*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 outcome[1-6]. 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 blood[7-9] (~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 controls[12]. 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 LLS[13] 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 organisms[18]. 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-hydroxymethylcytosine[19], and *DNTM3A* is a DNA Cytosine-5-Methyltransferase 3 that is involved in *de novo* methylation[20], 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 approach[27] 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 age[28], 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 field[32] 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 mouse[33], 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 (STRING[37]) 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-weighing 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 ANNOVAR[39] 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-weighed 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 advice 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 enhancers[41], 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 consortium[42] (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 promotor 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-signalling[45], 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

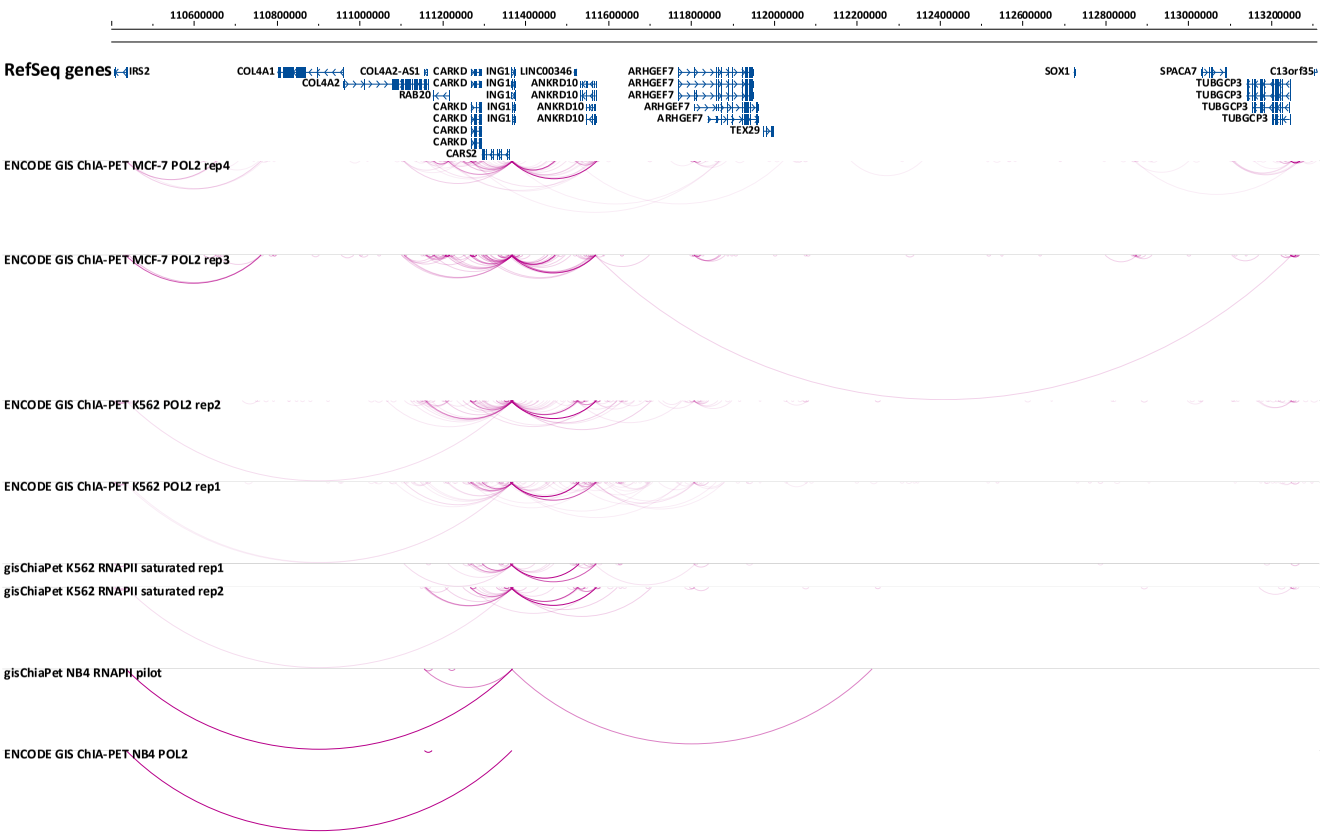


**Figure 1: Incorporating DNA-DNA interactions as a source of prior knowledge.** This figure illustrates the potential merit of incorporating ChIA-PET data created by the ENCODE[42] consortium for the prioritization of variants. Depicted is the 1-LOD-drop region on chr13q34 (**Chapter 5** of this thesis) in which a lot of crosstalk is seen for the intergenic region between *CARS2* and *IRS2* and other genes within this region, including the promotor 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* gene[48], 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 reads[49]. 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 cell[50]. 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 individuals[54-56], 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 findings[57]. 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 large[58], 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-lived[13]. 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 pathways[44]. The comprehensive assessment of ones 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 level[59], whereas aging of the neuromuscular system is marked by atrophy of muscle and neuronal cells[60]. 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 GenAge[61], 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 extent life span upon knockdown serve as input for further research into specific cohorts suitable for studying biological aging in humans.

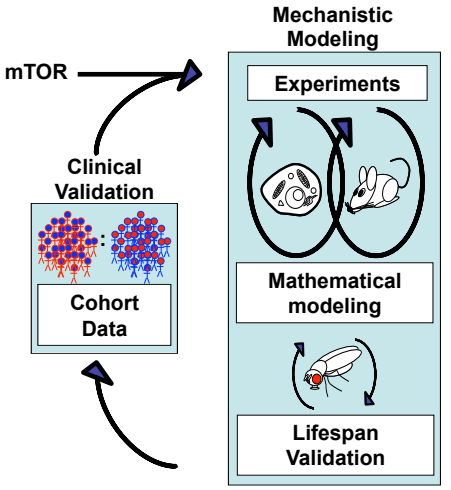


FIGURE 2: THE PARADIGM OF SYSTEMS BIOLOGY APPLIED TO AGING RESEARCH.

## 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 GenomEUtwin. *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).