FACULTY OF SCIENCES

FACULTY OF BIOSCIENCE ENGINEERING

FACULTY OF ENGINEERING AND ARCHITECTURE

HUMAN TRANSCRIPTION FACTOR ANALYSIS IN THE CONTEXT OF AGING

Word count: 29494

Simon Plovyt
Student number: 01204786

Promoter(s): Prof. Dr. Tim De Meyer

Supervisor(s): Prof. Dr. Hao Li and Dr. Jiashun Zheng

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Master of Science in Bioinformatics: Bioscience Engineering

Academic Year: 2017- 2018

GHENT UNIVERSITY

1 Acknowledgements

This project would not have been possible without Prof. Dr. Hao Li and Dr. Jiashun Zheng at the University of California, San Francisco (UCSF). Under the expert guidance of Hao Li, the Li lab grew over 50% in the last few months, shaping up into one coherent unit with world-class researchers at every pillar. Labs like these are single-handedly responsible for the global success and reputation of UCSF, next to providing an excellent incubator for research projects similar to the one documented in this manuscript.

Secondly, many thanks to Prof. Dr. Tim De Meyer for paving the way for this project and providing expert opinion on every aspect of my master's dissertation. I truly appreciate the trust and freedom given, also in cooperation with Ghent University.

Above all, the success of this project can be mainly attributed my parents, my grandparents and my brother for working to the best of their abilities to provide their unconditional support around the clock. I am extremely grateful for everything they have given me, and I will continue to do everything in my power to return the favor. No matter how far my endeavors may take me, there is no place like home.

Contents

2 Abbreviations

AUROC: area under the receiver operating characteristic

bp: basepairs

DBD: DNA-binding domain

DNA: deoxyribonucleic acid

FDR: false discovery rate

FIMO: find individual motif occurrences

FPR: false positive rate

FWER: family-wise error rate

FOXO: Forkhead box O

GO: gene ontology

HSP: heat-shock protein

IC: information content

IGF: Insulin-like growth factor

IIS: insulin and IGF-1 signaling

JAK: Janus kinase

LPS: lipopolysaccharide

MoSBAT: motif similarity based on affinity of targets

mtDNA: mitochondrial deoxyribonucleic acid

ncRNA: non-coding RNA

NHR: nuclear hormone receptor

NRF: nuclear respiratory factor

PCC: Pearson correlation coefficient

PFM: position frequency matrix

PLM: probe-level linear model

PPM: position probability matrix

PSAM: position-specific affinity matrix

PSSM: position-specific scoring matrix

PSWM: position-specific weight matrix

PWM: position weight matrix

qPCR: quantitative polymerase chain reaction

RC: reverse complement

RNA: ribonucleic acid

RNA-seq: RNA sequencing

ROS: reactive oxygen species

RPKM: reads per kilobase of exon per million mappable reads

scRNA-seq: single cell RNA sequencing

siRNA: small interfering RNA

STAT: signal transducer and activator of transcription

SVM: support vector machine

TF: transcription factor

TFBS: transcription factor binding site

TSS: transcription start site

UPGMA: unweighted pair group method with arithmetic mean

3 Abstract

Over time, life becomes increasingly vulnerable to death. This time-dependent functional decline known as aging goes hand in hand with a progressive loss in physiological integrity. Many pathologies accompany aging individuals, either as the cause, result, or accelerating factor. Cardiovascular diseases and cancer top the rankings of the leading causes of death with a combined estimated fraction of 46.8%, and remain largely inevitable to this day. Due to the scale and impact of aging, many research studies aim to take a step closer towards the symbolic fountain of youth. In this manuscript, a novel algorithm and corresponding web tool is introduced to reliably identify relevant transcription factors related to aging. With this methodology, gene expression analyses can be extended by returning information on the regulatory level. The identified transcription factors at the base of the changes in the transcriptome present a second, extremely valuable layer of insight. Validation of the implementation on a transcription factor perturbation data set revealed that even slight reductions in gene expression for a sole gene could be picked up with considerable accuracy. The proposed methodology was applied in the context of aging, but can easily be generalized to all human gene expression experiments. Essential and possibly causal aging-related transcription factors could be identified for which drugs may be developed. Additionally, during the development of the algorithm, several insights regarding theoretical binding of transcription factors were found. Besides analyzing the list of most relevant transcription factors, recurring motif patterns and functional categories could be added to the list of discoveries. Finally, a combinatorial analysis which effectively integrated several gene expression experiments provided strong evidence for the contribution of several recurring transcription factors to aging. Among others, the transcription factor POU2F1 (alias Oct-1) was consistently returned in aging experiments. POU2F1 is known to be related to *herpes simplex*, which recently has been linked to Alzheimer's disease, suggesting an involvement of POU2F1 in aging.

4 Abstract – Dutch

Ouderdom gaat gepaard met een steeds groter wordende vatbaarheid voor de dood. Deze tijdsafhankelijke functionele afname, 'veroudering' genoemd, gaat gepaard met een progressief verlies in fysiologische integriteit. Een variëteit aan aandoeningen vergezelt deze oudere levensvormen, ofwel als oorzaak, als gevolg, of als versnellende factor. Cardiovasculaire ziektes en kanker zijn de doodsoorzaken bij uitstek met een gecombineerd aandeel van 46.8%. Helaas zijn deze ziektes tot vandaag de dag de norm en zo goed als onvermijdbaar. De schaal en impact van verouderingsverschijnselen is enorm, waardoor omvangrijk onderzoek wordt opgestart met als doel een waardevolle stap te zetten richting het ontdekken van de symbolische fontein van het eeuwige leven. In dit manuscript wordt een nieuw algoritme toegelicht, samen met de bijhorende online-beschikbare technologie. Deze methodologie heeft als doel wetenschappers toe te laten om relevante transcriptiefactoren te identificeren in genexpressie contrasten. Met deze methode kan de hoeveelheid informatie die omvat wordt in de gebruikelijke genexpressie experimenten naar een ongezien niveau getild worden door middel van de informatie op het vlak van genregulatie weer te geven. Doordat genregulatie aan de basis van genexpressie ligt, is deze informatie van onschatbare waarde. Een validatie-experiment op een transcriptiefactor-verstoringsexperiment toont aan dat het voorgestelde algoritme in staat is om zelfs kleine reducties in genexpressie voor een enkel gen te registreren en de causale transcriptiefactor aan te duiden met een aanzienlijke accuraatheid. De methode werd toegepast in de context van experimenten gerelateerd aan veroudering, maar de achterliggende technologie kan probleemloos veralgemeend worden tot alle genexpressie experimenten met betrekking tot de mens. De geïdentificeerde transcriptiefactoren worden relevant geacht in het experiment en zijn daarom uitstekende doelwitten voor het ontwikkelen van anti-ouderdom drugs. Verder werden tijdens het ontwikkelen van het algoritme belangrijke inzichten verworven, waarvan de meerderheid betrekking heeft op de binding van transcriptiefactoren met DNA. Buiten het analyseren van potentieel belangrijke transcriptiefactoren worden ook terugkerende motiefpatronen en functionele categorieën besproken. Ten laatste werd een analyse voltooid waarin informatie uit verscheidene verouderingsexperimenten werd gecombineerd. Hieruit werd een lijst gevonden die de consensus aan belangrijke transcriptie factoren in een verouderingsproces weergeeft. Onder andere werd het sterke voorkomen van de transcriptiefactor POU2F1 (alias Oct-1) opgemerkt. POU2F1 is gekend om zijn link met het *herpes simplexvirus*, wat onlangs gerelateerd werd aan de ziekte van Alzheimer. Het laatste suggereert aldus een belangrijke connectie tussen de POU2F1 transcriptiefactor en veroudering.

5 Introduction

5.1 Aging

Living organisms are characterised by a time-dependent functional decline, broadly termed as 'aging' [López-Otín et al., 2013]. The importance of research in the domain of aging does not require explanation. Associated diseases such as arthritis (affects 49.7% of all adults over 65), cardiovascular diseases (37% of men and 26% of women 65 and older) and cancer (28% of men and 21% of women over age 65 are living with cancer) are unfortunately common and in many cases inevitable up until this day, as estimated by the above statistics provided by the Centers for Disease Control and Prevention (CDC) in 2014 [Sherry L. Murphy et al., 2017]. Heart disease and cancer top the rankings of the leading causes of deaths in the United States, where the average life expectancy at birth is estimated at 78.8 years [Sherry L. Murphy et al., 2017].

As outlined in the aims section, the main goal is the identification of transcription factors regulating the aging process. Therefore, in the first part of the introduction section, we focus on the molecular biology of aging, whereas in the second part we elaborate on transcription factors and the identification of candidate targets by means of motif scanning.

5.1.1 Aging on a molecular level

Many ancient tales document the search for the fountain of youth. However, perhaps the true answer can be found on a molecular level. The widely considered cause of aging is assumed to be the gradual accumulation of cellular damage. Next to aging, this same underlying process can also manifest itself under the form of cancer, emphasizing the causal similarities in both processes [López-Otín et al., 2013]. Uncontrolled cellular overgrowth or hyperactivity are therefore associated with aging. These processes often manifest as atherosclerosis and inflammation, amongst other pathologies [Blagosklonny, 2008].

In general, nine candidate hallmarks that categorize the aging process have been proposed [López-Otín et al., 2013]: genomic instability, telomere attrition, epigenetic alterations, loss of proteostasis, deregulated nutrient sensing, mitochondrial dysfunction, cellular senesence, stem cell and exhaustion, and altered intercellular communication. The corresponding article, 'The hallmarks of aging' by C López-Otín (2013), was used as a guiding line for this section. In the light of this manuscript, only the most relevant hallmarks are briefly introduced.

Genomic instability

The integrity and stability of deoxyribonucleic acid (DNA) is constantly being put to the test. Both exogenous factors such as physical, chemical and biological agents and endogenous factors, such as replication errors, malfunction of proteins, spontaneous hydrolytic reactions and

Figure 1: The nine hallmarks of aging by C López-Otín (2013). Graphical representation of the nine hallmarks of aging. Throughout the first part of the introduction, the publication by C López-Otín (2013) will be used as a guideline. The rights and credits of this image belong to the original authors.

reactive oxygen species (ROS) can interfere. Genetic damage can take different shapes and is often observed as either point mutations, chromosomal gains and losses, translocations, gene disruption or shortening of the telomeres. These harmful effects are ideally kept to a minimum and are usually dealt with by efficient repair mechanisms [Lord and Ashworth, 2012].

Genome instability can also be caused by defects in the nuclear lamina. Progerin, a prelamin A isoform, has been linked to human aging [Ragnauth et al., 2010]. Moreover, the formation of progerin is also detected together with telomere dysfunction in normal human fibroblasts, suggesting a relationship between telomere maintenance and progerin expression [Cao et al., 2011].

Telomere attrition

Eukaryotic chromosomes are almost always linear. Exceptions have been found in yeast mutants but are generally very rare [Naito et al., 1998]. In contrast to circular DNA, linear DNA has designated ends. During DNA replication, the DNA polymerases fail to completely replicate the terminal ends of the DNA molecules. As a result, a small fraction of the molecule is lost during every replication cycle. To reduce the effect of this progressive and cumulative loss, the cell provides a protective cap at the ends named telomeres. In addition, these telomeres make distinguishing between double-stranded fractions and chromosome ends possible. Because telomerase is not expressed in most mammalian somatic cells, exhaustion of telomeres leads to DNA damage and goes by the term 'replicative senescence'. Related to replicative senescence

is the Hayflick limit, i.e. the amount of times a somatic cell can divide before the telomeres are exhausted and replicative senescence can be observed [Hayflick and Moorhead, 1961]. Previous studies in mice have shown that shortened telomeres exhibit decreased lifespans. The reverse is also true in the sense that increased lifespans are achieved by lengthened telomeres [Armanios et al., 2009, Blasco et al., 1997, Rudolph et al., 1999, Tomás-Loba et al., 2008] and that aging could be reverted by the activation of the ribonucleoprotein telomerase, effectively adding telomere caps [Jaskelioff et al., 2010]. Telomerase reverse transcriptase (TERT or hTERT in humans) is a catalytic subunit of the enzyme telomerase and is the focus of many research studies [Ducrest et al., 2002]. Absence of hTERT is associated with the *Dyskeratosis congenita* syndrome, also called the Zinsser-Cole-Engman syndrome.

Loss of Proteostasis

Protein homeostasis, also termed proteostasis, tends to be associated with aging [Powers et al., 2009, Koga et al., 2011]. Unfolded, misfolded or aggregated proteins are known to be causal to age-related diseases, such as Parkinson's and Alzheimer's [Powers et al., 2009].

Additionally, the discovery that the mTOR inhibitor rapamycin has the potential to increase the longevity of middle-aged mice also sparked great interest in the research community [Blagosklonny, 2011, Harrison et al., 2009]. However, as shown in yeast, nematodes and flies, the lifespan-extending effect of rapamycin is dependent on autophagy induction [Bjedov et al., 2010, Rubinsztein et al., 2011], although similar supporting evidence does not yet exist with regard to mammalian aging. The effect of autophagy and proteostasis could be linked to neurogenerative disease in humans [Novack, 2017].

Deregulated Nutrient Sensing

The insulin and IGF-1 signaling (IIS) pathway is the most conserved aging-controlling pathway in evolution. Members of this pathway include the FOXO transcription factor family, it's targets, and the mTOR complexes. Clear evidence for an involvement in aging has been presented both in humans and in model organisms [Barzilai et al., 2012, Kenyon, 2010]. Moreover, caloric and general dietary restriction increases longevity in all investigated eukaryote species, including humans, and this effect is not to be underestimated [Colman et al., 2009, Fontana and Partridge, 2015, Bloomer and Lee, 2014]. The most relevant downstream effector of the IIS pathway is the FOXO transcription factor and is therefore the focus of many research studies in the field of aging [Kenyon et al., 1993, Slack et al., 2011]. A decreased ISS pathway is considered to be associated with aging and this assumption has been proven by decreasing growth hormone and IGF-1 levels in mice [Schumacher et al., 2008]. This observation can be explained intuitively by associating slower metabolism and slower cell growth with lower rates of cellular damage [López-Otín et al., 2013].

In addition to the ISS pathway, other nutrient-sensing systems exist such as the mTOR system for amino-acid concentrations, next to AMPK and sirtuins for low-energy states (high AMP and high

NAD+, respectively) [Houtkooper et al., 2010]. The AMPK and sirtuins systems sense nutrient scarcity and catabolism, whereas IIS and mTOR sense nutrient abundance and anabolism. The activity of mTOR is increased in older mouse hypothalamic neurons and downregulation of mTORC1 appears to be critical to mammalian longevity. Increased mTOR activity is found to contribute to age-related obesity and is reversed by injecting rapamycin to the hypothalamus. Shutting down TOR activity results in clear benefits during aging, however also has undesirable effects such as insulin resistance, impaired wound healing and more [Wilkinson et al., 2012].

Mitochondrial Dysfunction

In older individuals, the respiratory chain tends to have reduced efficacy. One of the causes is electron leakage, which results in diminished ATP generation [Green et al., 2011]. Moreover, mitochondrial dysfunction leads to increased reactive oxygen species (ROS). Multiple sources support a role for ROS in aging, although an unexpected relationship between increased ROS and prolonged lifespan in C. elegans and yeast was observed [Doonan et al., 2008, Mesquita et al., 2010, Raamsdonk and Hekimi, 2009]. Although ROS are generally damaging to the cell, data suggested that these reactive oxygen species trigger proliferation and survival pathways in the cell and produce a survival signal which has lead to the reconsideration of the role of these ROS in aging [Ristow and Schmeisser, 2011, Sena and Chandel, 2012]. However, the damaging effects remain largely undeniable.

Cellular Senescence

A popular way of identifying senescent cells is the use of markers that expose and quantify DNA damage, such as senescence-associated β-galactosidase (SABG). Quantification using this marker in liver tissue has shown an increase from 8% senescent cells to 17% for the comparison of young versus very old mice [Wang et al., 2009]. Similar conclusions were drawn for other tissues such as skin, spleen and lung, but not for heart, muscle and kidney tissue cells [Wang et al., 2009]. This observation suggests that different tissues operate on different 'tissue ages' and that age is therefore not to be generalized across different tissues. Removal of older cells by the immune system plays a major role in this process [Hoenicke and Zender, 2012, Kang et al., 2011, Xue et al., 2007].

Next to DNA damage, excessive mitogenic signaling is also associated to senescence. Oncogenic events and alterations could be linked to inducing or accelerating cellular aging [Gorgoulis and Halazonetis, 2010]. The most important pathways are the $p16^{INK4a}$/Rb and $p19^{ARF}$/p53 pathways [Serrano et al., 1997]. The INK4a/ARF locus encodes both $p16^{INK4a}$ and $p19^{ARF}$ and over 300 genome-wide association studies (GWAS) support the association between this locus and aging-related pathologies. Elimination of $p16^{INK4a}$ and p53 confirmed the hypothesis that these factors are responsible for pro-aging activity and physiological aging as a result [López-Otín et al., 2013]. However, more recent evidence shows a more complicated view. A slight increase in expression for $p16^{INK4a}$, $p19^{ARF}$ or p53 tumor suppressors results in a longer lifes-

pan, in contrast to their expected role in accelerating aging [Matheu et al., 2007]. Activity of INK4a/ARF and P53 tends to avoid the propagation of damaged cells and therefore suppresses aging and cancer, however when damage is pervasive, the regenerative capacity of tissues can be exhausted. In these extreme cases, the INK4a/ARF and p53 responses can accelerate aging and reduce longevity.

Epigenetic alterations

A variety of age-related epigenetic changes are causal to undesired cell functionality. These changes can involve, but are not limited to histone modification, methylation patterns or chromatin remodeling. The role of transcription factors is not to be underestimated with regards to proper expression and activity of epigenetic proteins such as methyltransferases, methylases, demethylases and histone acetylases, amongst others [Maleszewska et al., 2016, Sen et al., 2016].

Increased longevity in worms was shown by inhibition of histone demethylases for H3K27 by targeting key longevity pathways such as the insulin/IGF-1 signaling route [Jin et al., 2011]. Moreover, a number of studies in model organisms such as worms, yeast and flies show that members of the sirtuin family are involved in extending lifespan, and Sir2 for Saccharomyces cerevisiae in particular [Guarente, 2011].

Moreover, chromatin alterations are considered associated with aging because of the supporting evidence that loss-of-function (LOF) of the heterochromatin protein HP1α results in shortened lifespan, whereas overexpression extends lifespan in flies [Larson et al., 2012]. This observation was explained by the assumption that heterochromatin formation is important for chromosomal stability.

Finally, in aged cells, transcriptional noise is found to be increased and is associated with aberrant mRNA and miRNA production [Bahar et al., 2006, Liu et al., 2012, Smith-Vikos and Slack, 2012]. Comparisons between young and old cells reveal that aging-associated transcriptional signatures influence longevity by targeting longevity networks or by regulating stem cell behavior [Boulias and Horvitz, 2012, Toledano et al., 2012, Ugalde et al., 2011].

Altered intercellular communication

Increased inflammatory reactions in aged tissue, also termed 'inflammaging', is extremely prominent and strongly characterizes aging in mammals [Salminen et al., 2012]. Inflammatory pathways are linked to pathologies such as obesity, type II diabetes and atherosclerosis. Notably, these diseases are strongly correlated with human age [Tabas, 2009, Barzilai et al., 2012]. Malfunction of the immune system results in failure to remove pathogens and to clear malignant cells. This is partially caused by and accompanied by enhanced NF-κB activity and the occurrence of a dysfunctional autophagy system. These proinflammatory pathways result in increased production of interleukins such as IL-1β, tumor necrosis factors and interferons [Green

et al., 2011, Salminen et al., 2012]. The hypothesis of the importance of inflammatory pathways in aging is supported by a variety of studies [Magalhães et al., 2009, Lee et al., 2012]. Activity of the NF-κB signaling route is associated with aging and inhibiting this pathway causes phenotypic rejuvenation of skin tissue in transgenic mice in addition to restoring transcriptional signatures to these of young mice [Adler et al., 2007], emphasizing the importance of this pathway. Inflammation was also linked to aging through the mRNA decay factor AUF1 [Gratacós and Brewer, 2010]. This factor is involved in bringing the inflammatory response to a stop by degrading cytokine mRNA molecules. Inactivation of AUF1 in mice was found to accelerate aging by increasing inflammation, thus proving the correlation. In addition, AUF1 contributes to maintaining telomere length by positively influencing the expression of TERT, which is part of telomerase [Pont et al., 2012]. This observation emphasizes that a single factor is able to affect different aging hallmarks. Next to AUF1, a similar situation holds true for SIRT1, SIRT2 and SIRT6, which indirectly influences aging by limiting the expression of inflammation-related genes [Li et al., 2013]. In general, anti-inflammatory agents have proven to be effective in slowing down aging and healthy aging. This hypothesis is fully supported by longevity studies on mice and humans in which anti-inflammatory agents, such as aspirin, were applied [Rothwell et al., 2011, Strong et al., 2008], while also suggesting a role of the gut microbiome [Claesson et al., 2012, Ottaviani et al., 2011].

Relevance to this work

The general image of aging depicted by the described hallmarks strongly suggests an association between inter- and intracellular signaling events and the aging of cells and tissues. In particular, the importance of transcription factors in intracellular signaling pathways is undeniable and is the main focus of this thesis. As presented higher, transcription factors involved in extending or reducing the lifespan of mammals include the STAT protein family (signal transducers and activators of transcription), the FOXO family (Forkhead box O) [Martins et al., 2015], the NRF protein family (nuclear respiratory factor) [Zhang and Manning, 2015], and NF-κB signaling components, amongst others. This selection of components is highly relevant because their individual methods of action and signaling targets often cover and affect multiple aging hallmarks at once. Moreover, about 10% of currently prescribed drugs directly target the nuclear receptor class of transcription factors [Overington et al., 2006].

STAT factors are known to be involved in the well documented JAK/STAT pathway (Janus kinase / signal transducers and activators of transcription). A multitude of development and homeostasis signals are transduced throughout this pathway and cytokines and growth factors are part of the interaction. Important adjustments of the cell are precisely regulated by this pathway, such as immunity, cell division and death, tumor formation and general growth processes [Rawlings, 2004]. Although consisting of relatively few components [Aaronson, 2002, Kisseleva et al., 2002], malfunction of this signaling pathway is severe and can lead to a variety of diseases such as cancers, immune system malfunction, inflammation and skin conditions (e.g. psori-

asis) [Aaronson, 2002]. Expression levels of JAK-STAT signaling targets were found to be increased with age and more specifically, STAT3 was associated with cardiovascular pathologies in older mice [Jacoby et al., 2003]. The transcription factors STAT3 and STAT1 take effect on downstream genes in inflammatory responses together with NF-κB. A recent study on human kidney tissue highlights changes in inflammatory transcriptional patterns during aging [O'Brown et al., 2015].

FOXO proteins represent a conserved subfamily of transcription factors, which is known to affect longevity by regulating signaling processes including the insulin-like growth factor signaling route (IGF-signaling), amongst others. Invertebrate genomes include a sole FOXO gene, whereas mammal genomes have four FOXO genes: FOXO1, FOXO3, FOXO4 and FOXO6. These factors take effect in crucial cellular processes such as cell cycle, stress resistance, metabolism and apoptosis [Martins et al., 2015]. For example, the FOXO6 protein is closely involved in modifying histones and chromatin. Furthermore, the impact of FOXO6 is diminished during aging due to phosphorylation [Kim et al., 2014]. Additionally, FOXO1, FOXO3 and APOE were consistently revealed as longevity genes by several studies [Willcox et al., 2006, Anselmi et al., 2009, Flachsbart et al., 2009, Brooks-Wilson, 2013, Broer and Duijn, 2015]. The first aging hallmark to be described in animals was deregulated nutrient sensing through the insulin and IGF-1 signaling route (IIS) and is directly influenced by the FOXO proteins [Brunet et al., 1999, Dong et al., 2008]. Generally, FOXO proteins function as transcriptional activators and they tend to bind the consensus core motif TTGTTTAC.

Furthermore, the NF-κB signaling pathway has been proposed as one of the main mediators of aging. This route is activated by a variety of stress signals (genotoxic, oxidative, and inflammatory) and regulates expression of genes involved in cell senescence, inflammation, and apoptosis, next to cytokines and growth factors. Age-related pathologies such as Alzheimer's, diabetes or osteoporosis show increased activity of the NF-κB pathway [Tilstra et al., 2012]. Moreover, NF-κB activation is associated with other known lifespan regulators including, but not limited to, FOXO, SIRT, mTOR and the IIS pathway. Due to the prevalence NF-κB and the large amount of support for it's importance in aging, it presents a possible therapeutic target for increasing mammalian longevity [Tilstra et al., 2012].

5.2 Introduction to the methods and terminology

The general strategy of this analysis can be summarized as developing a methodology to scan the human genome with transcription factor motifs in order to derive potential targets of these transcription factors. The likelihood of these targets is characterized by a theoretical binding score derived from the genome scanning approach. If this method is proven to generate desirable results, the theoretical score vectors for the target genes can be integrated with fold change vectors of gene expression vectors by calculating the correlation. This correlation vec-

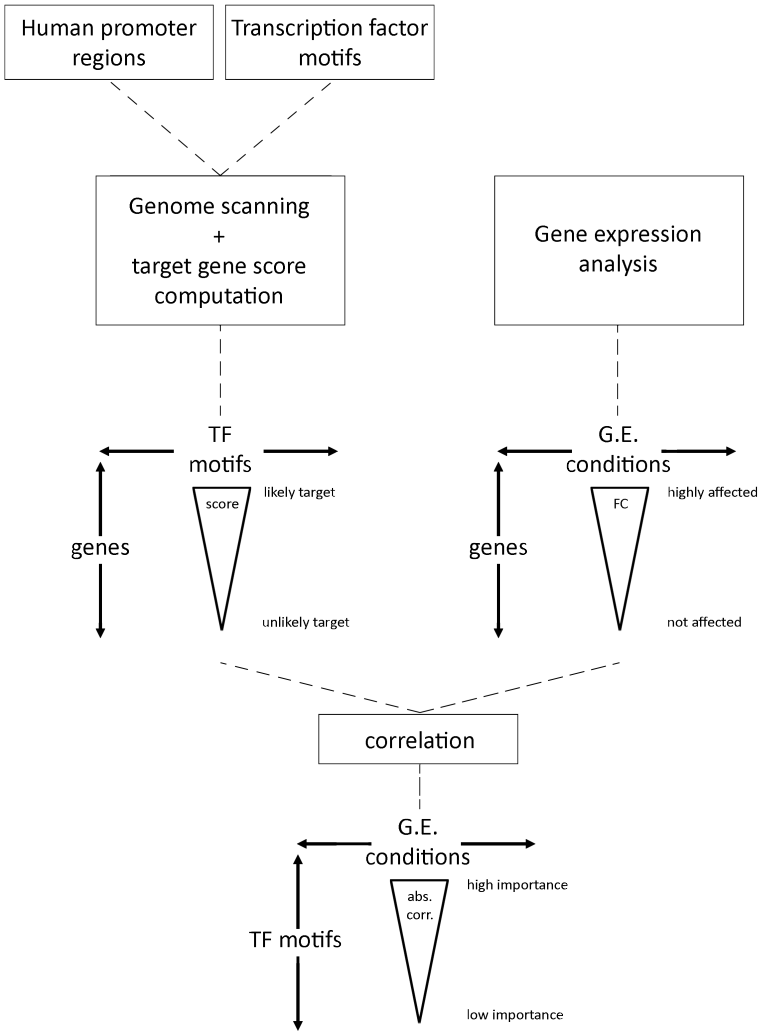


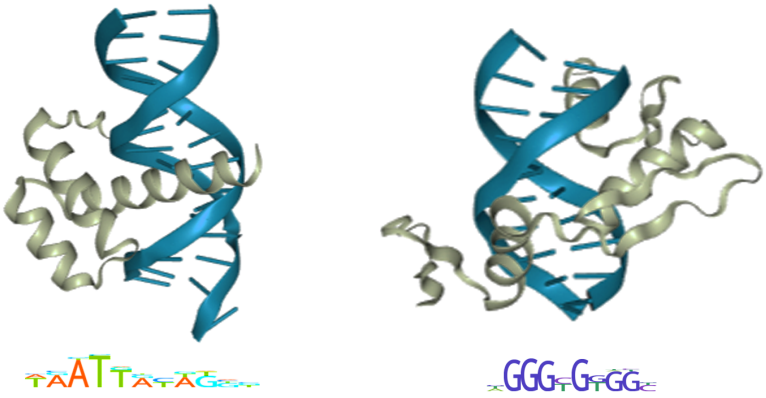
Figure 2: The general strategy of the proposed transcription factor analysis in the context of aging. This diagram depicts the approach proposed in this document. Transcription factor motifs are used to scan human promoter regions in order to compute a score that characterizes the likelihood of the gene being a target for the transcription factor. These score vectors per transcription factor motif are integrated with gene expression fold change data by measuring the correlation. As a result, highly correlating transcription factors (by motif) are considered to play an important role in the gene expression experiment and associated with the changes in fold change.

tor can be used to support conclusions drawn from gene expression data with the theoretical binding potency or the reverse in supporting theoretical transcription factor targets with gene expression data. Basically, from a vector of changes in gene expression, we can find a list of transcription factors that are likely involved in these changes (figure 2) and are therefore considered to be transcription factors potentially causal (or at least related) to the measured changes in the transcriptome. Thus, in theory, sorting the correlations between transcription factor score vectors and aging-related fold changes (e.g. young versus old) is expected to return crucial transcription factors (i.e. factors with the highest correlations).

