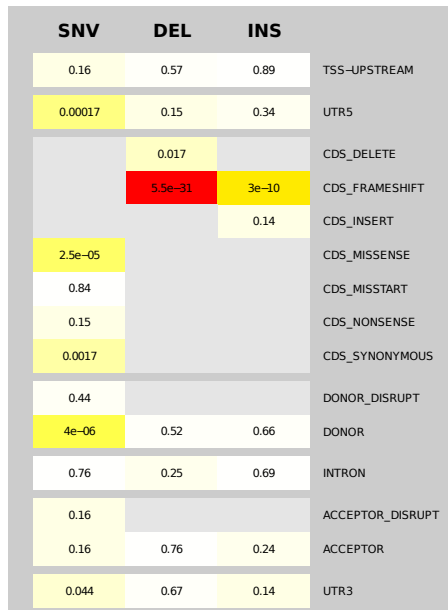


FIGURE 2: DEPLETION OF DISRUPTIVE VARIANTS IN GENOMES OF LONG-LIVED INDIVIDUALS. A heatmap displaying the results of all variant-categories created by cross tabulating variant-types (columns: SNV, DEL and INS) and variant-impacts (rows: TSS-UPSTREAM (Transcription Start Site and 7.5 kb upstream), UTR5 (UnTranslated Region at 5')), CDS_DELETE (in frame deletion), CDS_FRAMESHIFT (out of frame deletion or insertion), CDS_INSERT (in frame insertion), CDS_MISSENSE (amino acid substitution), CDS_MISSTART (start removed), CDS_NONSENSE (stop created), CDS_NONSYNONYMOUS (no change to protein), DONOR_DISRUPT (2 bp of essential splice donor site), DONOR (12bp of splice donor site), INTRON, ACCEPTOR_DISRUPT (2 bp of essential splice acceptor site), ACCEPTOR (8 bp of splice acceptor site), UTR3 (UnTranslated Region at 3')). The intensity of each cell represents the significance of the Wilcoxon Rank-Sum test computed on the difference in proportions of a particular variant-type annotated to a variant-category between the long-lived cases and the population controls. P-values are displayed in the cells. Cells are empty if no or to little data were available for testing. Note that the frameshift variants are most significantly depleted in the long-lived cases as compared to the random population controls.

significant differences (SNV: $p=0.157$, DEL: $p=0.575$ and INS: $p=0.887$, Figure 1, right column). Since total numbers of variants might also reflect the quality of alignment and depth of sequencing, we inspected the correlation between the proportions of variants annotated to the CDS and the total numbers of variants discovered in cases and controls (Supplemental Figure 1), but found no significant biases. Hence, compared to the general population, long-lived cases show a depletion of variation in the coding part of the genome.

When applying the testing to the more fine-grained annotations of the coding sequence, as provided by Complete Genomics²² we observe that the depletion of CDS variants in long-lived cases compared to population controls can be explained by a few categories in particular. DELs and INSs inducing frameshifts, and missense and synonymous SNVs were present in significantly lower proportions in the long-lived cases as compared to the population controls (Figure 2, Supplemental Table 1). In addition, SNVs residing in splice donor sites and the 5' untranslated regions (5UTR) displayed a similar depletion. Of the depleted variant categories, we expect the most disruptive variant categories to show the highest depletion in long-lived cases. To verify this, counts of frameshift DELs and INSs were re-analyzed, while normalizing for frame preserving DELs and INSs and counts of missense SNVs or SNVs residing in splice donor sites or 5UTR were normalized on counts of synonymous SNVs (Figure 3). Indeed frameshift DELs ($p = 1.84 \times 10^{-26}$) and INSs ($p = 2.60 \times 10^{-09}$) and SNVs residing in splice donor sites

($p = 4.51 \times 10^{-04}$) displayed an additional significant depletion on top of the general depletion of coding variation in long-lived cases compared to population controls.

High impact variants calls made with short-read sequencing platforms are associated with an increased false positive rate. To investigate the rates of truly reported high impact variants in long-lived cases and random population controls, we randomly selected 15 frameshift variants in each of the two cohorts and validated these using Sanger sequencing. Of the 15 assays for frameshift variants only observed in the long-lived cases, 12

returned good data, which confirmed the presence of seven (58.3%) frameshift variants (Supplemental Table 2). Whereas all of the 15 assays for frameshift variants observed in the population controls that could be successfully designed, only two (13.3%) validated the presence of its targeted variant (Supplemental Table 3). Thus, the ratio of falsely reported variants within the two small samples of high impact variants is considerable, and notably, highest amongst population controls. DNA of long-lived cases and population controls was sequenced on the same platform, be it at two different points in time (within

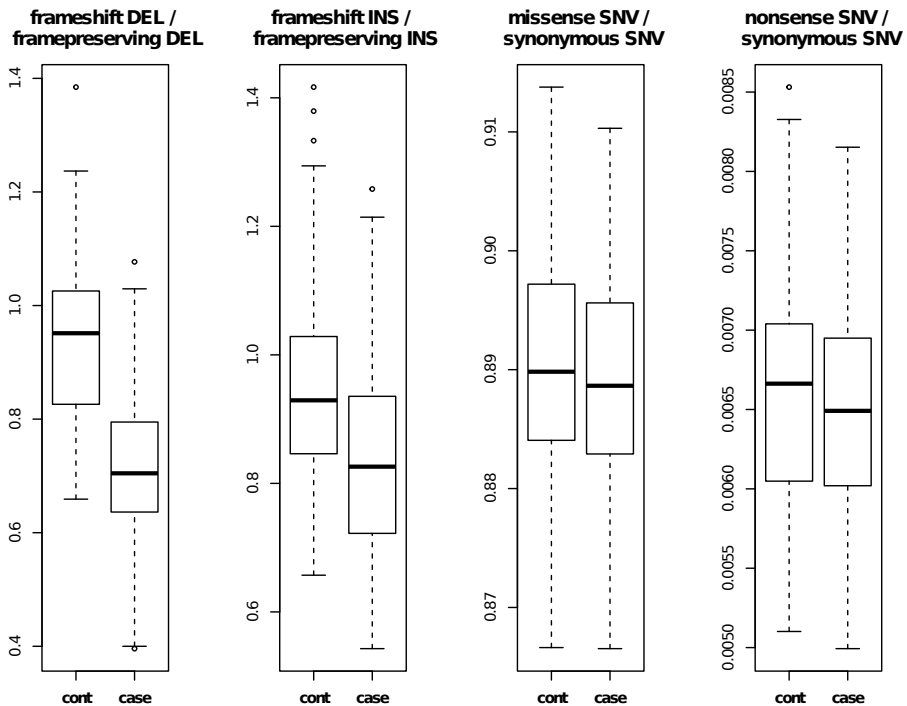


FIGURE 3: THE HIGHER THE IMPACT, THE MORE DEPLETED. When normalizing the counts in the more disruptive variant categories on those in the less disruptive variant categories of the same variant type, e.g. by normalizing counts on frame shifting DELs on frame preserving DELs, we confirm our previous findings of a depletion of the most disruptive variants in long-lived cases compared to those population controls. Frameshift DELs ($p = 1.84 \times 10^{-26}$) and INSS ($p = 2.60 \times 10^{-09}$) and SNVs residing in splice donor sites ($p = 4.51 \times 10^{-04}$) displayed an additional significant depletion on top of the general depletion of coding variation in long-lived cases compared to population controls.

2 years), possibly leading to a technical bias. From the validation experiment we conclude that the previously observed difference in prevalence of disruptive variants is most likely due to an elevated false discovery rate in the controls rather than a depletion of rare disruptive variants in the long-lived cases.

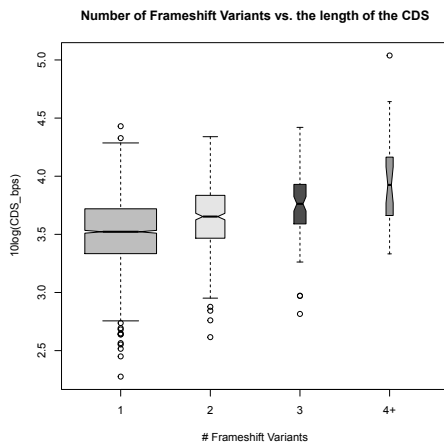


FIGURE 4: LONGER GENES ARE MORE LIKELY TO CATCH FRAMESHIFT VARIANTS. When plotting the length of the coding sequence as a function of the number of frameshift indels we observe a clear positive correlation.

3.3 Rare disruptive variants cluster at *TET2* and *DNMT3A* in nonagenarian genomes

Moving away from the whole genome depletion of variants, we next investigated whether genes are preferentially hit by disruptive variants as postulated in the second proposed genetic mechanism for human longevity. To investigate which genes are preferentially hit by the disruptive frameshift variants, irrespective of the study, i.e. in long-lived cases and in population controls, we collapsed the deletions and insertions to

gene annotations. This yielded a total of 2,193 unique deletions and 1,764 unique insertions in respectively 1,970 and 1,601 genes. Assuming a coding transcriptome of 18,000 independent transcript clusters, we used a resampling approach to assess the significance of the joint presence of the numbers of frameshift deletions and insertions per gene (Experimental Procedures 5.3). The 27 genes hit by at least four unique frameshift mutations are presented in Table 1 and jointly comprise 3.2% of the total number of frameshift variants observed. A strong trend between the length of the coding sequence and the number of frameshift variants present in genes in cases and controls jointly was observed (Figure 4), with the largest gene present in the genome, *TTN*, showing the most significant enrichment of frameshift variants. Hence, few relatively long genes accumulate multiple frameshift variants.

Next we investigated whether any of the 27 genes with four or more frameshift indels was preferentially hit by mutations unique to either the long-lived cases or the population controls. By again using a resampling approach, we assessed the significance of the observed number of private frameshift deletions and insertions present in each of the genes (Experimental Procedures 5.4). Interestingly, we note that the most significant gene-specific accumulations of frameshift variants occur in two genes hit in long-lived cases only: *TET2* and *DNMT3A* (Table 2). Other categories variant types, e.g. nonsense SNVs, confirmed the burden of disruptive variants in *TET2* and *DNMT3A* present in only long-lived cases (Table 3). In total, *TET2* was hit by six frameshift indels and

GeneSymbol	EntrezGeneID	DEL	INS	DEL + INS	p.perm
<i>TTN</i>	7273	7 (2/5/0)	2 (0/2/0)	9 (2/7/0)	$<1.00 \times 10^{-6}$
<i>DNAH14</i>	127602	8 (4/2/2)	0 (0/0/0)	8 (4/2/2)	$<1.00 \times 10^{-6}$
<i>FSIP2</i>	401024	1 (1/0/0)	6 (3/2/1)	7 (4/2/1)	$<1.00 \times 10^{-6}$
<i>LOC100506072</i>	100506072	3 (2/0/1)	4 (1/2/1)	7 (3/2/2)	$<1.00 \times 10^{-6}$
<i>TET2</i>	54790	4 (4/0/0)	2 (2/0/0)	6 (6/0/0)	5.00×10^{-6}
<i>SSPO</i>	23145	5 (2/0/3)	1 (1/0/0)	6 (3/0/3)	6.00×10^{-6}
<i>VPS13C</i>	54832	3 (0/3/0)	2 (1/1/0)	5 (1/4/0)	1.90×10^{-5}
<i>IL3RAY</i>	8218	0 (0/0/0)	4 (3/0/1)	4 (3/0/1)	9.90×10^{-5}
<i>SYNE1</i>	23345	0 (0/0/0)	4 (2/2/0)	4 (2/2/0)	9.90×10^{-5}
<i>UGGT2</i>	55757	0 (0/0/0)	4 (2/2/0)	4 (2/2/0)	9.90×10^{-5}
<i>SLFN12L</i>	100506736	1 (1/0/0)	3 (3/0/0)	4 (4/0/0)	1.27×10^{-4}
<i>ZBTB1</i>	22890	1 (1/0/0)	3 (2/1/0)	4 (3/1/0)	1.27×10^{-4}
<i>SPATA3E1</i>	286234	1 (1/0/0)	3 (2/1/0)	4 (3/1/0)	1.27×10^{-4}
<i>PNPLA7</i>	375775	1 (1/0/0)	3 (2/0/1)	4 (3/0/1)	1.27×10^{-4}
<i>HECTD4</i>	283450	1 (1/0/0)	3 (1/2/0)	4 (2/2/0)	1.27×10^{-4}
<i>PTCHD3</i>	374308	1 (0/0/1)	3 (2/0/1)	4 (2/0/2)	1.27×10^{-4}
<i>POLQ</i>	10721	2 (0/1/1)	2 (0/2/0)	4 (0/3/1)	1.36×10^{-4}
<i>NOTCH3</i>	4854	2 (0/2/0)	2 (1/1/0)	4 (1/3/0)	1.36×10^{-4}
<i>ADAM8</i>	101	2 (0/1/1)	2 (2/0/0)	4 (2/1/1)	1.36×10^{-4}
<i>NIN</i>	51199	2 (0/2/0)	2 (2/0/0)	4 (2/2/0)	1.36×10^{-4}
<i>ZNF469</i>	84627	2 (0/2/0)	2 (2/0/0)	4 (2/2/0)	1.36×10^{-4}
<i>MUC16</i>	94025	2 (0/1/1)	2 (1/1/0)	4 (1/2/1)	1.36×10^{-4}
<i>LMOD2</i>	442721	3 (0/3/0)	1 (0/1/0)	4 (0/4/0)	1.91×10^{-4}
<i>PIK3C2G</i>	5288	3 (3/0/0)	1 (0/1/0)	4 (3/1/0)	1.91×10^{-4}
<i>TNRC18</i>	84629	3 (2/1/0)	1 (1/0/0)	4 (3/1/0)	1.91×10^{-4}
<i>DNMT3A</i>	1788	4 (4/0/0)	0 (0/0/0)	4 (4/0/0)	2.11×10^{-4}
<i>ABCA10</i>	10349	4 (1/1/2)	0 (0/0/0)	4 (1/1/2)	2.11×10^{-4}

TABLE 1: THE 27 GENES ACCUMULATING AT LEAST 4 FRAMESHIFT VARIANTS. Counts of variants are given for DELetions and INSertions separately according to the following format: A (B/C/D) indicate respectively the total (A), private in case (B), private in control (C) and shared number of variants (D).

GeneSymbol	DELs	INSs	Totals	p_case	p_cont
<i>TET2</i>	4 (4/0/0)	2 (2/0/0)	6 (6/0/0)	0.0049	1
<i>DNMT3A</i>	4 (4/0/0)	0 (0/0/0)	4 (4/0/0)	0.019	1

TABLE 2: GENES WITH A PRIVATE BURDEN IN LONG-LIVED CASES. Within the top 27 genes accumulating at least 4 frameshift variants, *TET2* and *DNMT3A* exhibited a study specific preference. Noteworthy is that both these genes feature frameshift variants in only the long-lived cases.

Gene	Chrom	Start	End	Type	Ref	Alt	Impact	LLS	BBMRI
<i>TET2</i>	chr4	106155736	106155737	DEL	T	-	FRAMESHIFT	1	0
<i>TET2</i>	chr4	106155765	106155766	DEL	G	-	FRAMESHIFT	1	0
<i>TET2</i>	chr4	106156685	106156686	SNV	C	A	NONSENSE	1	0
<i>TET2</i>	chr4	106156758	106156758	INS	-	C	FRAMESHIFT	1	0
<i>TET2</i>	chr4	106157246	106157246	INS	-	A	FRAMESHIFT	1	0
<i>TET2</i>	chr4	106157781	106157782	DEL	G	-	FRAMESHIFT	1	0
<i>TET2</i>	chr4	106157913	106157914	SNV	C	T	NONSENSE	1	0
<i>TET2</i>	chr4	106158107	106158108	SNV	G	A	NONSENSE	1	0
<i>TET2</i>	chr4	106196212	106196213	SNV	C	T	NONSENSE	1	0
<i>TET2</i>	chr4	106196221	106196222	SNV	G	T	NONSENSE	1	0
<i>TET2</i>	chr4	106197352	106197353	DEL	A	-	FRAMESHIFT	1	0
<i>DNMT3A</i>	chr2	25463181	25463182	SNV	G	A	NONSENSE	2	0
<i>DNMT3A</i>	chr2	25463296	25463296	INS	-	A	NONSENSE	1	0
<i>DNMT3A</i>	chr2	25468153	25468154	DEL	G	-	FRAMESHIFT	1	0
<i>DNMT3A</i>	chr2	25468921	25468923	DEL	AC	-	FRAMESHIFT	1	0
<i>DNMT3A</i>	chr2	25469921	25469922	SNV	G	A	NONSENSE	1	0
<i>DNMT3A</i>	chr2	25469990	25469991	DEL	A	-	FRAMESHIFT	1	0
<i>DNMT3A</i>	chr2	25470930	25470931	DEL	G	-	FRAMESHIFT	1	0

TABLE 3: FRAMESHIFT AND NONSENSE MUTATIONS IDENTIFIED IN *TET2* AND *DNMT3A*, EXCLUSIVELY PRESENT IN LONG-LIVED CASES.

five nonsense SNVs and *DNMT3A* by four frameshift indels, two nonsense SNVs and a single nonsense insertion, all in the 218 genomes of long-lived cases only. Moreover, a look-up on the Exome Variant Server (<http://evs.gs.washington.edu/EVS>) in exome sequencing results in ~4,125 U.S. participants of European ancestry revealed that *TET2* and *DNMT3A* were hit with unique frameshift indels or nonsense SNVs with a significantly lower frequency (*TET2*: $N_{\text{disrupt_EVS}}=9$, OR: 24.2 95% CI: 9.0-67.0, $p=4.5\times 10^{-10}$; *DNMT3A*: $N_{\text{disrupt_EVS}}=7$, OR: 19.5 95% CI: 5.8-65.6, $p=1.9\times 10^{-6}$, Fisher's Exact tests, Supplemental Table 4).

Unlike the poor validation rates observed for frameshift variants sampled

from the whole genome, frameshift variants identified within *TET2* and *DNMT3A* in the long-lived were generally confirmed using Sanger sequencing (9 out of 10). A closer inspection of these Sanger sequencing results showed in general a much lower signal for the mutant allele as compared to the wild-type allele, an observation supported by the whole genome sequencing results for the frameshifting indels in *TET2* and *DNMT3A* (Table 4). This clear deviation from the 1:1 ratio (Experimental Procedures 5.6), as expected for heterozygous germ line variants, suggests that the identified variants are present in only a part of the measured cells. These results support the impression that the

Gene	Chrom	Start	End	Type	# Ref	# Alt	% Alt	p_{som}
<i>TET2</i>	chr4	106155736	106155737	DEL	30	11	26.8%	0.017
<i>TET2</i>	chr4	106155765	106155766	DEL	60	9	13.4%	2.7×10^{-7}
<i>TET2</i>	chr4	106156758	106156758	INS	33	10	23.3%	0.0047
<i>TET2</i>	chr4	106157246	106157246	INS	24	9	27.3%	0.034
<i>TET2</i>	chr4	106157781	106157782	DEL	36	13	26.5%	0.0083
<i>TET2</i>	chr4	106197352	106197353	DEL	47	22	31.9%	0.016
<i>DNMT3A</i>	chr2	25463296	25463296	INS	47	17	26.6%	0.0028
<i>DNMT3A</i>	chr2	25468153	25468154	DEL	34	10	22.7%	0.0035
<i>DNMT3A</i>	chr2	25468921	25468923	DEL	30	8	21.1%	0.0039
<i>DNMT3A</i>	chr2	25469990	25469991	DEL	34	21	38.2%	0.12
<i>DNMT3A</i> [§]	chr2	25470930	25470931	DEL	47	4	7.8%	0.0019

TABLE 4: NUMBER OF READS SUPPORTING THE REFERENCE AND ALTERNATIVE ALLELES OF FRAMESHIFT VARIANTS IN *TET2* AND *DNMT3A*. All Frameshift variants identified in the long-lived cases could be confirmed by Sanger sequencing except the variant marked by §. Since this non-confirmed variant had a relatively low % Alt of 7.84% it leaves the possibility that this variant may have gone undetected, as it was not present in a sufficient proportion of the sequenced cells.

long-lived cases, as compared to the younger population controls, have a higher prevalence of somatic frameshifting indels in *TET2* and *DNMT3A*.

Somatic mutations in *TET2* and *DNMT3A* have previously been associated with aging of hematopoietic stem cells (HSCs)²³, which is characterized by a skewing of progenitor cells towards the myeloid fate that compromises immune function and increases the risk for myeloid malignancies^{24,25}. Hence, we investigated whether carriership of the identified disruptive variants (Table 3) in long-lived cases was reflected by their blood cell composition. Whereas no signs of skewing in the blood cell composition was observed for the carriers of disruptive variants in *TET2* ($\beta=1.29$, 95% CI: -1.04-0.78, $p=0.78$), we observed that carriers with disruptive variants in *DNMT3A* have significantly higher granulocyte counts

than non-carriers ($\beta=1.29$, 95% CI: 0.24-2.43, $p=0.016$, Experimental Procedures 5.7). Since this may indicate an underlying risk for a compromised immune-capacity or hematopoietic malignancies, we compared the prospective survival of long-lived carriers versus long-lived non-carriers. A prospective survival analysis with a ten years follow-up did not indicate a significantly increased risk on mortality for the carriers of disruptive variants in either *TET2* ($N_{\text{tot}}=214$, $N_{\text{death}}=190$, HR=1.30, 95% CI 0.68-2.47, $p=0.424$) or *DNMT3A* ($N_{\text{tot}}=214$, $N_{\text{death}}=190$, HR=0.37, 95% CI 0.15-0.91, $p=0.031$, Experimental Procedures 5.8). In fact, a modest protective effect was observed for *DNMT3A* mutant carriers (Figure 5) and noteworthy, 4 out of the 9 carriers were still alive at our most recent census of 2012 at ages 99, 100, 104 and 105.

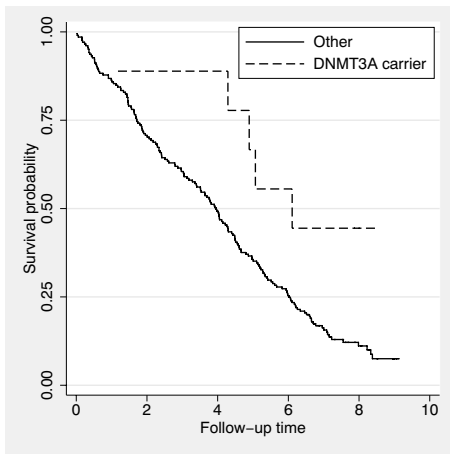


FIGURE 5: PROSPECTIVE SURVIVAL ON CARRIERS OF DISRUPTIVE VARIANTS IN *DNMT3A*. Kaplan-Meier curves for the long-lived cases carrying either a nonsense SNV or frameshift indel in *DNMT3A*, as compared to long-lived non-carriers.

4. Discussion

In the current study we analysed the genome of 218 independent nonagenarians for rare disruptive variants contributing to familial longevity. Although our sequencing study is the largest amongst the oldest old, we found no decisive evidence for either an excess or depletion of rare disruptive germ line variants to contribute to familial longevity. In contrast, we did observe and validate recurrent somatic variants in *TET2* and *DNMT3A*, exclusively present in the genomes of long-lived cases. Hence, we conclude that within this limited sample size, the characteristics most discriminative for the long-lived genome are acquired during life, which, to our current understanding, seem unlikely to constitute a heritable component predisposing to familial longevity.

The genomes of long-lived cases exhibited a gene-specific burden of rare

somatic disruptive variants from multiple categories in *TET2* and *DNMT3A*. Somatic mutations in *TET2* and *DNMT3A* were first reported in patients suffering from myeloid malignancies^{26,27}, but also appear in elderly exhibiting myelodysplasia without overt hematopoietic malignancies²³. This suggests that somatic mutations in *TET2* and *DNMT3A* in hematopoietic stem cells confer enhanced self-renewal and clonal expansion leading to an age-related myeloid lineage bias. Indeed significantly elevated levels of granulocytes were observed in carriers of somatic mutations in *DNMT3A*. Surprisingly, neither the carriers of somatic disruptive mutations in *DNMT3A*, nor in *TET2*, did exhibit a significantly increased mortality risk over 10 years time, while similar mutations have previously been associated with an increased risk of progression to and poor outcome of acute myeloid leukaemia (AML)²⁷. This either suggests that clonal expansion of the myeloid lineage in itself may not necessarily contribute to cancer risk in the highest ages, or alternatively, it may suggest that additional genetic factors, absent in long-lived, may be required for transforming into AML, in which case carriership may accelerate disease progression. Since these somatic mutations are typically found in elderly patients, it is reasonable to assume that the genetic burden at these loci should in effect be interpreted as markers of chronological age, rather than heritable factors underlying human longevity.

Assuming that the disruptive mutations in *TET2* and *DNMT3A* have been acquired during life, in absence of any overt malignancies, the question rises whether

these somatic variants in fact could have contributed to the observed extension in lifespan. Both *TET2* and *DNMT3A* are factors for epigenetic control^{28,29} and are thought to silence hematopoietic stem cell self-renewal to permit efficient hematopoietic differentiation^{30,31}. Therefore, loss of functionality in these genes is likely to underlie an enhanced self-renewal leading to the observed age-related myeloid lineage bias. This skewing towards the myeloid lineage is assumed to have adverse effects on immune functionality in normal healthy individuals, but in the oldest old the increase of the myeloid compartment might be compensative for the age-related decrease in naive T-cells, known as immuno-senescence³². Hence on condition that the enhanced self-renewal, instigated by somatic disruptive mutations in *TET2* and *DNMT3A*, leads to increased levels of competent immune cells, be it of the myeloid lineage though, might partly compensate for the age-related loss of immuno-capacity of the lymphoid compartment.

Initial analyses of the whole-genome sequencing data lead us to the false impression that the long-lived genome was characterized by a depletion of coding variation, most evidently present amongst SNVs residing in splice donor sites or indels leading to a frameshift. The prevalence of these disruptive variants per individual is generally very low, which indicates that these types of variants are generally not tolerated. This also explains the increased false positive rate amongst the variant calls of disruptive variants generally observed in sequencing studies, including the current one. Validation experiments

indicated that the few disruptive variants observed in the long-lived cases and population controls combined, were almost as likely to be erroneous as to be genuine and notably that the false positive rate was considerably higher amongst population controls. We therefore conclude that a genome-wide depletion of germ line disruptive variants in the genomes of long-lived individuals could not be decisively shown.

We conclude, that nonagenarian members of long-lived families have an increased prevalence of somatic disruptive variants in *TET2* and *DNMT3A*. Given their somatic origin, however, these variants seem unlikely to represent the heritable component of familial longevity. Previously, somatic mutations in these loci have been associated with risk on progression to^{33,34} and poor prognosis of AML^{27,35}. In the long-lived cases of our study, however, disruptive somatic variants in *TET2* and *DNMT3A* do not seem to compromise the 10-year survival. Implications of this finding are twofold. First, clinical risk assessments based on the mutational status of *TET2* and *DNMT3A* might not be accurate for the oldest old. Secondly, elderly carrying the somatic disruptive mutations in *TET2* and *DNMT3A* in absence of any overt malignancies may provide key insights in the factors most decisive for oncogenic transformation. Hence, the implications of somatic mutations in either *TET2* or *DNMT3A* for health in the oldest old remain illusive and therefore warrant more research into these key epigenetic loci.

5. Experimental Procedures

5.1 Study population

The Leiden Longevity Study⁴ is a family based study consisting of 421 Dutch Caucasian nonagenarian sibships and is designed to investigate the genetic determinants of human longevity. To maximally enrich for genetic signal predisposing to human longevity within the sample of sequenced genomes, we selected those sibships (N=218) displaying the most profound family history of excess survival³⁶. For each of these sibships, the DNA sequence of the genome of the sib with the highest age at censoring was determined using Next Generation Sequencing (Complete Genomics Inc.). As controls for our study, we employed sequencing data assayed on 100 individuals of Dutch Caucasian origin aged below 65 and collected by the Dutch Biobanking and Biomolecular Resources Research Infrastructure initiative^{20,21} (BBMRI). Participants of BBMRI are not selected for particular characteristics other than that they should reflect a random sample of the apparently healthy Dutch population.