5.2.1 Collecting human transcription factors and corresponding motifs

Transcription factors

Transcription factors function as regulators of genes, directly impacting the transcriptome. As mentioned before, mutations in these transcription factors and in their corresponding binding sites underlie many pathologies. Due to their importance, their structure and networks are conserved, although the method of action may differ per tissue (tissue-specific regulation). This observation can be attributed to the fact that TFs both work individually and as an ensemble. Therefore, transcription factors are often regulated by other transcription factors, effectively forming modules together with a variety of target genes [Lambert et al., 2018]. The conserved nature of the transcription factors is illustrated by the finding that factors of the same TF family share DNA-binding domains. Homeodomain (figure 3), basic leucine zipper (bZIP), basic helix-loop-helix (bHLH), C2H2-zinc finger (ZF, figure 3), and nuclear hormone receptor (NHR) are examples of conserved DNA-binding domains (DBD) and by definition crucial to the function of the TF. Knowledge of these DNA-binding domains allows for identification of transcription factors. Protein microarrays, DNA affinity purification-mass spectrometry, one-hybrid assays, DNase footprinting or mobility shift are experimental techniques that target DNA-binding proteins.



(a) DLX5 (homeobox) (b) KLF4 (zinc-finger)

Figure 3: 3D models of human TF-DNA interactions using typical DNA-binding domains. The DLX5 transcription factor interacts with DNA through the conserved homeobox domain (left). On the right, KLF4 interacts with DNA using a zinc-finger. The 3D models were accessed through the PDB (Protein Data Bank) website. The DNA-binding motif (downloaded from the HOCOMOCO archive) is presented below the 3D visualizations. The DNA-helix is highlighted in dark blue to allow for easy distinguishing from the transcription factor.

The 'futility theorem' suggests that the TFs are required to cooperate in order to achieve the desired specificity and effect [Wasserman and Sandelin, 2004]. Confirmation is found in the statement that very few proteins occupy most of their binding sites under regular physiological conditions, implying that a combination of TFs is required to bind. One of the exceptions is CTCF,

which occupies most of the ∼14000 binding sites [Lambert et al., 2018]. Plenty of other cases have been studied where TFs bind cooperatively as homodimers or higher-order complexes [Lambert et al., 2018]. Biochemical modeling and structural analyses indicate that TFs can influence transcription by altering chromatin to either promote binding of other TFs or non-DNA-binding cofactors. Other TFs, such as the Yamanaka factors KLF4, SOX2, POU5F1 and MYC, amongst others, can bind nucleosomal DNA and possibly recruit RNA polymerase, whereas others can directly recruit RNA polymerase without nucleosomal DNA binding. In contrast to activating transcription, TFs can also repress RNA polymerisation, and are then classified as repressors. Inhibitory transcription factors can exert their effect by disrupting RNA polymerase, altering DNA structure, binding activator TFs, recruiting proteins that degrade or bind mRNA, amongst other methods of operation [Lambert et al., 2018].

In the scope of the proposed analysis, a list of all human TFs is desirable. One reason is completeness and to enable calculate of how many TFs are not assigned any motif, whereas the second reason is filtering out non-human transcription factors. The human genome is predicted to contain about 1600 TFs (about 8% of all genes), with binding motifs for two-thirds of those [Lambert et al., 2018]. A collection of human transcription factors is publicly available in the TcoF-DB v2 archive (the database of transcription co-factors and transcription factor interactions, version 2) [Schmeier et al., 2016] and in the TRANSFAC database (transcription factor database, also termed geneXplain or TFClass) [Wingender, 1996, Matys et al., 2006, Wingender et al., 2012]. The latter offers both a restricted public and a subscription-based private service.

Motifs in biological sequences

The construction of a database of human transcription factors and their corresponding motifs is an extensive process and is crucial to the success of the analysis. The absence of true motif-transcription factor relationships can result in missing experimental information that could otherwise lead to crucial information to understanding the aging process, whereas the presence of false motif-transcription factor relationships generates wrong experimental information that compromises the outcome.

In general, biological sequence motifs are short (usually 6-12 bases), flexible recurring patterns in DNA that presumably exhibit a biological function [D'haeseleer, 2006, Lambert et al., 2018]. These motifs can be involved in processes at the RNA level, but in the context of this analysis the binding sites for transcription factors at the DNA-level are meant [D'haeseleer, 2006, Dhaeseleer, 2006]. Binding motifs for a TF are also frequently termed TF binding sites (TFBSs), whereas the domains of the transcription factor protein that bind the sequence are termed DNA-binding domains (DBDs). The binding site (or multiple binding sites) of a transcription factor is determined by techniques such as chromatin immunoprecipitation (ChIP) followed by massively parallel sequencing, ChIP-seq in short. This technique precipitates the protein together with the bound DNA fragment [Dhaeseleer, 2006]. Subsequently, the DNA fragments are purified and sequenced. Upon multiple alignment of the sequenced reads, the nucleotide count (A, C, G

or T) can be computed at every position. The positional nucleotide counts over the sequence length are summarized in a matrix called the position frequency matrix (PFM) [Bailey et al., 2006]. When the counts are converted to probabilities by using the total amount of sequences as a denominator in the division, the matrix is called a position probability matrix (PPM). Moreover, a third type of matrix representation can be used for a motif, named a position-specific scoring matrix (PSSM, also termed as position weight matrix (PWM) or position-specific weight matrix (PSWM)). The latter is generated by applying the following function to every element of a PPM [D'haeseleer, 2006].

$$M_{i,j,PWM} = \log_2(M_{i,j,PPM}/b_{nt})$$

In above function, i and j denote the position in the matrix (the nucleotide (row) and the motif position (column), respectively). b_{nt} is the background frequency for the respective nucleotide, which is often close to 0.25 (4 possible nucleotides) [D'haeseleer, 2006]. If the probability of observing a nucleotide at a certain position in the motif were zero, a log 2 transformation results in $-\infty$. To avoid this, pseudocounts are added to offset the value into the determined numeric range [Xia, 2012].

Additionally, graphical representation of a matrix in the form of a sequence WebLogo is common [Crooks, 2004]. Finally, presenting the motif matrix in the form of a consensus sequence, either with or without the use of regular expressions, is possible, but results in loss of information.

Calculating the information content of a motif position is a simple way to quantify how conserved the position is. This value ranges between 0 (all nucleotides equally probable) and 2 (one nucleotide fully conserved) and is calculated by the function below [D'haeseleer, 2006], where $f_{b,i}$ indicates the frequency of base b at position i. The summation symbol indicates that the calculation is summed up for all four nucleotides.

$$I_i = 2 + \sum_{b}^{A,C,G,T} f_{b,i} \log_2(f_{b,i})$$

Sources for accessing and downloading human transcription factor motifs are the HOCOMOCO database [Kulakovskiy et al., 2017], the motifs included in the popular HOMER tool [Heinz et al., 2010], and the elaborate CIS-BP archive [Weirauch et al., 2015], amongst others, although conversion to the more practical PPM form is likely required. Aggregating and curation of a human TF database is a complicated, labour-intensive process, as was recently detailed elsewhere [Lambert et al., 2018, Madsen et al., 2017a]. Intuitively, it is to be expected that low quality and non-human motifs are to be omitted, and exact redundancies should be removed [Lambert et al., 2018, Madsen et al., 2017a]. The curated and publicly available human TF database belongs to the department of Molecular Genetics at the University of Toronto [Lambert et al.,

2018], but was released very recently and is incomplete at the time of writing (summer 2018). Reasons vary from some TF-DNA interactions being unpublished to others being part of a paid subscription to TRANSFAC. Moreover, not all redundancies tend to be removed.

The basic alignment-based motif derivation algorithm described earlier is often substituted with more complicated motif discovery tools, relying on experimental data or statistical assumptions for better performance. Recent reviews reveal that rGADEM, MEME-ChIP [Bailey et al., 2015], HOMER [Heinz et al., 2010], and ChIPMunk [Kulakovskiy et al., 2010] are state-of-the-art amongst recent advancements, with rGADEM as the pinnacle [Jayaram et al., 2016].

5.2.2 Processing of the motifs before scanning

When combining different data sources, preceding research shows that a variety of problems needs to be solved [Lambert et al., 2018, Madsen et al., 2017a]. Some motifs present low overall information content and therefore low quality with regards to conservation, whereas others are redundant, either by being exactly similar or differing very slightly for the same TF. Another problem faced is the presence of motifs with low evidence (generated from a small amount of sequences).

Motif similarity and redundancy

As outlined below, solving the problem of redundancy requires a similarity measure and a cut-off value. Previous applications have shown that calculating a correlation score between PPMs is effective when followed by hierarchical clustering [Madsen et al., 2017a, Heinz et al., 2010]. For example, a tree height cut-off can be applied to group similar motifs and the redundant motifs in a cluster can be merged [Madsen et al., 2017a]. Finally, the edges of the motifs may be trimmed if these have an information content of 0.3 or lower [Madsen et al., 2017a, Mercier et al., 2011].

Several studies were carried out to investigate motif similarity measures [Gupta et al., 2007, Tanaka et al., 2011, Lambert et al., 2016, Mercier et al., 2011, Mahony and Benos, 2007], although the majority of the described implementations have limited practical applicability due to not or no longer being supported on the web, bad maintenance or no option for high-throughput usage.

The recently developed (2016) MoSBAT (motif similarity based on affinity of targets) tool starts by converting the motifs to position-specific affinity matrices (PSAMs) [Lambert et al., 2016, Foat et al., 2006]. PSAM is evaluated by scanning sequences and computing score vectors. Subsequently, the binding energy profile of the motif is obtained by taking the logarithm of these score vectors. Similarity (1 - distance) between energy profiles of two motifs is then calculated by computing the Pearson correlation coefficient (PCC). MosBAT is Unix-compatible and also available as an R package. Older tools such as the STAMP web tool (2007) [Mahony and

Benos, 2007] depend on Needleman-Wunsch (global) or Smith-Waterman (local) alignment of PSSMs. Upon aligning, distance metrics such as the PCC, the Kullback-Leibler information content, sum of squared distances or average log-likelihood ratio can be applied to calculate similarity. In practice, the PCC distance metric in combination with an ungapped Smith-Waterman alignment performs best in the case of STAMP [Mahony and Benos, 2007]. In addition, the TOMTOM (2006) and modified TOMTOM (2011) realizations (part of the MEME Suite) confirm the ungapped alignment approach, although based on offset p-values and validate the PCC and average log-likelihood ratio distance metrics, in addition to also adding functions based on the Pearson χ^2 and Fisher-Irwin exact tests to the collection [Gupta et al., 2007, Tanaka et al., 2011].

Benchmarking the above methods reveals that MoSBAT, TOMTOM and STAMP are comparable in terms of total area under the receiver operating characteristic curve (AUROC) [Lambert et al., 2016]. The Pearson's correlation coefficient computation method is implemented in the Biopython library for Python 3.

Clustering Motifs

Clustering of similarity scores in the context of biological motifs is mainly carried out by hierarchical clustering [Kankainen and Löytynoja, 2007, Madsen et al., 2017a]. The MATLIGN web-application (2007) clusters (PCC) similarity scores by applying an agglomerative hierarchical algorithm that recursively joins the two most similar clusters using the average distance for all motifs in a cluster [Kankainen and Löytynoja, 2007]. However, running MATLIGN in the browser tends to be quite slow and error-prone. Substituting with the common UPGMA (unweighted pair group method with arithmetic mean) agglomerative (bottom-up) hierarchical clustering method seems to be appropriate.

Merging Motifs

Using recent publications as a guideline [Madsen et al., 2017a], merging similar motifs instead of arbitrarily making a selection is advised. Averaging over the positions in the motif alignment is suggested by image content in the respective publication, but is yet to be specifically detailed [Madsen et al., 2017a, Kankainen and Löytynoja, 2007].

Trimming and filtering

The information content (IC) of a position in a motif signifies the conservation and thus summarizes the strength of the position based on the motif evidence. A position with IC of 2 is perfectly conserved (e.g. adenine in 100% of the cases equals 0% chance to observe cytosine, guanine or thymine), whereas an IC of 0 indicates equal probability for each nucleotide (25% chance for A, C, G and T). Positions with high uncertainty with regards to the nucleotide are not informative and can be removed if they are found at the edges of the motif. An appropriate IC cut-off for trimming is 0.5 [Madsen et al., 2017a]. Motifs with a length shorter than 4 are also filtered out, since non-informative hits would be observed all over the genome [Lambert et al., 2018, Madsen et al., 2017a].

5.2.3 Collecting known human transcription factor targets

Transcription factor targets resulting from these *de novo* approaches have been stored in public databases, such as the TRED [Jiang et al., 2007], ITFP [Zheng et al., 2008], ENCODE [Consortium, 2012], Neph2012 [Neph et al., 2012], TRRUST [Han et al., 2015] and Marbach2016 [Marbach et al., 2016] archives and can be readily accessed through the R *tftargets* library [Slowikowski, 2018]. Moreover, this collection of TF targets can be extended with the public version of the TRANSFAC database and the PAZAR project [Wingender, 1996, Matys et al., 2006, Portales-Casamar et al., 2007]. It must be noted that unfortunately, all TF-target gene databases are incomplete.

5.2.4 Experimental transcription factor target identification

Next to the experimental methods for transcription factor target identification outlined higher, the latter can also be derived *in silico* from genome scanning tools such as HOMER or the MEME Suite [Heinz et al., 2010, Bailey et al., 2015]. Recently, a surge in machine learning (ML) based TF-target identifiers was observed [Salekin et al., 2017, Bhardwaj, 2005, Holloway et al., 2006, Honkela et al., 2010, Qin and Feng, 2017, Karimzadeh and Hoffman, 2018]. These *in silico* approaches are the subject of this subchapter.

Promoter scanning with transcription factor motifs

In the context of this thesis, the ideal sequence scanning algorithm takes a motif and a list of sequences (e.g. the promoter sequences of all human genes) as input and returns a score for every sequence in the list. The returned score should accurately quantify the theoretical binding capacity of the TF (corresponding with the motif) for the scanned sequence. Secondly, the ideal method is interpretable, preferably simple, and easily fine-tuned to the specific application at hand.

In line with the simple scanning algorithm presented in the impactful REDUCE paper (published in 2001 by the P.I. of the Li lab) [Bussemaker et al., 2001], a simple and interpretable model of the binding affinity can be proposed. In a sliding window alignment between the motif and every position in the promoter sequence, a position-specific alignment score can be computed using the following formula:

$$S_i = \frac{f_1}{f_{01}} * \frac{f_2}{f_{02}} * ... * \frac{f_n}{f_{0n}}$$

With i in S_i as the position of the sliding window and n the motif length. f_x indicates the motif PPM's frequency of occurrence for the corresponding nucleotide in the sequence to be scanned, with x the aligned position of the motif PPM. f_{0x} is the background frequency of the aligned nucleotide (4 possibilities). This formula is very similar to the log-odds probability cal-

culation applied in the renowned HOMER motif scanning software [Heinz et al., 2010].

$$S_i = \log \frac{f_1}{f_{01}} + \log \frac{f_2}{f_{02}} + ... + \log \frac{f_n}{f_{0n}}$$

S_i as the result of this formula may be aggregated over the sequence by summing, either with or without applying a score cut-off beforehand. The sole difference of the suggested scoring formula with the HOMER formula presented above is the log transformation. The best state-of-the-art methods tend to rely on a variation of this scoring formula [Grant et al., 2011, Jayaram et al., 2016]. These models only take into account the motif and the sequence and is therefore an oversimplification of the molecular mechanisms involved.

Previous advancements in molecular biology show that, in contrary to the assumption of the simple scanning algorithm, the affinity of a transcription factor for a promoter sequence is not only determined by the TF-motif and the target sequence. For example, the TFBS is not necessarily located proximal to the TSS. Potential distal TFBS (e.g. far upstream the TSS) are also proven to impact the TF binding, together with the presence of other transcription factors. In addition to forming complexes with other transcription factors (*cis*-regulatory modules), also the presence of other co-factors impacts the binding [Wasserman and Sandelin, 2004]. Adverse secondary structure or other forms of chromatin inaccessibility may also be detrimental to transcription initiation. It must be noted that transcription initiation is not the only mechanism that controls gene expression [Wasserman and Sandelin, 2004]. Additionally, CpG dinucleotides and their methylation status play a key role in gene activity. In active regulatory sequences CpG is unmethylated, whereas generally speaking, up to 80% of CpGs are methylated on a cytosine. Furthermore, increasing evidence that promoters are bidirectional should be taken into account when scanning these regions. Finally, splicing, protein modification or mRNA degradation are all features that can affect gene expression.

Although the less complex model will likely underperform in terms of accuracy in predicting transcription factor binding, the simplicity and interpretability of this algorithm makes it a good starting point. If promising results are achieved regarding the ultimate goal of accurately identifying important transcription factors, the insights obtained by applying this 'glass box' method will likely lead to better selection and substitution by more accurate and appropriate scanning method. In the latter case, the initially developed method serves well as informal benchmark. Some of these state-of-the-art methods, next to HOMER, are briefly introduced below.

A recent comprehensive review by Jayaram et al. (2016) revealed that the comparison of individual TFBS predicton tools heavily favored the FIMO algorithm [Grant et al., 2011]. FIMO (find individual motif occurrences) excelled with an average sensitivity of 0.933, positive predictive value of 0.839, accuracy of 0.884 and false positive rate (FPR) of only 0.002 on the simulated data [Jayaram et al., 2016]. FIMO outperformed it's competitors for every tested metric. The Patser algorithm [Thomas-Chollier et al., 2008] claims second place with an average sensitivity

of 0.887, positive predictive value of 0.774, accuracy of 0.828 and FPR of 0.008 on the same simulated data [Jayaram et al., 2016]. FIMO is web-accessible and is part of the MEMESuite toolbox for motif discovery and searching [Bailey et al., 2006, Bailey et al., 2015]. As input, it takes a motif in MEME format in addition to a DNA sequence FASTA file. In line with the previous tools, this tool is limited to only motifs and sequences as input. Furthermore, it must be noted that nucleotide background scores are included in the MEME motif format. The FIMO implementation shares the calculation of log-odds scores with the previously discussed algorithms. However, additional complexity is introduced by converting the log-odds scores into p-values by using a dynamic programming approach under the assumption of a zero order background model [Grant et al., 2011]. Furthermore, the p-values are updated to q-values following the Benjamini and Hochberg false discovery rate (FDR) approach [Benjamini and Hochberg, 1995]. Both the forward (+) and reverse complement (-) sequences are scanned. Additionally, when priors are provided FIMO makes the transition from the log-odds scores to log-posterior odds, detailed elsewhere [Cuellar-Partida et al., 2011]. After filtering out non-significant motif hits, the retained hits are parsed to a variety of different file formats, such as HTML, XML, GFF and more.

Parallel to the publication of the review by Jayaram et al. (2016) an R package and corresponding publication was released, named TFBSTools under the Bioconductor bioinformatics framework [Tan and Lenhard, 2016]. This package was therefore not included in the comparison and is still under active development and updates are frequently released around this time of writing (June 2018). TFBSTools encapsulates a variety of motif tools, similar to the MEMESuite (e.g. motif alignments, motif discovery, sequence scanning, motif conversion, amongst other tools). However, due to the active development, the method of operation remains largely 'black box'.

Alternative approaches are less common at this time, but are gaining popularity and might pose improvements. Statistical machine learning approaches allow for incorporation of extra data such as chromatin accessibility, enhancers, general information on the cell, and other DNA characteristics. In theory, given the molecular mechanisms of TF-binding, this should greatly improve the accuracy. However, practical implications are that these methods require extra data and are therefore only applicable in situations where these data are available. The limitation to niche applications likely contributes to the unpopularity at this time. An example is the IMAGE algorithm (integrated analysis of motif activity and gene expression changes of transcription factors) in which the combination of motifs, enhancer data and promoters is fed into a ridge regression algorithm, resulting in highly accurate transcription factor binding predictions [Madsen et al., 2017b]. The success of this combination can be attributed to modeling the activity of a motif before linking it to changes in gene expression.

Other examples of learning approaches are DeepSNR (convolutional and deconvolutional neural network) to predict similar transcription factor binding sites by using ChIP-exo data as input [Salekin et al., 2017]; a kernel-based approach to classify DNA-binding proteins and non-

binding proteins using support vector machines (SVMs) [Bhardwaj, 2005]; applying SVMs to classify TF-targets using the position-specific scoring matrix (PSSM) form of a TF-motif [Holloway et al., 2006]; Gaussian processes and Bayesian inference to model TF expression profiles [Honkela et al., 2010]; ensemble methods in combination with SVMs using position weight matrices (PWMs) to predict TF-targets [Fan et al.,]; predicting TF-binding in new cell types using an artificial neural network trained on ChIP-seq results in other cell types [Karimzadeh and Hoffman, 2018]. None of these methods have been elected as state-of-the-art, although their potency in certain use cases is undeniable.

Method validation

Higher scoring target genes for a motif under study represent higher likelihood of TF binding in theory. Sorting the potential target genes by score ranks the genes from better theoretical targets to worse targets. For the transcription factors with known targets, one is able to verify whether the high scoring target genes correspond to known and verified targets, under the assumption that the TF-target databases are accurate. This presents an excellent use case for a Wilcoxon rank sum test. Given that the predicted scores are sorted in descending order (high scoring target genes correspond to low rank number), the requirements to statistically validate the approach is a significant p-value and a negative test-statistic. The negative test-statistic corresponds to a positive correlation between high scores and verified targets. In other words, under these conditions a high score would represent a high chance of being a true target for the transcription factor under study. This Wilcoxon rank sum test, also termed the Mann-Whitney U test, is conducted for every binding site of a transcription factor for which known targets are available. This approach is therefore subject to the multiple testing problem and the resulting p-value is adjusted to control for the false discovery rate (FDR).

5.2.5 Computational approach for identifying crucial transcription factors in gene expression experiments

Identifying crucial transcription factors

The vector of expression fold changes represents whether a gene is expressed more or less in the contrast under study (e.g. aged tissue versus young tissue). Secondly, the vector of scanning scores represents whether a gene is likely to be a target for a transcription factor (grouped by TF-DBS). If a change in gene expression for certain genes correlates with the likelihood of these genes to be a target of a specific transcription factor, it is likely that this transcription factor is involved in the condition contrasted by the gene expression experiment.

Transcription factors can either be transcriptional activators or repressors. Therefore, the absolute value of the gene expression fold changes must be taken as a measure of impact before calculating correlation. This way, highly impacted genes (high absolute fold changes) and high theoretical probability of the gene being a TF target (high scanning scores) go hand in hand and

would result in high positive correlation for crucial transcription factors in the gene expression experiment under study. Negative correlation scores would mean that unlikely targets of the transcription factor are highly impacted, which argues against the importance of the transcription factor.

Computing the correlation score can be done using either Pearson's correlation coefficient (PCC) or the nonparametric Spearman's rank correlation coefficient. The latter will likely be superior with the raw data, because nonparametric metrics usually perform better in cases with different scales. However, if appropriate data transformations are applied, PCC has the potential to do better as more information is captured.

Method Validation

Verifying the approach for finding relevant transcription factors can be done by studying a well-characterized gene expression data set for which the crucial transcription factors are known. If the approach returns these factors with high correlation together with other closely related factors, the method can be considered validated.

A study by Cusanovich et al. (2014) presents a data set in which 59 TFs are knocked down in a HapMap lymphoblastoid cell line [Cusanovich et al., 2014]. The knock-downs were achieved using siRNA (small interfering RNA) and are targeting transcription factors known to be involved in the immune response, which is closely related to aging. Quantifying the effect of the knock-downs by means of gene expression data was done using qPCR (quantitative polymerase chain reaction) and microarrays. The data are available with accession number GSE50588 in the GEO archive.

6 Aims and Strategy

6.1 Experimental setup

A step towards the ultimate goal of understanding and decelerating the aging process can be taken by uncovering essential transcription factor modules and designing novel drugs to effectively target the functional networks involved in this time-dependent degenerative process. Research at the Li lab at the University of California, San Francisco (UCSF) is being carried out to develop and optimize a large-scale and highly efficient drug screening methodology. The general strategy is briefly summarized below, while leaving out technical details in order to protect confidentiality.

The analysis starts with engineering a yeast system (*Saccharomyces cerevisiae*) in order to achieve mother cells capable of dividing and producing daughter cells that are in their turn unable to multiply. This modification implies that all observed daughter cells are offspring from the same mother cell. The designed system essentially provides us with a way of determining the replicative lifespan of the mother yeast cell by counting the number of daughter cells it has produced. Thus, the (final) daughter cell count is a proxy for the replicative longevity of the mother cell. Administration of drugs and determining the change in replicative lifespan is now possible and has the potential to reveal effective drugs or other perturbing agents, in addition to allowing the researchers to gain more insight in the aging process. However, the replicative lifespan is not equal to the chronological lifespan, although similar pathways are affected. The latter provides grounds to deem the yeast system suitable as a model system for the pathways shared with humans.

Parallel to the development of the drug screening procedure, potential drug targets must be predicted and data from previous research must be analyzed to be incorporated for more insight and to steer the analysis in the appropriate direction. Furthermore, aging data from other research groups, databases and consortia is gathered and stored in a web-accessible and aging-specific database (currently under development). This thesis concerns the analysis of human transcription factors and the prediction of their targets in the context of aging and therefore relates to the first part of this paragraph, more specifically by aiming at identifying potential drug targets in crucial transcription factors.