5.2 Data preprocessing and quality control

Complete Genomics performed whole genome sequencing (>30x), read alignment and variant calling for both the long-lived cases as population controls, though at different time points. To minimize the technical variance between datasets, raw sequencing data created on the LLS samples was reprocessed by Complete Genomics to match the version of the preprocessing pipeline used for calling variants in the genomes of the BBMRI participants. The quality of the resulting data was re-checked (Supplemental Figures 2-6) per study separately and in combination.

One of the population controls was excluded beforehand for its distant familial relationship with one of the nonagenarian cases. Another population control displayed excessive proportions of unique variants indicating either a potential contamination of the sample before

sequencing or a mixed ancestry of one of the BBMRI participants. Multidimensional scaling was performed with 10,000 randomly selected common SNVs (MAF \geq 5%), and did not indicate the presence of population substructure. In effect, all following comparisons reported in this paper have been performed using 218 nonagenarian cases (median age 93.7, $N_{\text{male}} = 82$ (37.6%)) and 98 population controls (median age 57, $N_{\text{male}} = 39$ (39.6%)).

5.4 Assessing the significance of a genic burden of frameshift indels

To assess the significance of the presence of $k_{j,D}$ unique frameshift deletions and $k_{j,I}$ unique frameshift insertions jointly giving rise to k_j unique frameshift mutations in gene j , irrespective whether observed in long-lived cases or population controls, the following resampling approach was used. Assuming a coding transcriptome of 18,000 independent transcript clusters, we determined the prior probabilities of a gene being hit by a frameshift deletion ($p_D = 2,193/18,000 = 0.122$) or a frameshift insertion ($p_I = 1,764/18,000 = 0.098$). To assess the empirical probability $P(K_j > k_{j,D} + k_{j,I} | p_D, p_I)$ we repeatedly resampled ($Z=1,000,000$) $k_{j,D}$ deletions and $k_{j,I}$ insertions with prior probabilities p_D and p_I and counted the number of times where the resampled numbers of frameshift variants k_j^s equaled or exceeded the number of observed frameshift variants k_j , yielding k_j^s . The estimated p -value is then obtained using:

$$\hat{P}(K_j > k_{j,D} + k_{j,I} | p_D, p_I) = \sum_s (I(k_j^s) + 1) / (Z + 1) \quad (1)$$

Computations were performed in R³⁷ and repeated with different random seeds to verify the stability of the sampling experiments.

5.5 Assessing the significance of a case or control specific genic burden of frameshift indels

When inspecting the repeatedly hit genes, we noted that some genes were hit by frameshift mutations exclusively present (private) in either

the long-lived cases or the population controls. To assess the significance of the preference of a gene for being hit by $k_{j,D}^p$ private frameshift deletions and $k_{j,I}^p$ private frameshift insertions jointly giving rise to k_j^p unique and exclusive frameshift mutations in gene j , all observed in either long-lived cases or population controls, the following resampling approach was used. First we determined the prior probabilities of a frameshift deletion to be exclusively observed in long-lived cases ($p_{D,case} = 814/2,193 = 0.370$) or population controls ($p_{D,ctr} = 1,122/2,193 = 0.512$) and a frameshift insertion to be exclusively observed in long-lived cases ($p_{I,case} = 896/1,764 = 0.508$) or population controls ($p_{I,ctr} = 729/1,764 = 0.413$). Note that these probabilities do not add up to one as some deletions and insertions are observed in both the long-lived cases as the population controls and thus are not exclusive to any of the two. Furthermore, let $k_{j,D}$ and $k_{j,I}$ respectively be the total numbers of unique frameshift deletions and unique frameshift insertions observed for a particular gene j . Then we assess the empirical probability $P_{priv}(k_j^p \geq k_{j,D}^p + k_{j,I}^p | p_{D,case}, p_{D,ctr}, p_{I,case}, p_{I,ctr})$ for a given gene j by repeatedly resampling ($Z=1,000,000$) $k_{i,1}$ deletions and $k_{i,2}$ insertions with prior probabilities $p_{D,case}, p_{D,ctr}, p_{I,case}$ and $p_{I,ctr}$ for respectively obtaining private deletions ($k_{i,d}^{p,s}$) and insertions ($k_{i,i}^{p,s}$) in cases and controls for each sampling and subsequently counted the number of times the number of sampled private mutations $k_{i,1}^{p,s}$ equaled or exceeded the observed number of private mutations k_i^p ($k_{i,i}^{p,s}$). The p -value was then estimated by:

$$\hat{P}(k_j^p > k_{j,D}^p + k_{j,I}^p | p_{D,case}, p_{D,ctr}, p_{I,case}, p_{I,ctr}) = \frac{\sum_s (I(k_j^{p,s}) + 1)}{(Z + 1)} \quad (2)$$

5.6 Somatic calls

Heterozygotic variant calls with read evidence deviating from the expected 1:1 ratio might point to the presence of a somatic variant that is present in part of the sequenced DNA. Alternatively, it might comprise either a sequencing error, or an under-sampling of a truly heterozygotic variant, which both can be modeled by employing Poisson distributions.

First we model the probability of sequencing errors explaining the observed disbalance in ratio's, by assuming an error rate $E = 1\%$ of reads falsely supporting a variant call. Hence, a Poisson model $P(\lambda, K)$ with mean $\lambda = E \times R_{tot}$ and $K = R_{var}$ is used to estimate the probability p_{hom} that a variant, called with reads R_{tot} of which at least R_{var} support the variant, is likely to comprise a homozygous reference variant with some noisy reads. Similarly, we employ a Poisson model with mean $\lambda = 0.5 \times R_{tot}$ and $K = R_{var}$, to estimate the probability p_{het} that the alternative allele, supported by R_{var} or less reads, is likely to be a truly heterozygotic variant of which the alternative allele is under-sampled relative to the reference. In case both these hypotheses are rejected, we may assume that the variant is indeed a somatic variant, thus: $p_{som} = \max(p_{hom}, p_{het})$.

5.7 Associations with granulocyte counts

Absolute counts of granulocytes in long-lived cases were computed by summing counts of neutrophils, eosinophils and basophils derived from whole blood cell counts. Differences in granulocyte counts between carriers of disruptive variants in *TET2* or *DNMT3A* were tested using a linear model as implemented in the *lm* package of the statistical language R³⁷:

$$G \sim \beta_1 \times age + \beta_2 \times sex + \beta_3 \times carrier \quad (3)$$

where the covariates *age* is provided in years, *sex* as either 1 (male) or 2 (female), *carrier* as either 0 or 1 to indicate carriership of a disruptive variant.

5.8 Associations with prospective survival

Associations with prospective survival were performed with the *Survival* package³⁸ of R³⁷ using an age at inclusion and sex-adjusted, left-truncated Cox proportional hazards model to adjust for late entry into the dataset according to age. Mortality analyses between carriers

and non-carriers of disruptive variants in *TET2* were performed using:

$$\lambda(t) \sim \lambda_0(t) \times \exp(\beta_1 \times \text{age} + \beta_2 \times \text{sex} + \beta_3 \times \text{carrier}) \quad (4)$$

where the covariates *age* designates age at inclusion and is provided in years, *sex* as either 1 (male) or 2 (female), *carrier* as either 0 or 1 to indicate carriership of a disruptive variant.

6. Acknowledgements

The research leading to these results has received funding from the Medical Delta (COMO), Pfizer Inc, and the European Union's Seventh Framework Programme (FP7/2007-2011) under grant agreement number 259679. This study was financially supported by the Innovation-Oriented Research Program on Genomics (SenterNovem IGE05007), the Centre for Medical Systems Biology and the Netherlands Consortium for Healthy Ageing (grant 050-060-810), all in the framework of the Netherlands Genomics Initiative, Netherlands Organization for Scientific Research (NWO), by Unilever Colworth and by BBMRI-NL, a Research Infrastructure financed by the Dutch government (NWO 184.021.007).

7. 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. Skytthe, A. *et al.* Longevity studies in GenomEUtwin. *Twin Res* **6**, 448-54 (2003).
4. 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).
5. Atzmon, G. *et al.* Clinical phenotype of families with longevity. *J Am Geriatr Soc* **52**, 274-7 (2004).
6. Terry, D.F. *et al.* Lower all-cause, cardiovascular, and cancer mortality in centenarians' offspring. *J Am Geriatr Soc* **52**, 2074-6 (2004).
7. Beekman, M. *et al.* Genome-wide association study (GWAS)-identified disease risk alleles do not compromise human longevity. *Proc Natl Acad Sci U S A* **107**, 18046-9 (2010).
8. Deelen, J. *et al.* Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. *Hum Mol Genet* (2014).
9. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-8 (2012).
10. Garinis, G.A., van der Horst, G.T., Vijg, J. & Hoeijmakers, J.H. DNA damage and ageing: new-age ideas for an age-old problem. *Nat Cell Biol* **10**, 1241-7 (2008).
11. Hoeijmakers, J.H. DNA damage, aging, and cancer. *N Engl J Med* **361**, 1475-85 (2009).
12. Clancy, D.J. *et al.* Extension of life-span by loss of CHICO, a Drosophila insulin receptor substrate protein. *Science* **292**, 104-6 (2001).
13. Sebastiani, P. *et al.* Whole genome sequences of a male and female supercentenarian, ages greater than 114 years. *Front Genet* **2**, 90 (2011).
14. Ye, K. *et al.* Aging as accelerated accumulation of somatic variants: whole-genome sequencing of centenarian and middle-aged monozygotic twin pairs. *Twin Res Hum Genet* **16**, 1026-32 (2013).
15. Holstege, H. *et al.* A Longevity Reference Genome Generated From the World's Oldest Woman. *Oral Presentation ASHG 2011* (2011).
16. Gierman, H.J. *et al.* Whole-Genome Sequencing of the World's Oldest People. *PLoS One* **9**, e112430 (2014).

17. Cash, T.P. *et al.* Exome sequencing of three cases of familial exceptional longevity. *Aging Cell* **13**, 1087-90 (2014).
18. 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).
19. 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).
20. Boomsma, D.I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet* **22**, 221-7 (2014).
21. The Genome of the Netherlands, C. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* (2014).
22. CG_website. http://media.completegenomics.com/documents/DataFileFormats_Standard_Pipeline_2.4.pdf.
23. Busque, L. *et al.* Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. *Nat Genet* **44**, 1179-81 (2012).
24. Beerman, I. *et al.* Functionally distinct hematopoietic stem cells modulate hematopoietic lineage potential during aging by a mechanism of clonal expansion. *Proc Natl Acad Sci U S A* **107**, 5465-70 (2010).
25. Beerman, I., Maloney, W.J., Weissmann, I.L. & Rossi, D.J. Stem cells and the aging hematopoietic system. *Curr Opin Immunol* **22**, 500-6 (2010).
26. Delhommeau, F. *et al.* Mutation in TET2 in myeloid cancers. *N Engl J Med* **360**, 2289-301 (2009).
27. Ley, T.J. *et al.* DNMT3A mutations in acute myeloid leukemia. *N Engl J Med* **363**, 2424-33 (2010).
28. Okano, M., Xie, S. & Li, E. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nat Genet* **19**, 219-20 (1998).
29. Mohr, F., Dohner, K., Buske, C. & Rawat, V.P. TET genes: new players in DNA demethylation and important determinants for stemness. *Exp Hematol* **39**, 272-81 (2011).
30. Trowbridge, J.J. & Orkin, S.H. Dnmt3a silences hematopoietic stem cell self-renewal. *Nat Genet* **44**, 13-4 (2012).
31. Moran-Crusio, K. *et al.* Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. *Cancer Cell* **20**, 11-24 (2011).
32. Franceschi, C., Bonafe, M. & Valensin, S. Human immunosenescence: the prevailing of innate immunity, the failing of clonotypic immunity, and the filling of immunological space. *Vaccine* **18**, 1717-20 (2000).
33. Jankowska, A.M. *et al.* Loss of heterozygosity 4q24 and TET2 mutations associated with myelodysplastic/myeloproliferative neoplasms. *Blood* **113**, 6403-10 (2009).
34. Ewalt, M. *et al.* DNMT3a mutations in high-risk myelodysplastic syndrome parallel those found in acute myeloid leukemia. *Blood Cancer J* **1**, e9 (2011).
35. Metzeler, K.H. *et al.* TET2 mutations improve the new European LeukemiaNet risk classification of acute myeloid leukemia: a Cancer and Leukemia Group B study. *J Clin Oncol* **29**, 1373-81 (2011).
36. Rosing, M.P. *et al.* Familial longevity is associated with decreased thyroid function. *J Clin Endocrinol Metab* **95**, 4979-84 (2010).
37. R-Core-Team. R: A Language and Environment for Statistical Computing. (2013).
38. Therneau, T. A Package for Survival Analysis in S. (R package version 2.37-7; <http://CRAN.R-project.org/package=survival>, 2014).

Supplemental Materials

Genome of the Netherlands Consortium members:

Analysis group: Morris A. Swertz^{6,7} (Co-Chair), Laurent C. Francioli¹, Freerk van Dijk^{6,7}, Androniki Menelaou¹, Pieter B.T. Neerincx^{6,7}, Sara L. Pulit¹, Patrick Deelen^{6,7}, Clara C. Elbers¹, Pier Francesco Palamara², Itsik Pe'er^{2,8}, Abdel Abdellaoui⁹, Wigard P. Kloosterman¹, Mannis van Oven¹⁰, Martijn Vermaat¹¹, Mingkun Li¹², Jeroen F.J. Laros¹¹, Mark Stoneking¹², Peter de Knijff¹³, Manfred Kayser¹⁰, Jan H. Veldink¹⁴, Leonard H. van den Berg¹⁴, Heorhiy Byelas^{6,7}, Johan T. den Dunnen¹¹, Martijn Dijkstra^{6,7}, Najaf Amin¹⁵, K. Joeri van der Velde^{6,7}, Jouke Jan Hottenga⁹, Jessica van Setten¹, Elisabeth M. van Leeuwen¹⁵, Alexandros Kanterakis^{6,7}, Mathijs Kattenberg⁹, Lennart C. Karssen¹⁵, Barbera D.C. van Schaik¹⁶, Jan Bot¹⁷, Isaäc J. Nijman¹, David van Enckevort¹⁸, Hailiang Mei¹⁸, Vyacheslav Koval¹⁹, Kai Ye^{20,21}, Eric-Wubbo Lameijer²¹, Matthijs H. Moed²¹, Jayne Y. Hehir-Kwa²², Robert E. Handsaker^{5,23}, Shamil R. Sunyaev^{4,5}, Mashaal Sohail^{4,5}, Fereydoun Hormozdiari²⁴, Tobias Marschall²⁵, Alexander Schönhuth²⁵, Victor Guryev²⁶, Paul I.W. de Bakker^{1,3-5} (Co-Chair);

Cohort collection and sample management group: P. Eline Slagboom²¹, Marian Beekman²¹, Anton J.M. de Craen²¹, H. Eka D. Suchiman²¹, Albert Hofman¹⁵, Cornelia van Duijn¹⁵, Dorret I. Boomsma⁹, Gonneke Willemssen⁹, Bruce H. Wolffenbuttel²⁷, Mathieu Plattee⁶, Steven J. Pitts²⁸, Shobha Potluri²⁸, David R. Cox^{28,34};

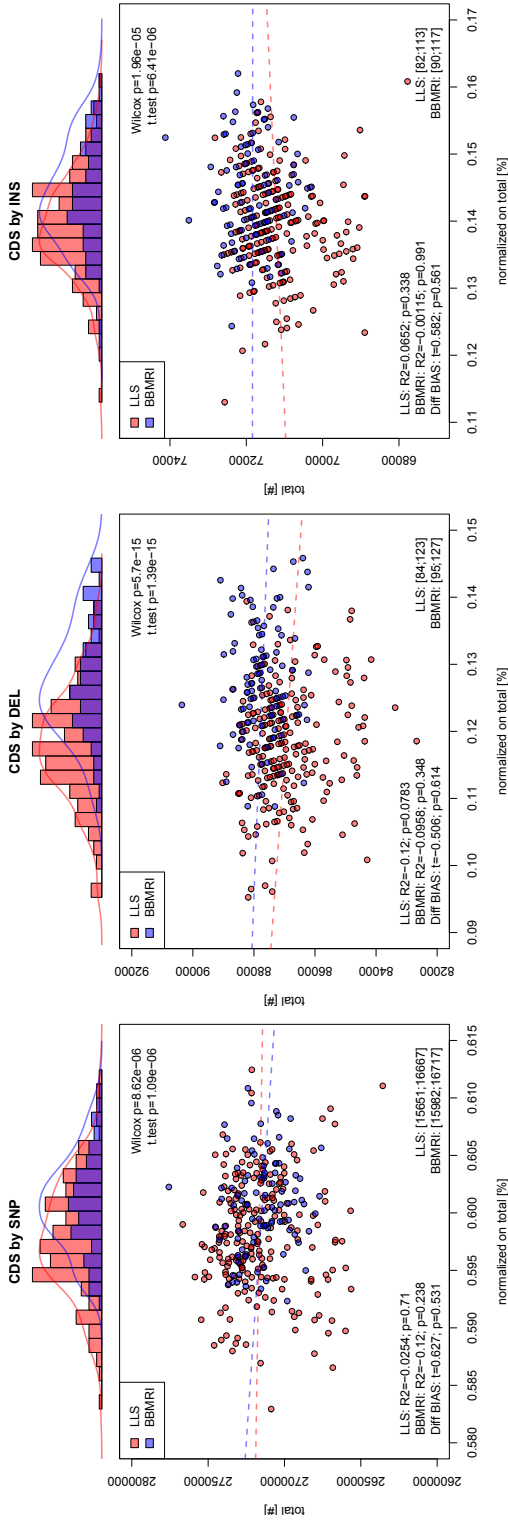
Whole-genome sequencing: Qibin Li²⁹, Yingrui Li²⁹, Yuanping Du²⁹, Ruoyan Chen²⁹, Hongzhi Cao²⁹, Ning Li³⁰, Sujie Cao³⁰, Jun Wang^{29,31,32}; Ethical, Legal, and Social Issues: Jasper A. Bovenberg³³

Steering committee: Cisca Wijmenga^{6,7} (Principal Investigator), Morris A. Swertz^{6,7}, Cornelia M. van Duijn¹⁵, Dorret I. Boomsma⁹, P. Eline Slagboom²¹, Gertjan B. van Ommen¹¹, Paul I.W. de Bakker^{1,3-5}

Affiliations:

- 1: Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands
- 2: Department of Computer Science, Columbia University, New York, NY, USA
- 3: Department of Epidemiology, University Medical Center Utrecht, Utrecht, The Netherlands
- 4: Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
- 5: Broad Institute of Harvard and MIT, Cambridge, MA, USA
- 6: Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
- 7: Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
- 8: Department of Systems Biology, Columbia University, New York, NY, USA
- 9: Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands
- 10: Department of Forensic Molecular Biology, Erasmus Medical Center, Rotterdam, The Netherlands
- 11: Leiden Genome Technology Center, Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
- 12: Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
- 13: Forensic Laboratory for DNA Research, Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

- 14: Department of Neurology, University Medical Center Utrecht, Utrecht, The Netherlands
- 15: Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands
- 16: Bioinformatics Laboratory, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Amsterdam Medical Center, Amsterdam, The Netherlands
- 17: SURFsara, Science Park, Amsterdam, The Netherlands
- 18: Netherlands Bioinformatics Centre, Nijmegen, The Netherlands
- 19: Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands
- 20: The Genome Institute, Washington University, St. Louis, MO, USA
- 21: Section of Molecular Epidemiology, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands
- 22: Department of Human Genetics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands
- 23: Department of Genetics, Harvard Medical School, Boston, MA, USA
- 24: Department of Genome Sciences, University of Washington, Seattle, WA, USA
- 25: Centrum voor Wiskunde en Informatica, Life Sciences Group, Amsterdam, The Netherlands
- 26: European Research Institute for the Biology of Ageing, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
- 27: Department of Endocrinology, University Medical Center Groningen, Groningen, The Netherlands
- 28: Rinat-Pfizer Inc, South San Francisco, CA, USA
- 29: BGI-Shenzhen, Shenzhen, China
- 30: BGI-Europe, Copenhagen, Denmark
- 31: Department of Biology, University of Copenhagen, Copenhagen, Denmark
- 32: The Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark
- 33: Legal Pathways Institute for Health and Bio Law, Aerdenhout, The Netherlands
- 34: Deceased



SUPPLEMENTAL FIGURE 1: A SENSITIVITY ANALYSIS ON THE OBSERVED DIFFERENCES IN PROPORTIONS OF VARIANTS ANNOTATED TO THE CDS BETWEEN LONG-LIVED CASES AND POPULATION CONTROLS WITH RESPECT TO THE OVERALL CALLING QUALITY, PROXIED BY OVERALL CALLING RATES. The proportions of variants annotated to the CDS are related to the total number of variants discovered within the long-lived cases (red) and random population controls (blue) for the variant types SNV, DEL and INS separately. Distributions of the percentages of variants annotated to CDS are displayed at the top of the figure for each variant type respectively. Test results for differences in these distributions are reported below (Wilcoxon Rank-Sum test and a Welch's t test). Raw data is also plotted in the scatter plot below to visually inspect the relation between proportions of variants annotated to the CDS (x-axis) and the absolute number of variants identified per sample. The correlation between the two, as illustrated by the dotted lines, is assessed per study and is compared between studies, reported in the low left corner. The minimal and maximal numbers of variants annotated to CDS are reported in the right lower corner for both studies separately. No significant biases were observed in these plots.

GERM LINE AND SOMATIC CHARACTERISTICS OF THE LONG-LIVED GENOME

Type	Location	Impact	BBMRI	LLS	stat.W	p
DEL	CDS	FRAMESHIFT	40	32	19378	5.53E-31
INS	CDS	FRAMESHIFT	31.5	28	15414.5	3.00E-10
SNV	DONOR		358	354	14145	4.05E-06
SNV	CDS	MISSENSE	7182	7131	13846	2.54E-05
SNV	UTR5		2637.5	2625	13510	1.67E-04
SNV	CDS	SYNONYMOUS	8051	8026	13034	1.75E-03
DEL	CDS	DELETE	44	45	8894	0.02
SNV	UTR3		18785.5	18762.5	12194	0.04
SNV	CDS	NONSTOP	10	9	11983	0.08
INS	UTR3		793	791	9560	0.14
INS	CDS	INSERT	34	34	9572.5	0.14
SNV	CDS	NONSENSE	53	52	11766	0.15
DEL	UTR5		73	71	11757	0.15
SNV	TSS- UPSTREAM		104542.5	104644	11746	0.16
SNV	ACCEPTOR	DISRUPT	20	21	9629	0.16
SNV	ACCEPTOR		1733	1730	11734	0.16
INS	ACCEPTOR		60	61	9802	0.24
DEL	INTRON		35412.5	35148.5	9824	0.25
INS	UTR5		74	73	11402	0.34
SNV	DONOR	DISRUPT	34	35	11261	0.44
DEL	DONOR		20	20	11170	0.52
DEL	TSS- UPSTREAM		3787.5	3767	10260	0.57
INS	DONOR		13	13	11015.5	0.66
DEL	UTR3		920	912	10363	0.67
INS	INTRON		28921.5	28725	10381	0.69
SNV	INTRON		1036737	1037931	10449	0.76
DEL	ACCEPTOR		100	99	10454	0.76
SNV	CDS	MISSTART	16	15.5	10838	0.84
INS	TSS- UPSTREAM		3199	3183.5	10575	0.89

SUPPLEMENTAL TABLE 1: MEDIAN COUNTS PER VARIANT CATEGORY. Median counts of variants observed per variant category in long-lived cases (LLS) and population controls (BBMRI). Comparisons with median counts < 10 for both long-lived cases (LLS) as population controls (BBMRI) were not considered (NONSTOP SNV). Differences between distributions of counts normalized on totals per gvarType were tested using the Wilcoxon Rank-Sum test.

AssayID	Chrom	Start	End	Type	Ref	Alt	Carrier
LLS_01	chr2	113479780	113479780	INS		C	FAILED
LLS_02	chr11	120198285	120198287	DEL	CT		Yes
LLS_03	chr19	16976264	16976264	INS		G	FAILED
LLS_04	chr17	62038698	62038698	INS		C	No
LLS_05	chr17	7323945	7323946	SUB	C	AA	Yes
LLS_06	chr1	151372104	151372104	INS		C	No
LLS_7	chr4	95496888	95496888	INS		A	No
LLS_08	chr2	27730169	27730169	INS		A	Yes
LLS_09	chr10	55568865	55568867	DEL	TG		Yes
LLS_10	chr17	5404003	5404004	DEL	A		FAILED
LLS_11	chr9	90500990	90500992	DEL	CT		Yes
LLS_12	chr9	134398412	134398412	INS		G	Yes
LLS_12	chr9	134398412	134398412	INS		G	Yes
LLS_13	chr13	113980131	113980135	DEL	AAAC		Yes
LLS_13	chr13	113980131	113980135	DEL	AAAC		No
LLS_14	chr7	76828864	76828867	SUB	GAC	AGGT	No
LLS_15	chr17	41174273	41174274	SUB	T	AA	No

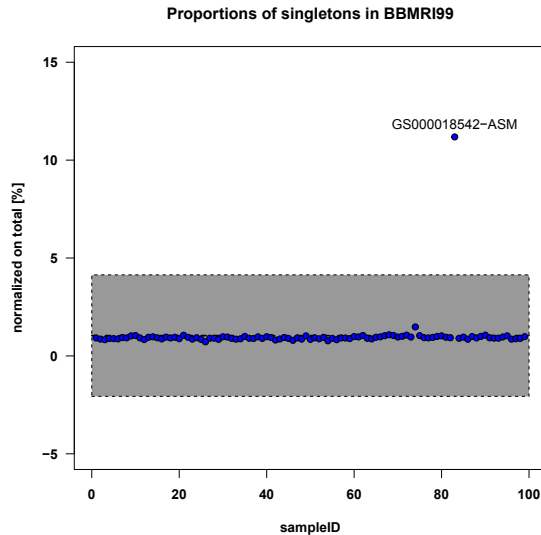
SUPPLEMENTAL TABLE 2: SANGER SEQUENCING EXPERIMENTS ON FRAMESHIFT VARIANTS IDENTIFIED WITHIN THE LONG-LIVED CASES. Of the 15 independent assays designed, 12 returned good data, which confirmed the presence of 7 variants.

AssayID	Chrom	Start	End	Type	Ref	Alt	Carrier
BBMRI_01	chr3	121208840	121208840	INS		T	No
BBMRI_02	chr2	11696893	11696897	DEL	GAAG		Yes
BBMRI_03	chr10	118305625	118305629	SUB	TCAC	GGACT	No
BBMRI_04	chr1	100207825	100207825	INS		T	No
BBMRI_05	chr22	37964284	37964285	DEL	G		No
BBMRI_06	chr7	134719554	134719554	INS		C	No
BBMRI_07	chr9	35738865	35738865	INS		A	No
BBMRI_08	chr13	97639501	97639501	INS		AAGAAGGCATCT	Yes
BBMRI_09	chr18	29122734	29122734	INS		G	No
BBMRI_10	chr18	55322554	55322555	SUB	C	AA	No
BBMRI_11	chr1	11008274	11008275	DEL	C		No
BBMRI_12	chr16	46695701	46695702	DEL	G		No
BBMRI_13	chr9	113457714	113457714	INS		A	No
BBMRI_14	chr4	122741747	122741748	DEL	A		No
BBMRI_15	chr17	7369290	7369291	DEL	C		No

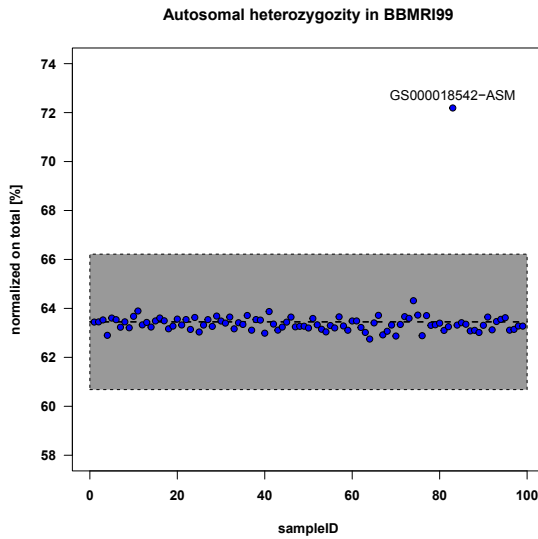
SUPPLEMENTAL TABLE 3: SANGER SEQUENCING EXPERIMENTS ON FRAMESHIFT VARIANTS IDENTIFIED WITHIN THE POPULATION CONTROLS. Of the 15 independent assays designed only 2 confirmed the presence of the targeted variant.

Gene	Chrom	Position	Type	Ref	Alt	Impact	Allele Counts (Alt Ref)
<i>TET2</i>	chr4	106156278	DEL	G		FRAMESHIFT	1 8251
<i>TET2</i>	chr4	106156687	SNV	C	T	NONSENSE	1 8599
<i>TET2</i>	chr4	106157504	DEL	C		FRAMESHIFT	1 8251
<i>TET2</i>	chr4	106157506	SNV	C	T	NONSENSE	1 8597
<i>TET2</i>	chr4	106157653	SNV	G	T	NONSENSE	1 8599
<i>TET2</i>	chr4	106157700	SNV	T	G	NONSENSE	1 8599
<i>TET2</i>	chr4	106157807	DEL	C		FRAMESHIFT	3 8251
<i>TET2</i>	chr4	106158113	DEL	G		FRAMESHIFT	1 8253
<i>TET2</i>	chr4	106158157	SNV	C	T	NONSENSE	1 8599
<i>TET2</i>	chr4	106158441	DEL	C		FRAMESHIFT	21 8233
<i>DNMT3A</i>	chr2	25459834	SNV	C	A	NONSENSE	1 8599
<i>DNMT3A</i>	chr2	25466830	DEL	T		FRAMESHIFT	1 8115
<i>DNMT3A</i>	chr2	25467468	SNV	G	C	NONSENSE	1 8599
<i>DNMT3A</i>	chr2	25468163	SNV	C	A	NONSENSE	1 8599
<i>DNMT3A</i>	chr2	25468917	DEL	TCGTACA		FRAMESHIFT	20 8234
<i>DNMT3A</i>	chr2	25469529	DEL	C		FRAMESHIFT	12 8226
<i>DNMT3A</i>	chr2	25471030	DEL	GGCT		FRAMESHIFT	69 8185

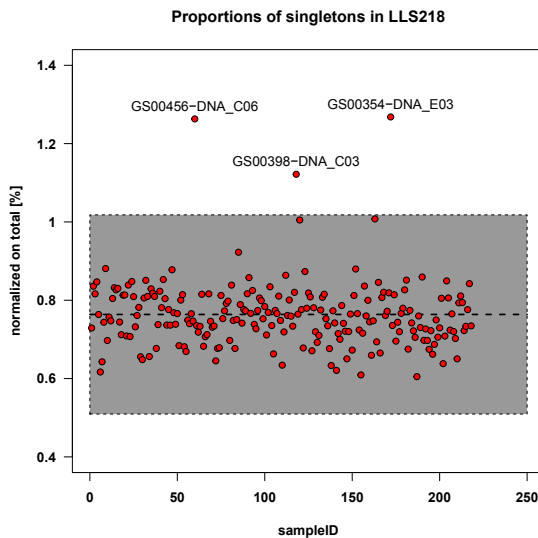
SUPPLEMENTAL TABLE 4: FRAMESHIFT AND NONSENSE VARIANTS IN *TET2* AND *DNMT3A* ON EXOME VARIANT SERVER. Variants were called against the reference transcript NM_017628.4 and NM_022552.4 for *TET2* and *DNMT3A* respectively.



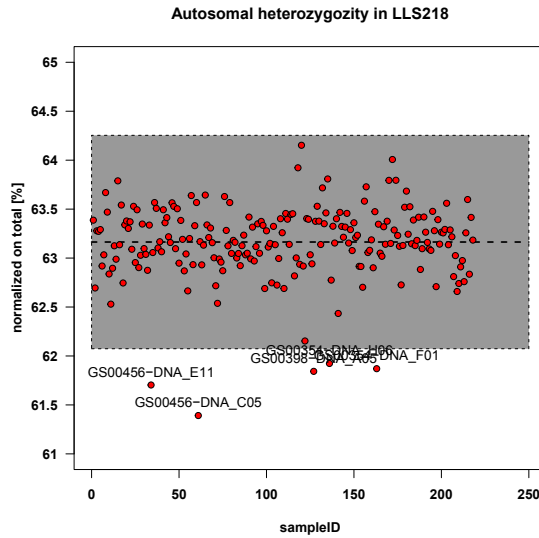
SUPPLEMENTAL FIGURE 2: SINGLETON CHECK BBMRI. Depicted are the proportions of singletons (SNVs unique for one sample) and overall numbers of identified SNVs per sample within the BBMRI study. The grey area marks the 3 SD thresholds, indicating that sample GS000018542-ASM has a disproportionately high number of variants not observed in the rest of the study, suggesting either a distinct ancestry or a sample contamination.



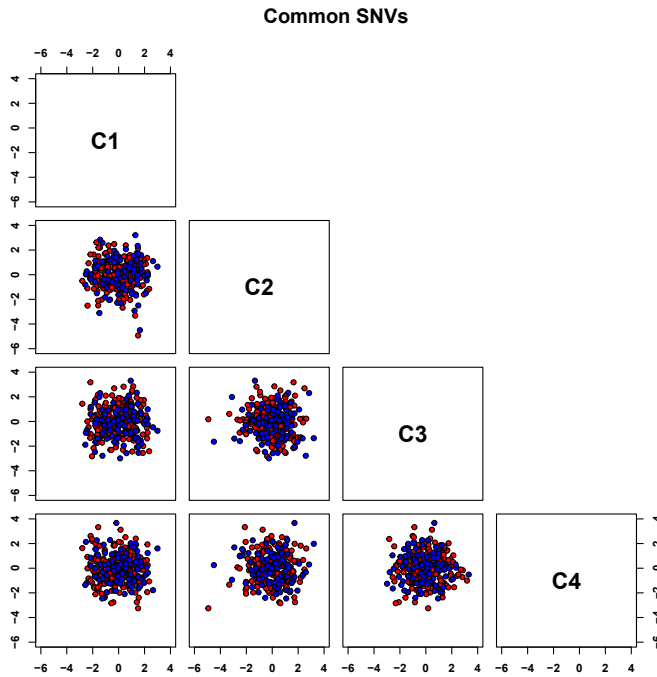
SUPPLEMENTAL FIGURE 3: AUTOSOMAL HETEROZYGOSITY BBMRI. The proportions of heterozygous SNV genotypes and overall numbers of identified SNVs per sample within the BBMRI study. Again the grey area marks the 3 SD thresholds, indicating again that sample GS0000018542-ASM exhibits a genomic make up that is very distinct from the remaining participants of the BBMRI study. Such an elevated heterozygosity again points to either a distinct ancestry or a sample contamination. Due to the consistent appearance of sample GS0000018542-ASM as a major outlier, we decided to remove it from further analyses.



SUPPLEMENTAL FIGURE 4: SINGLETON CHECK LLS. Depicted are the proportions of singletons (SNVs unique for one sample) and overall numbers of identified SNVs per sample within the LLS study. The 3 SD deviation of the expectation is indicated in grey. Slightly elevated proportions of unique SNVs are observed for GS00456-DNA_C06, GS00354-DNA_E03 and GS00398-DNA_C03.



SUPPLEMENTAL FIGURE 5: AUTOSOMAL HETEROZYGOSITY LLS. The proportions of heterozygous SNV genotypes and overall numbers of identified SNVs per sample within the LLS study. The 3 SD deviation of the expectation is indicated in grey. Slightly lowered proportions of heterozygous SNVs are observed for GS00354-DNA_H06, GS00354-DNA_F01, GS00456-DNA_E11, GS00456-DNA_C05 and GS00398-DNA_A05. Noteworthy is that none of the samples overlapped with the outliers that came forward in the singleton check. Various types of artefacts such as mixed ancestry, sample pollution, variation in total read depth or just biological variation might explain slight deviations in both the singleton and heterozygosity proportions. However, since outliers were not consistently picked up in both tests, we decided not to exclude any samples.



SUPPLEMENTAL FIGURE 6: MULTIDIMENSIONAL SCALING. MDS was performed with Plink using 10,000 randomly selected common SNVs ($MAF \geq 5\%$) to inspect the data for signs of differences in population substructure. Sample space was reduced to four dimensions and all combinations thereof are plotted. Long-lived cases are displayed in red, population controls in blue. No apparent substructure was observed.

Chapter 5:

A novel life span regulating locus at chr13q34 influencing serum triiodothyronine level

Erik B. van den Akker^{1,2}, Steven J. Pitts³, Joris Deelen^{1,4}, Matthijs H. Moed¹, Shobha Potluri³, H. Eka D. Suchiman¹, Diana van Heemst⁵, Jeanine J. Houwing-Duistermaat⁶, David R. Cox^{3†}, Marcel J.T. Reinders², Marian Beekman^{1,4}, P. Eline Slagboom^{1,4}

1. Molecular Epidemiology, Leiden University Medical Center, Leiden, The Netherlands
2. The Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands
3. Rinat-Pfizer Inc, South San Francisco, CA 94080, USA
4. Netherlands Consortium of Healthy Ageing
5. Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands
6. Medical Statistics, Leiden University Medical Center, Leiden, The Netherlands

† In memoriam

In preparation

1. Abstract

The genetic propensity for an attenuated thyroid function and lifespan extension co-occur in long-lived families and suggest the existence of a pleiotropic genetic mechanism underlying human longevity as well as attenuation of the thyroid function. Attenuation of the thyroid function is even more profound in long-lived sibships whose parents also exhibited an excess survival compared to their respective birth and sex cohorts, and thus suggests an enrichment of variants explaining the pleiotropic relation in sibships with this marked family history (FH(+)). Linkage analyses among all 415 nonagenarian sibships from the Leiden Longevity Study (LLS) identified suggestive linkage at chr13q34 (LOD=2.96) that was highly specific to the 239 FH(+) sibships (LOD=3.35). The FH(+) subset was characterized by significantly lowered levels of serum free triiodothyronine (fT3) and its interaction with prospective survival thereof. Subsequent fine mapping of the 1-LOD-drop 13q34 region by QTL analyses on serum fT3 levels using variants from whole genome sequencing data identified a fT3 lowering haplotype tagged by the rs9515460 minor C allele, that also associates with prospective survival. Hence, we hereby report chr13q34 as a novel locus harbouring variants underlying lifespan regulation as well as an attenuated thyroid function.

2. Introduction

Attenuation of the thyroid function has been implicated in life span extension in rodents¹⁻³, as well as in several large longitudinal human cohort studies⁴⁻⁶. Active thyroid hormone, triiodothyronine (T3), and its precursor, thyroxine (T4), are produced by the thyroid gland under influence of thyroid stimulating hormone (TSH), and are both key factors in the regulation of the basal metabolic rate and cardiac output. A mildly lowered thyroid function (subclinical hypothyroidism), marked by an increased serum level of TSH and a reduced level of free T4, has been associated with a lowered risk of cardiovascular disease from middle age onward⁶ and a decreased cardiovascular mortality in the oldest old⁴. Interestingly, first-degree relatives of long-lived persons have, compared to the general population, an increased probability of becoming long-lived themselves⁷⁻⁹, while exhibiting from middle age onward an attenuated thyroid function¹⁰ and an improved cardiovascular health⁷⁻⁹. Moreover, attenuation of the thyroid function seems to be more profound in long-lived families with a family history of excess survival, as compared to long-lived families without such a marked family history¹¹. These findings suggest a genetic mechanism for human longevity in which a life-long survival advantage, especially with respect to cardiovascular mortality, is promoted by attenuation of the thyroid function.

The relation of the active thyroid hormone itself with respect to cardiovascular mortality is less clear, possibly because serum T3 as a marker

integrates cues from both the endocrine system, through conversion of T4, as non-thyroidal routes¹². Parle *et al.* finds levels of free T4 (fT4) and free T3 (fT3) to inversely correlate with the level of TSH in hyperthyroid patients aged over 60 years, but observes a significant association with cardiovascular mortality only for TSH and not for fT4 nor for fT3¹³. In contrast, Gussekloo *et al.* finds significant associations with mortality for TSH, fT4 and fT3 in patients aged above 85, but only reports an inverse correlation between levels of TSH and fT4, and not for TSH and fT3⁴. More importantly, whereas lowered fT4 levels confer a protective effect, they find lowered levels of fT3 to convey an increased risk on mortality and argue that the association of fT3 with mortality can be explained by disability at baseline. Thus, whereas the findings of Parle *et al.* and Gussekloo *et al.* agree on the protective effect of high serum TSH levels, they disagree on the relation between thyroid parameters, thereby questioning the causality of fT3 in conveying the protective effect on life span regulation. Since the fT4 precursor also exhibits a very modest regulatory capacity, but is present in a much higher abundance, it raises the possibility that the life-prolonging effects of an attenuated thyroid function are either totally dependent on or directly transmitted through serum fT4. Genetic research on the main parameters of the thyroid system e.g. TSH, fT4 and fT3, may shed light on the causal mechanism relating an attenuated thyroid function to life span regulation.

The genetics of thyroid homeostasis, human longevity and interactions thereof

have been studied with varying degrees of success. The genetic influences on the phenotypic variation of the main parameters involved in thyroid metabolism have been shown to be moderate to strong (30-60%)^{11,14}. Furthermore, genome wide associations studies (GWASs) have identified over 60 genome-wide significant loci associated with serum levels of either TSH or fT4¹⁵⁻¹⁹. However, despite this strong evidence for genetic variation to influence the phenotypic variation in thyroid metabolism and the strong indications that attenuation of the thyroid system, at least with respect to TSH, leads to a prolonged life span, little evidence has been reported for genetic variations to affect both traits. An important exception is the work by Atzmon *et al.* reporting two common variants upstream of the TSH receptor gene (*TSHR*) to be enriched amongst the oldest old in the Ashkenazi Jewish population, as compared to middle aged controls, while also marking increased TSH serum levels²⁰. Neither of these two variants, however, was re-identified in any GWAS on either serum TSH levels¹⁵⁻¹⁹ or human longevity²¹⁻²⁴, suggesting considerable heterogeneity in both traits. In summary, whereas genetic approaches have been able to relate common variants to phenotypic variation in each of TSH, fT4 and life span regulation separately, little evidence exists to date for genetic factors influencing both the thyroid metabolism and human longevity.

Here we study the genomes of members of long-lived families of the Leiden Longevity Study (LLS) for the presence of rare variants that explain the pleiotropic interactions between an attenuated thyroid function and human life span regulation.

To this end, we focussed on nonagenarian sibships of the Leiden Longevity Study (LLS) with parents exhibiting the largest excess survival relative to their respective sex and birth cohort specific life expectancies¹¹. Since long-lived sibships with this marked family history (FH(+)) exhibit an even more attenuated thyroid function, as compared to long-lived sibships without such a marked family history (FH(-))¹¹, we hypothesize the FH(+) subset to be enriched for genetic determinants underlying familial longevity by attenuating the thyroid function. To this end, we performed affected sibling pair analyses to identify genomic regions exhibiting linkage for familial longevity in the whole study, and show that one of the suggestive signals (chr13q34) is explained by significant linkage among the FH(+) sibships specifically. Since the FH(+) sibships are characterized by significantly lower fT3 levels, as compared to FH(-) sibships, which moreover seems to affect their prospect of survival, we hypothesize the 13q34 locus to harbour variants underlying human lifespan regulation by lowering serum fT3. Subsequent QTL analyses for loci associating with serum fT3 employing variants from whole genome sequencing data residing in chr13q34, identified a serum fT3 lowering haplotype tagged by the minor C allele of rs9515460_T<C. This serum fT3 lowering haplotype was also shown to associate with prospective survival in a 10 years follow up. Hence, we hereby report chr13q34 as a novel locus harbouring variants contributing to lifespan regulation as well as an attenuated thyroid function.

3. Results

3.1 Demographic and thyroid characteristics

Demographic and thyroid characteristics in all 415 nonagenarians sibships of the Leiden Longevity Study (LLS), the 239 sibships enriched for family history of extended survival (FH(+)), the 176 non-enriched sibships (FH(-)) and the 214 unrelated nonagenarians from enriched families selected for whole genome sequencing (SEQ_FH(+)) are given in Table 1 (Experimental Procedures 1). Previously, we showed that a family history of excess survival, as indicated by a lower Family History Score, significantly associated with higher levels of serum TSH and lower levels of serum fT4 and serum fT3¹¹. Although all thyroid parameters show directions of correlation consistent with those reported earlier, the current dichotomization into two subsets of sibships yielded significant differences in baseline levels for serum TSH ($N_{FH(-)}=344$, $N_{FH(+)}=482$ $\beta=1.06$, 95% CI 1.00-1.12, $p=0.039$) and serum fT3 ($N_{FH(-)}=348$, $N_{FH(+)}=492$, $\beta=-0.10$, 95% CI -0.19--0.01, $p=0.026$), but not for serum fT4 ($p=0.17$, see Experimental Procedures 2 and 3.1). This suggests that in the current setting, we have most power to identify genetic determinants underlying lifespan regulation and an attenuated thyroid function through influencing either serum TSH or fT3 levels.

3.2 Prospective survival of study subsets and thyroid parameters

We compared the prospective survival above 90 years between the FH(-) and FH(+) subsets and between the FH(-) and

the SEQ_FH(+) subsets, but found no significant differences, indicating that a family history of excess survival has a negligible effect on the survival beyond age 90 (Experimental Procedures 3.2). Except for a suggestive association in the FH(-) subset, serum TSH levels did not display any significant associations with prospective survival beyond 90 years, neither in the whole study nor for any of the analysed subsets of the LLS (Table 2). However, when investigating the influence of the thyroid parameters fT4 and fT3 on prospective survival, we observed significant associations for both these thyroid parameters in the whole LLS (ALL) and some interesting interactions with the FH strata (Table 2). The association of fT4 with prospective survival in the whole study was explained by the protective effect of low serum fT4 levels specifically observed in the FH(-) subset. In contrast, the association of fT3 with prospective survival in the whole study was explained by the deleterious effect of low serum fT3 levels most profoundly observed in the FH(+) subset, confirming the observation of Gussekloo *et al.* that low serum fT3 becomes detrimental in the oldest old (Figure 1). Joint modelling of fT4 and fT3 with prospective survival in the whole study and its subsets indicates that these associations are independent (data not shown). From these findings we conclude that the FH(-) subset seems to exhibit relations between the thyroid parameters and prospective survival that are very similar to those reported by Gussekloo *et al.* for the oldest old in the general population¹.

¹ High TSH and low fT4 levels are beneficial, low fT3 levels are detrimental.

	ALL 415 sibships	FH(+) 239 sibships	FH(-) 176 sibships	SEQ_FH(+) 214 cases
Total # individuals	931 (37.4% male)	540 (36.7 % male)	391 (38.4% male)	214 (37.9% male)
Age at inclusion [years]	92.9 (91.5 – 94.8)	93.1 (91.5 – 95.1)	92.8 (91.4 – 94.7)	93.7 (91.8 – 95.5)
Follow-up time* [years]	3.4 (1.5 – 5.8)	3.4 (1.5 – 5.6)	3.4 (1.6 – 6.0)	4.0 (1.8 – 6.1)
Number of deaths [N, %]	863 (92.7%)	504 (93.3%)	359 (92.0%)	200 (93.5%)
Age at censoring [years]	97.1 (94.6 – 99.9)	97.2 (94.8 – 99.9)	97.0 (94.3 – 99.6)	97.9 (95.7 – 100.2)
TSH availability [N, %]	826 (88.7%)	482 (89.3%)	344 (88.0%)	206 (96.3%)
TSH [mU/L]	1.52 (0.98 – 2.42)	1.64 (1.03 – 2.61)	1.38 (0.93 – 2.15)	1.58 (1.04 – 2.43)
FT4 availability [N, %]	840 (90.2)	493 (91.3%)	347 (88.8%)	209 (97.7%)
FT4 [pmol/L]	15.9 (14.4 – 17.5)	15.8 (14.2 – 17.5)	16.1 (14.6 – 17.6)	15.9 (14.3 – 17.3)
FT3 availability [N, %]	840 (90.2%)	492 (91.0%)	384 (89.0%)	208 (97.2%)
FT3 [pmol/L]	4.0 (3.7 – 4.4)	4.0 (3.6 – 4.4)	4.1 (3.8 – 4.5)	4.0 (3.7 – 4.3)

* at February 2014

TABLE 1: BASELINE CHARACTERISTICS OF THE STUDIED STUDY SUBSETS. Baseline characteristics of the whole Leiden Longevity Study (ALL) and the currently defined subsets: FH(+): 239 nonagenarian sibships with a marked family history of an extended survival into old age; FH(-): 176 nonagenarian sibships without such a marked family history; SEQ_FH(+): 214 independent index cases selected from the 239 sibships exhibiting a marked family history (FH(+)) of whom the whole genome has been sequenced previously (**Chapter 4** of this thesis).

In contrast, the FH(+) subset, selected for their genetic propensity of excess survival, displays relationships between thyroid parameters and prospective survival that is governed solely by levels of serum FT3. Together these findings suggest

that family history, as reflected by the FHS, is not so much a marker of prospective survival beyond age 90, but instead it may indicate a distinct mechanism through which the thyroid function may affect life span regulation.

	ALL 415 sibships	FH(+) 239 sibships	FH(-) 176 sibships	SEQ_FH(+) 214 cases
TSH				
Ndeath [%]	767 [92.9%]	451 [93.6%]	316 [91.9%]	194 [94.2%]
HR [95% CI]	0.86 [0.80-1.21]	1.22 [0.94-1.58]	0.75 [0.55-1.01]	1.20 [0.78-1.85]
p	0.85	0.14	0.059	0.40
FT4				
Ndeath [%]	780 [92.9%]	462 [93.7%]	318 [91.6%]	197 [94.3%]
HR [95% CI]	1.04 [1.01-1.07]	1.00 [0.96-1.04]	1.09 [1.04-1.14]	0.99 [0.92-1.07]
p	0.013	0.94	1.4 × 10⁻⁴	0.827
FT3				
Ndeath [%]	781 [93.0%]	462 [93.9%]	319 [91.7%]	197 [94.7%]
HR [95% CI]	0.73 [0.64-0.83]	0.68 [0.56-0.82]	0.80 [0.65-0.97]	0.56 [0.41-0.77]
p	4.25 × 10⁻⁶	5.49 × 10⁻⁵	0.023	2.55 × 10⁻⁴

Table 2: Prospective survival on thyroid parameters. Prospective survival on TSH, fT4 and fT3 in a 10 years follow up performed in the whole Leiden Longevity Study (ALL) and the currently defined subsets: FH(+): 239 nonagenarian sibships with a marked family history of an extended survival into old age; FH(-): 176 nonagenarian sibships without such a marked family history; SEQ_FH(+): 214 independent index cases selected from the 239 sibships exhibiting a marked family history (FH(+)) of whom the whole genome has been sequenced previously (Chapter 4 of this thesis). HR indicates the hazard ratio for mortality.

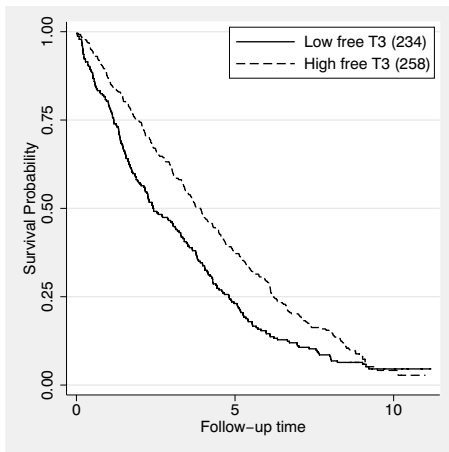


FIGURE 1: PROSPECTIVE SURVIVAL OF fT3 IN FH(+). Kaplan-Meier curves of fT3 in the 239 sibships with a marked family history of an extended survival into old age (FH(+)). A clear mortality risk is observed for subjects with low levels of fT3.

3.3 Linkage analysis identifies 13q34 as a longevity locus

To identify genomic regions harbouring variants predisposing to familial longevity and an attenuated thyroid function, affected sib pair analyses were performed on all nonagenarian sibships (ALL) and the FH(+) and FH(-) study subsets (Experimental Procedures 4). The locus displaying the largest disparity in linkage results between FH(+) and FH(-) in presence of suggestive linkage in the whole study was found at 13q34 (rs752342 at 120 cM; $LOD_{ALL}=2.96$; $LOD_{FH(+)}=3.35$; $LOD_{FH(-)}=0.28$, Figure 2 and Supplemental Figure 1). The 1-LOD-drop interval in the FH(+) subset is 8.36 cM at chr13:110,823,340-113,522,717 (GRCh37/hg19 coordinates) and harbours 16 RefSeq genes: (part of) *COL4A1*, *COL4A2*, *COL4A2-AS1*, *RAB20*, *CARKD*, *CARS2*, *ING1*, *LINC00346*,

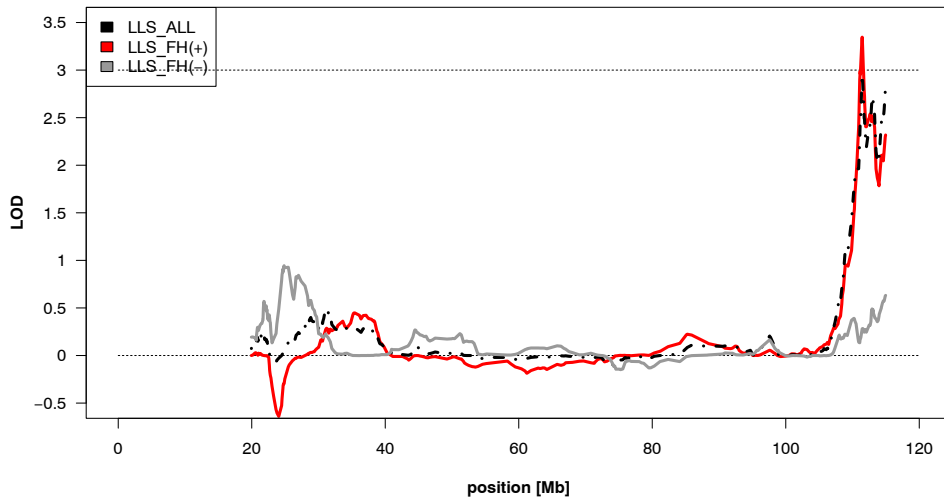


FIGURE 2: LINKAGE RESULTS AT CHR13Q34. Significant linkage for familiar longevity was established at locus 13q34 in the FH(+) subset. Linkage was computed using the 239 sibships with the most profound family history of longevity (FH(+): N=540, $LOD_{max}=3.35$ at rs752342, red), all 415 sibships (ALL: N=931, $LOD_{rs752342}=2.96$, black dotted), and the remaining 176 nonagenarian sibships without a marked family history (FH(-): N=391, $LOD_{rs752342}=0.28$, grey).

ANKRD10, *ARGHEF7*, *TEX29*, *SOX1*, *SPACA7*, *TUBGCP3*, *C13orf35* and (part of) *ATP11A* (Supplemental Figure 2 and Supplemental Table 1). Linkage at chr13q34 is highly specific to the FH(+) subset, which are more hypothyroidal as compared to the FH(-) subset, e.g. higher TSH and lower fT3, and exhibit a thyroid prospective survival profile solely depending on serum fT3. We therefore hypothesize that this locus contributes to a family history of excess survival by harbouring variation attenuating the thyroid function, through constitutively lowering levels of serum fT3.

3.4 Next Generation Sequencing identifies an intergenic variant under the linkage peak marking lowered serum levels of fT3

Thus far we have observed that nonagenarian sibships with a marked family history of excess survival (FH(+))

exhibit a deviating thyroid prospective survival profile, that seems to largely depend on only levels of serum fT3. To investigate whether any variants in the chr13q34 region could explain this deviating thyroid function observed in the FH(+) subset, we obtained the whole genome sequence of 214 unrelated index cases from the FH(+) subset, termed the SEQ_FH(+) subset (Table 1 and 2, Experimental Procedures 1.2). Depending on DNA availability, the sibling that showed the most extended life span was preferably included for sequencing to further enrich the SEQ_FH(+) study subset for longevity variants.