Although this transcription factor analysis is focused on the building blocks of the human genome, the described approach can be generalized to yeast due to the similarity of both eukaryotic genomes and the simplicity of the *S. cerevisiae* model system. The main goal of this analysis is to explore the possibilities in terms of finding crucial aging-related modules and to evaluate potential methodologies to develop and test drugs with the highest potential and experimental efficiency. In addition to the aging-related discoveries, the method can certainly be generalized beyond the scope of aging.

6.2 Aims and strategy

In this work, an attempt is made to develop a robust approach to identify relevant transcription factors involved in the changes observed in a gene expression experiment.

Every step towards our goal needs to be carefully broken down to guarantee the reliability of the building blocks for the desired algorithm. The first step in the pipeline to rank relevant transcription factors in a gene expression contrast is to identify the target genes of these transcription factors. A theoretical score describing the likelihood of these targets needs to be computed. Therefore, a list of transcription factors will be gathered from online databases in addition to known target genes of these TFs. Secondly, the DNA-binding motifs of these TFs will be collected. Preceding research indicated that careful curation of these motifs is necessary. We aim to introduce as little bias and as little erroneous data entries as possible, because the transcription factor motifs are absolutely crucial to the success of this analysis. Constructing a reliable transcription factor and motif database is the first aim of this work.

The second challenge of this analysis is to integrate the transcription factor DNA-binding motifs with the promoter regions in the human genome. More specifically, we aim to compute a score for every TF-promoter combination representing the theoretical binding affinity of the transcription factor as accurately as possible. The resulting score will be carefully examined in order to guarantee practical reliability, next to deriving molecular insights. Several genome scanning approaches were introduced earlier (Li lab implementation and FIMO, amongst others), and need to be evaluated to make sure the algorithm with the best performance is selected.

A third aim is linking heavily impacted genes in a gene expression contrast to transcription factors with high theoretical binding affinity for these heavily impacted genes. Transcription factors that score high for the genes with high changes in gene expression are searched for by calculating the correlation. Well-correlating TFs are assumed to be relevant to the gene expression experiment. Practical applicability of this proposed algorithm needs to be assessed and can be studied on validation data for which the transcription factors that cause gene expression changes are known.

If the implementation is found to be successful, further optimization may be done and the optimal algorithm can be applied to a couple of aging-related gene expression experiments. Following separate analysis of the data sets, we hope to be able to combine the observations across different experiments in order to gain novel insight.

Finally, we aim to launch a publicly available version of the implementation in order to translate this research project into direct practical value.

7 Results

7.1 Building a human transcription factor motif database

The consensus annotation of the human genome has been studied elaborately and is predicted to contain 1447 unique transcription factors [Zhang et al., 2014, Wingender et al., 2012]. Searching for transcription factor – motif relationships and constructing a collection in the form of position probability matrices (PPMs) from the JASPAR [Sandelin, 2004], HOCOMOCO [Kulakovskiy et al., 2017] and CIS-BP [Weirauch et al., 2015] archives yielded a total of 2771 PPM data entries. After a filtering step with two merged lists of human transcription factors obtained from TcoF-DB (v2) [Schmeier et al., 2016] and TRANSFAC, 2645 entries were kept (95%). These 2645 PPMs corresponded to 772 unique transcription factors. The remaining 675 transcription factors from the 1447 predicted TFs from the consensus annotation are not publicly assigned a motif in the above databases as of February 2018. 236 of the 772 unique transcription factors are only represented by a single motif (31%), whereas the remaining 536 (69%) are human transcription factors represented by more than one motif. The latter is equal to an average of 3.5 motifs per unique transcription factor.

7.1.1 Motif processing

Locating redundant motifs

Since many of the motifs are duplicates across databases or slight variations of one motif, a clustering and merging step was necessary to filter out non-unique motifs (figure 4). The clustering algorithm used was the agglomerative hierarchical UPGMA method [Sokal and Michener, 1958] and was applied to the TF-specific correlation matrix. This matrix contains the similarity scores between the different motif PPMs for one transcription factor, based on the Pearson correlation coefficient (PCC). Extensive visual inspection of the clusters revealed that calculating the similarity by computing the PCC between the motifs is surprisingly effective in grouping similar motifs, therefore eliminating the need to apply more complex algorithms. The merging step was performed after creating a multiple alignment of a cluster. Two approaches for the merging process were evaluated: unweighted and weighted merging, respectively.

Unweighted motif merging

This type of merging approach is characterized by simply averaging over corresponding positions of the different motifs in one cluster (split on a 0.5 tree height cut-off value), without incorporating any weight information. As a result, the average amount of motifs per unique transcription factor was brought down from 3.5 (2645 motifs for 772 unique TFs) to about 1.5 (1155 motifs for 746 unique TFs). Finally, motifs with a length shorter than 4 nucleotides were removed, which reduced the amount of unique TFs with one or more motifs from 772 to 746.

Different properties of the motifs are studied, both before and after the merging process (figure 5).

Weighted motif merging

In contrast to unweighted merging, in weighted merging, we do not simply average over the corresponding positions of the different motifs in one cluster, but we multiply the value in the position with the relative fraction of evidence for that motif (e.g. the amount of ChIP-seq sequences incorporated in the multiple alignment from which the motif was drawn).

The weighted approach was evaluated for the motifs in the JASPAR and HOCOMOCO archives, where the evidence counts are readily available. 745 motifs with corresponding evidence were listed. After clustering and weighted merging, 536 motifs for 502 unique transcription factors were returned. This corresponds to an average of 1.07 motifs per transcription factor. The main reason for the difference between the average motif per TF in the unweighted (average of 1.5 motifs per TF) and weighted (average of 1.07 motifs per TF) method is because the starting set of motifs is a lot smaller (2644 unweighted vs. 745 weighted). The direct comparison is therefore not fair and is improved upon in the next paragraph.

Unweighted vs. weighted merging

A fair comparison between the unweighted and weighted approach could be obtained by initializing both methods with the same motif subset. Only for the 745 motifs represented by evidence, the results of both the unweighted and weighted merging approaches were compared on different levels. Because only the merging process is different, the preceding clustering is not subject to change and both approaches will therefore result in the same amount of motifs (536), corresponding to the same amount of unique transcription factors (502). Properties that are expected to be affected by the merging process are information content, motif nucleotide probabilities, resulting motif consensus sequence and motif length after trimming, amongst others. Although the majority of the PPMs are exactly identical (354 out of 536 PPMs or 66%), the consensus sequences are merely equal in 423 of the 536 cases (79%). However, consensus sequences are subject to borderline cases, lose out on information and are thus representing the difference between the methods incorrectly. More telling is the correlation score between resulting PPMs from both approaches, which is computed to be equal to the very high value of 0.975. Regarding average information content, the distributions are extremely similar (figure 6) and the means of the average IC over the different motifs align as well (1.27 for the approach involving weights and 1.26 for the approach without). These results are grounds to formulate the assumption that weighted merging is not superior or substantially different by a clear margin. Therefore, spending additional resources on also gathering corresponding motif evidence for the CIS-BP database to perform weighted merging on the complete motif set does not outweigh the benefits.

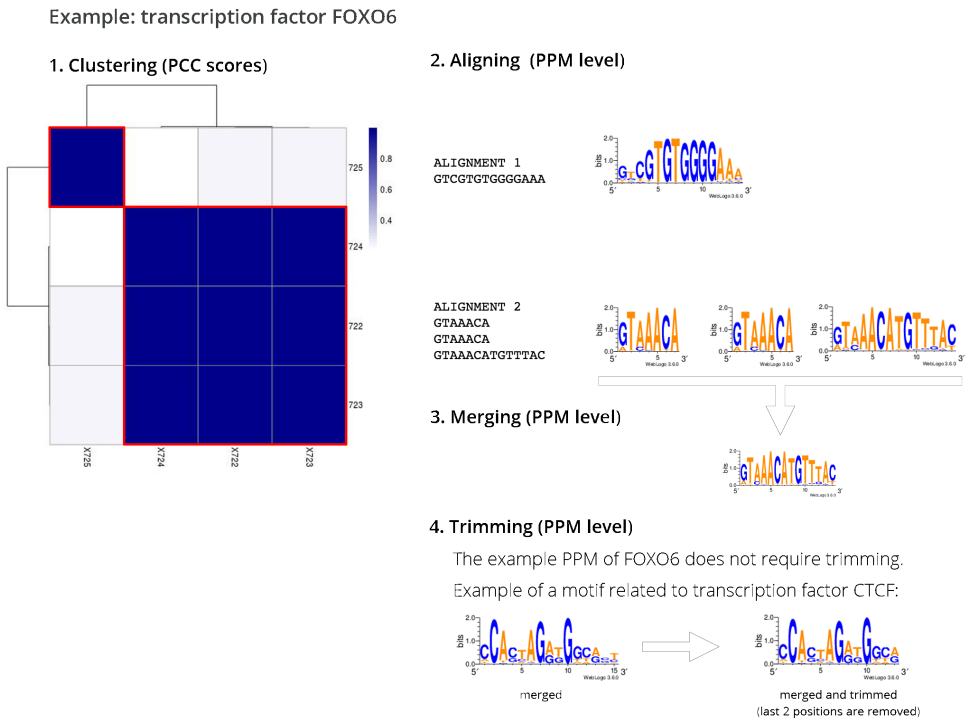


Figure 4: Motif processing for FOXO6. A schematic example of the processing of motifs related to transcription factor FOXO6. The choice for FOXO6 was made because of the relationship of the TF to the aging process (see introduction). During aging, the activity of FOXO6 is diminished due to phosphorylation, which affects transcriptional activity. FOXO6 is closely involved in modifying histones and chromatin and therefore able to generate age-related effects on a molecular level [Kim et al., 2014]. In this example, the motifs of FOXO6 are processed. First, the Pearson correlation coefficients (PCCs) are calculated between the different motifs in their position probability matrix (PPM) form. The correlation matrix is then subjected to the hierarchical clustering UPGMA algorithm. After clustering, the heatmap is plotted as shown in (1). Cluster information is extracted and the cluster trees are subdivided in separate clusters with a tree height of 0.5 as maximum cut-off. Next, a multiple alignment of the motifs contained in a cluster is computed (2). The motifs are padded to equal length (with blank values that will be disregarded during merging) and merged as one by averaging over the positions of the respective PPM (3). Finally the merged motif is trimmed by removing the positions with an information content lower than 0.5 on a scale of 2 to return the final PPM (4). Moreover, motif PPMs with a length lower than 4 nucleotides are removed (not shown in figure).

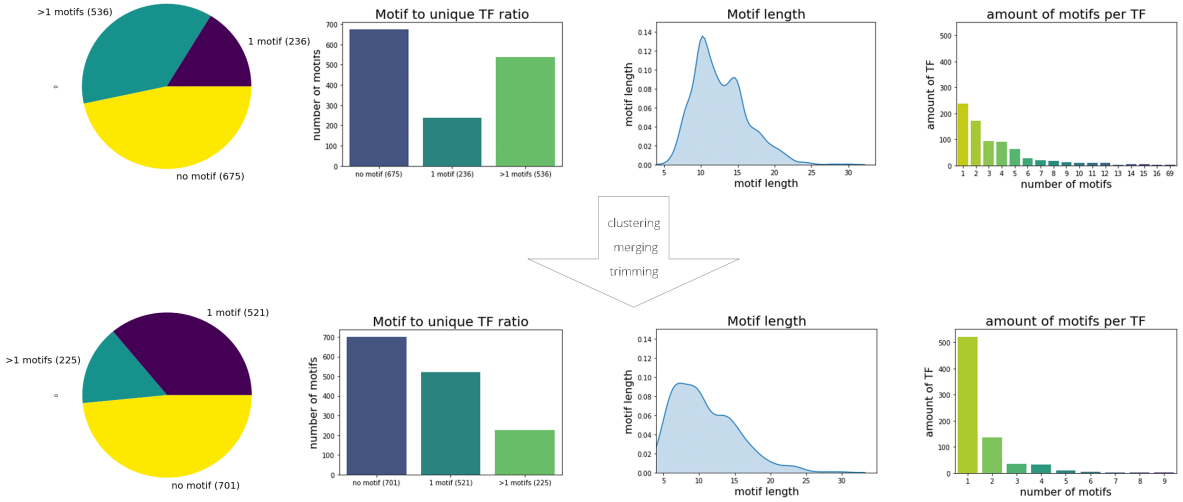


Figure 5: Comparison of motif properties before and after merging and trimming. From left to right, the ratio of transcription factors (total of 1447) without any motif, with one motif, and with more than one motif. Merging and trimming followed by filtering significantly reduces the average amount of motifs per unique transcription factor. On the second plot from the left, the same data are presented in a different way. The third plot from the left showcases the effect of trimming on the average motif length, as the distribution is shifted towards lower length. Lastly, the plot on the right represents the amount of motifs per transcription factors by number. After merging and trimming, there are fewer transcription factors with a high number of motifs.

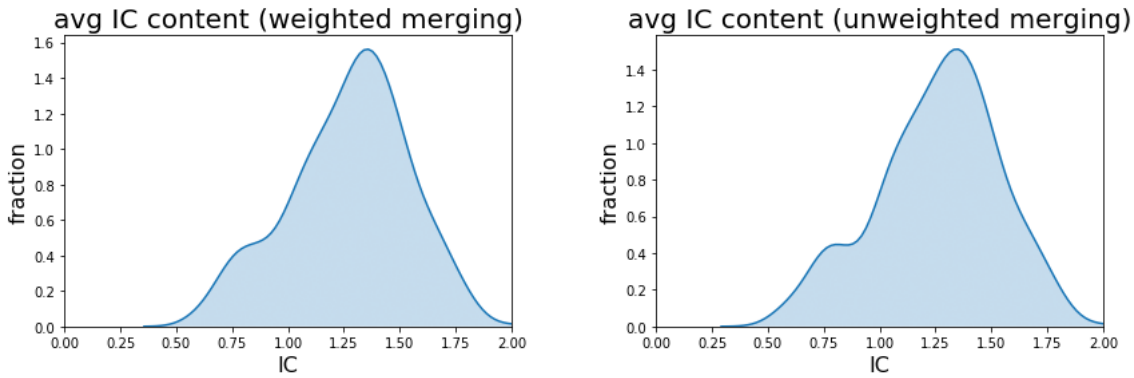


Figure 6: Comparison of average information content (IC) over all motif positions between the merging approach involving weights and the merging approach based on averaging. The distribution of average information content over all motif positions is compared between both approaches. The mean of merging with weights (left) is 1.27, whereas the mean of unweighted merging (right) is 1.26. In terms of shape and scale, both distributions are virtually identical.

7.2 Collecting targets of human transcription factors

After building the motif database, the target genes of the unique transcription factors are to be listed. The targets (based on in silico prediction or wet lab experiments) were found through the R package *tftargets* [Slowikowski, 2018], which allowed indirect access to the TRED [Jiang et al., 2007], ITFP [Zheng et al., 2008], ENCODE [Consortium, 2012], Neph2012 [Neph et al., 2012], TRRUST [Han et al., 2015] and Marbach2016 [Marbach et al., 2016] databases. Moreover, hits for human transcription factors in the public TRANSFAC [Wingender, 1996] and the PAZAR [Portales-Casamar et al., 2007] archives were parsed and added to the table. When applicable, the TF-target occupancy of the database was computed (figure 7). The number of transcription factors with one or more targets was calculated and listed (table 1). The union of the databases assigns one or more target genes to 687 of the 746 transcription factors in our motif database. In other words, 92% of the transcription factors in the motif database were linked to one or more target genes (table 1).

Table 1: TF targets per database. The database-specific amount of transcription factors with one or more target genes was calculated. As expected from data visualization beforehand (figure 7), the MARBACH data represents the largest fraction of transcription factors. The union of the 7 different databases yields 687 transcription factors with one or more targets, or 92% of the 746 factors in the motif database.

PAZAR	ITFP	TTRUST	ENCODE	TRED	MARBACH	TRANSFAC
149	280	443	106	119	547	137
union: 687						

7.3 *De novo* target prediction

7.3.1 Genome scanning

A concise exploratory analysis (data not shown) leads to the observation that the different transcription factor target databases exhibited little overlap. Therefore, a more elaborate transcription factor target study was required. Theoretical binding sites for each TF were found by scanning over a collection of human promoter regions. More specifically, the collection of human promoter regions contains 25492 canonical sequences characterizing the 5000 bp region upstream to a canonical transcription start site (TSS). These 25492 non-redundant sequences correspond to 18421 different genes. For every motif in the motif database (1155 total), all 25492 sequences are scored as described in the Materials and Methods section. The convolution-based computation method completed about 10 times faster in practice (on a CPU machine; most likely faster on a GPU machine) compared to a sliding window implementation based on standard Python for-loops.

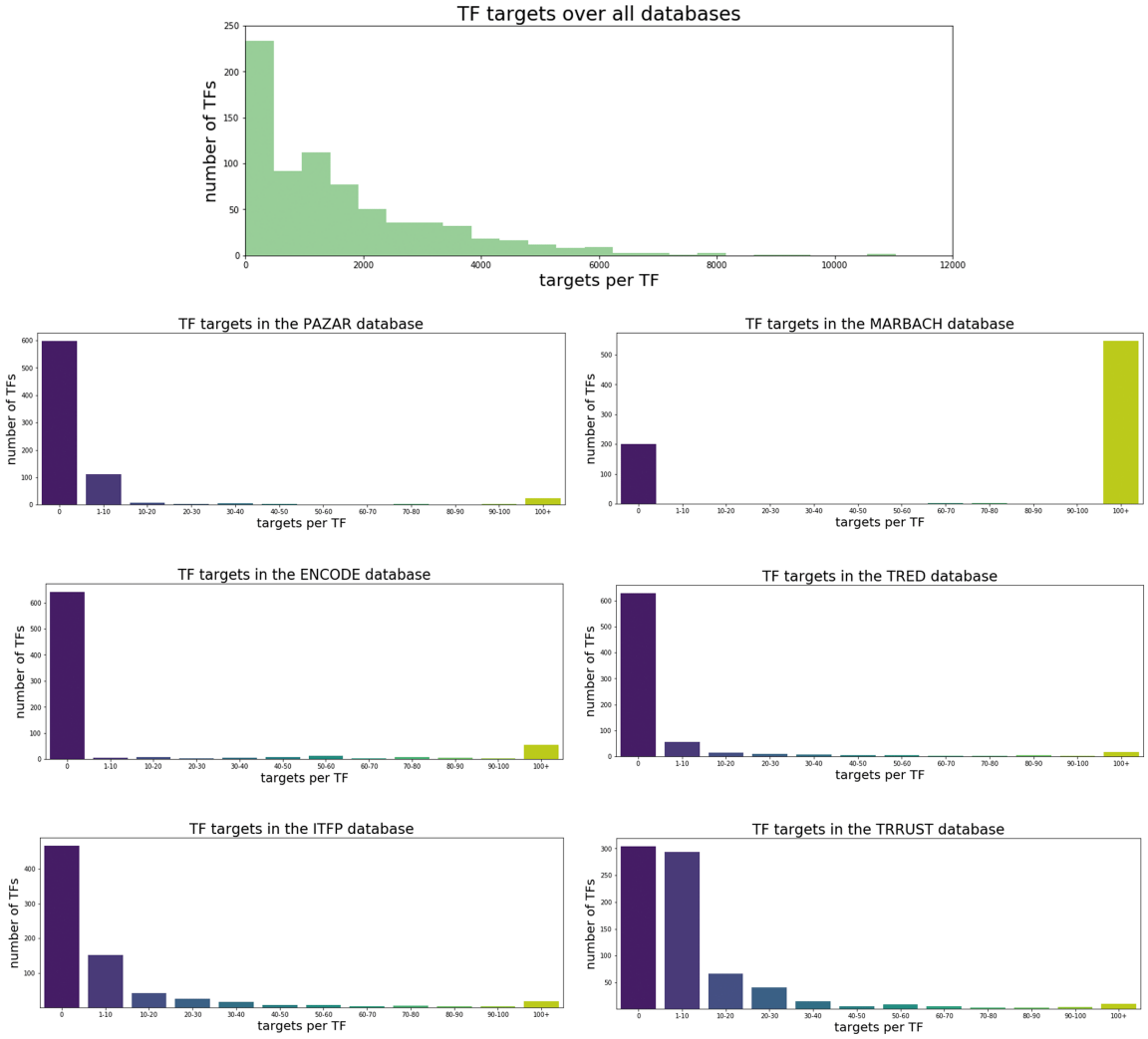


Figure 7: Occupancy of TF targets in different databases. The top plot represents the union of the TF targets contained in the other 6 databases. The 6 databases are the PAZAR, MARBACH, ENCODE, TRED, ITFP and TRRUST databases. The tissue specific nature of the NEPH2012 dataset is less suited to be represented in this way and is therefore left out, as well as the public TRANSFAC database. The MARBACH dataset strongly influences the top plot, as it contains the most transcription factor targets, especially transcription factors with more than 100 targets, up to thousands of targets.

7.3.2 Score analysis

The predicted upper bound of the scores was set to roughly 10^{10} and was never exceeded as expected. Because of the relationship of STAT3 and the aging process, the results are mainly demonstrated for STAT3 and the three motifs corresponding to this TF. Scanning upstream promoter sequences with motifs grouped by transcription factor returns a table with a total raw score over all sliding windows for each motif and for each potential target gene (table 2). The scanning results were elaborated on by including a column representing whether the gene is known to be a target to the particular transcription factor under study. Secondly, for each motif the motif score rank over all promoters (by descending order) and the corresponding z-score is

calculated (figure 8). Finally, for each candidate target gene, the best rank of all motif-specific ranks is also generated. A visual representation of the raw score distribution for the second motif of transcription factor STAT3 is made as an example (figure 11) and is characterized by a maximum score of 6.8e5 and minimum of 3e2, with a median of 7.5e3 and standard deviation of 2.8e4.

Table 2: Promoter scanning scores for the three motifs of STAT3. In this table, the aggregated total promoter scanning scores per motif are presented ('motif1', 'motif2' and 'motif3' columns). The 'id' column represents the promoter id in the hg19 version of the human genome. The corresponding HGNC-symbol is found in the 'gene' column. Moreover, a 'database confirmed' column is added to indicate whether the target gene is known to be a target of STAT3 and is therefore supported by at least one of the TF-target databases. Furthermore, the raw total scores are standardized (to z-scores) to account for the motif-length bias introduced by the scoring formula. Finally, also the ranks of the target genes are computed based on descending scores per motif. The results for the second DNA-binding motif of the STAT3 transcription factor are marked.

#	id	motif1	motif2	motif3	gene	Database confirmed	Zscore motif1	Rank motif1	Zscore motif2	Rank motif2	Zscore motif3	Rank motif3	Best rank
0	NM_014305	12103	677696	271527	TGDS	TRUE	-0.145	14915	4.325	0	3.167	114	0
1	NM_001136040	11000	603105	22643	CPSF7	TRUE	-0.864	20798	4.209	1	0.907	4377	1
2	NM_024811	11136	603101	22638	CPSF7	TRUE	-0.771	20231	4.209	2	0.907	4379	2
3	NM_017841	13459	602262	22331	SDHAF2	TRUE	0.653	6656	4.207	3	0.895	4454	3
4	NM_001321096	13073	563866	554501	SREBF1	FALSE	0.434	8852	4.142	4	3.816	24	4
5	NM_006545	14231	506506	140873	NPRL2	FALSE	1.073	3286	4.035	5	2.570	386	5
6	NM_001323619	14473	479698	91698	UNC45A	TRUE	1.200	2513	3.981	6	2.179	784	6
7	NM_001323621	14488	479451	90545	UNC45A	TRUE	1.207	2475	3.981	7	2.168	798	7
8	NM_001199861	13954	477644	578095	KCNAB2	FALSE	0.925	4328	3.977	8	3.854	23	8
9	NM_001039675	9281	469964	90256	UNC45A	TRUE	-2.142	24858	3.961	9	2.165	801	9

The total scores per target gene can be broken down into positional scores for every sliding window. In order to increase the signal-to-noise ratio, the value corresponding to the 99.9th percentile of a randomly sampled (without resampling) background distribution of scores was subtracted from the scores, with a minimum of zero upon subtraction (figure 10). Representative plots for the top motifs were generated, as well as for some of the examples related to aging, such as STAT3 and FOXO6 (figure 9). The scatter plots for the positional scores clearly show an enrichment in hits towards the right, which corresponds to the transcription start site (figure 9). This observation was further investigated by combining plots from multiple transcription factor motifs and by calculating a regional p-value based on the Poisson distribution computed over the entire promoter region (figure 13). Additionally, the p-value was calculated for individual motifs as well (figure 13).

The different motifs considered in this analysis are being referred to as the name of the transcription factor and an automatically incrementing motif index. For the most commonly returning motifs, the corresponding PPM and visual logo are stored in a table in the appendix (table 10). Importantly, this table contains the PPMs and logos relevant throughout the document.

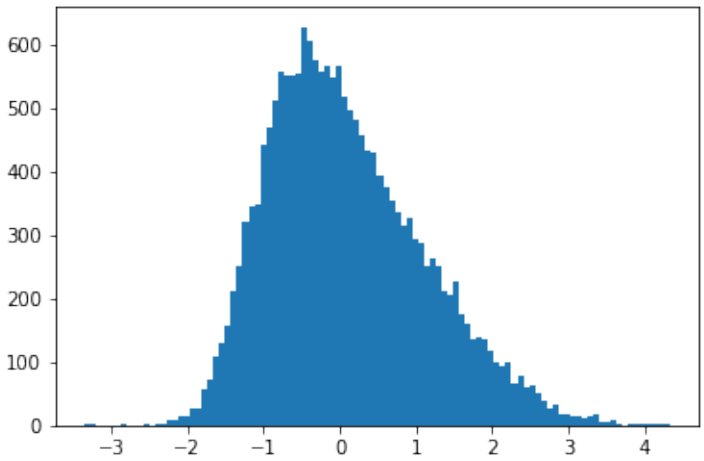


Figure 8: Z-score distribution of the raw scores for motif 2 of aging-related transcription factor STAT3. The raw total scores for every potential target gene of STAT3-motif2 were computed by scanning the 5000 bp upstream sequences and taking the sum for both the forward and reverse form of the motif. Because the scale of the scores is a function of the motif length, the scores were standardized by way of z-scores. The resulting distribution is plotted. A higher z-score is an omen for higher raw scores for the corresponding target gene and said gene is regarded as a gene with stronger interaction potential for the transcription factor under study. The mode (and mean) of the distribution is slightly negative, indicating that the majority of genes do not show strong theoretical binding capacity. However, the wider right tail is a prior for a few amount of genes with high z-scores and thus representing strong potential target genes.