Using this data, we performed two types of association tests with serum fT3 levels. First, we employed a sliding window of 25kb, 50kb, 100kb, and 150kb to bin and jointly associate common and rare variants in 13q34 ($N_{variants}=15,612$)

to fT3 levels using the Sequence Kernel Association Test (SKAT-O²⁵, Experimental Procedures 3.3). Though none of these tests achieved significance ($\alpha \leq 0.05$) after Bonferroni correction, it is noteworthy that highest significance was obtained using the broadest window size (150kb) at a genomic position roughly coinciding with the maximal linkage score (Supplemental Figure 3).

Common Single Nucleotide Variants (SNVs) (MAF $\geq 5\%$, N=5,997) residing in the 13q34 candidate region were tested for association with serum fT3 levels using a standard regression model (Experimental Procedures 3.4). Using this approach, we identified one independent association with serum fT3 levels that passed the Bonferroni correction (rs9515460, T/C, $\beta = -0.52$, 95% CI -0.73--0.30, $p = 3.43 \times 10^{-6}$, Supplemental Figure 4 and 5 and Supplemental Table 2). The identified SNV is in full disequilibrium with two neighbouring SNVs rs80043005 and rs74128254 and is situated between *TEX29* (266 kb) and *SOX1* (459 kb). Thus, within the 214 individuals selected for sequencing (SEQ_FH(+)), taken from sibships with a marked family history of excess survival (FH(+)), carriers of the rs9515460-C allele exhibit lower fT3 (Figure 3), potentially explaining the lower fT3 levels in the FH(+) subset as compared to the FH(-) subset.

3.5 Rs9515460 association with fT3 serum levels confined to siblings that contribute to linkage at chr13q34

Next we questioned whether the association of rs9515460 with serum level of fT3 is confined to sibling pairs contributing to the linkage at chr13q34

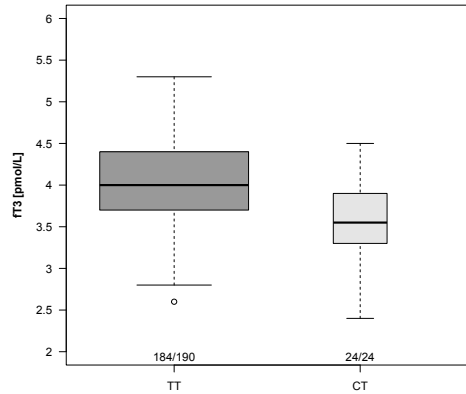


FIGURE 3: Rs9515460: A QTL FOR FT3 ON CHR13Q34. Within the individuals selected for sequencing (SEQ_FH(+)), the 24 rs9515460-CT carriers have a 12% lower serum level of fT3 as compared to the 190 rs9515460-TT carriers.

or is observed in the complete LLS nonagenarian cohort independent of IBD status or FH assignment. For this purpose, we performed Sequenom MassArray genotyping (Experimental Procedures 5) on rs9515460 in all 415 nonagenarian sibships in the LLS study. Sequenced genotypes of rs9515460 in the SEQ_FH(+) cohort were in perfect concordance with those obtained with Sequenom. Whereas evidence for association of fT3 with rs9515460 was validated in the whole FH(+) subset ($\beta = -0.23$, 95% CI -0.40--0.07, $p = 0.006$), such association was not found in the FH(-) subset ($p = 0.32$), nor in the whole LLS ($p = 0.21$). This either implies that the observed association of fT3 serum levels with rs9515460 in the FH(+) subset is a spurious finding, or alternatively, that not rs9515460 itself, but rare variants specific to the FH(+) subset and linkage disequilibrium with rs9515460 are causal for the lowered serum fT3 level.

To investigate whether the association of rs9515460 with fT3 can be explained by rare variants in partial linkage with

rs9515460 we employed the Identical By Descent (IBD) status at chr13q34. Nonagenarian sibships that have inherited identical strands of DNA from their parents (IBD2), while carrying the rs9515460-CT genotype should be enriched for fT3 lowering variants, as compared to sibships that inherited different strands of DNA (IBD0), but by coincidence carry the rs9515460-CT genotype. Thus to investigate whether there is a significant interaction between the rs9515460 SNV and IBD status at chr13q34, we selected those sibling pairs being IBD2 at the marker with the highest linkage signal (rs752342), while carrying the rs9515460-CT genotype. In total, 12 sib pairs fulfilled these criteria, 9 of which were included with 1 sibling in the SEQ_FH(+) subset. When stratified by rs9515460 genotype (CT or TT) and IBD status (IBD0 or IBD2) a significant lower fT3 level was observed in the 12 sibling pairs (IBD2-CT) as compared to the remaining sibling pairs in the whole LLS carrying CT, but not linking (IBD0-CT) or not carrying CT (IBD0-TT and IBD2-TT) ($\beta=-0.41$, 95% CI -0.62 - -0.21 , $p=6.33 \times 10^{-5}$, Figure 4). Since this shows that the fT3 lowering effect of the rs9515460-CT genotype is conditional on IBD status, these results indicate that not the rs9515460 SNV itself, but rare variants in linkage with rs9515460 explain the association with serum levels of fT3.

3.6 Variation at rs9515460 associates with prospective survival in nonagenarians

In population-based studies of elderly above 85 years^{4,26}, low serum fT3 associates with a poor prospect of survival, presumably

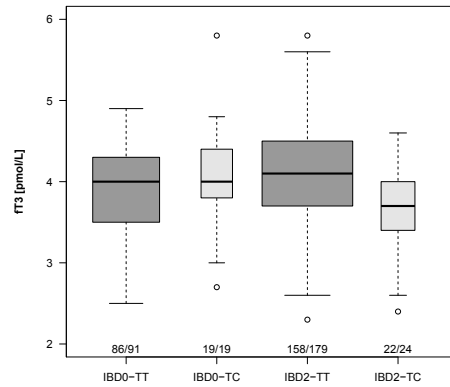


FIGURE 4: IBD × RS9515460 GENOTYPE INTERACTIONS FOR fT3. fT3 serum levels in nonagenarians of all 415 sibships stratified by the most contrasting IBD statuses (IBD0 or IBD2) and rs9515460 genotypes (CT and TT) were plotted for the nonagenarians for whom data on fT3 levels were available, indicated by the ratios at the bottom.

not so much as a causal factor but because it marks frailty among the elderly. Since the rs9515460-CT genotype seems to tag a fT3 lowering phenotype, we investigated whether carriage of the rs9515460-CT allele associates with a poor prospective survival beyond age 90 and again observed interesting interactions with the FH strata (Table 3). Whereas the rs9515460-CT genotype significantly associates with a poor prospective survival in the FH(+) subset (HR=1.47, 95% CI 1.11-1.95, $p=0.008$, Figure 5), it does not for the FH(-) subset ($p=0.98$). This can be explained by the fact that the rs9515460-CT genotype only marks a fT3 lowering haplotype enriched in the FH(+), but not in the FH(-) subset. This would imply that the association of rs9515460-CT with a poor prospective survival is also explained by rare variants in linkage with the rs9515460-CT allele.

To confirm whether the association of rs9515460 with mortality is also dependent on IBD status at chr13q34, we repeated

	ALL 415 sibships	FH(+) 239 sibships	FH(-) 176 sibships	SEQ_FH(+) 214 cases
rs9515460				
Ndeath [%]	848 [92.7%]	493 [93.9%]	355 [92.0%]	199 [93.4%]
HR [95% CI]	1.19 [0.98-1.46]	1.47 [1.11-1.95]	1.00 [0.75-1.34]	1.85 [1.19-2.88]
p	0.084	0.008	0.983	0.006

TABLE 3: PROSPECTIVE SURVIVAL ON RS9515460. Prospective survival on a QTL for fT3, rs9515460, in a 10 years follow up performed in the whole Leiden Longevity Study (ALL) and the currently defined subsets: FH(+): 239 nonagenarian sibships with a marked family history of an extended survival into old age; FH(-): 176 nonagenarian sibships without such a marked family history; SEQ_FH(+): 214 independent index cases selected from the 239 sibships exhibiting a marked family history (FH(+)) of whom the whole genome has been sequenced previously (Chapter 4 of this thesis). HR indicates the hazard ratio for mortality.

the survival analysis in the whole study, while stratifying for the most contrasting IBD statuses (IBD0 or IBD2, as indicated by rs752342). Indeed we observe that the association of rs9515460 and mortality is also confined to the sibling pairs that link on chr13q34 ($N_{\text{tot}}=313$, $N_{\text{death}}=274$, $HR=1.94$, 95% CI 1.23-3.04, $p=0.004$, Figure 6). Hence, both the associations of a lowered serum fT3 level as a poor prospective survival of rs9515460-CT carriers are conditioned on sibling pairs linking on chr13q34. This indicates that the rs9515460-CT genotype partially tags a fT3 lowering haplotype that conveys a poor survival prospect in late life. As hypothesized, chr13q34 appears to contribute to the familial history of excess survival by harbouring variation attenuating the thyroid function thereby influencing lifespan regulation, however, apparently not in a protective sense in the oldest old.

4. Discussion

In this paper we report chr13q34 as a novel locus harbouring genetic variants contributing to lifespan regulation as well as an attenuated thyroid function. Within this locus, we identified the rs9515460_T<C polymorphism as a QTL for serum levels of fT3, the unbound active thyroid hormone. Nonagenarian minor C allele carriers of rs9515460 exhibit a significantly lowered fT3 level, which may have contributed to their survival to the age of 90, as a moderately lowered fT3 level at middle age is assumed to be beneficial for cardio-metabolic health. Since the minor allele rs9515460-C carriers display higher mortality risk above the age of 90, we conclude that, in concordance with literature^{4,26}, low fT3 levels in exceptional old age may negatively influence survival.

Like previous genome wide linkage studies for longevity, we do not observe any overlap with any of the previously reported loci 4q25²⁷, 3p24-22, 9q31-34, 12q24²⁸, 6p12.1, 7q11.21, 14q22.1²⁹ or 14q11.2, 17q12-q22, 19p13.3-p13.11, and

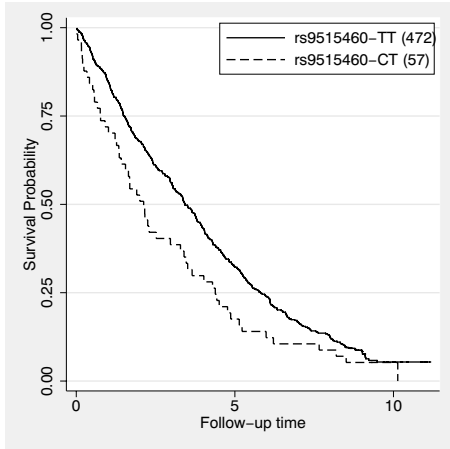


FIGURE 5: RS9515460 AND PROSPECTIVE SURVIVAL. Kaplan-Meier curves were drawn for rs9515460 genotypes in the 239 nonagenarian sibships with a marked family history of an extended survival into old age (FH+): Carriers of the rs9515460-CT genotype have an increased risk on mortality during a 10 years follow up, as compared to the TT carriers.

19q13.11-q13.32³⁰, suggesting that each of the studied populations have their private mechanisms leading to the longevity phenotype. In contrast to previous linkage studies, we have strong indications of the nature of the potential private mechanisms underlying the longevity phenotype in our study cohort. Previously we have shown that long-lived families with the most profound family history of excess survival (FH+) are characterized by an attenuated thyroid function. Together, these findings are a strong indication that chr13q34 harbours variants that constitutively attenuate the thyroid function and thereby promote survival, especially in middle age. Fine-mapping of the chr13q34 region was performed by associating each of the common variants under the linkage peak separately with ft3 and indicated one independent significant signal (rs9515460)

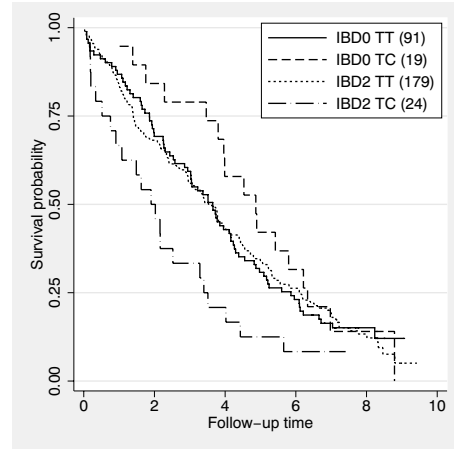


FIGURE 6: IBD × RS9515460 GENOTYPE INTERACTIONS FOR PROSPECTIVE SURVIVAL. Kaplan-Meier curves of rs9515460 genotypes in nonagenarians of all 415 sibships stratified by IBD status. IBD2-TC nonagenarians have an increased risk on mortality during a 10 years follow up, as compared to the TT genotype carriers or IBD0 sibling pairs.

situated between *TEX29* (266 kb) and *SOX1* (459 kb). The fact that this association with ft3 was conditional on IBD status indicated that the rs9515460 was probably not causal itself, but tagged a haplotype carrying the causal variants. Two aspects about the proximal genomic region are noteworthy. First, the rs9515460 is situated in the first intron of a putative protein-coding transcript (RP11-65D24.2) predicted by the GENCODE consortium on basis of a single spliced testis expressed EST. Secondly, nearby GWAS-hits are at 42kb and 82kb and respectively associate with age at menopause³¹ and age at menarche³², which themselves have been attributed to a gene upstream of *TEX29*, namely *ARHGEF7* (305kb). *ARHGEF7* is known to be an activator of *FOXO3a*³³, which is one of the most well studied longevity genes and has been shown to extend lifespan

upon disruption in multiple organisms^{34,35}. Overall, the rs9515460-C allele tags an fT3 lowering haplotype and its proximal genomic region seems to relate to hormone functioning.

As an alternative, we attempted to fine map the linkage region by associating groups of variants jointly with fT3 levels using the Sequence Kernel Association Test (SKAT) in conjunction with a sliding window for variant grouping. Benefits of SKAT lie in its ability to handle mixtures of rare and common variants with opposite signs of association, thus fitting our expectations with respect of the genetic heterogeneity in both human longevity and an attenuated thyroid signalling. The most significant signal ($p=0.0032$) was detected in a window neighbouring the one containing the highest observed linkage signal (~44.8kb) and coincided with the *ING1* gene and the first exons of the *CARS2* gene. However, none of the association tests performed with SKAT remained significant after correction for multiple testing, indicating that we might have insufficient power in the current study for performing joint association analyses of variants.

Since the FH(+) and FH(-) subsets show a very comparable prospective survival, we concluded that the genetic propensity to exhibit excess survival has no significant effect on the survival beyond age 90, which seems to be counterintuitive. Nonagenarian sibships with a genetic propensity to exhibit an excess survival were selected on basis of a so-called Family History Score (FHS), which expresses the mean survival advantage of the parents of a nonagenarian sibship relative to their respective sex and birth cohort specific life

expectancies. Hence, the FHS relates to the probability of the parents of a nonagenarian to outlive their sex and age matched peers derived from the general population, who typically did not live up to age 90, nor their age-equivalent of the oldest old. Hence, the genetic propensity to exhibit an excess survival relates to survival into old age, and therefore not necessarily to survival of old into oldest old, as these two periods in life history are characterized by very distinct disease specific mortality rates.

Nonagenarian sibships with the most profound family history of excess survival are enriched for an fT3 lowering haplotype that is assumed to be beneficial for survival into old age, but is apparently detrimental for survival of old into oldest age. This antagonistic pleiotropic effect of the fT3 lowering haplotype on human lifespan regulation may be explained by considering the thyroid axis in relation to blood pressure and its subsequent effect on cardiovascular mortality. Increased thyroid levels promote an increased heart rate and cardiac output, leading to an increased pulse pressure. Whereas a low systolic blood pressure is beneficial from middle age onward (65 to 84), as indicated by a lower risk on cardiovascular death, it becomes detrimental in the oldest old (age > 84)³⁶. Hence, a haplotype carrying variants constitutively lowering fT3 serum levels is expected to contribute to human longevity by transmitting its beneficial health effects prior age 90. The fact that many of the health parameters, that distinguish members of long-lived families from the general population, display such inverse health correlations, suggests that the hereby-proposed antagonistic pleiotropic

genetic mechanism for longevity might be common amongst long-lived families. Hence, the key to uncover the genetic basis for healthy ageing and human longevity lies in the knowledge on what conditions and life-timing health associated phenotypes actually confer their health benefits.

A limitation of the current study is that we thus far assumed that carriership of fT3 lowering haplotype confers a health benefit at middle age, especially with respect to the cardio-metabolic make up. To test whether this hypothesis holds, also in the general population, we first need to further characterize the fT3 lowering haplotype. Once a clearly defined haplotype or the causal variants on this haplotype have been established, we first will verify whether fT3 levels co-segregate with the presumed casual variants in the offspring of the studied nonagenarians. To verify that the causal variants indeed predispose to a beneficial cardio-metabolic health status at middle age, we will investigate whether carriership of the causal markers can explain the lowered incidence of cardio-vascular morbidity observed in the offspring of the studied nonagenarians⁹. A following step would be to generalize these observations to the general population in multiple cohorts of European decent, to validate that the chr13q34 region harbours genetic predispositions underlying a public mechanism for human longevity marked by low serum fT3 levels.

To conclude, we have performed in depth genetic analyses to disentangle the pleiotropic relation observed in long-lived families between the propensity to exhibit an excess survival and an attenuation of the thyroid function. Using linkage

analyses followed by QTL analyses for fT3 on NGS variants within the 1-LOD-drop interval, we were able to identify an fT3 lowering haplotype that might causally explain the attenuated thyroid function. Unlike previously reported longevity loci, that confer their beneficial effect either from middle age onward, e.g. *APOE-ε2* allele³⁷, or in late life only, e.g. *FOXO3A*³⁸⁻⁴⁰, we hereby report the chr13q34 locus, that exhibits an antagonistic pleiotropic relation with life span regulation. Whereas lowered fT3 levels contribute to cardio-metabolic health from middle into old age, it associates with a poor prospective survival in the oldest old. These findings collectively warrant further investigations into the mechanism how fT3 levels in specific and other ageing markers displaying inverse health associations in general contribute to human longevity and lifespan regulation.

5. Experimental Procedures

5.1 Study population: Leiden Longevity Study

5.1.1 Study Design: Families participating in the Leiden Longevity Study⁴¹ have at least two siblings meeting four inclusion criteria: (i) men are at least 89 years old and women are at least 91 years old, (ii) participants have at least one living brother or sister who fulfils the first criterion and is willing to participate, (iii) the nonagenarian sibship has an identical mother and father, and (iv) the parents of the nonagenarian sibship are Dutch and Caucasian. Using these criteria, a total of 421 nonagenarian sibships (N=944) have been recruited. For 415 sibships (N=931) genome wide SNP genotypes²² for at least two nonagenarian siblings were

available for the genetic linkage analyses.

5.1.2 Subset definitions: A so-called Family History Score (FHS)¹¹ was computed per sibship, which expresses the mean survival advantage of the parents of a nonagenarian sibship relative to their respective sex and birth cohort specific life expectancies. A threshold on the FHS at -1.05 was used to assign sibships to either the FH(+)(FHS≤-1.05, 239 sibships, N=540) or FH(-)(FHS>-1.05, 176 sibships, N=391) subset (Table 1). From the thus created FH(+) subset, 214 independent cases were selected for sequencing using the following criteria: (i) the availability of at least 5 µg of genomic DNA from whole blood for whole genome sequencing, (ii) the participation of children of one of the siblings for future research and (iii) the most extended lifespan compared to his/her siblings. Thus selected participants of the LLS were whole genome sequenced according to procedures fully described in **Chapter 4** of this thesis.

5.2 Thyroid serum parameters

Details regarding the measurement protocols for the thyroid serum parameters TSH, fT4 and fT3 have been described in full detail elsewhere¹⁰. Serum levels of TSH were log10 transformed in order to obtain an approximately normal distribution. Outliers in thyroid serum parameters were defined as observations deviating more than three standard deviations of the mean on basis of measurements performed in the whole population of nonagenarian participants in the Leiden Longevity Study (N=859).

5.3 Statistical analyses

5.3.1 Differences in thyroid parameters

between sub-populations: Differences in serum levels of the thyroid parameters TSH, fT4 and fT3 between sibships with a marked family history and without were tested using linear mixed models implemented in the *lme4*⁴² and *lmerTest*⁴³ packages of the statistical language R⁴⁴:

$$TP \sim \beta_1 \times age + \beta_2 \times sex + \beta_3 \times status + \mu_1 \times famID \quad (1)$$

Where *TP* indicates the level of a thyroid serum parameter, *age at inclusion* is provided in years, *sex* is provided as 1 (male) or 2 (female), *status* indicates the assignment to sibships with (*status* = 1) or without (*status* = 0) a family history of extended survival and *famID* indicates the family membership. Family membership was modelled using a random effect μ_1 to account for the phenotypic correlations observed between family members.

5.3.2 Associations with prospective survival:

Analyses were performed with the *Survival* package⁴⁵ of R⁴⁴ using an age at inclusion and sex-adjusted, left-truncated Cox proportional hazards model to adjust for late entry into the dataset according to age. Mortality analyses between different subsets of the study were performed using:

$$\lambda(t) \sim \lambda_0(t) \times \exp(\beta_1 \times age + \beta_2 \times sex + \beta_3 \times sel + \mu_1 \times famID) \quad (2)$$

Where the covariates *age* designates *age at inclusion* and is provided in years, *sex* as either 1 (male) or 2 (female), *sel* as either 1 or 2 to indicate study subset membership and *famID* indicates the family membership. Again, family membership was taken into account to correct for phenotypic correlations observed between family members and was done by supplying the term *cluster(famID)* in the Cox regression. Similarly, mortality analyses on the thyroid parameters were performed using:

$$\lambda(t) \sim \lambda_0(t) \times \exp(\beta_1 \times age + \beta_2 \times sex + \beta_3 \times TP + \mu_1 \times famID) \quad (3)$$

Where *TP* indicates the serum levels of TSH, fT4 or fT3. Finally, mortality analyses on the rs9515460 SNV were performed using:

$$\lambda(t) \sim \lambda_0(t) \times \exp(\beta_1 \times age + \beta_2 \times sex + \beta_3 \times SNV + \mu_1 \times famID) \quad (4)$$

Where *SNV* indicates the C allele dosage (0,1 or 2) of the rs9515460_T>C polymorphism.

5.3.3 Sequence Kernel Association

Test (SKAT-O): To jointly associate groups of genotype markers with serum levels of fT3, we employed the package SKAT⁴⁶ of the statistical language R⁴⁴ at default settings and used the following formula to describe the null hypothesis (containing covariates only):

$$fT3 \sim \beta_1 \times age + \beta_2 \times sex \quad (5)$$

5.3.4 QTL associations: Associations with single genotypic genotypes was done for common variants within the 1-LOD-drop linkage area (MAF \geq 5%, N = 5,480), using the *lm* function of the R package *stats*⁴⁴, using the following model:

$$fT3 \sim \beta_1 \times age + \beta_2 \times sex + \beta_3 \times SNV \quad (6)$$

Where the covariates *age* is provided in years and *sex* as 1 (male) or 2 (female). Genotypes were recoded to minor allele dosages: 0 (homozygous common allele), 1 (heterozygous) or 2 (homozygous rare allele). Genotype data was filtered on missingness (\leq 5%) and MAF (\geq 5%) with respect to the 208 out of 214 samples for which also data on fT3 levels was available. Obtained *p*-values were corrected for multiple testing (Bonferroni).

5.4 Linkage analysis

The Illumina 660Quad and Illumina OmniExpress arrays have been used for genotyping the participants of the Leiden Longevity Study. Details on data acquisition and pre-processing are have been described elsewhere²¹. For linkage analysis 12,000 equally spaced SNVs were selected that are genotyped on both arrays with a MAF>0.3 and a mutual $R^2 < 0.4$. MERLIN-0.10.2⁴⁷ was used to estimated the information content per genome and to estimate IBD probabilities in sibling pairs without parents. Nonparametric affected

sibling pair analyses were performed using a score test statistic for affected sibling pairs⁴⁸.

5.5 Sequenom MassArray

Genotyping

Genotyping of rs9515460 in the LLS study was performed using the Sequenom MassARRAY iPLEX Gold. Genotypes were successfully measured for 909 out of the 925 participants (98.3%). Sequenom genotypes in the 214 individuals with whole genome sequencing data were in perfect concordance with sequenced genotypes of rs9515460.

6. Acknowledgements

The research leading to these results has received funding from the Medical Delta (COMO), Pfizer Inc, and the European Union's Seventh Framework Programme (FP7/2007-2011) under grant agreement number 259679. This study was financially supported by the Innovation-Oriented Research Program on Genomics (SenterNovem IGE05007), the Centre for Medical Systems Biology and the Netherlands Consortium for Healthy Ageing (grant 050-060-810), all in the framework of the Netherlands Genomics Initiative, Netherlands Organization for Scientific Research (NWO), by Unilever Colworth and by BBMRI-NL, a Research Infrastructure financed by the Dutch government (NWO 184.021.007).

7. References

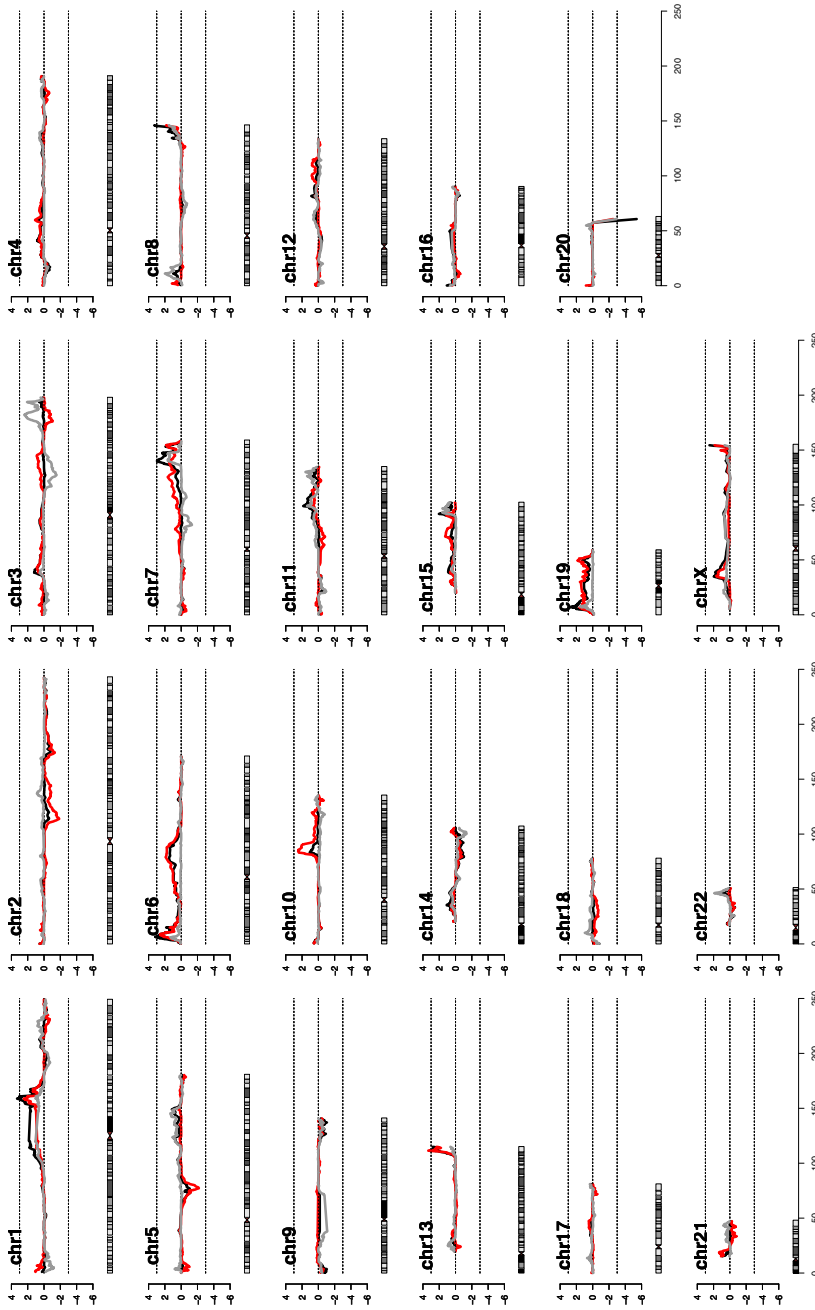
1. Ooka, H., Fujita, S. & Yoshimoto, E. Pituitary-thyroid activity and longevity in neonatally thyroxine-treated rats. *Mech Ageing Dev* **22**, 113-20 (1983).
2. Tatar, M., Bartke, A. & Antebi, A. The endocrine regulation of aging by insulin-like signals. *Science* **299**, 1346-51 (2003).
3. Brown-Borg, H.M., Borg, K.E., Meliska, C.J. & Bartke, A. Dwarf mice and the ageing process. *Nature* **384**, 33 (1996).

4. Gussekloo, J. *et al.* Thyroid status, disability and cognitive function, and survival in old age. *JAMA* **292**, 2591-9 (2004).
5. Atzmon, G., Barzilai, N., Hollowell, J.G., Surks, M.I. & Gabriely, I. Extreme longevity is associated with increased serum thyrotropin. *J Clin Endocrinol Metab* **94**, 1251-4 (2009).
6. Selmer, C. *et al.* Subclinical and Overt Thyroid Dysfunction and Risk of All-cause Mortality and Cardiovascular Events: A Large Population Study. *J Clin Endocrinol Metab*, jc20134184 (2014).
7. Atzmon, G. *et al.* Clinical phenotype of families with longevity. *J Am Geriatr Soc* **52**, 274-7 (2004).
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. 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).
10. 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).
11. Rozing, M.P. *et al.* Familial longevity is associated with decreased thyroid function. *J Clin Endocrinol Metab* **95**, 4979-84 (2010).
12. McIver, B. & Gorman, C.A. Euthyroid sick syndrome: an overview. *Thyroid* **7**, 125-32 (1997).
13. Parle, J.V., Maisonneuve, P., Sheppard, M.C., Boyle, P. & Franklyn, J.A. Prediction of all-cause and cardiovascular mortality in elderly people from one low serum thyrotropin result: a 10-year cohort study. *Lancet* **358**, 861-5 (2001).
14. Samollow, P.B. *et al.* Genetic and environmental influences on thyroid hormone variation in Mexican Americans. *J Clin Endocrinol Metab* **89**, 3276-84 (2004).
15. Porcu, E. *et al.* A meta-analysis of thyroid-related traits reveals novel loci and gender-specific differences in the regulation of thyroid function. *PLoS Genet* **9**, e1003266 (2013).
16. Rawal, R. *et al.* Meta-analysis of two genome-wide association studies identifies four genetic loci associated with thyroid function. *Hum Mol Genet* **21**, 3275-82 (2012).
17. Panicker, V. *et al.* A locus on chromosome 1p36 is associated with thyrotropin and thyroid function as identified by genome-wide association study. *Am J Hum Genet* **87**, 430-5 (2010).
18. Arnaud-Lopez, L. *et al.* Phosphodiesterase 8B gene variants are associated with serum TSH levels and thyroid function. *Am J Hum Genet* **82**, 1270-80 (2008).
19. Hwang, S.J., Yang, Q., Meigs, J.B., Pearce, E.N. & Fox, C.S. A genome-wide association for kidney function and endocrine-related traits in the NHLBI's Framingham Heart Study. *BMC Med Genet* **8 Suppl 1**, S10 (2007).
20. Atzmon, G., Barzilai, N., Surks, M.I. & Gabriely, I. Genetic predisposition to elevated serum thyrotropin is associated with exceptional longevity. *J Clin Endocrinol Metab* **94**, 4768-75 (2009).
21. Deelen, J. *et al.* Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. *Hum Mol Genet* (2014).
22. 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).
23. 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).

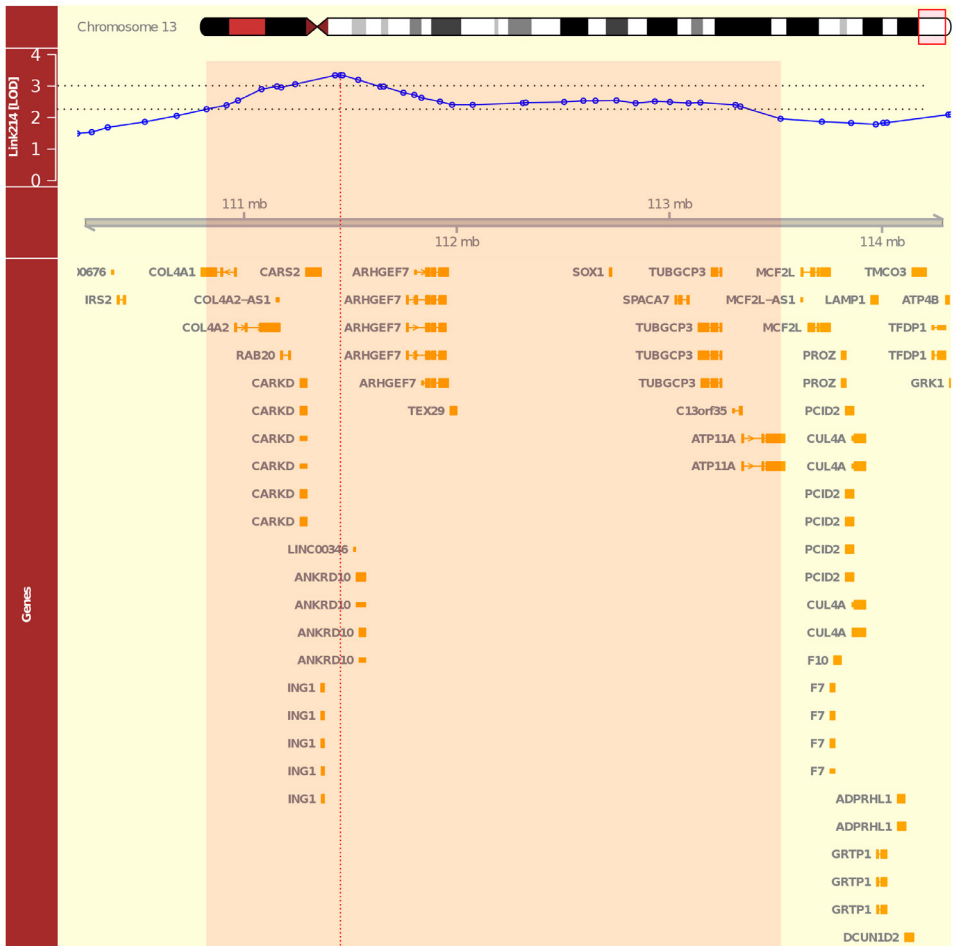
24. Sebastiani, P. *et al.* Genetic signatures of exceptional longevity in humans. *PLoS One* **7**, e29848 (2012).
25. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* **91**, 224-37 (2012).
26. Martin-Ruiz, C. *et al.* Assessment of a large panel of candidate biomarkers of ageing in the Newcastle 85+ study. *Mech Ageing Dev* **132**, 496-502 (2011).
27. 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).
28. Boyden, S.E. & Kunkel, L.M. High-density genomewide linkage analysis of exceptional human longevity identifies multiple novel loci. *PLoS One* **5**, e12432 (2010).
29. 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).
30. Beekman, M. *et al.* Genome-wide linkage analysis for human longevity: Genetics of Healthy Aging Study. *Ageing Cell* **12**, 184-93 (2013).
31. Stolk, L. *et al.* Meta-analyses identify 13 loci associated with age at menopause and highlight DNA repair and immune pathways. *Nat Genet* **44**, 260-8 (2012).
32. Elks, C.E. *et al.* Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies. *Nat Genet* **42**, 1077-85 (2010).
33. Chahdi, A. & Sorokin, A. Endothelin-1 couples betaPix to p66Shc: role of betaPix in cell proliferation through FOXO3a phosphorylation and p27kip1 down-regulation independently of Akt. *Mol Biol Cell* **19**, 2609-19 (2008).
34. 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).
35. Hwangbo, D.S., Gershman, B., Tu, M.P., Palmer, M. & Tatar, M. *Drosophila* dFOXO controls lifespan and regulates insulin signalling in brain and fat body. *Nature* **429**, 562-6 (2004).
36. Satish, S., Freeman, D.H., Jr., Ray, L. & Goodwin, J.S. The relationship between blood pressure and mortality in the oldest old. *J Am Geriatr Soc* **49**, 367-74 (2001).
37. Schupf, N. *et al.* Apolipoprotein E and familial longevity. *Neurobiol Aging* **34**, 1287-91 (2013).
38. Willcox, B.J. *et al.* FOXO3A genotype is strongly associated with human longevity. *Proc Natl Acad Sci U S A* **105**, 13987-92 (2008).
39. 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).
40. Anselmi, C.V. *et al.* Association of the FOXO3A locus with extreme longevity in a southern Italian centenarian study. *Rejuvenation Res* **12**, 95-104 (2009).
41. 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).
42. Bates, D., Maechler, M., Bolker, B. & Walker, S. lme4: Linear mixed-effects models using Eigen and S4. (R package version 1.1-6; <http://CRAN.R-project.org/package=lme4>, 2014).
43. Kuznetsova, A., Brockhoff, P.B. & Christensen, R.H.B. Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). (R package version 2.0-6; <http://CRAN.R-project.org/package=lmerTest>, 2014).
44. R-Core-Team. R: A Language and Environment for Statistical Computing. (R Foundation for Statistical Computing; <http://www.R-project.org/>, 2013).
45. Therneau, T. A Package for Survival Analysis in S. (R package version 2.37-7; <http://CRAN.R-project.org/package=survival>, 2014).
46. Lee, S., Miropolsky, L. & Wu, M. SKAT: SNP-set (Sequence) Kernel Association Test. (R package version 0.95; <http://>

- CRAN.R-project.org/package=SKAT, 2014).
47. Abecasis, G.R., Cherny, S.S., Cookson, W.O. & Cardon, L.R. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**, 97-101 (2002).
48. Houwing-Duistermaat, J.J., Uh, H.W. & van Houwelingen, H.C. A new score statistic to test for association given linkage in affected sibling pair-control designs. *BMC Proc* **1 Suppl 1**, S39 (2007).

Supplemental Materials



SUPPLEMENTAL FIGURE 1: GENOME-WIDE LINKAGE RESULTS ON THE SUBSETS OF THE STUDY. Genome wide linkage results for familial longevity using all 415 nonagenarian sibships of the LLS for which genetic data was available (N=931, black), the subset of sibships with a marked family history of an extended survival into old age (FH(+), red) and the remaining 176 sibships without such a marked family history (FH(-), grey). On chrom13q34, linkage highly specific for sibships with a profound family history is observed (rs752342, $LOD_{ALL}=2.96, LOD_{FH(+)}=3.35, LOD_{FH(-)}=0.28$).



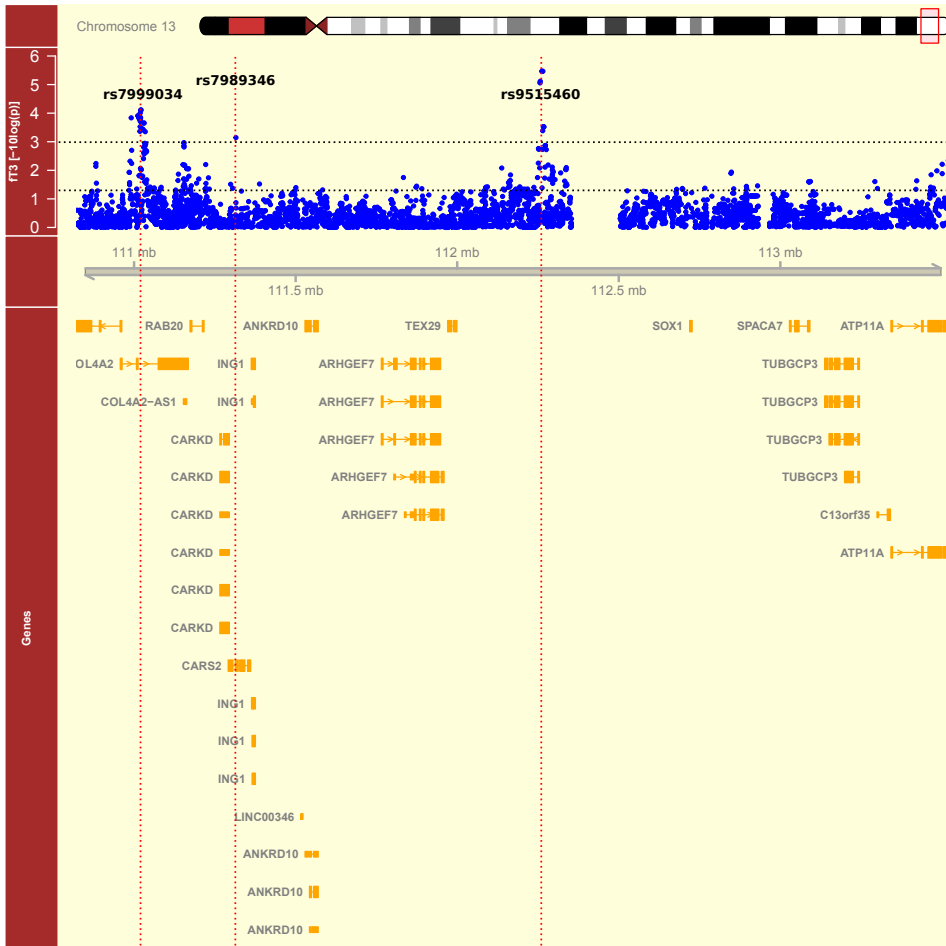
SUPPLEMENTAL FIGURE 2: AN OVERVIEW OF THE 1-LOD-DROP INTERVAL ON CHR13Q34. The red area indicates genes within the 1-LOD-drop interval around the top marker on chr13q34, rs752342, as determined in the FH(+) subset.

GeneSymbol	EntrezGeneID	TxAcc	chrom	strand	start	end
COL4A1	1282	NM_001845	chr13	-	110801310	110959496
COL4A2	1284	NM_001846	chr13	+	110959631	111165373
COL4A2-AS1	100874203	NR_046583	chr13	-	111154923	111160526
RAB20	55647	NM_017817	chr13	-	111175413	111214071
CARKD	55739	NM_001242882	chr13	+	111267807	111292342
CARKD	55739	NM_001242883	chr13	+	111267807	111292342
CARKD	55739	NR_040103	chr13	+	111267807	111292342
CARKD	55739	NR_040104	chr13	+	111267807	111292342
CARKD	55739	NM_001242881	chr13	+	111267931	111292342
CARKD	55739	NM_018210	chr13	+	111267931	111292342
CARS2	79587	NM_024537	chr13	-	111293757	111358480
ING1	3621	NM_198217	chr13	+	111364970	111373421
ING1	3621	NM_198218	chr13	+	111365610	111373421
ING1	3621	NM_198219	chr13	+	111365610	111373421
ING1	3621	NM_005537	chr13	+	111367359	111373421
ING1	3621	NM_001267728	chr13	+	111367784	111373421
LINC00346	283487	NR_027701	chr13	-	111516334	111522655
ANKRD10	55608	NM_017664	chr13	-	111530887	111567454
ANKRD10	55608	NR_104587	chr13	-	111530887	111567454
ANKRD10	55608	NM_001286721	chr13	-	111545039	111567454
ANKRD10	55608	NR_104586	chr13	-	111545039	111567454
ARHGEF7	8874	NM_001113511	chr13	+	111767624	111947542
ARHGEF7	8874	NM_001113512	chr13	+	111767624	111947542
ARHGEF7	8874	NM_145735	chr13	+	111767624	111947542
ARHGEF7	8874	NM_003899	chr13	+	111806061	111958081
ARHGEF7	8874	NM_001113513	chr13	+	111839173	111958081
TEX29	121793	NM_152324	chr13	+	111973015	111996594
SOX1	6656	NM_005986	chr13	+	112721913	112726020
SPACA7	122258	NM_145248	chr13	+	113030651	113089009
TUBGCP3	10426	NM_001286277	chr13	-	113139319	113242499
TUBGCP3	10426	NM_006322	chr13	-	113139319	113242499
TUBGCP3	10426	NM_001286278	chr13	-	113153121	113242499
TUBGCP3	10426	NM_001286279	chr13	-	113200796	113242499
C13orf35	400165	NM_207440	chr13	+	113301358	113338811
ATP11A	23250	NM_015205.2	chr13	+	113344643	113541482
ATP11A	23250	NM_032189.3	Chr13	+	113344643	113541482

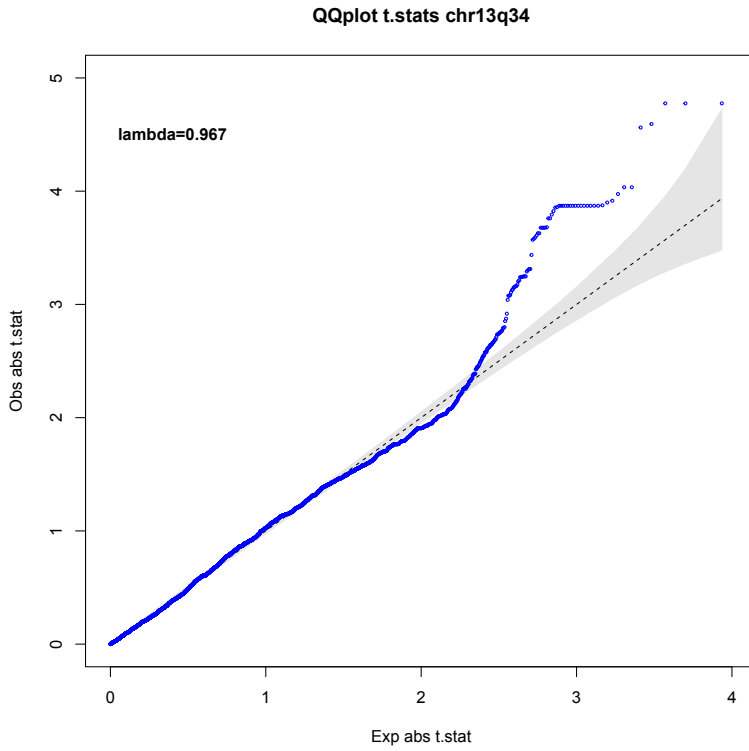
SUPPLEMENTAL TABLE 1: ACCESSION NUMBERS OF GENES WITHIN THE 1-LOD-DROP INTERVAL. Entrez Gene IDs and associated RefSeq transcript accessions are provided for genes situated in the 1-LOD-drop interval on chr13q34. The 1-LOD-drop interval was determined within the FH(+) subset.



SUPPLEMENTAL FIGURE 3: SEQUENCE KERNEL ASSOCIATION TEST (SKAT) IN THE 1-LOD-DROP INTERVAL ON CHR13Q34. Using the sequencing data in 214 nonagenarians, analyses were performed with SKAT for associating groups of SNVs under the linkage peak with FT3. Four different window sizes (25kb, 50kb, 100kb and 150kb) were used for grouping neighbouring variants according to a half overlapping tile pattern. Grouped variants were then submitted to a joint association analysis with FT3 adjusted for age and sex. Obtained results are indicated by the four tracks and display high consistency with respect to genomic position and significance. The highest significance after correction for multiple testing (Bonferoni) was observed using the 150kb windows ($p=0.00319$, 37 tests) for the variants positioned at chr13:111,348,340-111,498,339.



SUPPLEMENTAL FIGURE 4: ASSOCIATION ANALYSIS WITH COMMON VARIANTS IN THE 1-LOD-DROP INTERVAL ON CHR13Q34. Data on fT3 serum levels was available for 208 out of the 214 sequenced nonagenarian genomes. Variants were filtered on the minor allele frequency ($MAF \geq 5\%$) and call rate ($CR \leq 5\%$) in these 208 genomes and associated with fT3 serum levels using a sex and age adjusted linear regression. A total of 5,997 variants within the 1-LOD-drop region, as determined in the FH(+) subset were tested.



5

SUPPLEMENTAL FIGURE 5: QQPLOT OF ABSOLUTE T.STATISTICS OBTAINED WITH SINGLE MARKER ASSOCIATIONS WITH FT3 ON VARIANTS WITHIN THE 1-LOD-DROP REGION ON CHR13Q34.

chrom	position	allele	beta	se	tstat	pval	df
chr13	110991189	G/A	-0.251	0.065	-3.871	1.46E-04	204
chr13	111011400	A/T	-0.282	0.072	-3.915	1.23E-04	204
chr13	111015628	C/T	-0.198	0.051	-3.871	1.46E-04	204
chr13	111015780	C/A	-0.198	0.051	-3.871	1.46E-04	204
chr13	111015877	C/T	-0.198	0.051	-3.871	1.46E-04	204
chr13	111016124	A/C	-0.198	0.051	-3.855	1.56E-04	202
chr13	111016153	C/T	-0.199	0.051	-3.875	1.44E-04	203
chr13	111017045	C/T	-0.198	0.051	-3.871	1.46E-04	204
chr13	111017784	G/A	-0.201	0.052	-3.899	1.31E-04	204
chr13	111018009	G/C	-0.186	0.052	-3.580	4.33E-04	198
chr13	111018072	G/A	-0.198	0.051	-3.871	1.46E-04	204
chr13	111018132	T/C	-0.198	0.051	-3.871	1.46E-04	204
chr13	111018163	T/G	-0.198	0.051	-3.871	1.46E-04	204
chr13	111018235	C/A	-0.198	0.051	-3.871	1.46E-04	204
chr13	111018666	G/A	-0.199	0.051	-3.871	1.46E-04	202
chr13	111018729	C/G	-0.199	0.052	-3.823	1.76E-04	200
chr13	111018752	A/G	-0.198	0.051	-3.859	1.53E-04	203
chr13	111018909	C/G	-0.190	0.051	-3.681	2.98E-04	203
chr13	111019083	T/C	-0.198	0.051	-3.871	1.46E-04	204
chr13	111019278	C/A	-0.204	0.051	-3.975	9.77E-05	204
chr13	111019472	T/C	-0.198	0.051	-3.871	1.46E-04	204
chr13	111019508	C/T	-0.196	0.052	-3.793	1.96E-04	202
chr13	111019568	A/G	-0.198	0.051	-3.871	1.46E-04	204
chr13	111019895	T/C	-0.198	0.051	-3.871	1.46E-04	204
chr13	111019976	C/A	-0.198	0.051	-3.871	1.46E-04	204
chr13	111020878	G/A	-0.204	0.051	-4.034	7.73E-05	204
chr13	111021623	C/G	-0.204	0.051	-4.034	7.73E-05	204
chr13	111025118	A/G	-0.191	0.053	-3.628	3.61E-04	204
chr13	111026734	G/A	-0.191	0.053	-3.628	3.61E-04	204
chr13	111028978	G/A	-0.188	0.052	-3.609	3.86E-04	204
chr13	111029923	G/A	-0.194	0.052	-3.760	2.22E-04	204
chr13	111031180	G/A	-0.194	0.052	-3.760	2.22E-04	204
chr13	111034542	T/C	-0.180	0.050	-3.571	4.45E-04	202
chr13	111315249	G/A	0.178	0.052	3.437	7.17E-04	199
chr13	112256430	C/G	-0.516	0.113	-4.562	8.76E-06	203
chr13	112256996	G/A	-0.509	0.111	-4.593	7.64E-06	204

chr13	112261984	G/A	-0.517	0.108	-4.775	3.43E-06	204
chr13	112263091	C/T	-0.517	0.108	-4.775	3.43E-06	204
chr13	112264614	C/T	-0.338	0.094	-3.592	4.12E-04	203
chr13	112265213	G/A	-0.517	0.108	-4.775	3.43E-06	204
chr13	112266639	G/A	-0.320	0.087	-3.677	3.02E-04	204
chr13	112267153	A/C	-0.320	0.087	-3.677	3.02E-04	204
chr13	112267402	G/A	-0.320	0.087	-3.677	3.02E-04	204
chr13	112268411	C/T	-0.320	0.087	-3.677	3.02E-04	204

SUPPLEMENTAL TABLE 2: SNVs IN THE 1-LOD-DROP LINKAGE REGION. The 1-LOD-drop region was determined in the FH(+) subset, on chr13q34 with $p \leq 0.001$ in the sex and age adjusted regression with fT3 serum levels.

Chapter 6:

SATORi: An R package for generic access and handling of genomic data

Erik B. van den Akker^{1,2}, Marian Beekman^{2,3}, Joris Deelen^{2,3}, P. Eline Slagboom^{2,3} and Marcel J.T. Reinders¹

1. The Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands
2. Molecular Epidemiology, Leiden University Medical Center, Leiden, The Netherlands
3. Netherlands Consortium of Healthy Ageing

In preparation

1. Summary

SATORI is an R package offering standardized access to various types of big genomic datasets, enabling a rapid exploration, integration and mapping between different data types as well as to external genomic annotations. The package, vignette along with the example datasets used in this paper can be obtained from: bioinformatics.tudelft.nl/users/erik-van-den-akker

2. Introduction

A joint interpretation of life-science data is required for grasping the etiology of complex traits, however, this is challenging as the data types and quantities are ever increasing. The statistical platform R¹ provides some excellent tools for mapping and complex modeling of genomic data², and additionally offers a potent interface to several database engines³. However it lacks a uniform database design across all data sources, which would greatly ease the data integration necessary to solve complex life science problems.

Here we present an R package, called SATORi (Standardized Access To Omics in R), for accessing various big omics data types from R in a standardized way. SATORi organizes data by genomic location facilitating an easy annotation to genomic features, like genes, pathways or to other SATORi databases. Moreover, due to the standardized database design, SATORi provides generic ways for accessing genome-wide data enabling a rapid exploration of omics data, while keeping source code clear and understandable.

As a (running) example we applied SATORi to analyze methylation⁴ and SNP data⁵ assayed on two HapMap populations consisting of 30 trios of Caucasian (CEU) and Yoruban (YRI) origin. We show how easy data can be accessed and integrated with SATORi, while still having the power of the function set in R.

3. Implementation

The SATORi package employs the R plugin RSQLite to organize and access data by genomic location or measured entity ID, and subsequently, represents the associated data in a standardized and suitable format for performing analyses in R. SATORi introduces **GVARdb** objects to represent omics sources to R and different types of data sources are stored in specifically inherited subclasses of the **GVARdb** object. SATORi currently supports genotype data (**GWASdb**), methylation data (**METHdb**), and imputed genotype data (**impGWASdb**). Purpose-built constructors are defined for parsing raw data files of each supported data type to build the tables constituting the SATORi database (Figure 1A).

3.1 Accessing data stored in SATORi databases

Data in GVARdb objects can be accessed using the `getGVAR` function, which allows for intuitive queries mimicking normal matrix manipulation in R. For instance, entity IDs and sample IDs can be used to specify the composition of the output matrices.

*As an illustration, methylation data is stored as methylated and unmethylated signals per CpG site per sample (matrices *M* and *U* in Fig. 1B, respectively). A query using a specific CpG entity returns a vector containing the methylated and unmethylated signals for that CpG site for all samples:*

```
1> res_vc <- getGVAR(methdb,
  gvarID="cg00000292")
```

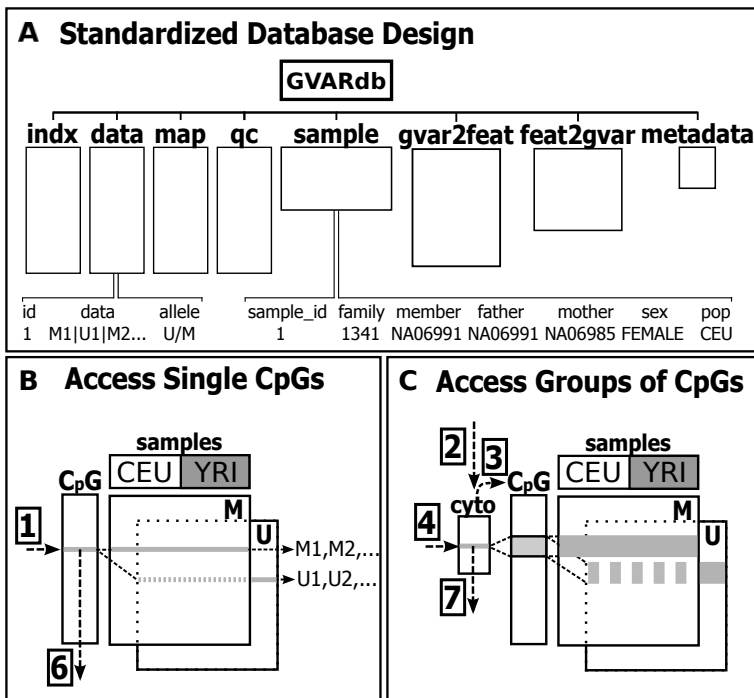


FIGURE 1: DESIGN AND FUNCTIONALITY OF SATORI. [A] An overview of the SQL tables composing a SATORI **GVARdb** object, including tables for storing data (data), mappings to the genome (map), mappings to genomic features (gvar2feat) and sample information (sample). Data points are collapsed per measured entity into single strings across all samples allowing a standardized storage and handling across data types (left). The sample table can be used to store for instance covariates or the familial relationships (right). [B] A database storing (for example) methylation data can be thought of as two big matrices containing the methylated (M) and unmethylated (U) signals for all measured CpGs for all samples. Using, for example, code snippet 1, one can access the data of a particular CpG, while with code snippet 6, one can find differentially methylated CpG sites. [C] Genomic features obtained from, for example, UCSC (cyto) are mapped to the database, after which data can be accessed by feature ID. See code snippets 2,3,4 on how to do this using SATORI.