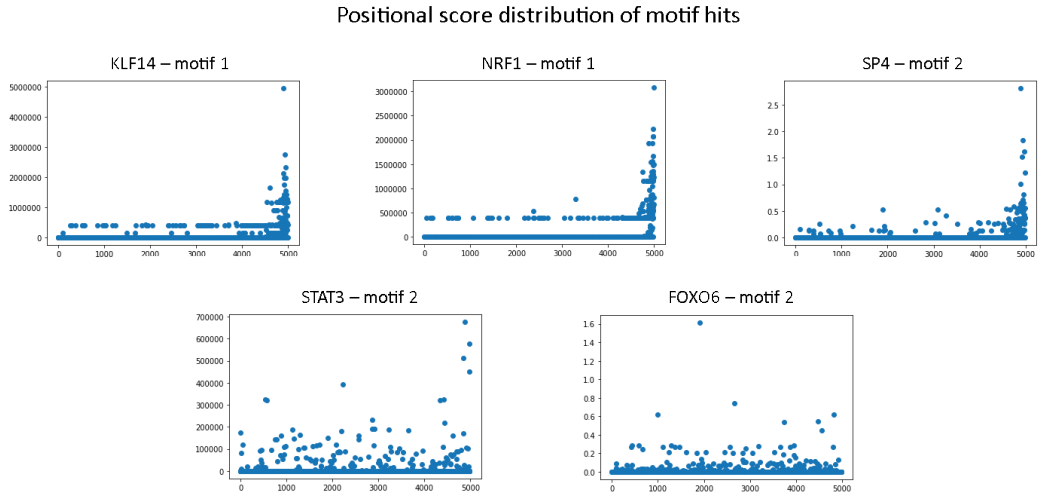


Figure 9: Positional distribution of motif hit scores for the 5 kb upstream promoter region. The 5 kb-long upstream regions relative to the transcription start site (TSS; position 5000) were scanned with 5 different motifs. The positional score distribution of motif hits scores for a variable amount (see Materials and Methods section) of top genes was summarized in scatter plots. Here, larger scores indicate stronger matches. The top three plots are the most effective motifs in terms of target gene prediction power. The bottom two motifs are familiar motifs used as examples throughout this document. For the well-performing motifs (the top level), the best matches (highest scores) are clearly located in close proximity of the TSS.

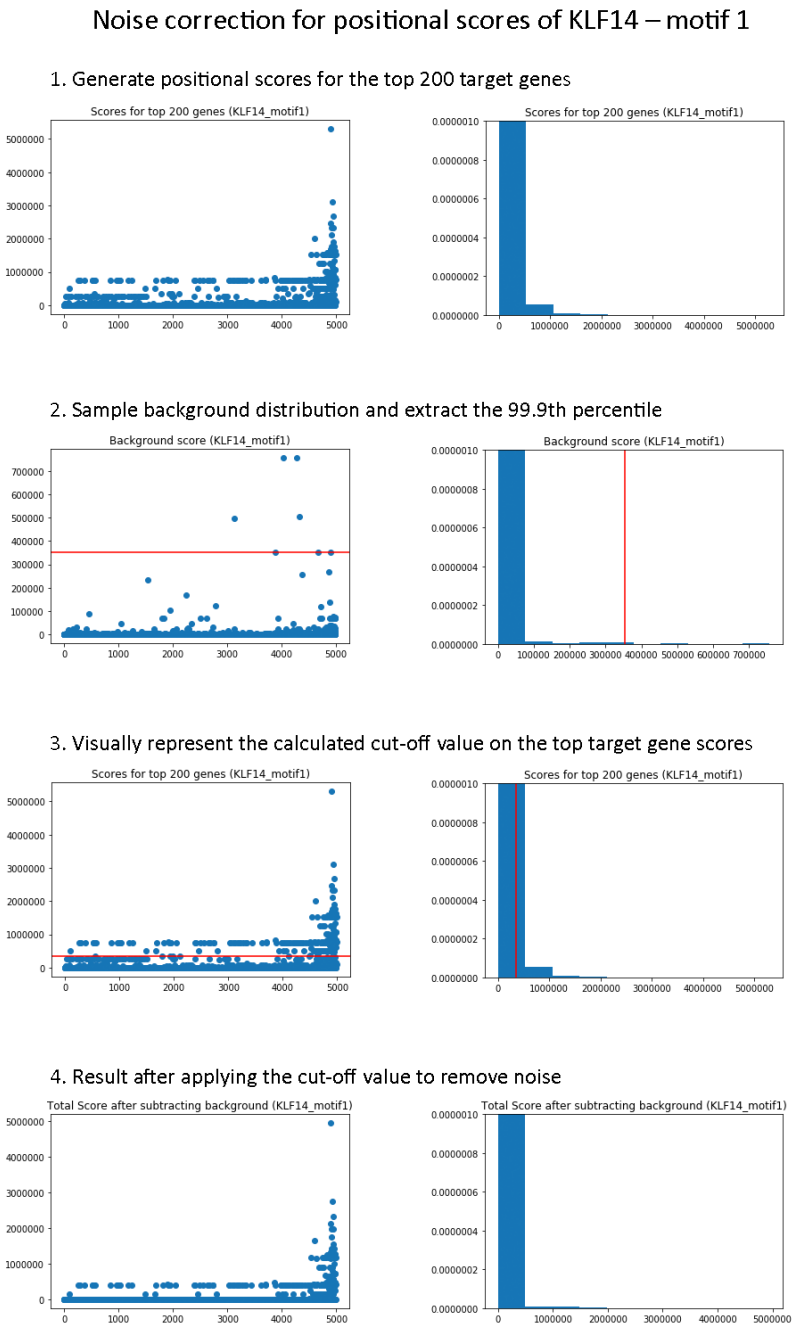


Figure 10: Step-wise diagram of the noise correction method applied to the positional scores. In this figure, the noise correction method is detailed for the example of the scores over the upstream promoter region for motif 1 of the KLF14 transcription factor. In the first step, the 200 best scoring target genes are gathered and motif hit scores over the entire upstream promoter region are presented, both in a scatter plot (top left) and in a histogram (top right) representing the distribution of the score magnitude. Position 5000 indicates the position of the transcription start site (TSS), whereas position 0 signifies 5000 bp upstream from the TSS. In the second step, the motif hits in the promoter regions of 500 randomly sampled genes are plotted. The value of the 99.9th percentile mark of the score distribution is computed. This value represents an upper bound of the motif hit scores for randomly samples target genes, and is therefore considered as the noise threshold. In the third step, the calculated noise threshold is shown for the 200 best potential target genes. In the fourth and final step, the noise correction is applied by subtracting the noise threshold (but with zero for negative values after subtraction).

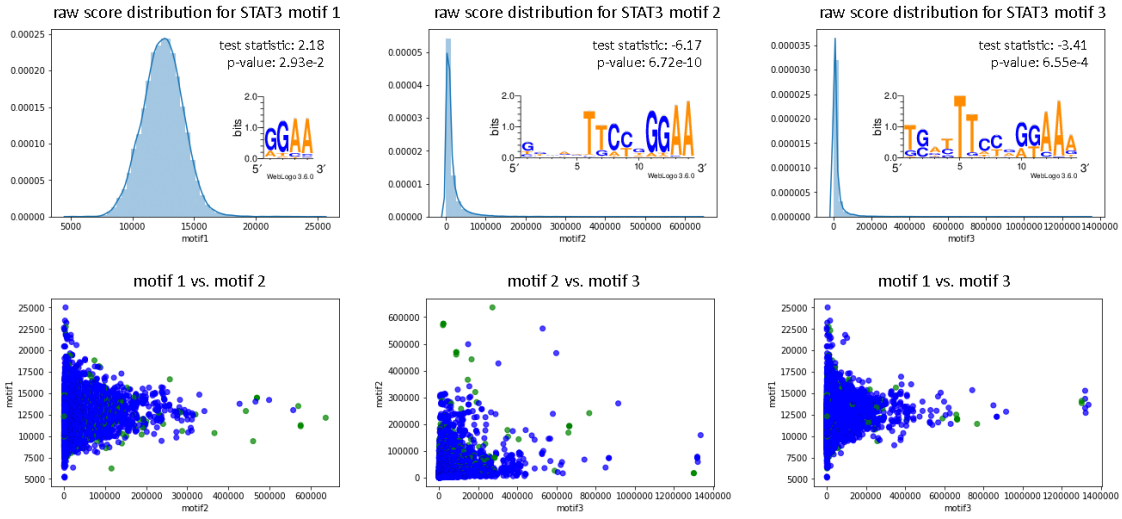


Figure 11: The raw total score distributions and scatter plots for the three motifs corresponding to the STAT3 transcription factor. The total scores are obtained after scanning the 5000 bp upstream regions of human genes with the motif under study. The visual motif representation is depicted, together with the effectiveness metrics in terms of target gene prediction. A negative test-statistic and significant p-value (less or equal than a critical value of 0.05) represents a significant association between high scores for database-verified target genes. The negative test-statistic indicates an enrichment of database support towards genes with higher scores and the p-value translates into the significance of this observation. Green data points in the scatter plots represent database-confirmed targets. Ideally, high-scoring target gene data points are expected to be green, which seems to be the case for motif 2 and 3. This observation is confirmed by their negative test-statistic and significant p-value.

7.3.3 Motif analysis

The resulting table for STAT3 was sorted based on the second motif (table 2). For this motif, the distribution of target genes confirmed by one or more databases seems to be a lot more dense towards the top scoring genes. This is an indication that the implemented scanning methodology is successful in identifying target genes. However, in order to draw a statistically sound conclusion, a Wilcoxon rank sum test is applied for every sorted motif-specific database support vector (table 3). In other words, the database information on the TF-target interactions is used to compare the scores between those with and those without independent support. A negative test statistic combined with a significant p-value measures a significant positive correlation between database support for target genes and a high scanning score. The corresponding motif is considered effective if and only if this condition is fulfilled (table 3). The best rank column is a representation of the smallest rank over the motifs for one transcription factor and is therefore a representation of the impact of the combination of the motifs. Similar to a sole motif, the test statistic and p-value is calculated and is added to the same table as the bottom row (table 3).

Because the Wilcoxon two-sample rank test was applied for every motif in the database, the test was to be conducted 1155 times. This test compares the scanning scores between known

Table 3: Target prediction effectiveness of the three motifs of the STAT3 transcription factor. For every motif, the derived scores for potential target genes were ranked. Subsequently, a Wilcoxon rank-sum test is applied to test if the database-verified targets are enriched towards the top of the ranked target gene list. This table lists the results of these Wilcoxon rank-sum tests used to determine the effectiveness of the DNA-binding motifs for the STAT3 transcription factor. A negative test-statistic indicates that more known targets are located towards the top of the ranking than towards the bottom. When this test-statistic is indeed negative and the p-value is below the Bonferroni corrected significance level, the motif is considered effective in predicting targets of the transcription factor.

motif	test-statistic	p-value	effective
motif1	2.178806	2.934612e-02	False
motif2	-6.172483	6.7222564e-10	True
motif3	-3.407745	6.550200e-04	True

versus unknown targets. This number of tests is subject to the multiple testing problem, which implies a correction regarding the significance threshold for the p-values was necessary. The combination of the motifs was not evaluated, which means 1155 is indeed the number of attempted tests at first sight. However, not all transcription factors are supported by at least one database. For these TFs, the test cannot be applied, effectively reducing the actual number of executed tests to 1095. The most common significance threshold for a p-value is 0.05, i.e. a 5% chance that the observation is witnessed under the null hypothesis. For 1095 tests, this threshold is corrected to reduce the number of false positives by way of the Bonferroni correction. α_{adj} was estimated to be 0.05/1095 or 4.5e-5.

633 of 1095 motifs (58%) were deemed effective in predicting target genes. For the remaining 462 motifs (42%), not enough statistical support was present to be considered effective after applying the multiple testing correction and under the constraint mentioned higher. In terms of transcription factors, 405 out of 686 TFs (59 %) have at least one significant motif after applying the very stringent Bonferroni correction. The remaining TFs (281 out of 686 or 41%) do not show at least one effective motif. Sorting by ascending p-value and test statistic produced an interesting list of motifs and their target-prediction capacity based on our scanning approach, with the most effective motifs on top (table 3). In addition, the information content (IC) of the motifs does not seem to relate to the test-statistic of the Wilcoxon rank test (figure 12).

Upon evaluating the top 20 transcription factors with the highest absolute test statistic, there appears to be enrichment for certain classes of TFs (table 8 in the appendix and figure 18). The KLF and SP prefixes are common and stand for Kruppel-Like Factors and 'Specificity Protein', respectively. Both are members of a family of transcription factors that bind GC/GT-rich boxes in the promoter, through 3 C_2H_2-type zinc fingers that are present at their C-terminal domains (figure 3). Using the GOrilla tool [Eden et al., 2009a] on the top 100 most effective (lowest p-value) motifs did not result in any significant enrichments, however.

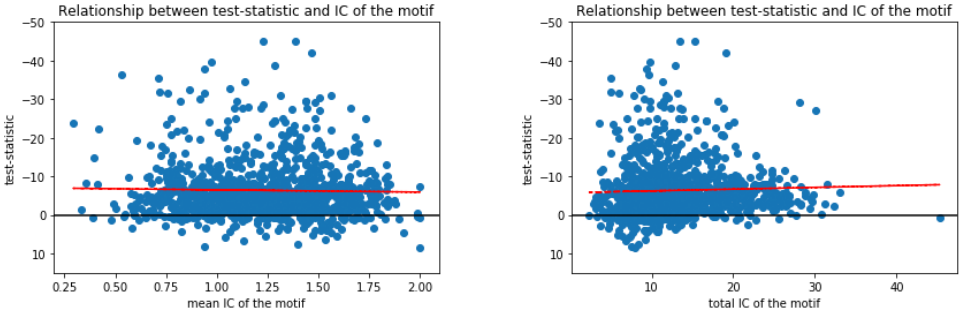


Figure 12: Scatter plot of the relationship between information content (IC) of a motif versus it's target-prediction effectiveness (measured by a strongly negative test-statistic). On the x-axis, the average (left) and total (right) information content (IC) of the motif is plotted. On the y-axis, the test-statistic corresponding to the Wilcoxon rank sum test with the database-verified targets. A more negative test-statistic represents better predictive power. The higher red line represents a linear fit, whereas the lower black line is a horizontal reference line at y = 0. In both plots, the red and black lines are close to being parallel and therefore no clear relationship between motif IC and predictive power is observed.

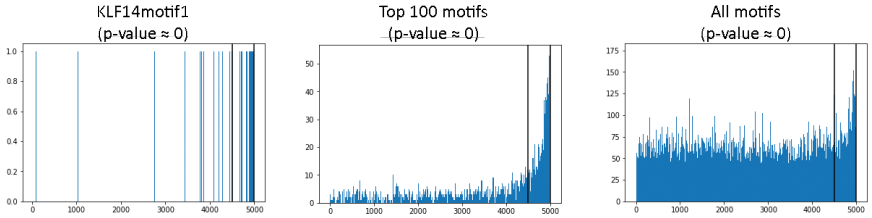


Figure 13: Regional analysis on the enrichment of motif hits in the 5 kb upstream region. The 5 kb promoter region was subdivided into 10 regions. Using the Poisson distribution, the 500 bp region upstream of the TSS (located at position 5000) was tested for an enrichment of large motif hit scores. In all three cases (most effective motif KLF14 - motif 1, the top 100 most effective motifs, and all motifs), the p-value of an enrichment is virtually zero. In other words, there is very strong statistical evidence for the hypothesis that motif hits in proximity of the TSS are more important than in tested regions that are more distal. This observation emphasizes the association between motif hits surrounding the TSS and motif-based prediction power.

7.3.4 Extending promoter regions

Because of the clear evidence in terms of enrichment around the transcription start site (TSS), the approach was extended by concatenating a 5000 bp long downstream region to the 5000 bp long upstream region to completely surround the transcription start site. The resulting motif-target gene scores were dealt with in a similar manner. The most significant motifs in terms of predicting power for target genes were generated and listed (appendix table 9). Although the ordering might differ, the top of the table is very similar to the upstream region table (Figure 16). Similar to the conducted upstream-only analysis, the scores were background corrected and summarized visually for the top motifs and aging-related examples (figure 15). From the figures it is clear that the region surrounding the TSS is a significant source of motif hits and the

quality of the matches around the TSS is a crucial factor in the target gene prediction power of the motif. The statistical significance of the surrounding zone is expressed in p-values (figure 17). For the motif with the highest predicting power (NRF1 - motif 1), the p-value based on the Poisson distribution is small enough to be virtually zero. Moreover, both the combination of the top 100 motifs and the combination of all motifs results in a rounded p-value of zero. Therefore, the enrichment around the TSS is therefore considered to be extremely significant and can be clearly observed from the plot. This suggests high efficiency for transcription factors binding around the TSS. In contrast, the motif with the least prediction power with regards to target genes (NOTO - motif 1) that had virtually no hits around the TSS resulted in a p-value as high as 0.999 (figure 17). Thus, the number of quality hits around the transcription start site tends to positively correlate with the predictive power.

In addition, splitting up the positional scores for the top target genes of NRF 1 - motif 1 in two plots, one plot for the 'known' targets that are verified by at least one of the database and one plot for the unsupported targets, reveals the positional differences in scoring between known targets and targets with high potential (figure 14).

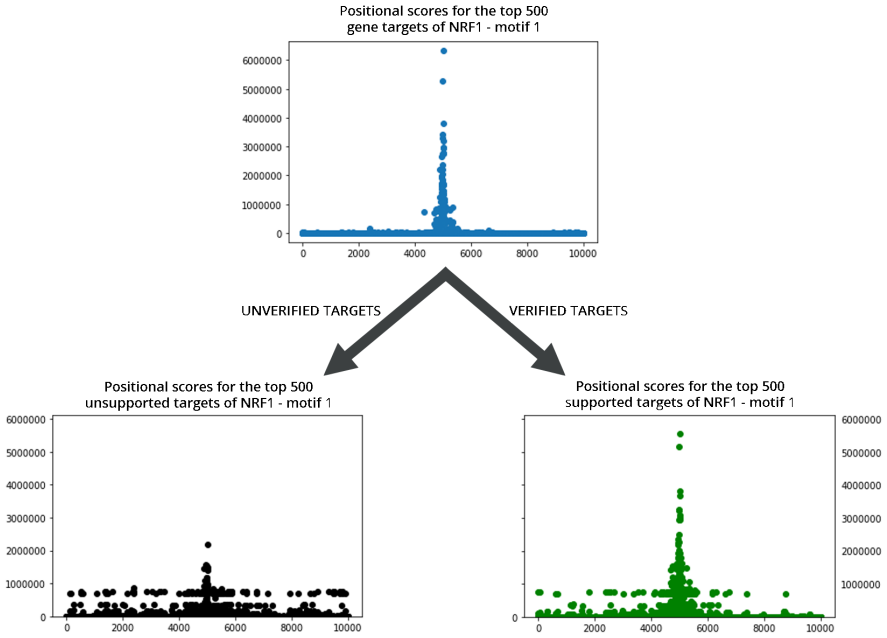


Figure 14: Difference in positional scores between verified and unverified database targets for NRF1 - motif 1. The x-axis values represent the position relative to the TSS, where the start of the TSS can be found at position 5000. Position 0 corresponds to 5000 bp upstream, and position 10 000 represents 5000 bp downstream of the TSS. The y-axis represents the scale (equal for all plots) of the scores. The top plot represents the score distribution of the 500 best scoring potential targets of NRF1 - motif 1. The split is made to distinguish the scores of the known targets (supported by at least one TF-target database) (bottom right in green) from the unverified targets (not supported by any TF-target database) (bottom left in black).

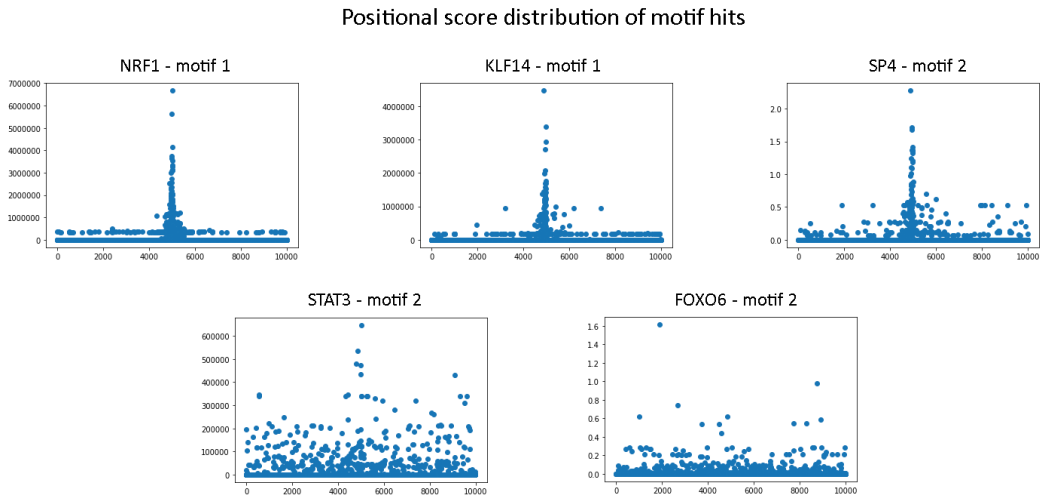


Figure 15: Positional distribution of motif hit scores for the 10 kb region surrounding the TSS. The 10 kb-long regions surrounding the transcription start site (TSS; position 5000) were scanned with 5 different motifs. The positional score distribution of motif hits scores for a variable amount (see Materials and Methods section) of top genes was summarized in scatter plots. Here, larger scores indicate stronger matches. The top three plots are the most effective motifs in terms of target gene prediction power. The bottom two motifs are familiar motifs used as examples throughout this document. For the well-performing motifs (the top level), the best matches (highest scores) are clearly located in close proximity of the TSS.

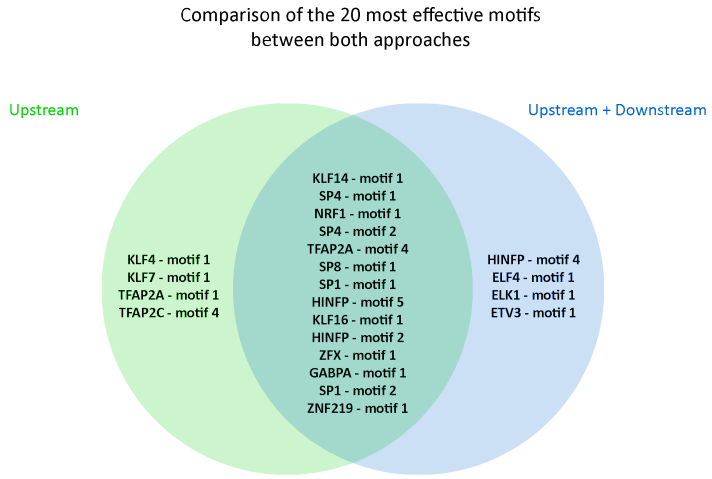


Figure 16: Comparison of the 20 most effective motifs between the analysis based on only the 5 kb upstream regions and the analysis based on the full 10 kb region surrounding the TSS. This figure represents the comparison between the top 20 most effective motifs in terms of target prediction power. The green half of the Venn Diagram depicts the top 20 most effective motifs based on only the 5 kb upstream region, whereas the blue half describes the top 20 most effective motifs for the full 10 kb region (5 kb upstream + 5 kb downstream) surrounding the transcription start site (TSS). The majority of transcription factor motifs are shared between both approaches.

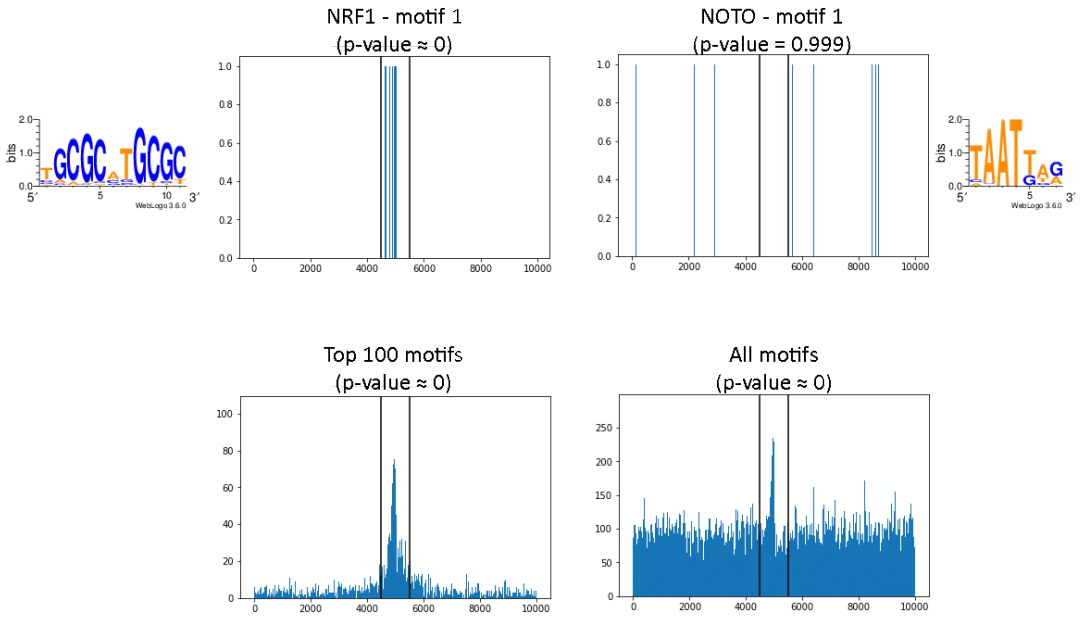


Figure 17: Regional analysis on the enrichment of motif hits in the 10 kb region surrounding the TSS. The 10 kb promoter region was subdivided into 10 regions. Using the Poisson distribution, the 1000 bp region surrounding the TSS (located at position 5000) was tested for an enrichment of large motif hit scores. In all three cases except NOTO - motif 1 (most effective motif NRF1 - motif 1 (top left), the top 100 most effective motifs (bottom left), and all motifs(bottom right)), the p-value for enrichment is virtually zero. In other words, there is very strong statistical evidence for the hypothesis that motif hits in proximity of the TSS are more important than in tested regions that are more distal. In the case of NOTO - motif 1, the least effective motif in the database, the p-value for the region around the TSS is virtually 1. This observation emphasizes the association between motif hits surrounding the TSS and motif-based prediction power.

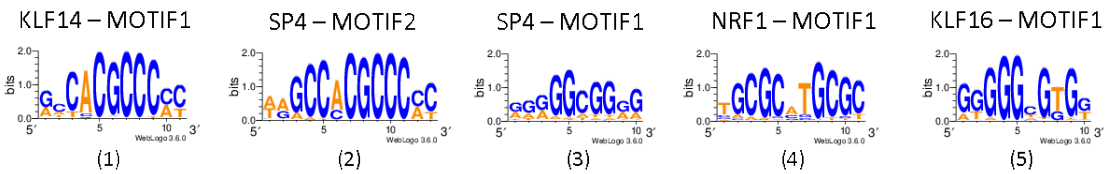


Figure 18: The top 5 most effective transcription factor DNA-binding motifs in terms of target gene prediction, in addition to their corresponding weblogo representation. In this figure, the weblogo's for the most effective motifs resulting from table 9 are shown. Interestingly, the best motifs consist mainly of a combination of guanine (G) and cytosine (C).

7.3.5 Score clustering

Additionally, the converted motif-specific z-scores and corresponding target genes were clustered with the CLUSTER 3.0 algorithm and visualized with TreeView 3.0 (appendix figure 33). As to be expected, similar motifs tend to cluster together as well as genes in the same GO categories (derived from manual inspection). Further manually investigating the resulting clusters is certainly interesting, but beyond the scope of this thesis.

7.3.6 Evaluating alternative scanning methods

The motifs were reformatted to the MEME motif format according to the guidelines on the website [Bailey et al., 2015]. In addition, the promoter sequences are poured into FASTA format in order to satisfy the FIMO input requirements.