Several wrappers for `getGVAR` have been defined to query data associated to particular (pre-defined) genomic features. For that, the SATORI database needs to be made aware of an entity-to-feature mapping (Fig. 1B), which can be created and stored using the `mapGVAR2FEAT` function.

For instance, suppose we want to access methylation data per cytoband. The genomic locations of cytobands can be downloaded

from UCSC to R and mapped to the **GVARdb** object using:

```
2> cyto <- getFEATfromUCSC
("cytoBand", "hg18")
3> mapGVAR2FEAT(methdb, cyto)
```

From now on, data can be queried using a specific cytoband designation, which would return a matrix containing the methylated and unmethylated signals for the CpG sites situated within the requested genomic feature for all samples (Fig. 1C):

```
4> res_mat <- getGVARbyFeat
(methdb,"chr1q21.3")
```

Conceptually, there is no difference between mapping to annotations downloaded from UCSC, to annotation databases in R storing transcript locations (Carlson, et al., 2013), or to other SATORI databases.

*For example, to map SNP data stored in a SATORI's **GWASdb** object to our methylation data within a 10kb region around the CpGs, also use:*

```
5> mapGVAR2FEAT
(methdb,gwasdb,flanking=10000)
```

From now on, data of SNPs positioned within 10kb from a CpG site can be queried with the same CpG IDs used to query the methylation data.

3.2 Operations on SATORI databases

Custom R functions can be applied to **GVARdb** objects using `dbGVARapply`, which applies a user-defined R function to each entry of the **GVARdb** object. This is done by sequentially loading and applying the function to chunks of the database (enabling parallel execution or handling big data sets).

*As an illustration, differentially methylated CpG sites (DM-CpGs) between the two HapMap populations can be found by applying a Wilcoxon signed-rank test to the **METHdb** object (Fig. 1B)*

```
6> res_wilcox <-
dbGVARapply(methdb,FUN=wilcox)
```

SATORI also allows applying functions to groups of entries using `dbFEATapply`. This function makes use of a previously stored mapping to iteratively load and analyze grouped data associated to genomic features.

For example, to investigate whether the previously identified DM-CpGs are confined to certain cytobands, we re-compute the number of DM-CpGs per region given a p-value threshold and report these as percentages per cytoband (Fig 1C):

```
7> res_p <- dbFEATapply
(methdb,FUN=perc,p=1e-5)
```

4 Application

Previous code illustrated the ease with which methylation and SNP data could be integrated using the SATORI package. To illustrate the seamless integration in R we investigate the relationship between CpGs and their neighboring SNPs within the CEU and YRI population by calculating CpG-SNP correlations.

*To perform this computation, we define a function that queries the data from a **METHdb** and a **GWASdb** given a CpG ID, and computes CpG-SNP pair correlations using all shared samples:*

```
# Find shared samples:
8> sampID <- intersect(colnames(methdb),
colnames(gwasdb))
# Get SNP data, for instance cpGID =
"cg00000292":
9> snp <- getGVARbyFeat
(gwasdb,cpGID,sampID)[[1]]
```

```
# Get the beta values (M/
(M+U+100)) from the methylated
data. This indicated to getGVAR
by the "as.B=T" argument
10> MR <- getGVAR
(methdb,cpgID,sampID,as.B=T)[1,]
# Compute (with pairwise.
complete observations)
11> cor(t(snp),MR,use="p")
```

Genome plots of chr21q22.11 (Figure 2) show that DM-CpGs coincide with significant CpG-SNP correlations in YRI, suggesting a potential crosstalk between the two data types on this location.

6

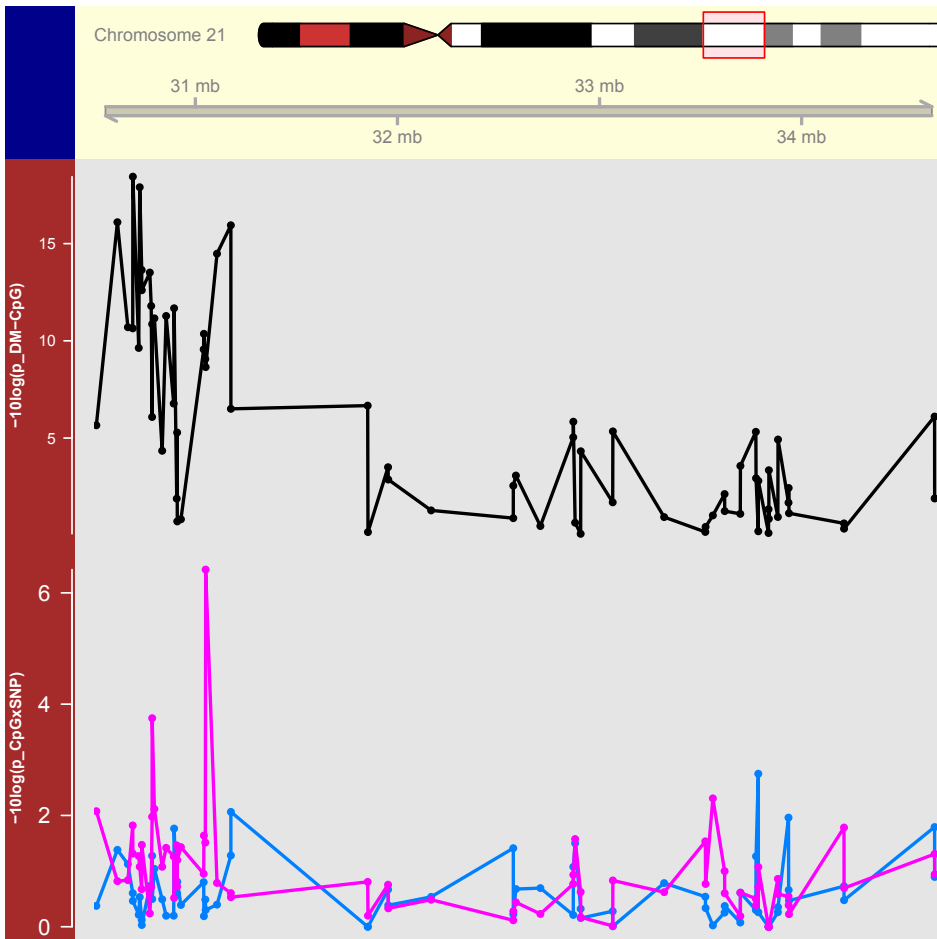


FIGURE 2: A VISUALIZATION OF THE RESULTS COMPUTED WITH AID OF SATORI ON CHR21Q22.11. In black, the significance ($-10\log(p)$) and relative positioning of DM-CpGs between the YRI and CEU populations are depicted. Below in pink (YRI) and blue (CEU) correlations between CpG levels and genotypes in cis are displayed, where each point represents the most significant correlation ($-10\log(p)$) with a SNP within 10kb. Detailed code for performing the analyses and drawing genome plots is provided in the supplemental materials.

5. Conclusions

With the examples in this paper we illustrated the merits of a generic access to various big data sources, which greatly simplifies integrative analyses of life science data. With SATORI the user can quickly explore data and test novel hypotheses by mapping the data to external annotations sources or applying user-defined functions to the data. Hence SATORI is a useful tool in the integrative analysis of omics datasets.

6. Acknowledgements

Funding: The research leading to these results has received funding from the Medical Delta (COMO) and the European Union's Seventh Framework Programme (FP7/2007-2011) IDEAL-ageing under grant agreement n° 259679. This study was supported by a grant from the Innovation-Oriented Research Program on Genomics (SenterNovem IGE05007), the Centre

for Medical Systems Biology, the Netherlands Consortium for Healthy Ageing (Grant 050-060-810), all in the framework of the Netherlands Genomics Initiative, Netherlands Organization for Scientific Research (NWO) and by Unilever Colworth.

7. References

1. R-Core-Team. R: A Language and Environment for Statistical Computing. (2013).
2. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9**, e1003118 (2013).
3. D. A. James & Falcon, S. RSQLite: SQLite interface for R. (2013).
4. Fraser, H.B., Lam, L.L., Neumann, S.M. & Kobor, M.S. Population-specificity of human DNA methylation. *Genome Biol* **13**, R8 (2012).
5. International HapMap, C. The International HapMap Project. *Nature* **426**, 789-96 (2003).

Chapter 7:

General Discussion

1. Main Aims

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 human aging. Human aging, its relation to health, and its effect for life span regulation is largely studied through biomarker or genetic research, but both are greatly hampered by the extreme complexity and heterogeneity of the studied trait. Development and application of novel methodology better capable of handling this complexity and heterogeneity is required for making any real advances in our understanding of human aging. Throughout this thesis we set out to devise novel strategies for data analysis, while adopting two concepts for data integration that are likely to improve both the robustness and interpretation of the obtained results: the joint analysis of omics data and incorporation of prior knowledge. In this chapter we will review the benefits of adopting novel methodology incorporating concepts of data integration in biomarker as well as genetic research for the advances in our understanding of human aging.

2. Main Findings

The first objective of this thesis was to develop methodology for the comprehensive and robust extraction of molecular biomarker profiles based on whole transcriptome expression data. In **Chapter 2**, we investigated the use of Protein-Protein Interaction (PPI) data as

a source of prior knowledge for grouping gene-expression data into comprehensive modules of functionally related genes and their potential to jointly serve as robust biomarkers. For this purpose, an algorithm for the detection of co-expressed PPI modules was developed and we show that its application yields highly reproducible modules of genes over six supposedly heterogeneous studies assayed on breast cancer outcome¹⁻⁶. Though cross-study prediction performances were on average slightly lower as compared to traditional methods for prospective signature construction, the gene composition of the identified modules was in broad agreement with the evidence for the underlying etiology of breast cancer reported in literature. Hence, the newly developed algorithm for construction of co-expressed PPI modules leads to robustly identifiable and comprehensive molecular biomarkers.

The high consistency with which the modules were detected across studies in **Chapter 2**, cleared the way for developing a statistical framework for a joint modular analysis of transcriptomic datasets, to further improve the robustness of the detected co-expressed PPI modules. In **Chapter 3**, the methodology for co-expressed PPI detection was put into a meta-analysis framework for module inference as well as for the subsequent module associations with the studied phenotype, in our case chronological age. Application of the improved algorithm to four transcriptomic data sets measured in blood⁷⁻⁹ (~2.500 samples) revealed five co-expressed modules of which the mean gene expression level associated with chronological age. Re-analysis in an

independent study^{10,11} (~3.500 samples) showed that the associations with chronological age replicated for four out of five identified modules, demonstrating the robustness of the presented method for biomarker extraction with respect to correlations with phenotypes. Remarkably, one of the modules contains the *ASF1A* gene, which has previously been identified as differentially expressed between members of long-lived families and controls¹². Moreover, using gene expression data of nonagenarians of the Leiden Longevity Study (LLS, ~50 samples)¹², we show that the *ASF1A* containing module also associates with prospective survival after age 90. Thus, expression of *ASF1A* and its co-expressed module members may constitute a novel robust biomarker for biological aging.

A second objective of this thesis was to develop methodology suited for the comprehensive and robust analysis of genetic variants coming from whole genome sequencing studies into human aging and longevity. In **Chapter 4**, we investigated analysis strategies that exploit prior knowledge on gene membership and impact for grouping and prioritizing coding variants. To assess the use of such gene-centric analysis strategies for identifying robust and interpretable genetic loci that affect human aging, we used Next Generation Sequencing (NGS) data on 218 long-lived cases from the LLS¹³ and 98 population controls^{14,15}. We first hypothesized that long-lived cases may have a genome-wide depletion of high impact protein-altering variants present in the germ line, either leading to a better functioning or more complete proteome,

or marking a high fidelity DNA repair system^{16,17}. On a genome-wide scale, we indeed observed that long-lived cases, as compared to the population controls, display a significant depletion of variants in the coding sequence and especially fewer of the highly disrupting frameshift insertions and deletions. Validation experiments using Sanger sequencing, however, could not underpin these findings. These experiments indicated that an excess of false positive disruptive variants in the control group contributed to the difference between cases and controls rather than a depletion of such variants in cases. This was likely caused by a technical bias in the sequencing of controls, which were measured at a later time point than cases, be it at the same Complete Genomics platform. The contribution of rare variants to familial longevity requires further research, ideally in larger studies than the ones performed here or by other groups in the field.

In **Chapter 4** we secondly hypothesized that long-lived cases may exhibit a gene specific enrichment of rare disruptive variants inhibiting gene functioning in line with knockout experiments leading to life span extension in model organisms¹⁸. Remarkably, we observed that long-lived cases carried a significant excess of frameshift deletions and insertions as compared to the population controls in two genes: *DNMT3A* and *TET2*. Notably, also other categories of disruptive variants, e.g. missense and nonsense SNVs, did support the genetic burden of disruptive variants at these two loci. The protein encoded by *TET2* is a methylcytosine dioxygenase that catalyses the conversion of methylcytosine

to 5-hydroxymethylcytosine¹⁹, and *DNMT3A* is a DNA Cytosine-5-Methyltransferase 3 that is involved in *de novo* methylation²⁰, which is essential for the establishment of DNA methylation patterns during development. Both encoded proteins are involved in myelopoiesis^{21,22}, and defects in these genes have been associated with several myeloproliferative disorders^{23,24}.

Sequence read evidence for the rare disruptive variants in both *DNMT3A* and *TET2* suggested that these variants are predominantly somatic, rather than germ line, and this observation was confirmed by Sanger sequencing experiments. Interestingly, similar somatic mutations in *TET2* and *DNMT3A* have previously been associated with an outgrowth of myeloid stem cells leading to myeloid dysplasia (MDS) and subsequent progression to^{23,24}, as well as outcome of acute myeloid lymphoma (AML)^{25,26}. Hence, in line with literature on the genetics of hematopoietic stem cell aging, the genetic burden at *DNMT3A* and *TET2* should in effect be interpreted as a potential marker of stem cell aging rather than a potential heritable factor underlying familial longevity.

In **Chapter 5**, we abandon the gene-centric analysis scope for grouping and prioritizing variants as employed in **Chapter 4**, and instead use a genomic convergence approach²⁷ to limit the analysis of NGS variants to those originating from genomic regions most likely to harbour determinants of human longevity. We obtain such regions by performing affected sibling pair analyses among all families from the LLS, while stratifying for the family history of excess survival (FH(+)), as we believed this

selection to further enrich for variants underlying human longevity. Importantly, the FH(+) families are characterized by an attenuated thyroid function as compared to long-lived families without such a marked family history (FH(-)), which thus suggests a pleiotropic relation between human longevity and attenuation of the thyroid function. The linkage analyses identified a 2.4 Mb region at chromosome 13q34 with significant linkage that was highly specific to the FH(+) subset. This finding indicates that sibs of the 239 long-lived sibships have inherited the identical strands of DNA from their parents significantly more often than would be expected by chance, thus implicating this locus to harbour variants underlying human longevity by attenuating the thyroid function.

We next employed NGS data assayed on 214 selected index cases, maximal one of each of the 239 FH(+) sibships, to further scrutinize the obtained 13q34 locus exhibiting significant linkage for familial longevity. To this end, we performed fine mapping by using a QTL analysis, i.e. genetic association analysis, employing the NGS genotypes and a relevant trait. Since the case-control comparison in **Chapter 4** had low power due to various reasons (e.g. phenotypic heterogeneity, binary trait), we used quantitative traits for fine mapping that mark the beneficial cardio-metabolic make-up of members of long-lived families. The FH(+) subset exhibited a significantly lower serum free triiodothyronine level, the active thyroid hormone itself (fT3), as compared to the FH(-) subset, which moreover seemed to affect the prospective survival in FH(+) differently as in the FH(-) subset. Hence, we employed fT3 as a trait

in the following QTL analyses for fine mapping the 13q34 locus and found the minor C allele of rs9515460 to mark an fT3 lowering haplotype, potentially explaining the attenuated thyroid function.

Thus far we concluded that rs9515460-C carriers exhibit an attenuated thyroid function, as indicated by their relatively low fT3 level and assume this to be beneficial to reach the age of 90 years. Whereas attenuation of the thyroid function is known to associate with a beneficial cardio-metabolic make up at middle age²⁸, low fT3 is also known to mark a poor prospective survival in the oldest old^{29,30}. Accordingly, nonagenarian sibships of the LLS carrying the fT3 lowering haplotype, tagged by rs9515460-C, also displayed a significantly poorer prospect of survival after age ninety as compared to the remainder of the study. This observation can be explained by considering the thyroid axis in relation to blood pressure and the relation with cardiovascular mortality thereof. Increased thyroid levels promote an increased heart rate and cardiac output, leading to an increased pulse pressure. Whereas a low systolic blood pressure is beneficial from middle age onward (65 to 84), as indicated by a lower risk on cardiovascular death, it becomes detrimental in the oldest old (age > 84)³¹. Hence, variants constitutively lowering serum fT3 levels are expected to contribute to longevity by transmitting their beneficial effects for cardiovascular health prior to age ninety.

In **Chapter 6**, an R package is presented facilitating the execution of some of the routinely encountered tasks in 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) and exemplify its use with publically available omics data sets.

To summarize, the research of this thesis provided the following insights in human aging. First, a number of gene networks changes their expression with age in such a consistent way that the phenotypic consequences can now be widely studied (**Chapter 3**). Secondly, we observed that a long life is not necessarily hampered by potentially premalignant somatic mutations in either *TET2* or *DNMT3A* (**Chapter 4**). Finally, attenuation of thyroid function as represented by low fT3, may be beneficial at middle age, but seems to contribute causally to increased mortality above 90 years (**Chapter 5**).

3. Integrative Omics in Biomarker Research into Human Aging

3.1 Module biomarkers in transcriptomics data analysis

In the first part of this thesis (**Chapter 2 and 3**), we show that the robustness and interpretability of molecular biomarkers for healthy aging extracted from transcriptome data sets can be improved by incorporating prior knowledge and adopting strategies for the joint analysis. Although many of the tested modules were significantly enriched for one or multiple functional Gene Ontology categories, a considerable number of modules did not

display any significant enrichment at all. This type of observation is currently under hot debate in the networking field³² and basically refers to the question whether such modules comprise novel knowledge or are more likely to represent artefacts of the employed method. Xue *et al.* show that such co-expression modules, in absence of any significant functional enrichment, are still consistently observed across multiple studies in human and even mouse³³, implying that these modules are not spurious findings and thus seem to constitute novel knowledge. Hence, results coming from network analyses that do not overlap with our current knowledge, are not spurious findings per se, but instead may point to novel contexts in which genes jointly perform a potentially unknown though apparently important cellular task.

Correlations in gene expression may arise as a result of a shared transcriptional program, implying functional relatedness, however, it might also arise due to a varying cell composition across samples. Since cell composition is known to vary with age and between breast tumours, we expect that parts of the recovered gene regulatory networks in fact represents changes in cell type composition. Indeed, some of the detected modules showing an association with the analysed phenotype were also enriched for particular cell types (**Chapter 2 and 3**), thus pointing to the potential presence of such confounders in our analyses. In contrast, such enrichments were not observed in the individual gene analysis (**Chapter 3**), which does not imply the absence, but merely the lack of power to detect such potential confounders. Once detected, modules enriched for a

particular cell type can be regarded as its biomarker and can subsequently serve as a surrogate variable for correcting the analysis (**Chapter 3**). Such an application of our module-based approach closely relates to deconvolution-based methods for correcting gene-expression data for blood composition^{34,35}, with the distinction that our method would not rely on calibration data sets to appoint genes *a priori* for creating surrogate variables. In effect, a modular analysis does not solve the problem of shifting cell type compositions confounding association analyses, but as opposed to an individual gene analysis, a modular analysis does have the power to discover and provide opportunities to adjust for such potential confounders.

The nature of the employed PPI resource is greatly influencing the outcome of network-based computations. For instance, West *et al.*³⁶ refer to the network positions of transcription factors (TFs) as *peripheral* with respect to the cellular signalling hierarchy, which can be explained by the fact that they do not consider DNA-protein interactions in their network. Hence, the choice of the PPI resource employed (STRING³⁷) is likely to have affected the exact composition of the obtained modules in **Chapters 2 and 3**. However, the aim of the developed method was not to infer complete collections of functional relationships, but merely to infer sufficient numbers required for improving the interpretability and robustness of the obtained modules. Hence, the presented modules are not necessarily exhaustive overviews of all genes involved in particular cellular functions, but instead represent clusters of genes with a tight

functional coherence as judged by the intersection of the co-expression data and the employed PPI resource.

4. Integrative Omics in the Genetics of Human Aging

4.1 Strategies for the analysis of whole-genome sequencing data

In the second part of this thesis (**Chapter 4 and 5**), we show that the robustness of NGS data analysis can be improved by adopting either a consecutive use of genetic data sources or by applying strategies incorporating prior knowledge for the grouping and prioritization of variants. The availability of NGS data raises the opportunity for investigating novel genetic variants and their potential relation with the aging phenotype in an unbiased genome-wide approach. However, in both **Chapter 4 and 5** we apply rigorous filtering of variants prior to the analysis, which may appear counterintuitive. Reasons for the stringent filtering relate to the statistical difficulties encountered when analysing NGS discovered variants. Newly discovered variants come in great numbers, though often only exhibit low frequencies in the general population, conveying very limited power in association tests. Yet, especially the analysis of these very rare variants is most interesting, as they have *a priori* the highest probability of conferring a profound impact on the phenotype. By limiting the number of tests, through aggregating (**Chapter 4**) and filtering (**Chapter 4 and 5**) individual variants, we gain additional power in the remaining association tests enabling research of the

role of rare genetic variation in the rate of aging and human longevity.

As an alternative to the stringent filtering performed in **Chapter 4**, one could also gain additional power in aggregate association tests by down-weighting unimportant variants, rather than discarding them, as is done in for instance the Sequence Kernel Association Test (SKAT)³⁸. Since additional power might be gained by also including the signal of lower impact variants, we investigated the application of SKAT in the data presented in **Chapter 4**. Various scenarios were investigated, based on inclusion of all or only particular variant types (SNV, deletion and insertion), expected impacts (missense, nonsense, non-stop etc.) and aggregations per gene or predefined sets of candidate genes taken from literature. SKAT applied per gene indicated that the baseline scenario using all coding protein-altering SNVs (missense, nonsense, nonstop and misstart) had more power over several scenarios in which the analysis was limited to subsets of high impact variants. Furthermore, the baseline scenario using only SNVs had comparable power to the scenario in which all variant types were included. Noteworthy, additional filtering of missense SNVs using ANNOVAR³⁹ decreased power in all scenarios, though generally exhibited comparable rankings of top genes to scenarios in which missense filtering was not applied. Both the gene-based as the gene set-based results were generally driven by a single rare variant (MAF~1%), but not by common (MAF>=5%) or multiple singleton observations. This common variant generally exhibited lower allele frequencies in the long-lived cases

as compared to the population controls. Hence, despite the variant weighing based on allele frequencies included in SKAT, contributions of common or singleton SNVs are still ignored. Moreover, in the absence of ready-applicable priors for the different variant categories, contributions of rare high impact variants are totally neglected. In effect, a joint frequency-weighted association analysis using SKAT in the current study does not provide additional benefits with respect to the interpretation or robustness of the obtained results over an association analysis based on individual variants.

When limiting aggregate association tests to the use of high impact rare variants only (**Chapter 4**) in so called Rare Variant Association Studies (RVAS)⁴⁰, one becomes especially vulnerable to bias by sequencing errors. Error rates increase with both increasing impact and increasing rareness of variants, yet we pursued an RVAS favouring singleton observations with a gene disrupting impact. An important motivation for this approach was that extensive genotyping experiments with the Sequenom Mass Array platform (data not shown) following the whole genome sequencing of the long lived cases, generally displayed a near perfect concordance with the NGS genotypes (by Complete Genomics), irrespective of the allele frequency or predicted impact of the assayed SNV. When extending the validation experiments to frameshift indels, initially only within *TET2* or *DNMT3A*, we found comparably high concordance rates, which led us to the false impression that all frameshift indels were identified with high confidence by our whole genome sequencing data. It was

when validation experiments on random subsets of frameshift indels originating from the whole genome were performed when we first learned that this is generally not the case. Hence, we strongly advise against conducting RVAS studies within a whole-genome framework, and instead advice to perform RVAS studies within a gene-specific context only, with the additional remark that re-sequencing of the identified disruptive variants must not be omitted.

Thus far, most of the attention in NGS experiments has gone to variants in the coding domain, due to their ease of inference and interpretation, thus providing a logical starting point for our research into the analysis of NGS variants in **Chapter 4**. However, as most of the established associations with common SNVs in GWASs generally seem to coincide with regulatory domains, like enhancers⁴¹, rather than with coding domains, it seems reasonable to assume that the same holds for rare variants coming from NGS experiments. Therefore, in this thesis we also pursued strategies for analysis of rare intergenic variants in which expected impact on the basis of gene annotations could not serve as filtering criterion. For instance, in **Chapter 5** we limited the number of variants employing the results of a genome-wide linkage study into familial longevity as a source of prior information. In conclusion, the incorporation of additional omic data sources to reduce and guide the number of tested hypotheses is highly recommended, and is even more required when extending the analysis to rare variants in intronic and intergenic regions.

4.2 Future prospects for integrative omics into the genetics of human aging

Both the gene-centric (**Chapter 4**) as the purely data-driven approach (**Chapter 5**) for prioritizing and analysing NGS variants have merits and in an ideal setting both approaches are applicable. The usefulness of an additional gene-centric approach for the interpretation of the NGS results in (**Chapter 5**) is illustrated when incorporating prior information on regulatory domains, such as ChIA-PET data created by the ENCODE consortium⁴² (Figure 1). ChIA-PET is a protocol for capturing the 3D physical proximity between distant DNA domains for elucidating for instance enhancer-promotor interactions, thus relating non-coding domains to genes. Unlike other conformation capturing techniques, ChIA-PET includes a Chip step, to specifically enrich for distant DNA interactions associated with one particular species of protein only. Multiple ChIA-PET experiments in the NB4 cell line on POL2, provide evidence for a physical interaction during transcription between the intergenic region of *CARS2/ING1*, coinciding with the maximum linkage signal and the promoter of *IRS2*, residing outside the linkage area. Interestingly, disruption of the *IRS2* homologs in the fruitfly *D. melanogaster*⁴³ was found to induce longevity, suggesting that the intergenic region of *CARS2/ING1* contains regulatory elements required for the transcription of *IRS2*, which is perturbed in members of long-lived families. These results show that a gene-centric or even candidate approach for analysing NGS data into human longevity is sensible, however,

should not be limited to protein-altering variants only. Hence, potential future extensions will focus on a gene-centric methodology that incorporates prior knowledge for including variants residing in regulatory domains as to improve the NGS analyses on strong candidate genes for human longevity.

5. Evidence for Hallmark Aging Processes

López-Otín *et al.*⁴⁴ described nine hallmark processes consistently observed to co-occur with aging and this thesis provides evidence for some of these processes to play a role in human aging processes as indicated by molecular aging processes in whole blood. In **Chapter 3**, a robust age-associated co-expressed module was identified reflecting a decline in transcription of ribosomal proteins, relating to the hallmark process of “loss of proteostasis” (Figure 1 Introduction). This observation also provides indirect evidence for “deregulated nutrient sensing”, as ribosomal expression is under control of mTOR-insulin-signalling⁴⁵, a pathway responsible for tuning the basal metabolism to the availability of nutrients. Corroborating evidence for the involvement of both processes in aging is given in **Chapter 5**. The main genetic leads for a successfully slowed aging are *CARS2*, a gene required for translation of novel proteins, *ING1*, an inhibitor of growth or *IRS2*, an insulin receptor. Hence, the hallmark aging processes “loss of proteostasis” and “deregulated nutrient

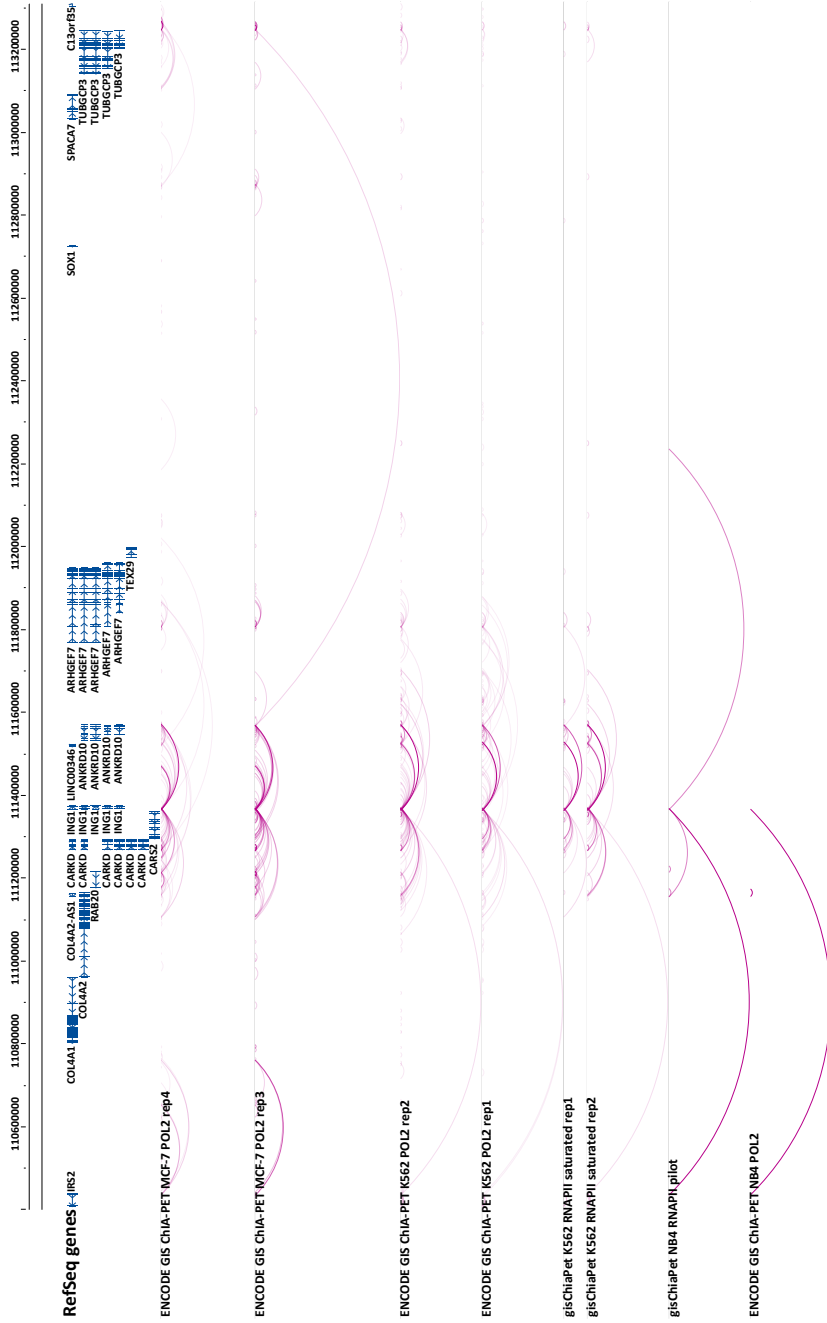


FIGURE 1: INCORPORATING DNA-DNA INTERACTIONS AS A SOURCE OF PRIOR KNOWLEDGE. This figure illustrates the potential merit of incorporating ChIA-PET data created by the ENCODE²² consortium for the prioritization of variants. Depicted is the 1-LOD-drop region on chr13q34 (**Chapter 5** of this thesis) in which a lot of crosstalk is seen for the intergenic region between *CARS2* and *IRS2* and other genes within this region, including the promoter of *IRS2*.

sensing” seem to play a role in the aging of whole blood.

The most significant age-associated co-expressed PPI module detected in **Chapter 3** was enriched for “T-Cell Activation” and down-regulated with age, probably reflecting the hallmark processes “senescence” or “stem cell exhaustion” within the lymphoid compartment. Interestingly, the somatic mutations in *TET2* and *DNMT3A* described in **Chapter 4** are associated with an outgrowth of myeloid stem cells^{19,22}, which thereby gradually displace the lymphoid compartment, and could therefore explain the observed age-associated down-regulation of genes involved in “T-Cell Activation” in **Chapter 3**. The *TET2* and *DNMT3A* mutations also provide indirect evidence for another hallmark process of aging, namely: “epigenetic alterations”. *TET2* and *DNMT3A* are both important epigenetic factors involved in DNA methylation^{20,46} and mutations in these genes have been associated with aberrant methylation levels^{19,25}, and progression to^{23,24} and prognosis of AML^{25,26}. There is an on-going discussion whether such epigenetics-modifying gene mutations are somehow responsible for causing the widespread genomic instability observed in MDS and AML^{25,47}, which also happens to be a hallmark of aging. This discussion is triggered by the observation of widespread genomic instabilities in mice upon knockout of the *DNMT1* gene⁴⁸, like *DNMT3A* a gene involved in DNA methylation, which notably was grouped in an age associated module enriched for DNA integrity identified in **Chapter 3**. Hence, this thesis provides evidence for the

involvement of the hallmark aging process “senescence”, “stem cell exhaustion”, “epigenetic alterations” and perhaps “genomic instability” in the aging of whole blood.

6. Outlook for Research in the Molecular Biology of Aging

According to the general expectation, future research into the genetics of complex traits will increasingly be based on NGS technology. However, NGS based advances into the genetics of complex traits in general and aging in specific have been fairly modest, which can be partly attributed to the limitations of the currently employed sequencing technology based on short reads⁴⁹. With the rapid advancements in sequencing technology, it is expected that the complications for assembly and subsequent variant calling arising due to limited length of reads will be effectively negated as that in the near future the need for a reference genome will be omitted. Another development in sequencing technology is the decreasing quantities required to serve as template in the sequencing protocols, and currently enables genotyping of DNA and quantification of RNA species of biomaterials derived from a single cell⁵⁰. Especially for research into aging of tissues with a very heterogeneous cell type composition, like whole blood, these developments are expected to shed many new insights. Though improvements in sequencing technology will alleviate many of the complicating factors of variant (**Chapter 4 and 5**) or expression (**Chapter 3**) analyses related to the certainty of the

data, it does not solve any of the problems arising due to the heterogeneity of the aging phenotype. Thus in the prospect that these advances will create even more data points, but not necessarily in more individuals, the need for incorporating techniques for data integration into the analyses of omics data into aging will only grow.

With the completion of large international efforts aimed at meticulously scrutinizing the functional elements in DNA and the interactions thereof^{42,51}, many exiting opportunities arise for the advanced interpretation of genomic variants in their genomic context. The next step is to assess the phenotypic effects in case such molecular circuits are perturbed. Data generation initiatives in large human bio-banks, like the BBMRI-NL BIOS consortium (<http://www.bbMRI.nl/en-gb/activities/rainbow-projects/bios>) are aimed at facilitating this link by collecting deep phenotypic information and multiple omics data sources all assayed in the same individuals. Jointly, these data resources would be of great value for translating omics findings into human aging on the molecular level to effects of health on the organismal level, ideally providing a mechanistic insight into the molecular drivers of human aging.

Family-based study designs are very valuable for studying biomarkers and the genetics for human aging, as they provide the means for controlling the considerable amounts of unwanted biological variation present in assayed omics data sets. For instance, since the genome of every individual contains many unique highly disruptive variants, each genome is said to

have a high “narrative potential” as many of these variants could provide a compelling story how the variant would influence a particular trait^{52,53}. Hence, the existing literature on NGS analyses on longevity, based on the genomes of few exceptionally long-lived individuals⁵⁴⁻⁵⁶, should be interpreted with extreme caution. Checks for co-segregation patterns across multiple carefully selected families with a deep or wide genealogy would greatly reduce the likelihood of reporting such false positive findings⁵⁷. Family based designs offer additional benefits for research into the genetics of complex traits, especially whenever complex traits exhibiting low or modest heritabilities are studied, as is for instance the case for human longevity. The fact that only ~25% of the variation in human life span is expected to be caused by genetic variations in the population at large⁵⁸, makes selection of suitable research subjects for research purposes into the genetics of human longevity challenging, but imperative. Information on family history can be exploited to select individuals with a genetic propensity to become long-lived¹³. To maximally exploit this concept, historical data will be explored in search for families, which have long-lived members in several generations. The currently living descendants can then be investigated for genetic variants in common. In effect, investigation of multiple members of long-lived families may contribute to determine the causality of genetic variants.

The identification of biomarkers for aging is predominantly studied in cross-sectional study design employing data sources assayed on a single time-point

on a single tissue. In order to get a better understanding of the systems dynamics that lead to human aging, it is imperative that repeated omics measures within the same individuals in longitudinal studies are available in large sample sizes. So-called systems approaches are performed for studying model organisms to assess systems-wide responses to perturbations across multiple systemic levels or tissues, at multiple time points, using multiple omic platforms, while collecting multiple phenotypic read out parameters. Ideally, this approach would be applied for studying human longevity, by collecting data within multiple members of families with a genetic propensity to become long-lived and in age and environmentally matched controls, as to identify the molecular drivers into a successfully decelerated aging.

Aging is a heterogenic phenotype caused by multiple functionally independent molecular pathways⁴⁴. The comprehensive assessment of one's overall state of aging, therefore, probably requires multiple independent biomarkers of biological aging. For instance, cardiovascular health and aging is marked by factors such as blood pressure and total cholesterol level⁵⁹, whereas aging of the neuromuscular system is marked by atrophy of muscle and neuronal cells⁶⁰. Hence, methodology is required allowing to analyse omics data sources in the perspective of a multitude of unrelated biomarkers of biological age.

Resources consistently collecting and curating lists of longevity genes established in model systems, such as GenAge⁶¹, can be incorporated to serve as prioritization or interpretation tools for omics analyses into human aging (Figure

2). Publically available transcriptomic resources assayed in multiple organisms can be exploited for estimating conserved gene regulatory networks, similar as is done in **Chapter 2 and 3** of this thesis. Genes within this network can then be prioritized by their proximity to already established longevity genes within this conserved gene regulatory network. Thus obtained novel candidate genes for longevity can be validated in knockdown experiments in for instance *C. elegans* or *D. melanogaster*. Genes that extend life span upon knockdown serve as input for further research into specific cohorts suitable for studying biological aging in humans.

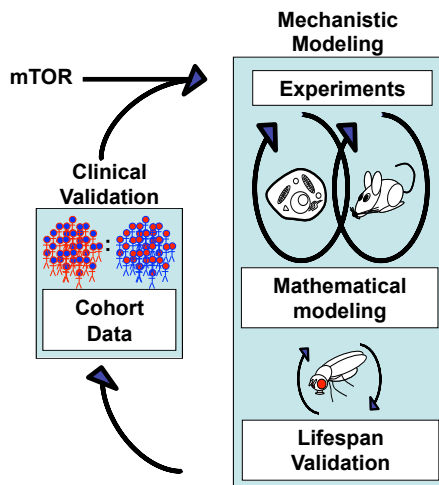


FIGURE 2: THE PARADIGM OF SYSTEMS BIOLOGY APPLIED TO AGING RESEARCH.

7. Conclusion

We have demonstrated the relevance of incorporating concepts of data integration for the comprehensive and robust analysis of omics datasets for molecular pathways driving human aging. Though we were

able to robustly assess the presence of certain hallmark processes of aging to occur in human blood, we could mostly only speculate on the causality and the significance of the crosstalk between the different aging processes. To further explore both the aspects of causality and crosstalk, and thus to get a deeper understanding of the systems biology of human aging, large-scale systems approaches are needed that assay multi-level omics data in family-based and large population-based studies, preferably across different tissues and time points. Methodology for data integration is essential in the analysis of these rich data sources required for the elucidation of the molecular pathways driving human aging.

8. References

1. Desmedt, C. *et al.* Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res* **13**, 3207-14 (2007).
2. Loi, S. *et al.* Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics* **9**, 239 (2008).
3. Miller, L.D. *et al.* An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci U S A* **102**, 13550-5 (2005).
4. Pawitan, Y. *et al.* Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res* **7**, R953-64 (2005).
5. Wang, Y. *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671-9 (2005).
6. Schmidt, M. *et al.* The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res* **68**, 5405-13 (2008).
7. Goring, H.H. *et al.* Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet* **39**, 1208-16 (2007).
8. Inouye, M. *et al.* An immune response network associated with blood lipid levels. *PLoS Genet* **6**, e1001113 (2010).
9. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423-8 (2008).
10. Boomsma, D.I. *et al.* Genome-wide association of major depression: description of samples for the GAIN Major Depressive Disorder Study: NTR and NESDA biobank projects. *Eur J Hum Genet* **16**, 335-42 (2008).
11. Jansen, R. *et al.* Sex differences in the human peripheral blood transcriptome. *BMC Genomics* **15**, 33 (2014).
12. Passtoors, W.M. *et al.* Transcriptional profiling of human familial longevity indicates a role for ASF1A and IL7R. *PLoS One* **7**, e27759 (2012).
13. 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).
14. Boomsma, D.I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet* **22**, 221-7 (2014).
15. The Genome of the Netherlands, C. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* (2014).
16. Garinis, G.A., van der Horst, G.T., Vijg, J. & Hoeijmakers, J.H. DNA damage and ageing: new-age ideas for an age-old problem. *Nat Cell Biol* **10**, 1241-7 (2008).
17. Hoeijmakers, J.H. DNA damage, aging, and cancer. *N Engl J Med* **361**, 1475-85 (2009).
18. Kenyon, C.J. The genetics of ageing. *Nature* **464**, 504-12 (2010).

19. Ko, M. *et al.* Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature* **468**, 839-43 (2010).
20. Okano, M., Xie, S. & Li, E. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nat Genet* **19**, 219-20 (1998).
21. Challen, G.A. *et al.* Dnmt3a is essential for hematopoietic stem cell differentiation. *Nat Genet* **44**, 23-31 (2012).
22. Moran-Crusio, K. *et al.* Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. *Cancer Cell* **20**, 11-24 (2011).
23. Jankowska, A.M. *et al.* Loss of heterozygosity 4q24 and TET2 mutations associated with myelodysplastic/myeloproliferative neoplasms. *Blood* **113**, 6403-10 (2009).
24. Ewalt, M. *et al.* DNMT3a mutations in high-risk myelodysplastic syndrome parallel those found in acute myeloid leukemia. *Blood Cancer J* **1**, e9 (2011).
25. Ley, T.J. *et al.* DNMT3A mutations in acute myeloid leukemia. *N Engl J Med* **363**, 2424-33 (2010).
26. Metzeler, K.H. *et al.* TET2 mutations improve the new European LeukemiaNet risk classification of acute myeloid leukemia: a Cancer and Leukemia Group B study. *J Clin Oncol* **29**, 1373-81 (2011).
27. 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).
28. Selmer, C. *et al.* Subclinical and Overt Thyroid Dysfunction and Risk of All-cause Mortality and Cardiovascular Events: A Large Population Study. *J Clin Endocrinol Metab*, jc20134184 (2014).
29. Martin-Ruiz, C. *et al.* Assessment of a large panel of candidate biomarkers of ageing in the Newcastle 85+ study. *Mech Ageing Dev* **132**, 496-502 (2011).
30. Gussekloo, J. *et al.* Thyroid status, disability and cognitive function, and survival in old age. *JAMA* **292**, 2591-9 (2004).
31. Satish, S., Freeman, D.H., Jr., Ray, L. & Goodwin, J.S. The relationship between blood pressure and mortality in the oldest old. *J Am Geriatr Soc* **49**, 367-74 (2001).
32. Lage, K. ASHG Network Session. (2013).
33. Xue, Z. *et al.* Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500**, 593-7 (2013).
34. Shen-Orr, S.S. *et al.* Cell type-specific gene expression differences in complex tissues. *Nat Methods* **7**, 287-9 (2010).
35. Gong, T. *et al.* Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS One* **6**, e27156 (2011).
36. 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).
37. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* **39**, D561-8 (2011).
38. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* **91**, 224-37 (2012).
39. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
40. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A* **111**, E455-64 (2014).
41. Maurano, M.T. *et al.* Systematic localization of common disease-

- associated variation in regulatory DNA. *Science* **337**, 1190-5 (2012).
42. Bernstein, B.E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
 43. Clancy, D.J. *et al.* Extension of life-span by loss of CHICO, a Drosophila insulin receptor substrate protein. *Science* **292**, 104-6 (2001).
 44. Lopez-Otin, C., Blasco, M.A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194-217 (2013).
 45. Laplante, M. & Sabatini, D.M. mTOR signaling at a glance. *J Cell Sci* **122**, 3589-94 (2009).
 46. Mohr, F., Dohner, K., Buske, C. & Rawat, V.P. TET genes: new players in DNA demethylation and important determinants for stemness. *Exp Hematol* **39**, 272-81 (2011).
 47. Wakita, S. *et al.* Mutations of the epigenetics-modifying gene (DNMT3a, TET2, IDH1/2) at diagnosis may induce FLT3-ITD at relapse in de novo acute myeloid leukemia. *Leukemia* **27**, 1044-52 (2013).
 48. Brown, K.D. & Robertson, K.D. DNMT1 knockout delivers a strong blow to genome stability and cell viability. *Nat Genet* **39**, 289-90 (2007).
 49. Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nat Genet* **44**, 623-30 (2012).
 50. Junker, J.P. & van Oudenaarden, A. Every cell is special: genome-wide studies add a new dimension to single-cell biology. *Cell* **157**, 8-11 (2014).
 51. Bernstein, B.E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**, 1045-8 (2010).
 52. Goldstein, D.B. *et al.* Sequencing studies in human genetics: design and interpretation. *Nat Rev Genet* **14**, 460-70 (2013).
 53. Dewey, F.E. *et al.* Clinical interpretation and implications of whole-genome sequencing. *JAMA* **311**, 1035-45 (2014).
 54. 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).
 55. Ye, K. *et al.* Aging as accelerated accumulation of somatic variants: whole-genome sequencing of centenarian and middle-aged monozygotic twin pairs. *Twin Res Hum Genet* **16**, 1026-32 (2013).
 56. Sebastiani, P. *et al.* Whole genome sequences of a male and female supercentenarian, ages greater than 114 years. *Front Genet* **2**, 90 (2011).
 57. MacArthur, D.G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469-76 (2014).
 58. Skytthe, A. *et al.* Longevity studies in GenomeEUtwin. *Twin Res* **6**, 448-54 (2003).
 59. Wilson, P.W. *et al.* Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837-47 (1998).
 60. Vandervoort, A.A. Aging of the human neuromuscular system. *Muscle Nerve* **25**, 17-25 (2002).
 61. de Magalhaes, J.P. & Toussaint, O. GenAge: a genomic and proteomic network map of human ageing. *FEBS Lett* **571**, 243-7 (2004).

Chapter 8:

Nederlandse samenvatting

1. Introductie

1.1 Waarom onderzoek naar veroudering?

Volgens de voorspellingen van het Centraal Bureau van de Statistiek zal ook in de aankomende decennia de levensverwachting van de algemene Nederlandse bevolking onveranderd blijven toenemen. Helaas zullen deze gewonnen jaren doorgaans niet allen in goede gezondheid worden doorgebracht. Immers, naarmate we ouder worden neemt het risico op het krijgen van velerlei zeer algemeen voorkomende ziekten, zoals verscheidene vormen van kanker, diabetes mellitus type II en hart- en vaatziekten toe. Naar verwachting zal de aankomende vergrijzing van onze samenleving gepaard gaan met een toename van dergelijke leeftijdgerelateerde ziekten. Fundamenteel onderzoek naar de oorzakelijke samenhang tussen de verschillende aspecten van gezondheid en het verouderingsproces is dan ook noodzakelijk.

1.2 Wat is veroudering?

Veroudering wordt gekenmerkt door een geleidelijke maar onherroepelijke afname van controle en dus functionaliteit over alle organisatorische lagen van het menselijk lichaam. In tegenstelling tot hoe de meeste andere karakteristieken van ons lichaam tot stand komen, zoals bijvoorbeeld oogkleur of lichaamslengte, is veroudering niet vastgelegd in ons DNA, maar is het een resultaat van een passief proces. De jarenlange blootstelling aan stochastische schade bronnen, zoals ultra violet licht of vrije radicalen, zorgt voor een

opstapeling van onvolkomenheden in de cel, de kleinste levende bouwsteen van ons lichaam, en ondermijnt daarmee geleidelijk zijn efficiëntie en incasseringsvermogen. Deze slijtage slag vindt plaats in al onze cellen verspreid over al onze weefsels en vergroot daarmee op den duur de vatbaarheid van ons lichaam voor allerlei kwalen en ziekten. Dientengevolge stelt de opstapeling van onvolkomenheden in onze cellen daarmee een limiet aan de duur van ons leven.

1.3 Langlevendheid: de sleutel tot het verouderingsonderzoek

Alhoewel het verouderingsproces zelf onafwendbaar lijkt, is de snelheid waarmee dit optreedt niet hetzelfde voor elk individu. Zo hebben directe familie leden van negentig jarigen, vergeleken bij de algemene bevolking, een grotere kans om zelf ook deze respectabele leeftijd te bereiken. Opvallend hierbij is dat de directe nazaten van deze langlevenden al op middelbare leeftijd minder kans hebben op het krijgen van leeftijd gerelateerde kwalen en ziekten, zoals diabetes mellitus type II of een te hoge bloeddruk. Enerzijds suggereert dit dat zowel de snelheid van veroudering als de uiteindelijke levensverwachting ten minste gedeeltelijk wordt bepaald door erfelijke componenten in ons DNA. Anderzijds suggereert dit dat de erfelijke componenten die deze levensverwachting beïnvloeden ook het risico op het krijgen van leeftijd gerelateerde ziekten verkleinen. Hieruit volgt dat we door een nauwgezette bestudering van de factoren die de snelheid van het verouderingsproces bepalen, we

ook de factoren kunnen identificeren die algemeen ten grondslag liggen aan diverse leeftijd gerelateerde ziekten.

1.4 Onderzoek naar de factoren van veroudering

In het humane verouderingsveld volgt men veelal twee onderzoeksstrategieën. Enerzijds wordt er gezocht naar factoren die vroegtijdig inzicht moeten geven in de mate en aard van veroudering. Het verouderingsproces verloopt bij iedereen anders en deze zogenaamde “biomarkers van biologische veroudering” meten ieder een ander aspect van dit proces. Anderzijds wordt er met verschillende technieken gekeken naar de genetische basis van factoren die de snelheid van veroudering lijken te beïnvloeden. We kijken hiervoor in het DNA van langlevenden en onderzoeken of bepaalde veranderingen in het DNA (variëaties) vaker of minder vaak voorkomen in de algemene bevolking. Voor beide strategieën wordt veelal gebruik gemaakt van zogenaamde “omics” meetmethoden, die er op gericht zijn om in één enkel experiment een zo compleet mogelijk beeld te geven van alle veranderingen die de cel op dat moment ondergaat. De cel reguleert zichzelf op vele organisatorische niveaus en voor ieder niveau bestaat er een “omics” methode die met behulp van duizenden tot enkele miljoenen metingen de staat van de cel vastlegt. Een grote uitdaging ligt momenteel in het onderling relateren en interpreteren van de data verkregen met verschillende van deze omics platforms. Het werk in dit proefschrift legt zich toe op het ontwikkelen van methoden voor de geïntegreerde analyse van “omics” data

bronnen, teneinde een beter en completer beeld te krijgen van de moleculaire biologie van veroudering. Twee strategieën voor data integratie zijn toegepast, namelijk: 1) de gecombineerde analyse van meerdere data bronnen, 2) de incorporatie van voorkennis uit externe informatie bronnen.

2. Integratieve Analyse van Genexpressie Data Bronnen

De instructies voor het maken van alle bouwstenen van de cel, de eiwitten, liggen besloten in ons DNA, dat veilig afgeschermd ligt in de celkern. Als nieuwe eiwitten worden aangemaakt, wordt eerst het daarbij behorende stukje DNA, het gen, gekopieerd. Door deze kopietjes voor ieder van naar schatting ongeveer 18.000 unieke genen te kwantificeren kunnen we een inzicht krijgen welke genen op dat moment door de cel gebruikt worden. Deze metingen van de zogenaamde expressie van genen wordt onder andere gebruikt voor het classificeren van tumoren van borstkanker patiënten (**hoofdstuk 2** in dit proefschrift) of om veranderingen in het gebruik van genen met toenemende leeftijd te karakteriseren (**hoofdstuk 3** in dit proefschrift). In het eerste gedeelte van dit proefschrift ontwikkelen we een methode voor een integratieve analyse van meerdere genexpressie data bronnen om zo de zekerheid en interpretatie mogelijkheden van onze bevindingen te vergroten.

Uit eerdere studies is gebleken dat in het gebruik van genexpressie metingen voor onderzoek naar borstkanker tumoren of veranderingen met leeftijd er

problemen optreden met betrekking tot de interpretatie en reproduceerbaarheid van de bevindingen. Dit is enerzijds toe te schrijven aan technische aspecten van de metingen, bijvoorbeeld de lage signaal-ruis verhouding, maar anderzijds wordt dit ook veroorzaakt door de heterogeniteit tussen de onderzochte individuen en de complexiteit van de onderzochte lichaamskarakteristieken. Om zowel de reproduceerbaarheid als de interpretatie van deze genexpressie analyses te vergroten hebben we een nieuwe integratieve methode ontwikkeld. In deze methode gebruiken we meerdere studies in een gecombineerde analyse om de kracht en consistentie van onze bevindingen te bevorderen. Daarnaast incorporeren we ook informatie over welke genen hun taken gezamenlijk uitvoeren in de cel, afkomstig uit speciaal voor dit doeleinde opgerichte databases en gebruiken ook dit om de consistentie van onze observaties te verifiëren. De incorporatie van dergelijke gen-gen interacties geeft bovendien voordelen bij de interpretatie van de resultaten. De ontwikkelde methode vereenvoudigt de analyse en interpretatie door systematisch te zoeken naar groepen van genen, met een nauw onderling verbonden biologische functionaliteit, gen modules genaamd, waarvan alle genen een onderlinge consistente relatie met betrekking tot de onderzochte lichaamskarakteristieken vertonen.

In **hoofdstuk 2** passen we de ontwikkelde methodologie toe op data van borstkanker tumoren en demonstreren we de hoge mate van interpreteerbaarheid en reproduceerbaarheid van de verkregen

gen modules. Alhoewel het gebruik van gen modules niet leidt tot een verbetering ten opzichte van de gangbare classificatie van de borstkanker tumoren, komt de samenstelling van de gevonden gen modules wel sterk overeen met genen die al eerder gevonden zijn in borstkanker onderzoek. We concluderen dat toepassing van de voorgestelde methode de data van naar schatting ongeveer 18.000 onafhankelijke metingen naar enkele reproduceerbare en goed interpreteerbare moleculaire biomarkers kan reduceren.

In **hoofdstuk 3** breiden we de voorgestelde methode verder uit door zowel het samenstellen van de modules, als de daaropvolgende stap waarin hun relatie tot karakteristieken van het lichaam wordt onderzocht, wordt uitgevoerd met meerdere gen expressie data bronnen tegelijkertijd. We passen de vernieuwde methode toe op vier genexpressie data bronnen in het bloed van bijna 2.500 individuen en analyseren de gemiddelde expressie van de genen in een module voor consistente veranderingen met leeftijd. We identificeren vijf modules waarvan de expressie gezamenlijk verandert met chronologische leeftijd. In een onafhankelijke studie waarin gen expressies zijn gemeten in het bloed van ongeveer 3.500 individuen bevestigen we de door ons gevonden relaties met leeftijd voor vier van de vijf modules. Een van de gerepliceerde modules bevat het gen *ASF1A*, waarvan we eerder al hebben aangetoond dat dit in andere hoeveelheden tot expressie komt in het bloed van directe nazaten van langlevenden ten opzichte van de algehele bevolking. Bovendien laten we zien dat de genexpressie van het

ASF1A gen alsmede het gemiddelde van de gehele gen module indicatief is voor de levensverwachting van ouderen boven de negentig jaar, die we gedurende 10 jaar gevolgd hebben. De *ASF1A* module lijkt dus een interessante kandidaat als nieuwe moleculaire biomarker voor biologische veroudering.