The sequential runtime of FIMO is 4 times longer than our own implementation using convolutions (12 days compared to 72 hours) and when restricted to upstream sequences relative to the TSS. Sequentially scanning the entire 10 kb regions surrounding the TSS indeed requires FIMO to run for 15 days, however FIMO could be run in parallel to effectively reduce the runtime to four days. FIMO estimates the significance of a motif hit and only the significant motif hits are retained. Resulting from this significance cut-off, merely 51% of motif-promoter sequence combinations have at least one hit. The average amount of motif hits per motif-promoter sequence combination is 1.5, whereas the average amount of hits is 2.5 if the cases without any motif hits at all are disregarded. Downstream processing of the motif scores was done the same way as our very own implementation (Li lab) to avoid minimal bias.

Similar to the positional scores of our own implementation (figure 15), the positional scores over the full 10 kb sequence surrounding the TSS are plotted for the exact same motifs for the sake of comparison after undergoing the same processing steps beforehand (figure 19).

Although the signal-to-noise-ratio seems to be in favor of FIMO (figure 19), conclusions on the superiority of either genome scanning algorithm could not be drawn based on these visual observations, but can be derived from a validation data set with appropriate metrics. Because the ultimate performance metric is the performance of the integrated application with gene expression data to infer important transcription factors, the comparison will be elaborated on in the next section.

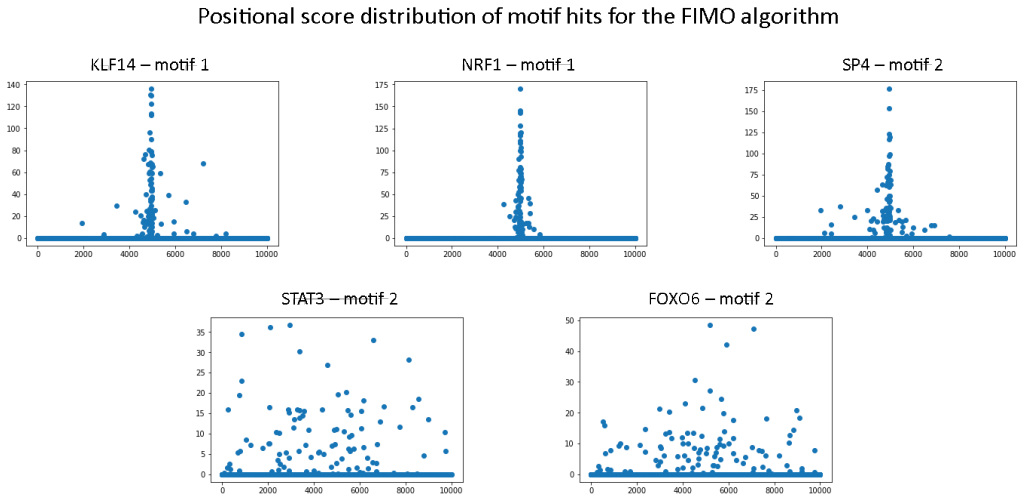


Figure 19: Positional distribution for FIMO-generated scores in the best scoring 500 full-range (10 kb) sequences and per motif. Similar to positional scores plotted for our own implementation, the FIMO-generated positional scores are plotted for the 500 best scoring sequences and the same motifs. Before plotting, the signal-to-noise ratio was increased by subtracting a score threshold representative of the background values. The x-axis represents all 10 000 positions in the 10 kb sequence surrounding the TSS, which itself is located at position 5000. The y-axis measures the strength of the motif hit. The top 3 motifs are the most effective motifs in terms of predicting targets, whereas the bottom two motifs are the aging-related examples used throughout this document. Similar to the analysis for the Li lab score vectors, strong motifs are characterized by motif-hit enrichment surrounding the TSS. This figure serves as comparison to figure 15.

7.4 Identifying crucial transcription factors in gene expression data

7.4.1 Method validation

The gene expression data was downloaded on June 13, 2018. These data are part of a study by Cusanovich et al. (2014) and concerns 59 different knock down experiments in which a different TF is knocked down in every experiment. 203 samples are available in the GEO archive, suggesting that some experiments are done in duplicate or triplicate. Upon filtering with the TFs represented in our database, the data corresponded to knock down information on 41 different TFs, which is still quite substantial as a sample pool. To achieve fold changes, the quantile-normalized and RUV2-adjusted gene expression counts were divided by the supplied background values. Next, an optional $\log 2$ transformation can be applied and is added to the performance comparison.

Subsequently, the correlation scores were computed between the motif score vectors and the gene expression fold change vectors. This results in a matrix in which the rows correspond to the TF motifs, whereas the columns refer to the gene expression conditions under study. The correlation values indicate the strength of the 'match' between the TF and the experimental

condition. The TF-motifs are sorted by descending correlation and the results are listed per condition, which equals the respective TF-perturbation experiment in the case of the validation experiment. Only the largest correlating motif occurrence of every TF is kept. Next, the perturbed TF is located in the sorted motif list for the corresponding perturbation experiment. The closer the TF is found to the top in the respective perturbation experiment, the better the performance of the algorithm. The results were summarized over the 41 different perturbation experiments and the performance was measured using three metrics: sum of ranks, median rank, and median rank percentage (table 4). The latter, median rank percentage, is the most interpretable, although the regular median rank is what determines the performance in practice.

The top 20 best performing implementations, i.e. the best combination of vector transformations, genome scanning algorithm, and correlation measure, is listed (table 4). In addition, a graphical representation of the impact of every parameter in the algorithm is included in figure 21. Based on the presented data, several conclusions could be drawn. Generally, a $\log 2$-transformation of the input gene expression fold changes appears to be mandatory. Secondly, the FIMO algorithm tends to be superior to our own implementation, which is clearly indicated by the absence of our implementation in the top 20 best performing implementations (table 4). Regarding the superior correlation metric, Pearson's correlation coefficient tends to edge over Spearman's rank correlation coefficient, although by a small margin. Surprisingly, in the very best implementations, the 5 kb upstream-only based genome scanning runs tend to perform better than the full-range (10 kb) scans. This is an indication that including the region 5 kb downstream of the transcription start site (TSS) is disadvantageous in the case of the identification of transcription factors relevant to the gene expression contrast. Finally, no clear tendencies could be observed for which motif score vector transformations are beneficial, including the incorporation of known TF-target interactions from databases. The best implementation takes $\log 2$-transformed gene expression fold changes as input, and applies $\log 2$-transformed upstream-only FIMO scores in order to calculate the Pearson correlation scores. The performance algorithm is detailed by the following performance metrics: median rank of 204 (top 27% in a database of 746 unique TFs), and a cumulative rank sum of 9216. In other words, in 50% of the perturbation experiments the causal transcription factor was ranked in the top 27% of important transcription factors. Experiment-specific results can be found in table 5. This implementation will be considered default from here on.

It is important to put the above results in perspective, because the results for this validation data set can not be generalized to every input data set. Intuitively, the algorithm is expected to underperform on this particular dataset. This statement can be explained by the fact that the TF perturbation contrasts are merely slight reductions in a sole transcription factor gene product. The adjective 'slight' was used here, because the efficiencies of the knock-down-caused perturbations are generally low. More specific, the fold changes of the targeted transcription factors range between 0.83 and 1.03, which is considered to be rather inefficient. Figure 20

shows that the more efficient the knock-down (lower FC), the higher the algorithm sensitivity. In other words, the sensitivity of the algorithm to pick up a minor reduction in gene product for a sole transcription factor was tested. In regular, strong gene expression contrasts (high FC difference over multiple genes), the results are expected to be substantially better.

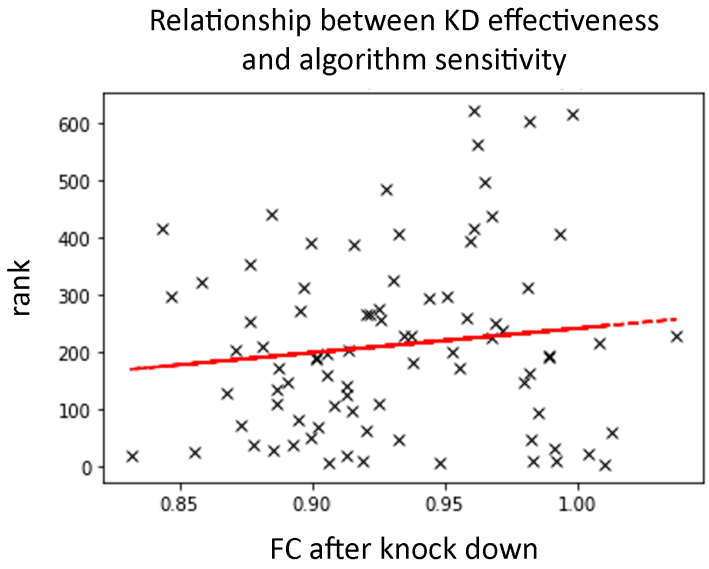


Figure 20: Algorithm sensitivity in function of knock down efficiency for the best performing method. On the y-axis, the rank of the causal TF in the perturbation experiment is presented (lower equals better performance), whereas on the x-axis, the FC of the perturbed gene after knock down is measured (lower FC equals higher efficiency of the knock down). Every data point in the graph presents a perturbation experiment. In red, a linear trendline is drawn. When the knock down efficiency is higher (lower FC), the algorithm is able to identify the causal transcription factor better (lower rank). The implementation used is the $\log 2$-transformed upstream-only FIMO-based scores on $\log 2$-transformed fold changes. The correlation metric used is Pearson's correlation coefficient.

Table 4: A comparison between the validation results of the top 20 different implementations of the TF identification algorithm. Based on the validation data, the best 20 implementations in terms of median rank are shown. The best performing implementation is shown on top and the other algorithms follow in ascending order. The first column (gene expression - log 2) describes whether the gene expression fold changes were log 2-transformed. The second column (correlation) shows which correlation metric is applied. The third column (algorithm) details if either the FIMO or Li lab genome scanning scores were used. The fifth and sixth columns (log 2 and z-score) describes whether the respective transformation was applied on the motif score vector. The 'database TF-targets' column indicates whether information on the known TF-target gene interactions was used in order to update the score. The last three columns measure the performance of the implementations on the validation data. In addition to this table, figure 21 presents a graphical comparison.

Gene Expression	**Correlation**	**Motif Scanning**					**Results**		
		Algorithm	Range	Score transformation					
log2	Pearson's or Spearman's	FIMO or Li lab	Fullrange (10 kb) or upstream-only (5 kb)	log2	Z-score	database TF-targets	Median Rank percent	Median rank	Rank sum
TRUE	Pearson	FIMO	5 kb	TRUE	FALSE	FALSE	27	204	9216
TRUE	Pearson	FIMO	5 kb	FALSE	TRUE	TRUE	27	205	9093.5
TRUE	Pearson	FIMO	5 kb	FALSE	TRUE	FALSE	27	205	9093.5
TRUE	Pearson	FIMO	5 kb	FALSE	FALSE	FALSE	27	205	9094.5
TRUE	Pearson	FIMO	5 kb	TRUE	TRUE	TRUE	28	206	10438.5
TRUE	Pearson	FIMO	10 kb	FALSE	FALSE	FALSE	28	207	9559
TRUE	Pearson	FIMO	10 kb	TRUE	TRUE	FALSE	28	207	9559
TRUE	Pearson	FIMO	10 kb	FALSE	TRUE	FALSE	28	207	9559
FALSE	Spearman	FIMO	10 kb	TRUE	FALSE	FALSE	28	212	11903
TRUE	Pearson	FIMO	5 kb	TRUE	TRUE	FALSE	29	213	9267
TRUE	Spearman	FIMO	10 kb	TRUE	FALSE	FALSE	29	213	10499.5
FALSE	Spearman	FIMO	10 kb	FALSE	FALSE	FALSE	29	213	11969
FALSE	Spearman	FIMO	10 kb	TRUE	TRUE	FALSE	29	213	11963
FALSE	Spearman	FIMO	10 kb	FALSE	TRUE	FALSE	29	213	11969
TRUE	Spearman	FIMO	5 kb	TRUE	TRUE	FALSE	29	220	10013.5
TRUE	Spearman	FIMO	10 kb	TRUE	FALSE	TRUE	29	220	10479
TRUE	Spearman	FIMO	10 kb	FALSE	TRUE	TRUE	30	222	10513
TRUE	Spearman	FIMO	10 kb	TRUE	TRUE	TRUE	30	222	10513
TRUE	Spearman	FIMO	5 kb	TRUE	FALSE	FALSE	30	226	10069.5
TRUE	Spearman	FIMO	5 kb	FALSE	TRUE	TRUE	30	226	10060.5

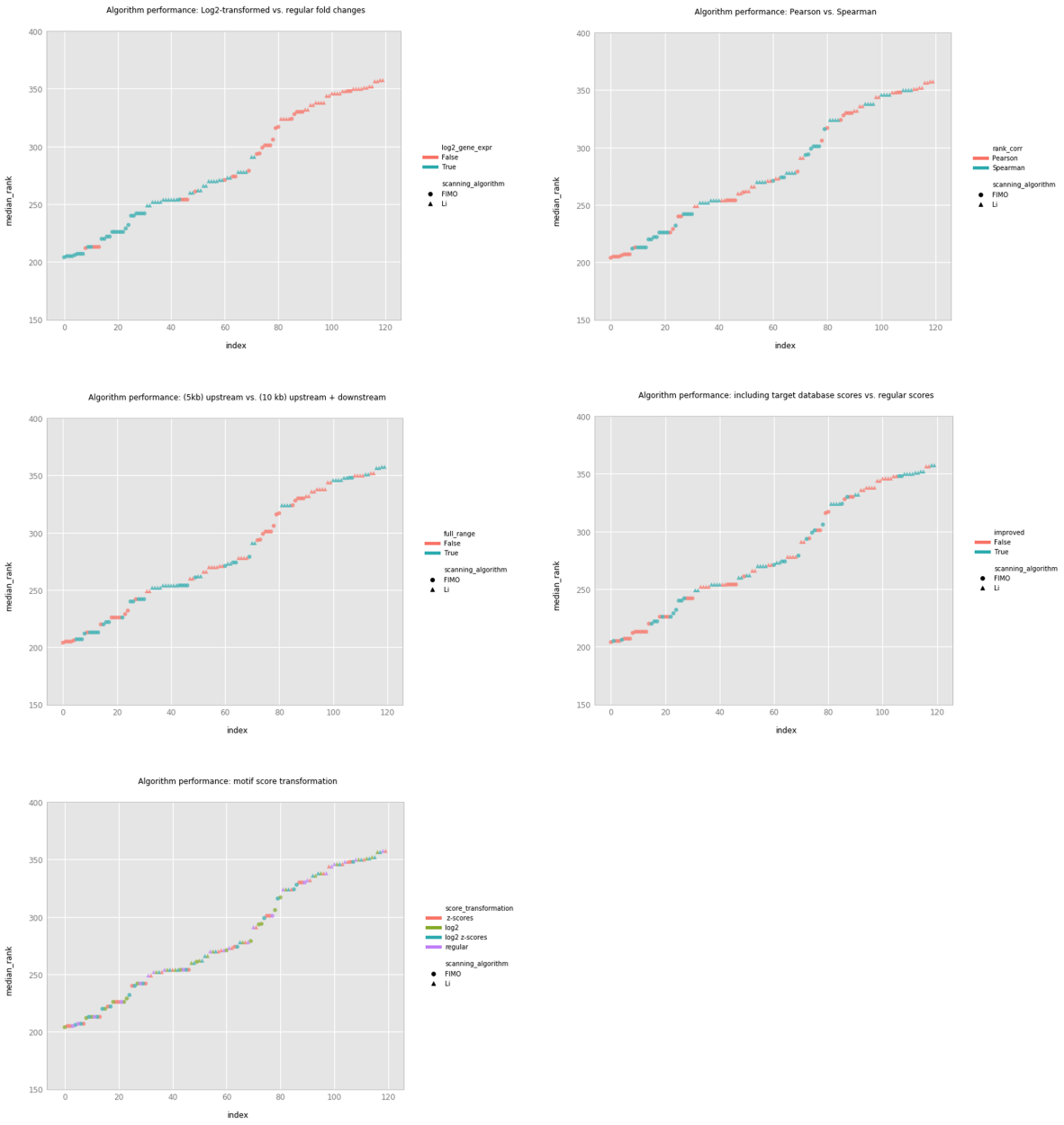


Figure 21: Performance of all possible algorithm implementations on the validation data set. On the y-axis, the median rank of the perturbed TF in the importance ranking is presented, which can be interpreted as algorithm sensitivity (lower equals better performance). The x-axis values are chosen as such that the algorithms are ranked by decreasing performance. On every subplot, a different parameter is color coded. On the top left, the effect of a log 2-transformation on the gene expression fold changes is shown. The top right subplot, the different correlation metrics are color-coded. The two middle plots show the effect of the length of the scanned sequence (left), and the effect of incorporating known TF-target interactions from the TF-target databases (right), respectively. The last plot (bottom left) measures the impact of different transformations on the motif scanning score vectors. In every subplot, FIMO-based implementations are marked by circles, whereas Li lab-based implementations are presented as triangles.

Table 5: Experiment-specific validation rankings for the best performing implementation. The gene expression data of 41 TF-perturbation experiments were fed into the algorithm, producing sorted lists of highly relevant transcription factors for every experiment. The perturbed TF is causal to these changes and should ideally be found towards the top of the ranked list, where the correlation with the gene expression data is higher. This table describes the sensitivity of this approach by returning the ranks of the perturbed TF in the corresponding experiment. The higher up in the rankings (smaller rank), the better the performance. The implementation used is the $\log 2$-transformed upstream-only FIMO-based scores on $\log 2$-transformed fold changes. The correlation metric used is Pearson's correlation coefficient. The entries are sorted by descending rank. The algorithm performs best for PAX5, and weakest for POU2F2.

TF perturbation experiment	Perturbed TF rank	Perturbed TF rank %
PAX5	10	1%
KLF13	17	2%
SP1	37	5%
E2F1	48	6%
TAF1	82	11%
TFDP1	97	13%
E2F4	108	14%
SP3	113	15%
IRF5	125	17%
JUND	135	18%
IRF4	147	20%
TFDP2	153	21%
ESRRA	157	21%
NFKB2	168	23%
IRF3	173	23%
BATF	180	24%
NR2F6	186	25%
CEBPG	190	25%
CLOCK	195	26%
CEBPZ	196	26%
RXRA	204	27%
ARNTL2	226	30%
RELA	229	31%
USF1	249	33%
NR3C1	251	34%
TFE3	259	35%
RELB	266	36%
YY1	266	36%
STAT2	271	36%
FOXA3	275	37%
STAT6	298	40%
SREBF2	312	42%
E2F6	316	42%
IRF9	323	43%
IRF7	358	48%
IRF8	386	52%
POU2F1	405	54%
TCF12	414	55%
HOXB7	414	55%
NFYC	415	56%
POU2F2	562	75%

7.4.2 Developing a web application

The fully functional web application is hosted on the Li lab servers. At the time of writing, the web application is available to members of the lab, although will be made publicly available in the future. The runtime for one gene expression contrast is less than a minute when a GOrilla GO analysis is included. This analysis tests the enrichment of the GO categories represented by the top TFs versus all tested TFs. If the automated GO analysis is excluded, the algorithm processes a gene expression contrast in a matter of seconds. An example of a screenshot including output results can be found in figure 22. Next to a ranking of the transcription factor motifs in the contrast under study, the corresponding motifs and correlation scores are also presented. The raw correlation data is also available for download. Results are available with the same URL until 30 days after the computation. Instructions on how to best apply the tool are detailed on the front page.

7.5 Application to Aging research

Having proven the effectiveness of the algorithm, we then proceeded by applying the pipeline to aging-related gene expression contrasts. This way, the method can be validated a second time by assessing the performance for TFs known to be related to these contrasts, in addition to deriving novel insights. In this section, the results were interpreted for the following three aspects. First, the TFs for which the promoter-binding scores correlate well with changes in gene expression were analyzed. Secondly, GOrilla GO analyses analyses were applied to identify potential GO enrichment in the most relevant TFs. Finally, the details of recurring subpatterns of DNA-binding motifs were evaluated.

7.5.1 Aging in the human frontal cortex of the brain: study 1

The study by Lu et al. (2004) compares gene expression in the human frontal cortex of the brain between two age groups, a group of people of $\leq$ 42 years and people aged $\geq$ 73 years. Gene-wise standardized expression values were downloaded from the NCBI database with series number GSE1572. Significance analysis of microarrays (SAM) software was used to compare young and aged groups to determine the list of genes with a high FC (> 1.5) and median false discovery rate (FDR) smaller than 0.01 [Lu et al., 2004]. This filtering step returned 463 significant probes that meet the FC requirement. However, only 166 probes (Affymetrix HG-U95Av2 oligonucleotide array) could be mapped to a gene symbol (36%). Although this dataset only consists of 166 genes after preprocessing, these genes are the ones that capture most information about the strong contrast that this aging gene expression experiment presents. The web application produces output in less than 1 minute, GOrilla query included (figure 22).

Figure 22: Web interface for the results of the first aging-related study on the transcriptome in cells of the human frontal cortex of the brain, accompanied with the GOrilla GO analysis results. This figure includes a screenshot of the results of the correlation analysis (left), and a zoomed barchart of the top 20 most relevant TF-motifs to the gene expression contrast under study (top right). The contrast under study is aged vs. young human frontal cortex gene expression from Lu et al. (2004). Next to the visual representation of the ranking, the correlation values can also be extracted from the table, in addition to the motifs WebLogo corresponding to the transcription factor. Finally, also the GOrilla GO analysis diagram is included (bottom right), in which the enriched categories are highlighted in yellow and the significant p-values are listed under the respective category.

The list of highly relevant transcription factors is led by POU2F1 (motif 5). Interestingly, motif 3 of POU2F1 also takes the third place in the ranking. One would expect to see high similarity in these both high-scoring motifs. However, these motifs are not very similar and therefore this is supporting evidence for the importance of POU2F1 in the aging contrast in the human frontal cortex. Furthermore, the same observation could be made for GATA1, although lower

ranked (motif 2 ranked 31st, motif 6 ranked 33rd and motif 4 ranked 44th), and SOX9 (ranked 28th and 50th). Moreover, 5 transcription factors in the top 20 are from the FOX protein family. Other remarkable TFs that are ranked in the top 50 are TCF15, PRDM6, BATF3, TAL1, ERG, and interferon regulatory factors (IRF5, IRF3, IRF6). These factors are all proven to be related to aging or inflammation in a broader context (see Introduction). In the top ranked motifs, the motif subpattern [C or G]AAA[C or G] or it's reverse compliment [G or C]TTT[G or C] is often represented amongst the top TFs, which suggests the importance of binding to genes on this compatible binding site in the context of human aging in the frontal cortex of the brain. One potential explanation is that genes involved in this gene expression contrast share this TF-binding motif in their biological sequence.

The inherent correlation calculation is also accompanied with an automatic GOrilla GO analysis (figure 22). This analysis tests the enrichment of the GO categories represented by the top TFs versus all tested TFs. Enriched categories are related to establishing monopolar cell polarity (p-value = 6.17e-4), maintenance of monopolar cell polarity (p-value = 6.17e-4), and establishment of (epithelial cell) apical/basal polarity (p-value = 6.17e-4). The gene basis of this enrichment are the TCF15 and FOXF1 transcription factors.

7.5.2 Aging in the human frontal cortex of the brain: study 2

The second study by the same research group Lu et al. (2014) is a more elaborate aging-related experiment in which data from 12 young (<40yr), 9 middle aged (40-70yr), 16 normal aged (70-94yr), and 4 extremely aged (95-106yr) tissue samples were analyzed. The microarray data were downloaded with accession number GSE53890 on June 22, 2018. Subsequently, the data were preprocessed in similar fashion to the first experiment and the resulting fold changes were $\log 2$-transformed preceding to applying our algorithm. All the different stages of aging were put in contrast relative to gene expression values from young patients. Therefore, the different contrasts are: extreme age vs. young, normal age vs. young, and finally also middle aged vs. young. This was done separately for male and female samples, resulting in 6 contrasts total. The samples from young individuals are the reference group in every sample.

Due to the abundance of data presented by these 6 gene expression contrasts, the description of the results will be limited to the most essential observations. For the top 20 TFs for every condition, a table was produced that indicates the ranking of the top 20 TFs of the condition under study in the other conditions (not under study). This way of presenting the data captures the most information possible for only 20 TFs. However, due to the large size, these tables are stored in the appendix (table 11, 12 and 13), whereas the GO analyses can be found in appendix (table 14). Finally, the motifs logos of the top 20 TFs for every condition are also listed (table 15).

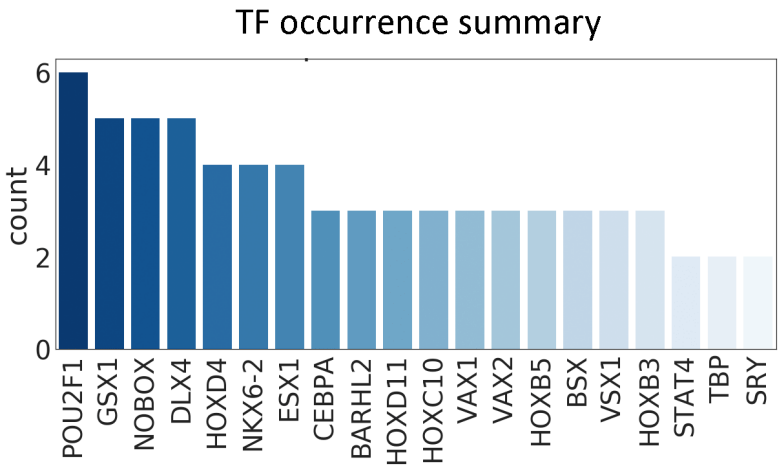


Figure 23: Summary of TF occurrence over the top 20 most relevant factors of all different contrasts described by the second aging study. In the second study by Lu et al. (2014), different stages of aged cells are compared with young cells (<40yr) in frontal cortex tissue of the human brain. The presence of the different transcription factors in the top 20 most relevant transcription factors over the different contrasts in this experiment is summarized in this histogram. The counts are represented by the y-axis, and the top TFs on the x-axis.

Here, a variation to common Venn diagrams is presented that highlights similarities and differences between gender and successive stages of the aging process (middle aged (40-70yr), normal aged (70-94yr), and extremely aged (95-106yr)) (figure 24). Secondly, a histogram representing the occurrence of the motifs over all six contrasts (3 for male and 3 for female samples, see higher) was generated (figure 23).

The TF with the strongest presence over the different conditions is POU2F1 with 6 occurrences. This means that over the 6 conditions, on average the TF occurs in the top 20 of every experiment. In fact, figure 24 shows that POU2F1 is the only TF that is present in the top 20 of each aged vs. young contrast. Next to POU2F1, the factors GSX1, NOBOX, DLX4, and HOXD4 populate the top 5 of most present transcription factors over the different contrasts.

In addition to the differences in the 20 most relevant factors between successive stages of aging, the Venn diagram also highlights the differences between gender (figure 24). When restricting the comparison to the top 20 TFs, the gender differences seem to increase with increasing age (few differences for younger age, and many for advanced age). However, it must be noted that this might be because contrasts in earlier aging-stages relative to young samples are not strong enough to rank the aging-related TFs reliably and consistently.

The motif table (table 15) reveals that for the mid-aged vs. young gene expression contrast, the recurring motif subpatterns align with the subpatterns in the first study. [G or C]AAA[G or C] is here observed as GAAA which is evidently considered to be similar. However, for the other contrasts, this exact motif rarely returns. Within these other contrasts, the motif subpatterns are similar and the pattern (T)AATT or a slight variation is observed.

Gene ontology enrichment analyses did not result in much additional insight (table 14). Only for the mid-age vs. young (male) and extreme-age vs. young (female) contrasts an enrichment with a p-value below the significance threshold was observed. The mid-age vs. young gene expression experiment originating from male tissue returned cellular amide metabolic process, negative regulation of smooth muscle cell differentiation, and diversification and recombination of T-cells as enriched GO categories. Extreme-age vs. young originating from the female transcriptome returned embryonic organ morphogenesis and embryonic skeletal system morphogenesis as enriched categories.

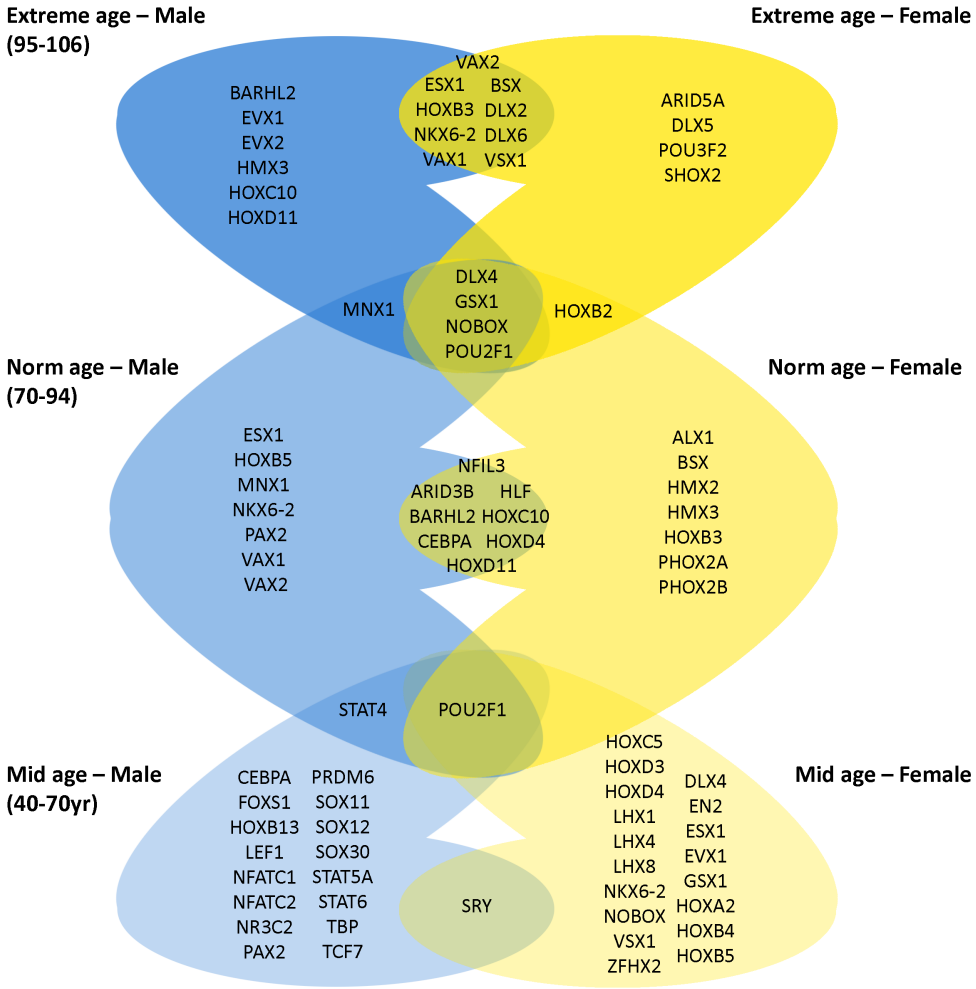


Figure 24: Multiple-way Venn diagram for the top 20 most relevant transcription factors over the 6 different aging-stage contrasts described by the second experiment of Lu et al. (2014). In this Venn diagram, the top 20 transcription factors for every old vs. young contrast are compared between the different successive stages of aging, in addition to the gender comparison. The different stages of aging were put in contrast with young cells (<40yr). The extreme aging contrasts (95-106yr) are shown on top, the normal age (70-94yr) vs. young contrast in the middle, and finally the mid age contrast (40-70yr) on the bottom. Besides the comparison between successive stages, both genders (male on the left and female on the right) are also compared. Overlapping sets indicate shared TFs in the top 20. Surprisingly, POU2F1 is the only transcription factor shared between the top 20 of every contrast.

7.5.3 Aging and rejuvenation in human skin cells

The RNA-seq gene expression experiment carried out by Chang AL et al. (2013) included 5 aged skin samples before and after broadband light treatment (BBL) (10 aged samples total), and 5 young skin samples. Broadband light treatment is practiced in dermatology centers and is considered to rejuvenate skin cells and to be beneficial when dealing with skin blemishes such as small facial veins, acne, wrinkles, and other aging-related conditions. The corresponding experimental data with accession number GSE39170 were downloaded on July 7, 2018. Before converting the RPKM (reads per kilobase of transcript, per million mapped reads) to fold changes, the raw expression levels were averaged within the sample groups. Subsequently, three different contrasts were created by division: untreated vs. BBL treated aged gene expression, untreated vs. young, and BBL treated vs. young. Subsequently, these data were converted to $\log 2$ fold changes and fed to the web application. The results were analyzed similarly to the aging experiments in the frontal brain cortex. In addition, the top 20 transcription factors are compared in a Venn diagram between the three different contrasts (figure 25). Finally, the top 20 TFs per condition were also formatted as a table, and the ranks of the mentioned TFs in the other conditions are included (table 6).

Before analyzing these results it must be noted that although a highly correlating TF can be shared between conditions, the effect of these transcription factors might not be the same. Because of the cooperative working of TFs and different circumstances on a cellular level, the same TF could target different genes or active a gene in one condition and repress gene activity in another condition. The approach only identifies the TFs involved. However, for the effects on their target genes we refer to the original publication.

The top TFs shared between the 3 different conditions are CEBPA, NR3C2, POU2F1, and STAT4. Secondly, between both old vs. young contrasts, the transcription factors ISL1 (48), LEF1 (159), NFATC2 (186), STAT6 (190), NFATC4 (200), and STAT5A (101) are shared. In the last sentence, the rank is listed for the experiment in which the TFs are not ranked within the top 20 (treated old vs. untreated old experiment) (table 6). Because these factors are not shared with the treated old vs. untreated old experiment, these factors are presumably more relevant to aging than they are to rejuvenation, as is indicated by their higher ranks for these factors. Moreover, between the BBL treated old vs. young experiment and the treatment vs. no treatment experiment, no factors are shared (as of the top 20). The transcription factors shared between untreated old vs. young and the rejuvenation experiment (treated old vs. untreated old) are NKX6-2 (40), SOX30 (31), SOX4 (57), and SRY (30). Again, between parentheses the rank in the BBL treated old vs. young experiment is indicated (table 6). The ranking in the latter is not particularly low, indicating that these are not entirely unrelevant to the third experiment and should not be treated as such. The remaining transcription factors mainly attributed to one experiment are EN2, MNX, NOBOX, PRDM6, SOX13 and VAX2 for untreated old vs. young. CUX1, ETS2, EUROG1, NFATC1, NFATC3, NFE2L1, NR3C1, PGR, STAT3, and UBP1 for BBL treated old vs.

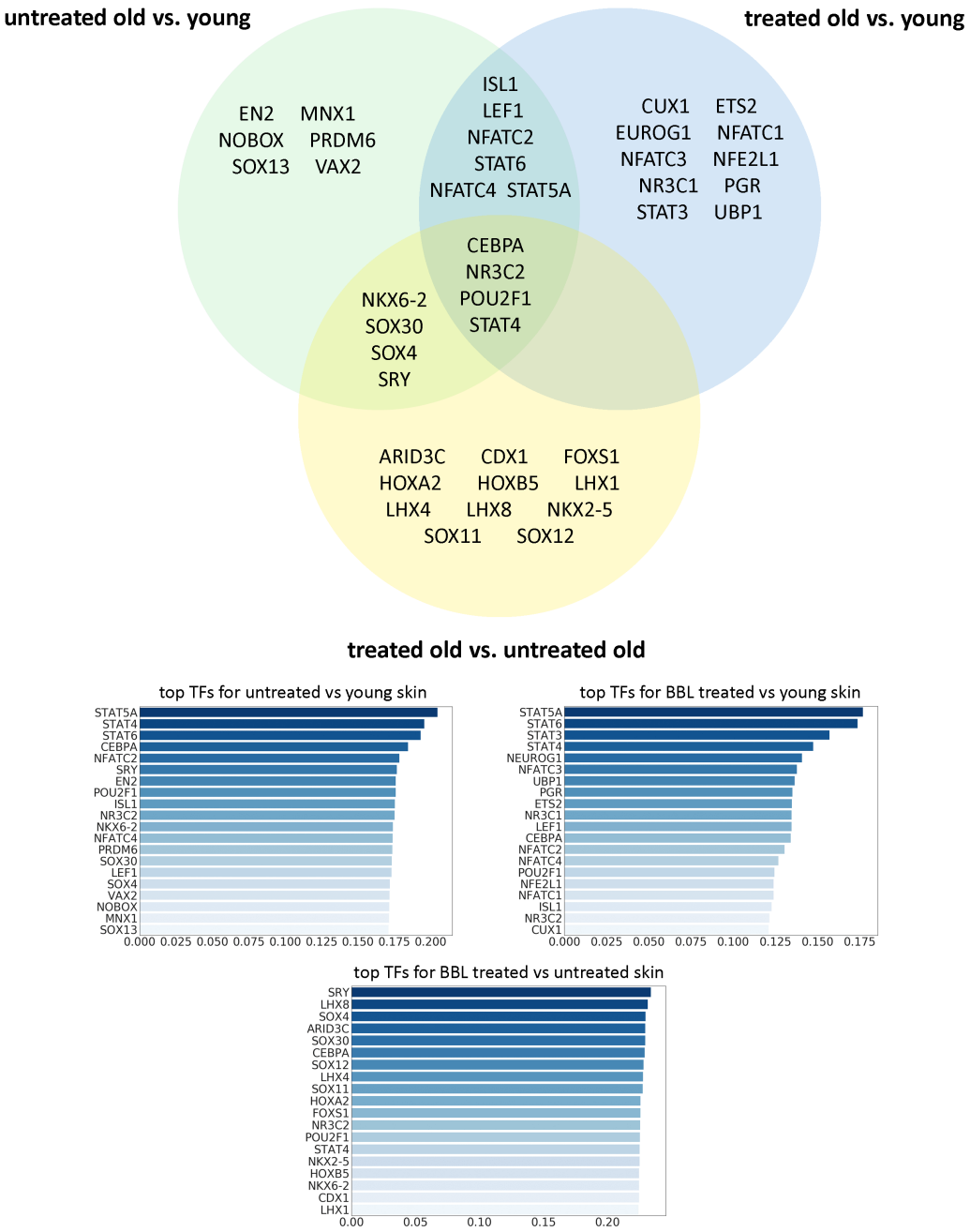


Figure 25: Visual representation of the top 20 TFs relevant to each gene expression contrast in the rejuvenation experiment. On the bottom, the 20 most relevant transcription factors to the different aging-related contrasts are depicted. On top, a Venn diagram representation is added. The different contrasts are untreated old vs. young (left), BBL treated old vs. young (right), and BBL treated old vs. untreated old (middle).

young. Lastly, the TF genes ARID3C, CDX1, FOXS1, HOKA2, HOXB5, LHX1, LHX4, LHX8, NKX2-5, SOX11, and SOX12 are only assigned to the rejuvenation experiment. The rankings in the other experiments can be retrieved in table 6.

In terms of recurring motif subpatterns (table 16), the pattern [A or T]AA[A or T] and it's reverse compliment seem to be present throughout the top factors of the BBL treated old vs. untreated old experiment. Genes with this DNA-binding site therefore seem to be more likely

Table 6: The top 20 transcription factor motifs in one contrast in addition to their respective ranks in the other two contrasts. This table lists the ranks of the top 20 transcription factors in one gene expression contrast in the other two gene expression contrasts. This representation allows to see how distinct two gene expression contrasts are in terms of the transcription factor relevance. The TF signatures of both contrasts are similar if the set of top-ranked transcription factors in one contrast is also well-ranked in the other contrast.

	BBL treated old vs. young skin	BBL treated old vs. untreated old skin	Untreated old vs. young skin
1.	STAT5A_motif2	101	1
2.	STAT6_motif2	190	3
3.	STAT3_motif1	412	124
4.	STAT4_motif2	15	2
5.	NEUROG1_motif1	144	59
6.	NFATC3_motif1	207	23
7.	UBP1_motif1	498	346
8.	PGR_motif1	296	208
9.	ETS2_motif2	517	295
10.	NR3C1_motif1	380	268
11.	LEF1_motif1	159	15
12.	CEBPA_motif1	6	4
13.	NFATC2_motif1	186	5
14.	NFATC4_motif1	200	12
15.	POU2F1_motif8	13	8
16.	NFE2L1_motif1	165	154
17.	NFATC1_motif4	152	62
18.	ISL1_motif1	48	9
19.	NR3C2_motif3	12	10
20.	CUX1_motif5	104	60

	Untreated old vs. young skin	BBL treated old vs. untreated old skin	BBL treated old vs. young skin
1.	STAT5A_motif2	101	1
2.	STAT4_motif2	15	4
3.	STAT6_motif2	190	2
4.	CEBPA_motif1	6	12
5.	NFATC2_motif1	186	13
6.	SRY_motif4	1	30
7.	EN2_motif1	38	22
8.	POU2F1_motif8	13	15
9.	ISL1_motif1	48	18
10.	NR3C2_motif3	12	19
11.	NKX6-2_motif1	18	40
12.	NFATC4_motif1	200	14
13.	PRDM6_motif1	124	35
14.	SOX30_motif1	5	31
15.	LEF1_motif1	159	11
16.	SOX4_motif1	3	57
17.	VAX2_motif1	41	47
18.	NOBOX_motif1	25	53
19.	MNX1_motif1	58	41
20.	SOX13_motif1	33	34

	BBL treated old vs. untreated old skin	BBL treated old vs. young skin	Untreated old vs. young skin
1.	SRY_motif4	30	6
2.	LHX8_motif1	59	32
3.	SOX4_motif1	57	16
4.	ARID3C_motif1	62	36
5.	SOX30_motif1	31	14
6.	CEBPA_motif1	12	4
7.	SOX12_motif1	52	26
8.	LHX4_motif1	73	46
9.	SOX11_motif1	44	21
10.	HOXA2_motif1	71	31
11.	FOXS1_motif1	90	45
12.	NR3C2_motif3	19	10
13.	POU2F1_motif8	15	8
14.	POU2F1_motif2	72	25
15.	STAT4_motif2	4	2
16.	NKX2-5_motif1	37	22
17.	HOXB5_motif1	60	27
18.	NKX6-2_motif1	40	11
19.	CDX1_motif1	109	56
20.	LHX1_motif1	123	75

to be involved in the rejuvenation mechanism caused by BBL treatment. Moreover, in terms of the comparisons of both treated and untreated old vs. young, a similar observation as the aging experiments in the human brain cortex can be made. The subpattern [C or G]AAA[C or G] and it's corresponding reverse compliment are by far the main TFBS observed in the top 20, similar to what could be derived from the previous experiments.

As expected, the GOrilla GO analysis mainly shows similar results for both untreated vs. young and BBL treated vs. young as the majority of signatures (and thus also top TFs) are shared. Many aging-related GO categories are enriched, such as the STAT cascade (p-value = 3.15e-4), calcineurin-NFAT signaling cascade (p-value = 1.87e-6), inositol phosphate-mediated signaling (p-value = 1.87e-6), cytokine production (p-value = 1.64e-4), and inflammatory response (p-value = 7.22e-4), amongst others. The actual difference between the BBL treated skin and

untreated skin is better characterized by the corresponding contrast and GO analysis. Surprisingly (see Discussion), the significantly enriched categories are motor neuron axon guidance (p-value = 1.37e-4), spinal cord development (p-value = 2.69e-4), neural tube formation (p-value = 6.17e-4), glial cell proliferation (p-value = 6.17e-4), and cardiac ventricle formation (p-value = 7.29e-4) as depicted in figure 26.

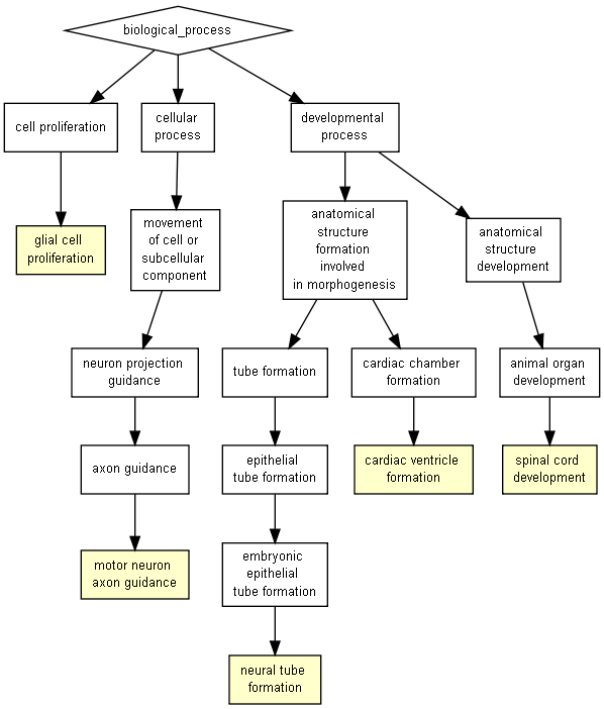


Figure 26: GOrilla GO analysis enrichment results diagram for the rejuvenation contrast on the transcription factor level. In this image, the GOrilla diagram for the significantly enriched categories is shown (highlighted in yellow) for the rejuvenation contrast (BBL treated old skin vs. untreated old skin).

7.5.4 Combinatorial analysis

Although drawing conclusions from a single experiment is a possibility, drawing conclusions from a combination of the experiments is undoubtedly the superior approach. In a combinatorial analysis, low quality data can be easily identified and left out by using other datasets as benchmark. A better picture of the importance of the top 20 TFs of every aging-related contrast can be generated by presenting a histogram indicating the occurrence of the transcription factors over all experiments (figure 27). Secondly, clustering the correlation scores of the TF-motifs in every condition is desirable in order to find (dis)similarities between experiments (figure 28).

Over 10 aging experiments, the histogram (figure 27) clearly shows POU2F1 as the transcription factor with the strongest presence over the different experiments, emphasizing it's importance. Other factors with a strong presence are CEBPA, NKX6-2, NOBOX, the STAT family (STAT4, STAT5A, and STAT6), GSX1, and DLX4, amongst others.

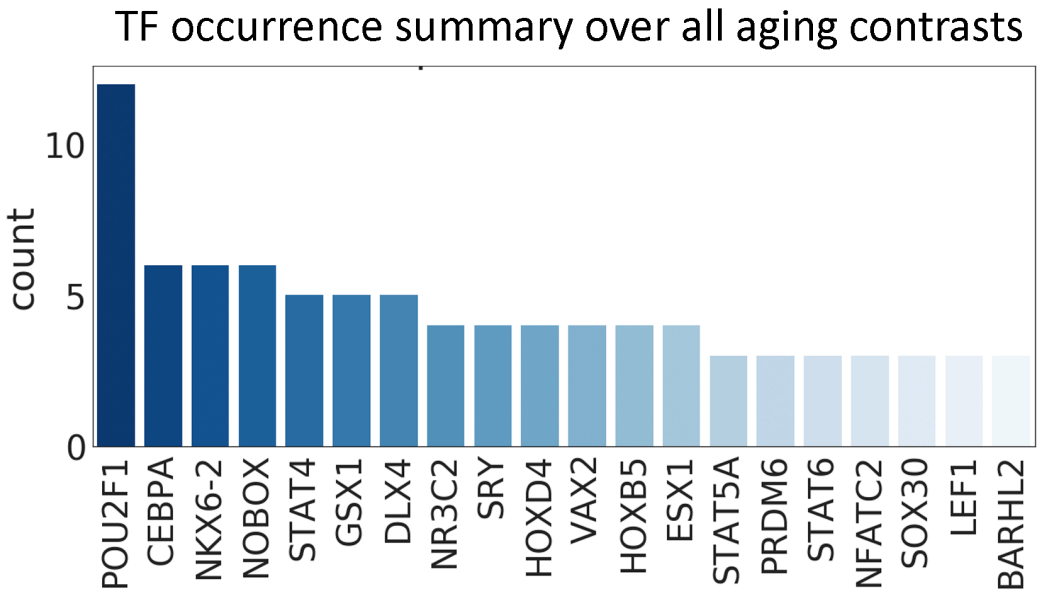


Figure 27: Histogram of the TF-motif presence in the top 20 TF-motifs over all aging-related gene expression contrasts. The occurrence of every TF-motif in the top 20 TF-motifs over all 10 aging-related gene expression experiments is counted and presented in a histogram. The counts are represented by the y-axis, and the top TFs on the x-axis.

The cluster of the correlation values was visualized by TreeView 3.0 (figure 28). The intensity of the colors represent the strength of the correlation. All contrasts are aged vs. young, except for the BBL treated old skin vs. untreated old skin contrast (surrounded by a red box in figure 28). Generally, with the exception of the control (inflammatory bowel disease), all aging-related correlation patterns are very similar, which is an indication that the BBL treated old skin vs. untreated old skin experiment resembles an old vs. young experiment when the transcriptome is analyzed on the level of transcription factors, i.e. the same TFs are relevant in the vast majority of cases. A second observation is that the BBL treated old skin vs. young skin contrast still has high similarity with general old vs. young experiments, which is an indication that the same transcription factors active in the BBL treatment vs. no treatment are also relevant to natural aging as expected, with very few exceptions. However, the effect of the broadband light treatment cannot be evaluated on the transcription factor level, due to the fact that these well-correlating TFs can change their method of action in each experiment (repression or activation, and different set of target genes). Moreover, untreated old skin vs. BBL treated skin, and untreated old skin vs. young skin are neighboring experiments and exhibit high similarity in terms of tree distance. A probable explanation can be found in the fact that the treated and untreated samples come from the same patient. Interestingly, for the gene expression experiments that study the different stages of aging, high similarity is found in the relevant transcription factors. For every stage but the extreme age stage (95-106), the male and female samples are neighboring. In the extreme age stage, the stage difference tends to matter less than the difference in gender, as the male and female experiments are no longer neighbor-

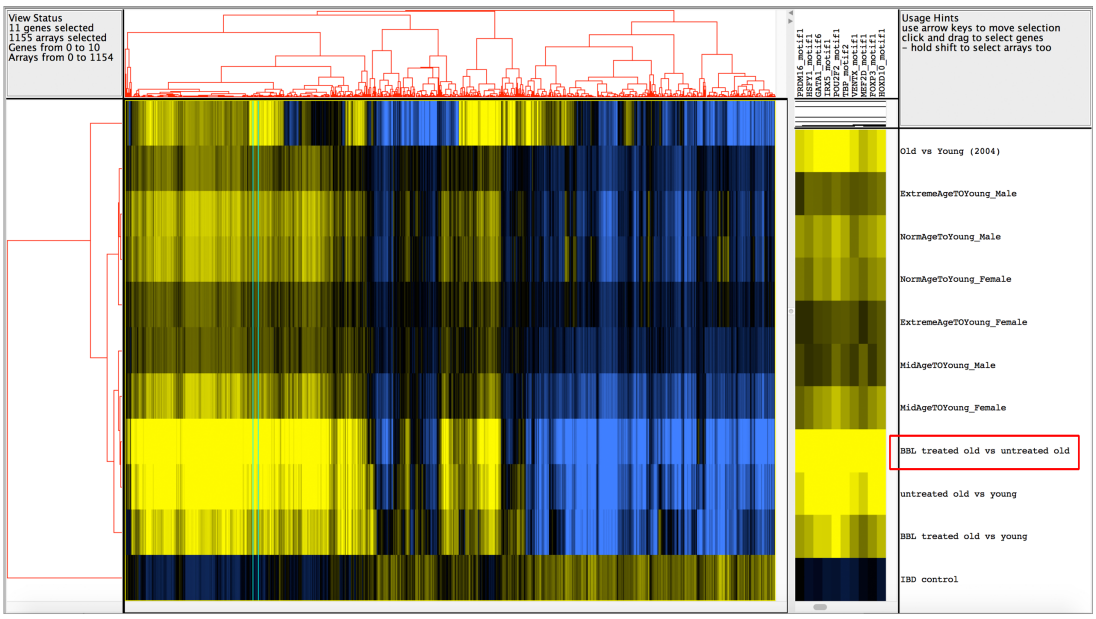


Figure 28: Clustering results of the transcription factor motif correlation values with all aging-related gene expression contrasts. Rows correspond to the different aging-related contrasts, whereas the columns represent the TF-motifs. The bottom row is a control experiment, more specifically the correlation vector for a gene expression contrast of inflammatory bowel disease (case vs. no case), and is the only row not directly related to aging. Yellow pixels mark positively correlating TF-motifs (columns) for the respective contrast (rows), and blue pixels indicate negative correlation. Color intensity measures the strength of the correlation (strong yellow for strongly correlating combinations, and strong blue for very bad correlation values). Correlation values close to zero are assigned a dark color. All contrasts are aged vs. young, except for the BBL treated old skin vs. untreated old skin contrast (surrounded by a red box). This cluster was generated using CLUSTER 3.0 and visualized using TreeView 3.0.

ing, but instead are located close-by the preceding stage of the same respective gender. The last observation is an indication for a gender difference in terms of transcription factors for extreme lifespans, although it must be mentioned that merely 4 samples are included that share the origin of extremely aged tissue.

As implied higher, similar transcriptional signatures on the TF level does not necessarily mean similar gene expression signatures. This assumption provided grounds to investigate the gene expression signatures on both the regulatory (TF) level and the gene expression level. In addition to clustering the transcription factor motifs of the second experiment by Lu et al., the gene expression data were also clustered separately (figure 30). As a reminder, in this experiment three successive stages of aging for male and female patients were placed in contrast with gene expression data from young patients. Although the gene expression patterns tends to be somewhat different for the majority of genes in all six contrasts, the same TFs seem to be relevant (high similarity for TF signatures). Similar regulatory patterns for different, but functionally related, gene expression levels is an indication that these TFs are indeed responsible for the changes in gene expression. In figure 29, a similar representation was made for the rejuvenation experiment. Here, the gene expression data is exhibits even lower similarity between contrasts, whereas the transcriptional signature is near identical. Additionally, the

observation that the control TF signature (inflammatory bowel disease in figure 28) disagrees with the aging signatures, is reason to believe that our methodology did not present a bias in favor of the TF signature observed for the aging experiments. Dissimilarity in gene expression patterns and the absence of a bias towards aging-related regulatory patterns are arguments to prove that these TF signatures are in fact typical for aging. This illustrates the effectiveness of the proposed methodology.