3. Integratieve Analyse van DNA Sequentie Data Bronnen

Metingen van variaties in de volgorde en samenstelling van ons erfelijk materiaal noemen we genetische databronnen. Vaak gaat het hierbij over zeer kleine variaties waarin slechts een enkele elementaire bouwsteen van het DNA, een nucleotide, vervangen is door een andere. Genetische data bronnen kunnen echter onderling sterk verschillen in de resolutie en schaal waarop de metingen aan het DNA zijn verricht. Genetische data bronnen zijn er in vele verschillende soorten en maten en verschillen onderling sterk in de resolutie en schaal waarop de metingen aan het DNA zijn verricht. De genetische data bron met de hoogst mogelijke resolutie wordt sequentie data genoemd, omdat deze letterlijk de volgorde van alle basenparen waaruit ons DNA is opgebouwd opsomt. Deze methode rapporteert per persoon gemiddeld ongeveer 3,2 miljoen veranderingen verspreid over het gehele genoom ten opzichte van een "gemiddeld" genoom, dat gebruikt wordt als referentie. De kracht van deze methode is echter ook meteen zijn grootste zwakte. Hoe onderscheiden we de veranderingen die van belang zijn, van alle miljoenen overige veranderingen

die waarschijnlijk geen enkele of slechts zeer beperkte consequenties hebben? In het tweede gedeelte van dit proefschrift richten we ons op het ontwikkelen van innovatieve methoden voor de analyse van sequentie data bronnen.

De consequentie van een variant in het DNA voor het functioneren van een cel heeft voornamelijk te maken met de positionering ten opzichte van stukken erfelijk materiaal die coderen voor de eiwitten in de cel. Genetische varianten die de eiwitcode veranderen, of zelfs geheel verstoren hebben over het algemeen een grotere kans om het functioneren van de cel te beïnvloeden. In **hoofdstuk 4** van dit proefschrift incorporeren we informatie uit externe bronnen, die voorspelt in welke mate een genetische variant een gen verstoort. We passen dit analyse kader toe op een sequentie databron bestaande uit variaties gemeten in 218 langlevenden uit de Leiden Langleven Studie (LLS) en 98 controles uit de algemene Nederlandse bevolking. Gebruikmakende van de voorspellingen of een variatie al dan niet een gen verstoort, hebben we genetische mechanismen onderzocht die gezonde veroudering en langlevendheid zouden kunnen bevorderen.

Zo hebben we onderzocht of bepaalde genen in langlevenden vaker dan verwacht geraakt zijn door zeer ernstig versturende variaties. Uit onderzoek in dier modellen is namelijk gebleken dat als specifieke genen worden verstoord, dit kan leiden tot een significante verlenging van hun levensverwachting. We hebben echter geen bewijs gevonden dat dergelijke genen de lange levensduur in de onderzochte

families verklaarden. Wel vinden we in de hoogbejaarden veel zeer verstorende varianten in de genen *TET2* en *DNMT3A*. Beide genen vervullen een belangrijke rol bij de differentiatie van bloed stamcellen in de verschillende soorten bloed cellen en zijn verstorende varianten in deze genen geassocieerd met een afwijkend bloedbeeld.

Diepere inspectie van de metingen van de meest verstorende varianten in *TET2* en *DNMT3A* liet zien dat waarschijnlijk slechts een klein gedeelte van alle cellen in het bloed van de onderzochte hoogbejaarden deze varianten draagt. Dit is een indicatie dat deze varianten vermoedelijk tijdens het leven zijn ontstaan en dit soort varianten worden somatische mutaties genoemd. In eerdere studies is van dergelijke somatische mutaties in *TET2* en *DNMT3A* al aangetoond dat deze samengaan met het onevenredig uitgroeien van het myeloïde bloedcompartiment en zelfs het risico op het ontwikkelen van Acute Myeloïde Leukemie verhogen (AML). Door de hoogbejaarde dragers van deze genafwijkingen 10 jaar lang te volgen konden we constateren dat de levensverwachting van dragers in het geheel niet wordt aangetast. Deze bevinding impliceert dat bejaarde dragers van dergelijke somatische mutaties een nog hogere leeftijd kunnen bereiken zonder een vorm van bloedkanker te ontwikkelen.

Onze zoektocht naar genetische factoren voor langlevendheid zetten we voort in **hoofdstuk 5** van dit proefschrift door een alternatieve strategie voor het prioriteren van variaties uit sequentie data toe te passen. We doen dit ten eerste door

informatie uit een additionele genetische data bron in onze analyses te incorporeren. Voor dit doeleinde gebruiken we de resultaten van een genomwijde koppeling analyse (linkage scan) naar familiare langlevendheid, waarin wordt gekeken naar de overervingspatronen binnen langlevende families. We passen deze techniek toe op de langlevende families uit de LLS, daarbij ook rekening houdende met de gemiddelde levensduur van de ouders van deze families. Gebruikmakende van dit statistisch kader voor linkage analyse, tonen we aan dat de genomische regio chr13q34 significant vaker dan verwacht in een identieke samenstelling wordt overgeërfd binnen de langlevende families en dat dit effect het sterkst is in de families waarvan de ouders ook zelf langer dan gemiddeld hebben geleefd. De resultaten van deze analyse doet vermoeden dat chr13q34 erfelijke factoren herbergt die bijdragen aan familiare langlevendheid.

Tentweede incorporeren we additionele informatie over bloedspiegels die karakteristiek zijn voor langlevenden, maar ook binnen deze groep een onderscheid maakt tussen langlevenden met langlevende ouders, FH(+), en langlevende families met ouders met een gemiddelde levensverwachting FH(-). Eerder onderzoek heeft namelijk aangetoond dat langlevende families waarvan de ouders ook langer dan gemiddeld hebben geleefd (FH(+)) een tragere schildklier functie hebben dan langlevende families met ouders met een gemiddelde levensduur (FH(-)). Dit suggereert dat de erfelijke componenten voor een tragere schildklier

functie onafhankelijk bijdragen aan langlevendheid.

Vervolgens gebruiken we de reeds in **hoofdstuk 4** van dit proefstuk beschreven sequentie data van de participanten uit de LLS voor het nader bestuderen van genomische regio chr13q34, waarvan we nu vermoeden dat het genen herbergt die betrokken zijn bij zowel een trage schildklierfunctie als langlevendheid. Aangezien deze regio nog steeds meerdere duizenden variaties telt in onze sequentie data passen we nogmaals een strategie toe ter prioritering. We doen dit door op systematische schaal voor iedere variatie te bepalen of dragerschap van deze variatie op chr13q34 in de sequentiedata samen gaat met laag fT3 een marker van de trage schildklierfunctie. Deze analyse laat zien dat dragers van het zeldzame C allel van de variant rs9515460 een veel lagere bloedspiegel van het ongebonden schildklierhormoon hebben. Mogelijkerwijs verklaart deze variant dus zowel de trage schildklier functie als de familiare langlevendheid.

In **hoofdstuk 6** van dit proefschrift presenteren we een nieuw R pakket, genaamd SATORi, waarmee vele van de integratieve berekeningen aan genetische data in dit proefschrift gedaan zijn. In dit pakket zijn enkele routines geïmplementeerd, die enkele van de meest uitgevoerde stappen in een integratieve analyse zouden moeten vergemakkelijken. Het nut en gebruik van dit pakket wordt geïllustreerd aan de hand van enkele publiekelijk beschikbare omics data sets.

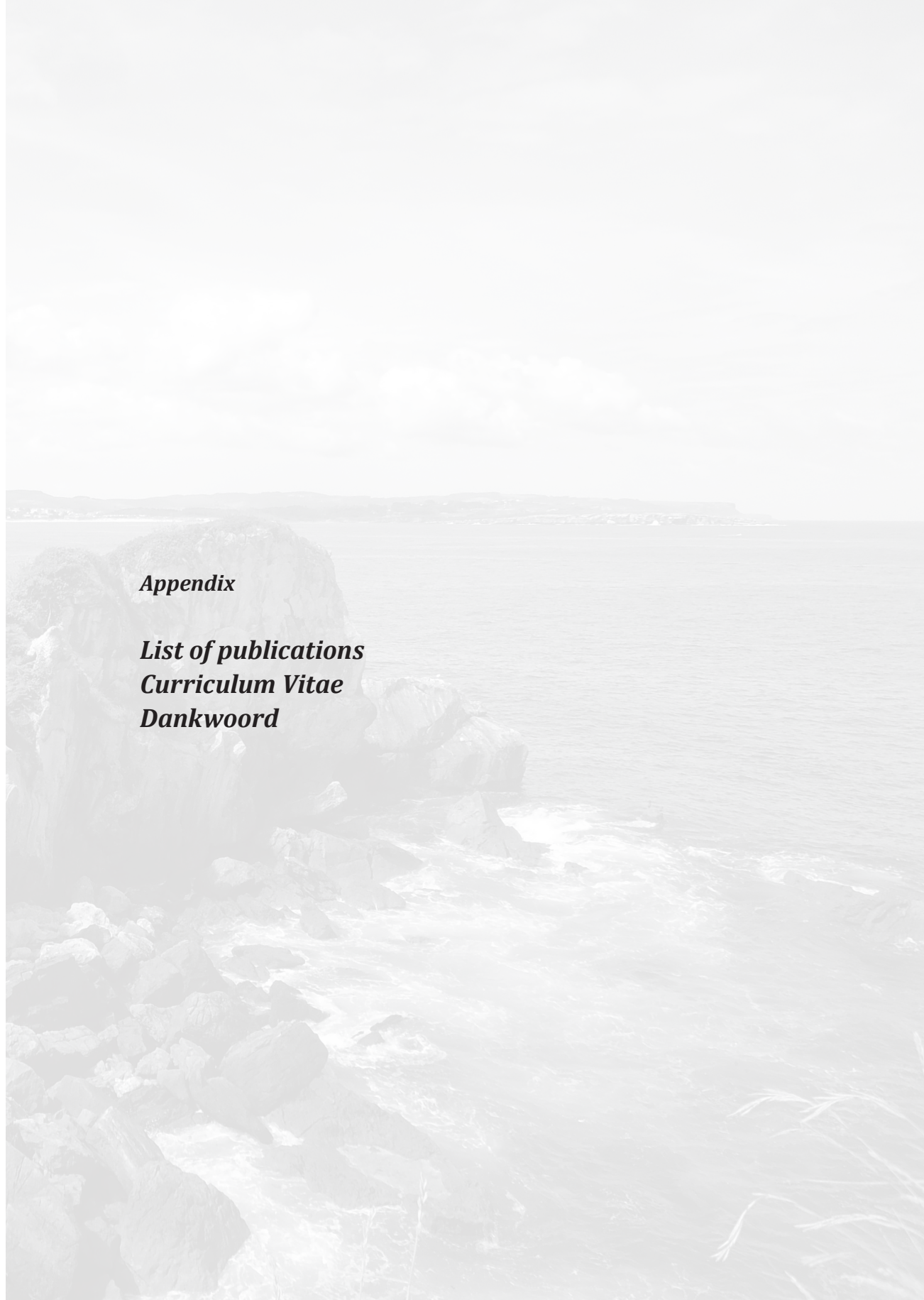
4. Conclusie en Toekomst

Het onderzoek in dit proefschrift is er op gericht om nieuwe integratieve analyse methoden te ontwikkelen en toe te passen, teneinde een beter en completer beeld te krijgen van de moleculaire biologie van veroudering. Het proefschrift bevat dus resultaten van twee soorten. Ten eerste heeft dit proefschrift integratieve methodologie opgeleverd voor een gecombineerde analyse van meerdere genexpressie datasets. Toepassing daarvan heeft enkele robuuste en interpreteerbare moleculaire profielen opgeleverd, die verder onderzocht kunnen worden voor hun potentie als biomarker voor biologische veroudering. Ten tweede hebben we gewerkt aan methodologie voor een integratieve analyse van genetische componenten die predisponeren voor een vertraagd verouderingsproces. Dit laatste bleek moeilijker dan verwacht, niet alleen door de omvang en complexiteit van de geanalyseerde sequentie data, maar ook doordat er van het verouderingsproces op hoge leeftijd zelf weinig relatief weinig bekend is. Desalniettemin hebben we toch enkele zeer interessante observaties gedaan met betrekking tot het verouderingsproces of de mogelijke vertraging daarvan. Enerzijds vinden we in de genomen van gezonde ouderen een hoge frequentie van varianten die tijdens het leven zijn ontstaan, die vermoedelijk de samenstelling van het bloedbeeld beïnvloeden. Anderzijds vinden we dat de genetische locatie chr13q34 waarschijnlijk bijdraagt aan langlevendheid door de schildklier functie iets te matigen.

Om verder progressie in het verouderingsveld te boeken is niet alleen meer onderzoek naar integratieve analyses nodig, maar kan er ook veel winst geboekt worden met de manier waarop data gemeten wordt. Zo wordt verwacht dat vernieuwingen in de experimentele methodologie van sequentie metingen de kwaliteit van zowel de genetische als de genexpressie datasets sterk zal verbeteren. Tevens kan de waarde van de gemeten data nog sterk vergroot worden door een nog slimmere experimentele opzet te kiezen. Niet alleen kan bijvoorbeeld de invloed van de interindividuele variatie in de data sterk gereduceerd worden door dezelfde mensen herhaaldelijk te meten. Ook kan door middel van een uitgebreide fenotypering of inspectie van de sterfte en geboortecijfers van de familie over meerdere generaties uitgezocht worden in welke familie de genetische component voor langlevendheid het sterkst aanwezig is. Beide methoden worden momenteel in onze groep toegepast in de acquisitie van nieuwe data over veroudering. Een andere trend waarvan het verouderingsveld kan profiteren is dat er steeds meer grote omics datasets publiekelijk beschikbaar worden gemaakt, zoals bijvoorbeeld door het biobank initiatief BBMRI. Met behulp van deze data bronnen kan steeds beter onderzocht worden wat de samenhang is tussen de verschillende soorten metingen zonder dat daar eigen data gecreëerd voor hoeft te worden. Een uitdaging voor de nabije toekomst ligt in het identificeren van gecombineerde biomarker profielen, die rekening houden met meerdere aspecten van veroudering, bijvoorbeeld

cognitieve of cardiovasculaire, en zo dus een getrouwer beeld geeft van de algemene biologische leeftijd. Ook ligt er een uitdaging in het efficiënt incorporeren van informatie over veroudering uit dier modellen teneinde onze zoektocht naar de moleculaire biologie van menselijke veroudering te versnellen.

In dit proefschrift hebben we het belang van data integratie technieken gedemonstreerd voor het doen van onderzoek naar de moleculaire biologie van veroudering. Alhoewel we met behulp van integratieve data analyses enkele interessante aspecten van veroudering hebben ontdekt blijven er echter nog vele hiaten bestaan in ons begrip over veroudering in relatie tot gezondheid. Idealiter zouden we onze kennis over veroudering het beste kunnen verdiepen door meer omics data met meerdere soorten meetmethoden in meerdere weefsels op meerdere tijdstippen in dezelfde personen te meten. De ontwikkeling van integratieve analyse technieken zal onmisbaar zijn voor de analyse van dergelijke rijke en complexe data bronnen en zal in de toekomst dan ook een doorslaggevende rol spelen in het onderzoek naar de moleculaire biologie van veroudering.



Appendix

List of publications

Curriculum Vitae

Dankwoord

List of Publications

EB van den Akker, B Verbruggen, BT Heijmans, M Beekman, JN Kok, PE Slagboom, MJT Reinders. Integrating protein-protein interaction networks with gene-gene co-expression networks improves gene signatures for classifying breast cancer metastasis. *Journal of integrative bioinformatics* 2011; 8, 188.

S Babaei, **EB van den Akker**, J de Ridder, MJT Reinders. Integrating protein family sequence similarities with gene expression to find signature gene networks in breast cancer metastasis. *Pattern Recognition in Bioinformatics: Springer* 2011; pp. 247-259.

ML Sampietro, S Trompet, JJW Verschuren, RP Talens, J Deelen, BT Heijmans, RJ de Winter, RA Tio, PAFM Doevendans, SK Ganesh, EG Nabel, HJ Westra, L Franke, **EB van den Akker**, RGJ Westendorp, AH Zwinderman, A Kastrati, W Koch, PE Slagboom, P de Knijff, JW Jukema. A genome-wide association study identifies a region at chromosome 12 as a potential susceptibility locus for restenosis after percutaneous coronary intervention. *Human molecular genetics* 2011; 20, 4748-4757.

J Deelen, M Beekman, HW Uh, Q Helmer, M Kuningas, L Christiansen, D Kremer, R van der Breggen, HED Suchiman, N Lakenberg, **EB van den Akker**, WM Passtoors, H Tiemeier, D van Heemst, AJ de Craen, F Rivadeneira, EJ de Geus, M Perola, FJ van der Ouderaa, DA Gunn, DI Boomsma, AG Uitterlinden, K Christensen, CM van Duijn, BT Heijmans, JJ Houwing-Duistermaat, RGJ Westendorp, PE Slagboom. Genome-wide association study identifies a single major locus contributing to survival into old age; the APOE locus revisited. *Aging cell* 2011; 10, 686-698.

PE Slagboom, M Beekman, WM Passtoors, J Deelen, AAM Vaarhorst, JM Boer, **EB van den Akker**, D van Heemst, AJM de Craen, AB Maier, M Rozing, SP Mooijaart, BT Heijmans, RGJ Westendorp. Genomics of human longevity.

Philosophical Transactions of the Royal Society B: Biological Sciences 2011; 366, 35-42.

WM Passtoors, Judith M Boer, JJ Goeman, **EB van den Akker**, J Deelen, BJ Zwaan, A Scarborough, R van der Breggen, RHAM Vossen, JJ Houwing-Duistermaat, GJB van Ommen, RGJ Westendorp, D van Heemst, AJM de Craen, AJ White, DA Gunn, M Beekman, PE Slagboom. Transcriptional profiling of human familial longevity indicates a role for ASF1A and IL7R. *PLoS one* 2012; 7, e27759.

J Deelen, HW Uh, R Monajemi, D van Heemst, PE Thijssen, S Böhringer, **EB van den Akker**, AJM de Craen, F Rivadeneira, AG Uitterlinden, RGJ Westendorp, JJ Goeman, P E Slagboom, JJ Houwing-Duistermaat, M Beekman. Gene set analysis of GWAS data for human longevity highlights the relevance of the insulin/IGF-1 signaling and telomere maintenance pathways. *Age* 2013; 35, 235-249.

M Beekman, H Blanché, M Perola, A Hervonen, V Bezrukov, E Sikora, F Flachsbarth, L Christiansen, AJM Craen, TBL Kirkwood, IM Rea, M Poulain, JMRS Valensin, MA Stazi, G Passarino, L Deiana, ES Gonos, L Paternoster, TIA Sørensen, QT, Q Helmer, **EB van den Akker**, J Deelen, F Martella, HJ Cordell, KL Ayers, JW Vaupel, O Törnwall, TE Johnson, S Schreiber, M Lathrop, A Skytthe, RGJ Westendorp, K Christensen, J Gampe, A Nebel, JJ Houwing-Duistermaat, PE Slagboom, C Franceschi. Genome-wide linkage analysis for human longevity: Genetics of Healthy Aging Study. *Aging Cell* 2013; 12(2):184-93.

RC Slieker, SD Bos, JJ Goeman, JV Bovée, RP Talens, R van der Breggen, HED Suchiman, EW Lameijer, H Putter, **EB van den Akker**, Y Zhang, JW Jukema, PE Slagboom, I Meulenbelt, BT Heijmans. Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array. *Epigenetics & chromatin* 2013; 6, 26.

K Ye, M Beekman, EW Lameijer, Y Zhang, MH Moed, **EB van den Akker**, J Deelen, JJ Houwing-Duistermaat, D Kremer, SY Anvar, JFJ Laros, D

- Jones, K Raine, B Blackburne, S Potluri, Q Long, V Guryev, R van der Breggen, RGJ Westendorp, PAC t Hoen, J den Dunnen, GJB van Ommen, G Willemsen, SJ Pitts, DR Cox, Z Ning, DI Boomsma, PE Slagboom. Aging as accelerated accumulation of somatic variants: whole-genome sequencing of centenarian and middle-aged monozygotic twin pairs. *Twin Research and Human Genetics* 2013; 16, 1026-1032.
- N Bomer, W den Hollander, YFM Ramos, SD Bos, R van der Breggen, N Lakenberg, BA Pepers, AE van Eeden, A Darvishan, EW Tobii, BJ Duijnsveld, **EB van den Akker**, BT Heijmans, WMC van Roon-Mom, FJ Verbeek, GJVM van Osch, RGHH Nelissen, PE Slagboom, I Meulenbelt. Underlying molecular mechanisms of DIO2 susceptibility in symptomatic osteoarthritis. *Ann Rheum Dis* 2014; 204739.
- J Deelen, M Beekman, HW Uh, L Broer, KL Ayers, Q Tan, Y Kamatani, AM Bennet, R Tamm, S Trompet, DF Guðbjartsson, F Flachsbar, G Rose, A Viktorin, K Fischer, M Nygaard, HJ Cordell, P Crocco, **EB van den Akker**, S Böhringer, Q Helmer, CP Nelson, GI Saunders, MA, K Andersen-Ranberg, ME Breen, R van der Breggen, A Caliebe, M Capri, E Cevenini, JC Collerton, S Dato, K Davies, I Ford, J Gampe, P Garagnani, EJC de Geus, J Harrow, D van Heemst, BT Heijmans, FA Heinsen, JJ Hottenga, A Hofman, B Jeune, PV Jonsson, M Lathrop, D Lechner, C Martin-Ruiz, SE McNerlan, E Mihailov, A Montesanto, SP Mooijaart, A Murphy, EA Nohr, L Paternoster, I Postmus, F Rivadeneira, OA Ross, S Salvioli, N Sattar, S Schreiber, H Stefánsson, DJ Stott, H Tiemeier, AG Uitterlinden, RGJ Westendorp, G Willemsen, NJ Samani, P Galan, TIA Sørensen, DI Boomsma, JW Jukema, IM Rea, G Passarino, AJM de Craen, K Christensen, A Nebel, K Stefánsson, A Metspalu, P Magnusson, H Blanché, L Christiansen, TBL Kirkwood, CM van Duijn, C Franceschi, JJ Houwing-Duistermaat, PE Slagboom. Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. *Human molecular genetics* 2014, ddu139
- N van Leeuwen, M Beekman, J Deelen, **EB van den Akker**, AJM de Craen, PE Slagboom, LM 't Hart. Low mitochondrial DNA content associates with familial longevity: the Leiden Longevity Study. *Age* 2014, 1-8.
- JJ Houwing-Duistermaat, Q Helmer, B Balliu, **EB van den Akker**, R Tsonaka, HW Uh. Gene analysis for longitudinal family data using random-effects models. *BMC proceedings* 2014. 8, S88.
- W den Hollander, YF Ramos, SD Bos, N Bomer, R van der Breggen, N Lakenberg, R Sliecker, R Luijk, EW Tobii, BJ Duijnsveld, **EB van den Akker**, BT Heijmans, PE Slagboom, RG Nelissen, I Meulenbelt. Genome wide DNA methylation profiling of osteoarthritic articular cartilage. *Osteoarthritis and Cartilage* 2014. 22, S40-S41.

Curriculum Vitae

Erik Ben van den Akker was born on June 11, 1981, in 's-Hertogenbosch, the Netherlands. In 2006 he obtained his Bachelors degree in Microbiology at the Fontys Hogescholen Eindhoven. His Bachelor thesis project titled 'Copy Number Variations in Immune Related Genes' was conducted at the Leiden University Medical Centre at the department of Human Genetics. He continued his education with a Master in Life Science & Technology, a shared program between Delft University of Technology and Leiden University from which he graduated *cum laude* in 2009. This time, his thesis work titled 'CRM Finding in Higher Eukaryotes' was conducted within The Delft Bioinformatics Lab, headed by Professor dr. Marcel Reinders at the Delft University of Technology. After graduation, he stayed as a PhD student within the Delft Bioinformatics Lab in a shared project with Professor dr. Eline Slagboom of the section Molecular Epidemiology at the Leiden University Medical Centre. His PhD work was funded by the Medical Delta in a project aimed at developing integrative algorithms for a comprehensive and robust analysis of omics data sources, for elucidating the determinants of healthy aging and longevity within the Leiden Longevity Study. The results of this research are outlined in this thesis. Currently he is employed as a post-doctoral researcher within the section of Molecular Epidemiology and still is a member of the Reinders group. His current work focuses on developing integrative algorithms for the analysis of cross-species omics data on aging and development created within the IDEAL consortium.

Dankwoord

Het is zo ver! Er ligt een proefschrift! Met veel plezier kijk ik terug op mijn promotie traject en zou bij dezen graag de mensen willen danken die op enige wijze hebben bijgedragen aan deze zeer leuke en leerzame periode in mijn leven.

Geachte Professoren Slagboom en Reinders, beste Eline en Marcel, het vergt moed, vertrouwen en doorzettingsvermogen om nieuwe bruggen te slaan tussen twee wetenschappelijke velden. Dank voor jullie onvoorwaardelijke vertrouwen, de vele leuke wetenschappelijke discussies en bovenal ook voor de vrijheid die jullie me lieten om problemen naar eigen inzicht op te lossen. Geachte Dr. Beekman, beste Marian, fijn zo'n rots in de branding! Dank voor al je advies, wetenschappelijk of over het leven daarbuiten, je geduld en gezelligheid. Dank ook aan Professor Kok en Dr. Heijmans voor de vele nuttige discussies.

Collega's van de sectie Moleculaire Epidemiologie in Leiden, dank voor de leuke tijd, de vele gezellige borrels en wetenschappelijke discussies. Dank aan de analisten en in het speciaal aan Nico, Wesley en Eka voor het vele werk in het lab ter validatie van mijn bevindingen. Dank aan Inge en Caroline voor de administratieve ondersteuning en ook dank aan mijn studenten, Bas, Renske en Adam voor jullie inzet en bijdragen. Graag wil ik mijn huidige en vorige kamergenoten bedanken voor de vele gezellige uurtjes samen ploeteren aan onze proefschriften. Dank aan Elmar voor de vele interessante papers en trein discussies. Speciale dank aan Joris, Matthijs en Eric-Wubbo voor alle hulp bij de analyses en het fungeren als sparringpartner.

Collega's van de sectie bioinformatica in Delft, dank voor alle verfrissende inzichten, de leuke wetenschappelijke discussies en de gezellige borrels. In het bijzonder dank aan Marc, Erdogan en Thies voor de vele vakoverstijgende discussies.

Graag zou ik ook mijn dank voor een fijne samenwerking willen uitspreken aan mensen van buiten de secties waarop ik werkzaam ben. Allereerst de naaste collega's van sectie Medische Statistiek in het LUMC; Professoren Houwing-Duistermaat en Goeman, Erik (van Zwet), Roula, Szymon, Hae-Won en Ramin dank voor jullie advies en uitleg. Ook dank aan de naaste collega's van de sectie Pattern Recognition aan de TUDelft voor advies, uitleg en de gezellige tijd. Dank aan Dr. Jansen, Dr. Willemsen en Professoren Boomsma en Penninx van het VUmc voor de fijne samenwerking. Hierbij speciale dank aan Rick voor je snelle maar gedegen analyses en kritische vragen. Ook dank aan Joost, Tina en Jelle van het WUR voor de leuke discussies binnen IDEAL. Special thanks to Dr. Pitts and Dr. Potluri of Rinat-Pfizer for your continuous efforts and collaboration on investigating the genetics of human longevity.

Uiteraard ook dank aan mijn familie en schoonfamilie voor jullie onvoorwaardelijke steun, begrip en vertrouwen. Als laatste en meest belangrijke zou ik graag mijn lieve Manon willen bedanken. Dank voor je oneindige geduld, je begrip en steun. Op de vraag: "Wanneer ben je nu eindelijk eens klaar?", zeg ik: "Hopelijk nooit". Dat we samen nog veel van onze zoon Hugo mogen genieten!