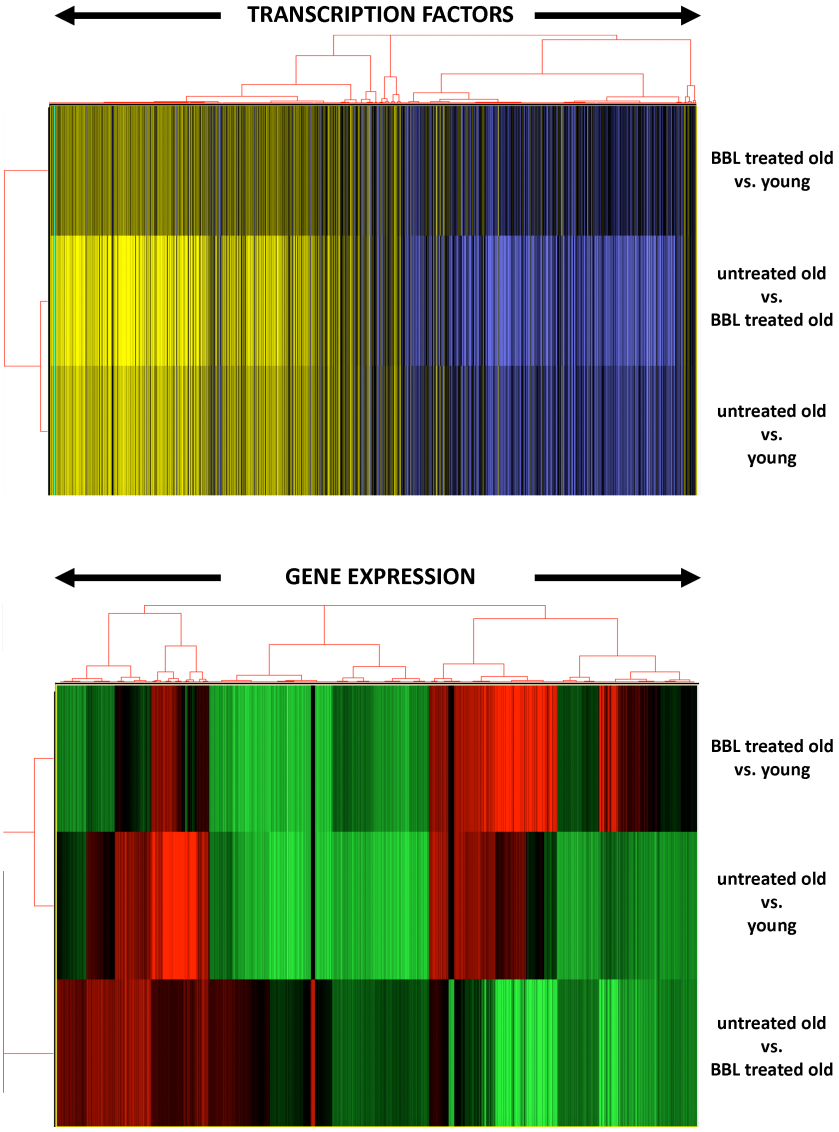


Figure 29: Comparison between the gene expression patterns and the resulting transcriptional signatures for the contrasts involved in the rejuvenation experiment. The bottom cluster illustrates the gene expression values for the three contrasts of the rejuvenation experiment. Horizontally, the genes are shown, whereas vertically the contrasts can be distinguished. On top, the resulting regulatory signatures are clustered. Here, the transcription factor motifs are shown horizontally, whereas the contrasts can again be distinguished vertically. The correlation scores used to cluster the transcription factor motifs were derived from the methodology proposed in this manuscript. The regulatory signatures tend to be extremely similar in all three contrasts. However, the gene expression patterns (on which the former observation is based) are rather dissimilar. Both clusters were generated by CLUSTER 3.0 and visualized with TreeView 3.0.

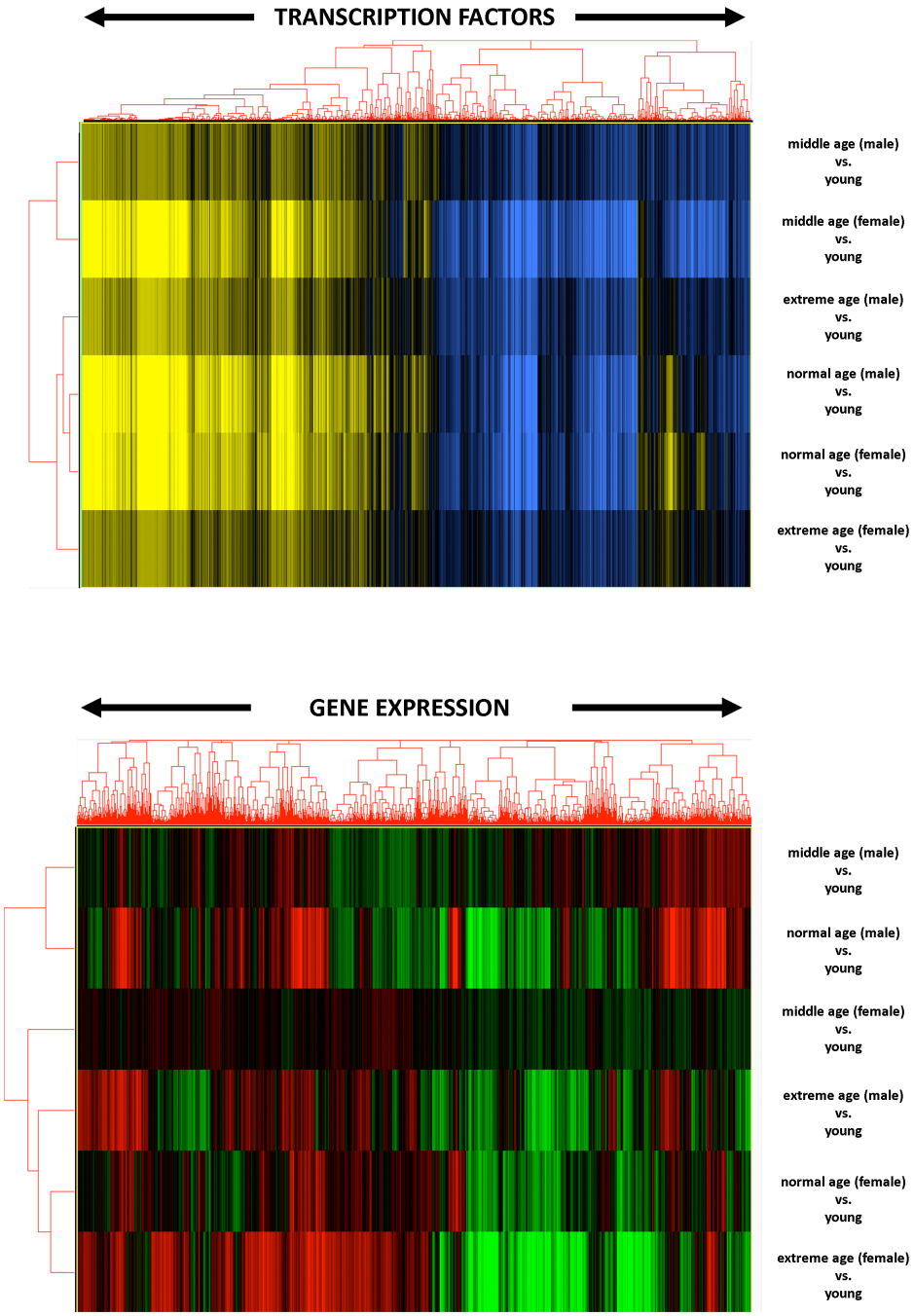


Figure 30: Comparison between the gene expression patterns and the resulting transcriptional signatures for the contrasts involved in the second aging experiment by Lu et al. The bottom cluster illustrates the gene expression values for all six contrasts of the second aging experiment by Lu et al. Three successive stages of aging for male and female patients were placed in contrast with gene expression data from young patients. Horizontally, the genes are shown, whereas vertically the contrasts can be distinguished. On top, the resulting regulatory signatures are clustered. Here, the transcription factor motifs are shown horizontally, whereas the contrasts can again be distinguished vertically. The correlation scores used to cluster the transcription factor motifs were derived from the methodology proposed in this manuscript. The regulatory signatures tend to be extremely similar in all three contrasts. However, the gene expression patterns (on which the former observation is based) exhibit fewer similarity. Both clusters were generated by CLUSTER 3.0 and visualized with TreeView 3.0.

8 Discussion

This manuscript describes a successful attempt to develop a robust approach able to identify the transcription factors involved in a gene expression experiment. To the best of our knowledge, no similar tool is available (as of August 2018) that is able to identify crucial transcription factors with similar sensitivity and simplicity. With simplicity, we imply general ease-of-use in addition to high speed and very few input requirements. Given that this tool does not rely on additional input information besides gene expression fold changes, the use case of this algorithm is extendable to any human gene expression study. Moreover, the integrated gene ontology analysis provides, together with the list of important transcription factors, extra interpretability that can lead to insight that would otherwise be missed.

The results on the validation data set prove that the algorithm is able to pick up small reductions in a sole gene with considerable accuracy (the perturbed gene is ranked in the top 27% or higher for 50% of the experiments). The consistency of the results across the aging case-studies (see further) suggests that the in-practice performance is way better for strong gene expression contrasts and that the validation experiment merely presents the lower bound of the algorithm performance. However, understanding the limitations inherent to the proposed methodology is extremely important regarding interpretation and will be discussed in this section, in addition to discussing the results.

8.1 Collecting human transcription factors and corresponding motifs

To begin with, from the 1447 predicted unique transcription factors in the human genome, only 746 were attributed at least one motif (52%). This means that the other 701 transcription factors (48%) are not present in the motif database. As a result, they will not be included in the subset against which the human promoter sequences are scanned. These factors are therefore excluded from downstream analyses, i.e. the correlation score computation. Thus, their effect, either in combination with different transcription factor or as a sole contribution to the changes in gene expression level, is disregarded. It must be noted that incomplete results favor the other motifs, because the output of the algorithm is a ranking, which is by definition a relative comparison (all vs. all). Fortunately, updating the TF-motif combinations in the database is straightforward.

Secondly, the approach fully depends on the assumption that the motifs in the original databases are correct. Given the good results for the validation experiment, this assumption can generally considered to be acceptable. During the course of this project, the human transcription factor database was released [Lambert et al., 2018]. Benefits of using the motifs from this dataset is that these motifs are curated and that corresponding evidence is provided. At the start of this analysis, the database was not yet made publicly available, but ideally should be added to the

collection of motifs to further improve our pipeline. In case of high motif similarity, the benefit of the doubt should be given to the data entries from Lambert et al. because of the curated nature of the database.

In terms of motif processing, calculating similarity between two motifs was initially done using the BioPython implementation. Manual inspection reveals that this method aligns motifs well on the PWM level, but that the comparison does not translate well to the PPM level. Our own implementation, which is rather simple compared to the introduced methods, performs surprisingly well on the PPM level. The PCC is a proven metric for PPM similarity calculation, although the alignment constraint has shown to be a necessary adjustment for good performance in this analysis. Aligning motifs on the level of consensus sequences would result in a major loss of information and is therefore not an option.

Subsequently, merging of the motifs was done using the unweighted merging approach due to the limitation that evidence (leading to the weights) is not explicitly listed for all motifs. Although assessing the differences between the weighted and unweighted merging showed that the variations are small, the weighted approach remains the right approach intuitively as the advantage is given to better supported motifs. The motif curation work by Lambert et al. (2018) combined with (weighted) motif merging on manually investigated clusters has the potential to be the best state-of-the-art transcription factor-motif database by a large margin and is assumed to improve downstream algorithms like ours. Preceding motif merging, the similar motifs are assigned to a cluster. The cluster tree was split using an arbitrary cut-off that was proven to be effective in practice [Madsen et al., 2017a]. This cut-off correctly separated the clusters in the majority of manually investigated cases, although some splits could be improved on. Manual cluster assignment and splitting, either unsupervised or guided by results of automated clustering, could potentially improve the quality of further used motifs.

Finally, but most importantly, the DNA-binding motifs of the transcription factors are derived from a variety of *in vitro* experiments. Inference of TF-DNA binding in these experiments is done under different cellular circumstances. This phenomenon of tissue-specific TF-gene interactions can be explained by the dependence of the interaction on many features, e.g. chromatin accessibility, the presence or absence of other TFs, formation of regulatory protein complexes consisting of multiple TFs, and more.

8.2 Collecting targets of human transcription factors

The genes that are regulated by the transcription factors derived from *in vitro* experiments were downloaded from several different databases. The target genes for the transcription factors in these databases show extremely little overlap between the databases. Similar to the case of the motifs (which are derived from these known TF-gene interactions), this observation can be partially explained by the tissue-specific activity of the transcription factors, subject to the

cellular circumstance in which the experiment took place. Although the databases tend to disagree, the union was taken to form one dataset and was subsequently applied to determine the effectiveness of the genome scanning approach. Based on these 'known' targets, the Li lab genome scanning approach showed good performance. The results of the rank-sum tests, using this prior knowledge in terms of target data, can be used to filter out non-informative motifs. As expected, the shorter motifs with less average IC were considered less effective than longer, high IC motifs. Additionally, given the reliability of the target databases, providing feedback on the preceding motif gathering and processing approach is a major bottleneck.

8.3 *De novo* target prediction by genome scanning

The problem regarding the effect of cellular conditions on TF-gene interactions directly translates into complicated TF-promoter binding score estimation. The lack of incorporation of information on the cellular circumstances is a double-edged sword. The disadvantage is that a bias is introduced in favor of the transcription factor - promoter interactions for which the *in vivo* behavior is similar to the *in silico* prediction. For example, transcription factors for which the interaction with the target gene largely depends on other parameters (e.g. the presence or activity of other transcription factors, eu/heterochromatin, methylation, etc.) are modeled less accurately, which creates a bias in favor of transcription factors that solely depend on the strength of potential binding sites. However, the performance on the validation data set shows that the final implementation is valuable, making the fact that these experiment-specific parameters are not required an advantage. The latter is important, because the absence of additional input requirements makes the approach applicable to a broad spectrum of gene expression experiments. Ideally, if the data were available, we would compute tissue-specific TF-target binding scores and select the appropriate binding scores for the gene expression contrast under study.

Both FIMO and the Li lab implementation are characterized by similar positional distributions regarding motif matching scores (figure 15 and 19). However, the FIMO approach results in slightly higher signal-to-noise ratio's, most likely resulting from removing insignificant matches in the filtering step. Generally, adding the downstream region to the scanned sequence does not increase the performance, which implies that most information determining the DNA-binding capacity could be extracted from the upstream region. This observation was expected, because in the majority of cases the regulatory elements are located upstream of the gene. Moreover, the region around the TSS is highly enriched for transcription factor DNA-binding sites (figure 17). However, caution is advised when drawing the conclusion that strong TF-target gene interactions are very likely characterized by binding sites close to the TF. The latter can be explained by stating that this analysis is biased towards interactions proximal to the TSS (within 5 kb). In addition, the fact that TF-TF interactions and distal regulatory elements (e.g. affecting chromatin structure) are not modeled is also of utmost importance.

Interestingly, the motifs with the best predicting power (table 18) almost exclusively consist of the G or C nucleotides. Additional explanatory analysis (data not shown) revealed that there is no preference in favor of this type of motifs, other than that the vast majority of these transcription factors are from the specificity protein (SP) / Krüppel-like factor (KLF) family. These families bind GC/GT-rich boxes in the promoter, through 3 C_2H_2-type zinc fingers that are present at their C-terminal domains (figure 3). These motifs are thus characteristic to the zinc-finger domains for which the TF-target interactions apparently can be predicted well.

Improvements to these genome scanning methods evidently can be found in introducing additional parameters to better describe the experimental circumstances (e.g. including other omics data). One option is exploring the field of supervised machine learning, although every application comes with specific requirements and puts more emphasis on correct training labels. In the case of target gene prediction based on biological sequences, the training labels directly correspond to the TF-target interactions present in the databases. The effectiveness of these machine learning methods will depend on the tissue-specificity and general reliability (i.e. quality) of these databases, in addition to the amount of data present (i.e. completeness). Another example of a potential improvement is factoring in the location of the motif match in the promoter region (e.g. the application of kernel-based methods) or perhaps studying the combinations of motif matches. Finally, adjusting the aggregation method of these scanning scores has great potential, next to combining the results of the different DNA-binding motifs of a TF into one score per TF. These suggestions are worth exploring in future research, because the performance of the final implementation stands or falls on the effectiveness of the TF-target scores.

8.4 Identifying crucial transcription factors in gene expression data

The claim that the downstream region does not contribute essential information to the algorithm performance is supported by the benchmark results on the validation data. Moreover, the superiority of the FIMO method can likely be attributed to the q-value-based significance cut-off before score aggregation. Furthermore, assessing the performance of Pearson's correlation coefficient over Spearman's correlation score is challenging, but a disadvantage of Spearman's correlation score can potentially be found in the fact that the separation between noise and signal TF-gene interaction scores is less clear due to the non-parametric nature of the formula. For example, in the case of few strong TF-target scores, the effect of these hits is almost negligible in a pool of noise when a non-parametric metric (i.e. ranking in this case) is applied.

The data presented in the comparison table (table 4) shows that the perturbed TF is placed near the top with considerable reliability (top 27% or better for the median experiment). However, the knock-downs used in the validation experiments only show a reduction of expression of the target gene ranging from about 17% to even an increase of 0.03% in expression level. Given that

these knock-downs do not fully restrict the expression of the gene and are only limited to the one gene targeted by siRNA, the performance of the algorithm is strongly expected to improve in gene expression conditions where multiple genes are affected and the absolute values of fold changes are larger. In other words, this validation experiment measures the sensitivity of the algorithm to register a very slight reduction in the gene product of a sole transcription factor. Therefore, when applied to more distinct contrasts in gene expression, the results are expected to be more evident. To prove this statement, the approach was be put to the test by applying it to more distinct contrasts such as aging-related gene expression experiments (old versus young). Globally known and proven assumptions applicable to the condition under study were used to verify the approach, in addition to deriving novel insights. The results are discussed in the next section.

Furthermore, having the perturbed TF not returned as the largest correlating transcription factor does not necessarily prove that the algorithm is less effective. When a transcription factor is knocked down, compensation mechanisms may be activated. Moreover, when a TF interacts with other TFs, a malfunction of one TF may directly translate into different behavior of the co-operative TFs, which in their turn impact a specific set of overlapping target genes. Therefore, the effect of a knock-down of one transcription factor is often observed from the signatures of other transcription factors which are part of the same transcriptional unit. If this hypothesis indeed holds true, a second application for this approach can surface: perturbing one transcription factor (similar to how the validation data set was obtained) would return a list of transcription factors affected by the perturbation in the causal transcription factor. In other words, one can identify transcription factors related to the perturbed TF in the specific tissue corresponding to the gene expression experiment, without requiring wet lab experiments (although wet lab experiments remain necessary to validate the hypothesis). This way, transcription factors either cooperating with the perturbed TF or regulated by the perturbed TF may be identified. A major advantage of this approach would be that also indirect targets are unveiled, since gene expression values are the base of the connection here, not TF-DNA interactions. Direct targets can potentially be verified by integration with tissue-specific ChIP-seq data. Revealing TF-TF interactions has the potential to uncover important networks, although is still subject to the limitation that the theoretical genome scanning method can only be applied for TF-DNA interactions in which the *in silico* predictions hold true in practice.

In addition, measuring enrichment of gene ontology categories of top TFs versus all tested TFs is certainly interesting and can provide more insight. However, the difficulty is that the number of TFs in our database is rather low (746), leading to low statistical power to detect enriched categories, if any. Ideally, enriched categories on the level of transcription factors are expected to correspond to enriched categories on the level of the measured genes in the gene expression contrast.

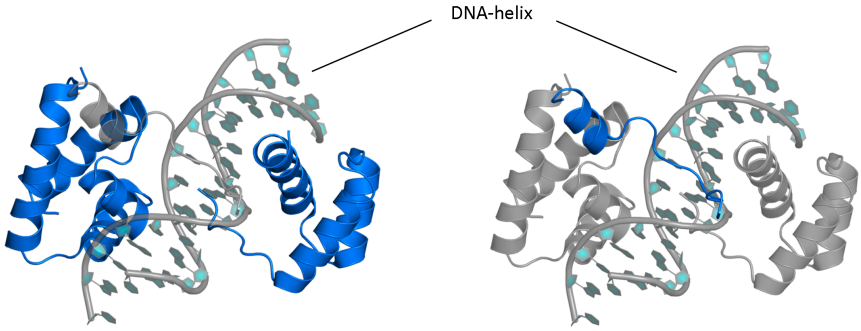
Further research regarding the integration of additional data sources can be suggested on the level of the gene expression data. For example, transcription factors can be weighted based on their relevance in the affected pathways. Relevant pathways can be inferred from strongly impacted genes on the gene expression level. Additionally, the amount of gene product of the relevant TFs could also provide more context for inference.

Finally, it is evident that because this approach builds upon the strengths of the genome scanning methods, it also builds upon the weaknesses (discussed higher). Generally speaking, in order to obtain the highest statistical power for the discoveries made by the algorithm, it is advised to combine the output of multiple studies on similar gene expression contrasts to get the best results in terms of the ground truth. Importantly, it must be noted that even though the same transcription factors may be considered relevant in similar gene expression contrasts, it does not automatically imply that the method of action of these TFs is perfectly equal.

8.5 Applications to Aging research

Throughout the applications to aging gene expression experiments, the presence of POU2F1 (alias OCT-1) is undeniably strong. From literature, we know that POU proteins are eukaryotic transcription factors characterized by the POU domain. This domain (about 150 amino acids) was derived from the pituitary-specific Pit-1, octamer-binding proteins Oct-1 (POU2F1) and Oct-2, and the neural Unc-86, hence the name POU (Pit-Oct-Unc) [Latchman, 1999]. The POU2F1 transcription factor is known to bind the octamer motif 5'-ATTTGCAT-3' (figure 31) and regulates genes for snRNA, as well as genes for histone H2B and immunoglobulins. In addition, the POU2F1 transcription factor modulates transcription transactivation by N3C1, AR and PGR [Segil et al., 1991]. Moreover, information on UniProt regarding the DNA-binding motif of POU2F1 (5'-ATTTGCAT-3') states that POU4-related factors preferentially bind two distinct half-sites, GCAT and TAAT, separated by a spacer region (not conserved) of 0, 2, or 3 nucleotides. The vast majority of POU4-related TFs are expressed at high levels in the brain, which could partly explain the high occurence of the TAAT subpattern in the results of the analysis on the aging-related gene expression data originating from the frontal cortex of the brain. Due to the high presence of POU2F1, the GCAT pattern is also observed, but is less abundant.

Because similarity in motif yields a similar ranking for the corresponding TFs, other POU proteins are also expected to perform well. Going back to the obstacle of the algorithm that the methodology is solely motif-based, in contrast to being function-based or accounting for tissue specificity, further adjustments are necessary to filter out these false positives. False positives (TFs with high theoretical binding capacity for the impacted genes, although not functionally related to the study) can potentially be filtered out (or weighted) using prior knowledge on the functional pathway(s) involved in the experiment. Incorporating networks to which the highly impacted genes in the gene expression contrast belong could possibly take the proposed algorithm to an unprecedented next level.



(a) POU domain, class 2, transcription factor 1 (POU2F1)

(a) POU domain, class 2, associating factor 1 (POU2AF1)

Figure 31: 3D structure of POU2F1 and the associated factor during binding of a DNA molecule. On the left, POU2F1 (highlighted in dark blue) is modeled while interacting with DNA, whereas on the right, POU2AF1 is highlighted. D. Chasman et al. (1999) obtained this 3D structure by utilizing X-ray diffraction with a resolution of 3.2Å.

Evidence for the relevance of POU2F1 (alias OCT-1) regarding aging is not abundant, but exists [Wu et al., 2014, Infante et al., 2014]. However, in a publication earlier this year, a link between Alzheimer's disease and herpesvirus has been confirmed [Readhead et al., 2018]. For POU2F1, it is known that malfunction of the TF is detrimental for immediate early gene expression at low multiplicities of herpes infection [Nogueira et al., 2004]. Herpes simplex virus infections are therefore arrested in Oct-1 (POU2F1) deficient cells. The method of action is the formation of a multiprotein-DNA complex by POU2F1, including the viral transactivator protein VP16 and HCFC1 and thereby enabling the transcription of the viral immediate early genes. This analysis supports the hypothesis of POU2F1 as a link between aging and Alzheimer's disease, although indirectly through the herpes simplex virus. Because herpes simplex virus type 1 (HSV1) leads to an increased risk of developing Alzheimer's disease, POU2F1 activity may increase with increasing age. However, according to the expression values of POU2F1 and POU2AF1 over the discussed contrasts (table 7), there is no clear evidence for an increased expression with increasing age. If POU2F1 is in fact involved, the method of action of the transcription factor must change, as the TF is not shown to be differentially expressed. For example, evidence for phosphorylation of the POU2F1 gene product in response to genotoxic stress was detailed elsewhere [Kang et al., 2009]. In BBL-treated skin tissue, a reduction in the expression level of POU2F1 is observed (table 7). Moreover, Oct1 is related to the Yamanaka factors (Oct3/4, Sox2, Klf4, c-Myc) which are known to play an important role in the proper functioning of stem cells throughout the entire lifespan of the individual. Stem cells are partially responsible for the rejuvenation capacity of the tissue.

The GOrilla GO analyses accompanying the TF ranking must be interpreted with caution. The total set of TFs contains about 746 unique TFs. This means that the top 10/20/50 TFs are tested versus all 746 TFs. In addition, GOrilla only recognizes about 90-95% of these TFs. Furthermore, in the list of top TFs, some TFs can be present multiple times because one TF can have multiple

Table 7: $\log 2$**-transformed changes in expression level for POU2F1 and POU2AF1 in the discussed experiments.** The expression values of POU2F1 and the associating factor POU2AF1 were traced back to the published values. The fold changes for the gene expression contrasts were $\log 2$-transformed and listed. The first 'Old vs. Young' experiment corresponds to the first study by Lu et al. (2004). The next 6 contrasts correspond to the second study by Lu et al. (2014). Finally, the last three contrasts are part of the rejuvenation experiment. Negative values indicate a reduction in expression level compared to the reference experiment in the contrast (indicated by the second part of the contrast name). Vice versa, positive values indicate an increase in expression relative to the reference sample group.

Contrast	POU2F1	POU2AF1
Old vs. Young	N.A.	N.A.
Middle aged vs. Young (M)	-0.005	-0.021
Middle aged vs. Young (F)	0.009	0.019
Normal aged vs. Young (M)	0.015	-0.004
Normal aged vs. Young (F)	0.018	0.035
Extreme aged vs. Young (M)	-0.006	0.006
Extreme aged vs. Young (F)	0.020	0.103
BBL treated old vs. Untreated old	-0.172	0.141
BBL treated old vs. Young	-0.197	0.697
Untreated old vs. Young	-0.025	0.556

motifs in the top ranking. GOrilla removes multiple occurrences, but this reduces the number of TFs in the top. The combination of these factors results in rather low statistical power and is expected to result in GO term associations based on few evidence. An example of a plausible enrichment are the GO categories relating to establishing and maintenance of cell polarity for the first aging-related study by Lu et al. The latter observation was verified in multiple studies [Florian and Geiger, 2010, Soares et al., 2013]. In contrast, the significantly enriched categories for the broadband light (BBL) treated old skin vs. untreated old skin are the following: motor neuron axon guidance, spinal cord development, neural tube formation, glial cell proliferation, and cardiac ventricle formation (figure 26). These GO associations are unlikely to originate from skin cell gene expression data. A possible explanation is the tissue-specificity of the transcription factors higher up in the ranking. These might in fact be related to the effects of the BBL treatment, but their actual function might be different in skin cells versus other cells. Therefore, it is advised to manually inspect GO associations and to verify the GO categories with GO term enrichment in significant differentially expressed genes. In other words, the significantly enriched GO categories on the gene expression level ideally correspond to the enriched categories on the TF (regulatory) level. The previously mentioned suggestion of filtering (or weighting) TFs based on function using known pathways would likely circumvent this problem.

Lastly, an explanation regarding the molecular changes at the basis of the rejuvenation experiment is desirable. We resort to the original publication, which concludes that the gene expression data of 2265 coding and noncoding RNAs were altered, 1293 of which became 'rejuvenated' after BBL treatment. In other words, the expression level of 1293 RNA molecules became more similar to their expression level in youthful skin [Chang et al., 2013]. 1112 genes with consistent changes in expression only occurring in BBL-treated samples but not in either untreated young or untreated aged samples were found. Visual representation of these gene expression changes due to treatment with pulsed light can also be found in figure 29 (clustered). Changes in gene expression are thus well characterized, although changes on the regulatory level of transcription factors are less apparent (figure 29). Including the inflammatory bowel disease gene expression contrast (versus healthy individuals) as a control in the clustering proved that the observed TF signatures for aging are not a coincidence, but that the regulatory pattern for old skin vs. BBL treated old skin in fact does resemble an aging signature. In addition, both the untreated old vs. young and the BBL treated old vs. young contrasts show similar patterns. These observations can possibly be explained by the hypothesis that the same transcription factors are relevant when comparing both treated and untreated old tissue with young tissue, although with different strength. For example, imagine a set of transcription factors highly functional in untreated old vs. young, and also functional, but less, in the BBL treated old vs. young contrast. In this case, we will observe a similar pattern as both contrasts are functional and therefore relevant. However, when comparing untreated old vs. BBL treated, the contrast in functional strength would again result in a good correlation score and yield good ranking, generating a similar pattern. According to this theory, the regulatory pattern of BBL treated old skin corresponds to signatures in between these of untreated old skin and young skin, which is not unlikely. Another option is that the functional strength of certain TFs remains the same, but that their method of action differs (e.g. interacting with other TFs, activation instead of repression, etc.). These hypotheses are potential examples of the detection of 'TF reprogramming' phenomena with the proposed methodology.

8.6 Conclusion

Although the proposed methodology is accompanied by some limitations, the outlined future improvements have the potential to take this algorithm to an unprecedented level of performance and applicability. Moreover, the current state of the algorithm already showed good performance on a validation data set, effectively registering slight changes in sole genes. In addition, the algorithm was proven to validate prior knowledge regarding aging-related experiments, next to inferring new insights. Transitioning from the gene expression level to the regulatory (TF) level on human data can now be accomplished in a fast and effective way, although caution when interpreting results is advised.

9 Materials and Methods

9.1 Conventions and general scripting and software tools

The main scripting language used throughout this analysis is Python, while also applying R and Perl for specific cases. If Python 3.6 is not used for an application, it will be clearly indicated. The majority of tasks are directly executed on one of the Li lab servers. The processor model is Intel Xeon E5-2620 v2 (6 cores @ 2439 MHz) and the total memory capacity is 264 GB.

Transcription factor genes and their corresponding protein products will both be referred to with the HGNC-symbol. When a method of action of a transcription factor is mentioned, we evidently assume the activity of the gene product.

9.2 Collecting human transcription factors and corresponding motifs

The motif-TF relationships were downloaded on Feb. 15, 2018, from three databases. The archive linked to the HOMER tool could be tracked back to the JASPAR database regarding human data [Heinz et al., 2010]. 537 position probability matrices (PPMs) were downloaded from the JASPAR database. Secondly, the HOCOMOCO database provides us with 270 data entries and finally the CIS-BP archive holds the largest collection, from which 1964 PPMs were downloaded [Weirauch et al., 2015] [Kulakovskiy et al., 2017]. Duplicates across these 3 databases were removed. If required, the conversion from PFM to PPM is made as described in the introduction (see 5.2.1).

A second verification step was performed by overlapping two lists of human transcription factors, downloaded on Feb. 16, 2018. The first gene list was obtained from the TRANSFAC database (also referred to as the geneXplain or TFClass database [Wingender, 1996] [Wingender et al., 2012]) and was merged with the second list obtained from the TcoF-DB (v2), providing a combined collection of 8746 human transcription factors, some of which present under multiple names. This collection was used as a filter for the motif-TF relationships described in previous paragraph. In other words, only the TF-motif combinations were selected for which the existence of the TF is verified by a human TF database.

9.2.1 Removing duplicates and further motif processing

For transcription factors with multiple motifs, a Pearson correlation coefficient (PCC) was used to calculate the pairwise similarities between the motifs. The scoring function was implemented in Python 3.6 and extends the regular PCC implementation to return the optimal relative offset for an ungapped alignment of two motifs. For every possible offset that results in

a different motif alignment, the correlation score is calculated between every aligned position in the two PPMs and averaged to provide a single score for the alignment under study. The best score out of all alignments is kept and indicates the best alignment. Secondly, a constraint was added in order to only consider alignment scores for matrices that overlap at least half of the minimal length of the PPM pair. This eliminates alignments where the Pearson correlation coefficient (PCC) is only based on very few aligning positions in the motifs. Unlike other approaches, in this implementation the correlation scores were calculated between the probability matrices instead of between consensus sequences in an attempt to preserve more of the information. This was done both for the forward matrix and also for the reverse complement of the matrix. The highest scoring form in terms of correlation was kept, together with the relative offset required to achieve the alignment score. The resulting correlation matrices were clustered using the *pheatmap* and *hclust* functions for R. The algorithm used was the agglomerative hierarchical UPGMA method [Sokal and Michener, 1958]. Cluster information was extracted and fed back to Python. Only clusters with a tree height smaller than 0.5 were selected. This threshold of 0.5 for clustering TF motifs was proven to be successful elsewhere [Madsen et al., 2017b].

PPMs in one cluster are considered to be representing the variability of one single motif and should therefore be aligned and merged as one PPM for better representation. If the clustering threshold was chosen correctly, the PPMs should be similar and the choice of algorithm creating the multiple alignment should have minimal impact on the resulting clusters. After extracting the stepwise information to rebuild the clusters, the alignment is made for every cluster in the same iterative process. The initial alignment is created by aligning the two most similar motifs with the optimal offset. Next, the most similar PPM to one of the previously aligned motifs in the existing alignment is added, without merging the other motifs in the alignments. The latter is not necessary due to the high similarity following clustering. Only when the entire aligning process is finished, the motif PPMs are merged into a single motif. This merging process is done by padding sequences to the same length and averaging over the positions, while only including the PPMs if they represent the position under study. Blank values created by padding the sequences are thus not taken into account in the merging process. A second interesting approach is applying weights based on the amount of evidence for a certain PPM. Weights in the form of counts (i.e. number of sequences for motif detection) were downloaded from the JASPAR and HOCOMOCO archives, but were not available for the CIS-BP archive. A subset of PPMs that were supported by such evidence is taken. Next, a merging step both with and without taking into account the weights was executed. The weights are expressed as fractions before the multiplication with the corresponding positional elements of the PPMs.

For example, consider two motifs: motif 1 and motif 2, given that motif 1 resulted from 20 observations and motif 2 from 80 observations. The probability of observing adenine in the first position of both motifs is 0.9 and 0.8, respectively. In unweighted merging, the average value of 0.85 is used. However, in the case of weighted merging, the relative fraction of evidence is

multiplied with the probability of observing adenine. The merged probability would therefore equal 0.9 x 20/(20+80) + 0.8 x 80/(20+80) or 0.82.

The comparison between both motifs was made based on length, information content, and overall similarity. Only if the difference between the two approaches is negligible no weights for the CIS-BP archive are required.

The motif PPM resulting from the clustering, aligning, and merging process are trimmed after computing the information content (IC) of the edges using the function below [D'haeseleer, 2006].

$$I_i = 2 + \sum f_{b,i} \log_2(f_{b,i})$$

$f_{b,i}$ indicates the frequency of base b at position i. A value of 2 is an indication for perfect conservation, whereas the occurence of two of the four nucleotides, both occuring 50%, results in a value of 1 bit. When all four nucleotides occur equally often, the information content is zero [D'haeseleer, 2006]. The IC cut-off used for trimming was 0.5. Resulting motifs with a length shorter than four nucleotides were removed.

9.3 Collecting targets of human transcription factors

The list of transcription factors in our database resulting from the previous section was updated with synonyms of the TFs, parsed from the public TRANSFAC collection, before consulting other public databases to retrieve known targets [Wingender, 1996]. Through the R package *tftargets* [Slowikowski, 2018], the TRED [Jiang et al., 2007], ITFP [Zheng et al., 2008], ENCODE [Consortium, 2012], Neph2012 [Neph et al., 2012], TRRUST [Han et al., 2015] and Marbach2016 [Marbach et al., 2016] databases were consulted. Secondly, both the public TRANSFAC table and the PAZAR [Portales-Casamar et al., 2007] projects were added independently of the *tftargets* R package. The gathering and conversion of the data entries took place in the week of February 26, 2018. The target lists from different sources are stored in different sets.

9.4 *De novo* target prediction

9.4.1 Genome scanning

Theoretical binding sites for every TF are found by using the gathered motifs to scan over the human genome. More specifically, a collection of human promoter regions, characterized as 5000 bp upstream of the transcription start site, was downloaded from the UCSC genome browser on March 18, 2018 (version GRCh37/hg19), together with the corresponding annotation file.

After removing redundant sequences (i.e. doubles, exact copies), the genome sequences were scanned using both forward and reverse complement instances of the motif subset generated by the weighted merging process discussed above. Scores should represent binding affinity and are therefore a measure for free energy. The scoring system should be interpreted as the sum of the individual scores for every sliding window alignment between the motif and the 5000 bp long promoter region. Finally, the sum of the forward and the reverse complement motif scores is made. The scoring formula for every position detailed in the introduction:

$$S_i = \frac{f_1}{f_{01}} * \frac{f_2}{f_{02}} * ... * \frac{f_n}{f_{0n}}$$

With i in S_i as the position of the sliding window and n the motif length. The value of i ranges from 1 until 5000 bp (upstream sequence length) minus n (the motif length) and plus 1. f_x indicates the motif PPM's frequency of occurrence for the corresponding nucleotide in the sequence to be scanned, with x the aligned position of the motif PPM. f_{0x} is the background frequency of the aligned nucleotide, and is computed as the frequency of that particular nucleotide over all of the 5000 bp long sequences. Since the computation of this score is a multiplication, a probability f_x equal to 0 is detrimental. In order to avoid this, probabilities of 0 and 1 are converted to 0.01 and 0.99, respectively.

An efficient implementation of the sliding process was made by making use of two-dimensional convolutions (using the *scipy.signal.convolve2d* implementation), effectively avoiding the use of slow for-loops in Python.

It is important to note that the formula above is a function of the motif length and that motifs of different size will result in different score scales. From this follows that the scores for one motif cannot be compared with scores for a motif of different length.

Both the total scores for the forward and reverse complement of the motif scans are computed and aggregated as follows:

$$S = \sum_{i=1}^{m} S_i(forward) + \sum_{i=1}^{m} S_i(reverse.complement)$$

In this formula m denotes the last sliding window, which is equal to 5000 minus the motif length n. The value of m translates to the last sliding window. The choice was made to sum up the scores without applying a positional threshold, because thresholding is arbitrary and the signal to noise ratio in this experiment is generally very high. A rough estimate of the total maximum achievable score can be made using the formula below.

$$S = 2 * (5000 - n) * \frac{1}{0.25} * \frac{1}{0.25} * ... * \frac{f_n}{f_{0n}} = (5000 - n) * 2^{2n+1}$$

In this calculation, the first multiplication with a value of 2 indicates the summation of the resulting scores from both forward and reverse complement PPM forms. 5000 minus n is the total amount of sliding windows, with n the length of the motif. The last part of the formula implies a perfect alignment (probability of 1) with the sequence for every position of window, which is theoretically impossible given that background frequencies of 0.25 were assumed for every nucleotide. Therefore, this formula is heavily flawed and only provides an estimate of a very generous upper bound for the motif scores in function of the motif length n. As an example, a motif of length 10 results in a maximum score of roughly 10^{10}.

9.4.2 Score analysis

Before totaling the scores, a positional score analysis was made. A histogram over the 5000-n sliding windows can be generated to gain more insight in where the motif targets with the highest binding potential are located in the promoter sequence. Data on these positions was collected as follows. For every TF motif, 500 promoter regions were sampled at random and their positional scores per sliding window were calculated. These scores served as background distribution. The 99.9th percentile from the distribution of positional scores was extracted and used as a representation of the sought-after noise threshold. After sorting potential target genes by descending total score (sum of scores over all sliding windows per promoter), a variable amount of top genes was taken and the value of the 99.9th percentile from the background distribution is subtracted from the positional score, subject to the constraint of zero as minimum adjusted score. Hence, the score of the majority of positions is zero and allowed for a histogram of non-zero values. The variable amount of top genes was decided by a simple, but flexible cut-off: the gene must have an aggregated total score bigger than the value of the 99.9th percentile of the background distribution, limited to a maximum of 500 genes.

The positional scores were analyzed both for a combination of motifs and for individual motifs. In order to find regions enriched for motif hits, the promoter region was subdivided in 10 regions and the enrichment of motif hits in the region was assessed by computing a p-value for that region using a Poisson distribution parametrized using the information of the entire promoter region.

9.4.3 Motif analysis

As mentioned above, the resulting total motif scores cannot be compared between motifs with different lengths. One solution is to standardize by computing z-scores. Secondly, the scores were sorted and assigned a rank. Finally, an extra boolean feature column was added for every target gene, which indicates whether the interaction between transcription factor and target gene is supported by at least one TF-target database. The distribution of true values, equaling support for a TF-target interaction, across the sorted motif-specific target gene scores could be

assessed by a Wilcoxon rank sum test. Moreover, a column storing the best rank over all the motifs was added and the same rank sum test was applied to this best rank column in order to evaluate the motif combination. Corresponding test statistics and p-values were recorded. Because a large number of tests is carried out, the multiple testing problem is faced. In order to reduce false positives at the 95% confidence level, the Bonferroni correction is applied to the significance cut-off α for the p-values, with n the amount of tests:

$$\alpha_{adj} = 0.05/n$$

The Bonferroni correction controls the family-wise error rate (FWER) at $\leq \alpha_{adj}$, meaning that the probability of making one or more false discoveries is smaller than α_{adj}. This method is known to be rather stringent compared to the Benjamini-Hochberg method [Benjamini and Hochberg, 1995], however we particularly aim to obtain robust results.

9.4.4 Score clustering

Next to motif-specific rank sum tests, unsupervised learning in the form of clustering was performed to gain more insight in the results. For example, transcription factors clustering together might target a similar subset of genes. Two clustering algorithms were applied: one where only the z-scores of all motifs were clustered (symmetric), as well as a second clustering algorithm on the z-scores together with the target genes (asymmetric). The first clustering ran in Rstudio using the pheatmap and hclust functions on the correlation matrix, resulting from pairwise PCC calculations between motif-specific z-score columns. The second clustering applies the CLUSTER 3.0 program [de Hoon et al., 2004] in order to find relationships between raw motif-specific z-scores and target genes. Furthermore, the output of CLUSTER 3.0 was connected to TreeView 3.0 [Page, 1996] for visualization. The parameters of the CLUSTER algorithm were determined to apply the Pearson correlation score as distance measure and pairwise complete-linkage as hierarchical clustering method.

9.4.5 Extending promoter regions

The interesting results for the upstream sequences sparked our interest to also analyze the downstream sequences. The hg19 version of the human genome was downloaded from UCSC genome browser website on April 16th, 2018. The genomic data were subdivided per chromosome and was therefore concatenated through the command line and indexed with the samtools *faidx* method. Following indexing, DNA substrings could be queried using the *pyfaidx* library for Python. For every upstream sequence, the ending position was taken and 5000 base-pairs were added to select the downstream sequence. Here, the remark should be made that when the chromosome coordinates are given for the reverse complement (RC) the downstream

sequences must be extracted accordingly (5000 bp before the upstream sequence) and complemented. The downstream sequences were then concatenated to the upstream sequences to form a 10 000 bp sequence around the TSS. Subsequently, the sequences were subjected to the same scanning algorithm described as for the upstream sequences. Additionally, the predictive power of the motifs in terms of transcription factor targets was also assessed in the same way as for the upstream promoter regions.

9.4.6 Evaluating alternative scanning methods

The pinnacle of state-of-the-art genome scanning methods as reviewed by Jayaram et al. (2016) appeared to be the FIMO tool [Grant et al., 2011]. FIMO is part of the MEME suite online platform [Bailey et al., 2015] and can be run as a command line tool. FIMO was downloaded on June 19, 2018, and is part of version 5.0.1 of the MEME suite. Because of the close relationship with the MEME suite, FIMO takes input motifs in the MEME format and the promoter sequences in FASTA format. Appropriate conversions were made, following the documentation on the MEME suite website. FIMO parameters were set to default and was ran on all motifs and the same promoters similar as for our own implementation (Li lab). In contrary to fast convolution-based implementation made earlier, FIMO is factor 4 slower and is estimated to have a total runtime of 12 days on the upstream sequences. When run on the full range around the TSS (10 kb by extending 5 kb upstream with 5 kb downstream), the runtime of FIMO was estimated to be 15 days. Fortunately, adaptations regarding parallelization could be made in order to effectively suppress the required runtime. Conducted preliminary research on the 5 kb upstream sequences decided in favor of an additional full range (10 kb) run of FIMO. Similar to the Li lab implementation, the aggregation of the FIMO scores was achieved by making the sum of the scores over the entire sequence.

Improvements to the score vectors were explored by incorporating information from known TF-targets. The score for a known target was updated to the maximum score in the score vector and compared using the same pipelines.

9.5 Identifying relevant transcription factors in gene expression data

9.5.1 Correlation analysis

The score vectors for every TF-motif were aggregated and stored in one matrix (M_{TF}). Rows represent target genes, whereas columns represent the different motifs. Fold change vectors (either log transformed or not) from gene expression analyses are reshaped into a similar format (M_{GE}). Rows again represent target genes, whereas columns represent different conditions instead of different TF-motifs. Rows (target genes) that are represented in both matrices are removed in an effort to equalize the amount of rows. The correlation between theoretical TF

binding scores (M_{TF}) and fold changes (M_{GE}) is computed and results in a new matrix. The non-parametric Spearman's rank correlation coefficient is used initially as a correlation measure because of the scale difference [Szczepańska, 2011], although will be compared to Pearson's correlation coefficient together with transformations on the input data. However, the disadvantages of using Spearman's correlation coefficient are not to be underestimated. For example, in case only a minor number of targets are available, and those have high FCs as well as high scores, the overall correlation will still be low as only ranks are taken into account. However, Pearson's correlation score would be clearly high. Regarding p-values, the standard p-value for Pearson's correlation coefficient may be inaccurate, but this does not present a problem as the method focuses on the R-value. The correlation scores are computed in a pairwise manner for all combinations of the columns of both matrices.

$$C_{i,j} = corr(M_{TF_{:,i}}, |M_{GE_{:,j}}|)$$

C indicates the resulting correlation matrix, with the TF-motifs as rows (i), the gene expression conditions as columns (j) and the correlation coefficient as value $C_{i,j}$. The correlation coefficient is thus computed pairwise between the columns of both matrices. However, the transcription factor can act as an activator or as a repressor and therefore the absolute values of the gene expression matrix are taken before the correlation computation. Sorting a gene expression condition column by descending correlation returns the crucial transcription factors on top. If target genes are present more than once (different TSSs), the best correlation score is withheld.

In order to prove the superiority of one correlation measure over the other (nonparametric Spearman vs parametric Pearson), the results are compared both with and without a $\log 2$ transformation preceding the conversion to z-scores for the motif score vectors.

9.5.2 Method validation

The described method was applied to a gene expression dataset generated by Cusanovich et al. (2014), accessible with the GSE50588 code on the GEO website. In this experiment, 59 TFs were knocked down using siRNA in a HapMap lymphoblastoid cell line [Cusanovich et al., 2014]. The knocked down TFs are involved in the immune response and are therefore closely related to aging (see introduction). Expression of the genes was quantified using qPCR and microarrays. These data were last updated on May 2, 2018 and were downloaded on June 13, 2018. The raw expression values are $\log 2$ transformed, quantile normalized (for all microarrays together), and further adjusted for batch effects with the RUV2 method [Gagnon-Bartsch and Speed, 2011]. Subsequently, the values were converted to fold changes using values from the control experiments. An optional $\log 2$ transformation on the fold changes can be applied next. For the TF knock-down experiments done in multiple, the experiment with the most efficient knock-down is kept (largest drop in expression after perturbation) if the expression of

the perturbed TF is measured. Because the expression level of the perturbed TF is not always present in the measured data, the experiment with median validation performance (in terms of allowing identification of the causal transcription factor (see further)) is chosen in order to guarantee objectivity (i.e. preventing bias).

Besides gene expression data, the variations on motif scanning scores are imported. The motif vectors originate from either the FIMO algorithm or our own scoring formula (Li lab). Different transformations such as $\log 2$ and/or z-score transformation are applied and are added to the comparison. Scores for genes that are not measured in the gene expression data are not relevant and are subsequently left out. The pairwise correlation is computed between the gene expression condition and the motif vector, effectively producing a score for the 'match' between the TF (by motif) and the perturbation experiment. Sorting these scores by descending ranks the best matching TFs to this perturbation experiment on top.

Knowledge of which TF is perturbed in every condition enables validation of the approach: if the algorithm returns the perturbed transcription factor (or very closely related TFs) as a crucial (highly ranked after sorting for descending correlation) factor consistently and for every condition, the method can be considered validated. Quantifying the sensitivity is done by continuous measures such as taking the median of the ranks (after sorting by correlation) of the perturbed transcription factors in their respective conditions. A median rank higher up the ladder (closer to 0) would mean that the crucial transcription factor are indeed placed close to the top of the sorted list. Secondly, the ranks are summed over the different experiments in order to obtain one more performance measure.

The metrics produced by this validation experiment allowed to assess the performance of the different genome scanning methods (FIMO versus our own implementation (Li lab)). In addition, other parameters such as the promoter scanning range (5 kb upstream versus the extended 10 kb upstream + downstream region) were evaluated in combination with transformations on the motif scores ($\log 2$ and/or z-score transformation). Moreover, the default correlation implementation in the non-parametric Spearman correlation formula was substituted with the parametric correlation coefficient suggested by Karl Pearson. Finally, an attempt was made to further improve the genome scanning scores by incorporating the information on known TF-target interactions. More specifically, the score for known TF-target interactions was updated to the maximum score in the vector in order to integrate the information in the TF-target databases. The best performing algorithm in terms of ranking relevant TFs in a gene expression experiment will be implemented as default in the web application.

9.5.3 Developing a web application

In order to support future research both in the Li lab and worldwide, a web tool was created to quickly and reliably find important transcription factors in a gene expression condition. The implementation was made in Django 2.0.5 (web framework for Python 3) and linked to a MySQL database. This application includes an optional gene ontology analysis through GOrilla [Eden et al., 2009b]. For the latter, the enrichment in the top correlating transcription factors is compared with a background list of all tested TFs by way of the publicly available online interface. Besides a visual representation including the TF-motifs, the raw correlation data can also be downloaded for cases where further processing is desirable.

9.6 Applications to Aging research

9.6.1 Aging in the human frontal cortex of the brain: study 1

An intriguing experiment by Lu et al. (2004) compares gene expression in the human frontal cortex of the brain between two age groups, a group of people of $\leq$ 42 years and people aged $\geq$ 73 years. Gene-wise standardized expression values were downloaded from the NCBI database with series number GSE1572 on June 22, 2018. Significance analysis of microarrays (SAM) software was used to compare young and aged groups to determine the list of genes with a high FC (> 1.5) and a median false discovery rate (FDR) smaller than 0.01 [Lu et al., 2004]. Subsequently, the fold changes were log 2 transformed and run through the algorithm and corresponding web application. The top transcription factors, corresponding motifs, and results of the GOrilla GO analysis were all carefully investigated.

9.6.2 Aging in the human frontal cortex of the brain: study 2

Next to aging in the frontal brain cortex (2004), Lu et al. also conducted a second, more recent gene expression experiment (2014) in which data from 12 young (<40yr), 9 middle aged (40-70yr), 16 normal aged (70-94yr), and 4 extremely aged (95-106yr) were analyzed [Lu et al., 2014]. RNA of prefrontal cortical brain samples with tissue pH > 6.5 and postmortem interval < 20 hrs was hybridized to microarrays and the probe-level linear model (PLM) platform was used in combination with the SAM software to perform 2-group comparisons of young adult vs. aged. In similar fashion to the previous analyses, the data were accessed through the GEO archive (GSE53890) on June 22, 2018, and the conversion to log 2 fold changes was made and connected to the web interface of our algorithm. In the same line of the previous experiment, the top transcription factors, corresponding motifs, and results of the GOrilla GO analysis were all carefully investigated. Moreover, a multiple-way Venn diagram was created in addition to a histogram summarizing the top transcription factors over different contrasts.

9.6.3 Aging and rejuvenation in human skin cells

The RNA-seq gene expression experiment carried out by Chang AL et al. (2013) included 10 female skin biopsies: 5 young volunteers (age <30) and 5 aged volunteers (age >50). Following an initial biopsy of untreated skin in the older patients, the patients have received three courses of monthly BBL treatment (broadband light) before taking 5 new skin samples. The reads per kilobase of exon per million mappable reads (RPKM) data were downloaded from GSE39170 on July 7, 2018, and converted into log 2 fold changes for three contrasts: untreated vs. young, treated vs. young, and treated vs. untreated. Subsequently, these data were fed to the web application and results were analyzed similarly to the aging experiment in the frontal brain cortex. In addition, the top transcription factors are compared in a Venn diagram between the three different contrasts. Top ranked transcription factors, significant GOrilla GO enrichment, recurring motifs subpatterns, and finally also the effectiveness of the treatment were examined.

9.6.4 Combinatorial analysis

To evaluate variance over different aging-related experiments, both biological and technical, the analyses can be integrated. Consistencies over similar experiments were sought for, next to identifying remarkable inconsistencies. The transcription factor correlation scores generated per gene expression contrast are clustered together using the CLUSTER 3.0 and TreeView 3.0 software tools. Moreover, the top transcription factors per experiment could be summarized in a histogram.

10 References

[Aaronson, 2002] Aaronson, D. S. (2002). A road map for those who dont know jak-stat. *Science*, 296(5573):1653–1655.

[Adler et al., 2007] Adler, A. S., Sinha, S., Kawahara, T. L., Zhang, J. Y., Segal, E., and Chang, H. Y. (2007). Motif module map reveals enforcement of aging by continual nf- b activity. *Genes and Development*, 21(24):3244–3257.

[Anselmi et al., 2009] Anselmi, C. V., Malovini, A., Roncarati, R., Novelli, V., Villa, F., Condorelli, G., Bellazzi, R., and Puca, A. A. (2009). Association of thefoxo3alocus with extreme longevity in a southern italian centenarian study. *Rejuvenation Research*, 12(2):95–104.

[Armanios et al., 2009] Armanios, M., Alder, J. K., Parry, E. M., Karim, B., Strong, M. A., and Greider, C. W. (2009). Short telomeres are sufficient to cause the degenerative defects associated with aging. *The American Journal of Human Genetics*, 85(6):823–832.

[Bahar et al., 2006] Bahar, R., Hartmann, C. H., Rodriguez, K. A., Denny, A. D., Busuttil, R. A., Dollé, M. E. T., Calder, R. B., Chisholm, G. B., Pollock, B. H., Klein, C. A., and et al. (2006). Increased cell-to-cell variation in gene expression in ageing mouse heart. *Nature*, 441(7096):1011–1014.

[Bailey et al., 2015] Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The meme suite. *Nucleic Acids Research*, 43(W1).

[Bailey et al., 2006] Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006). Meme: discovering and analyzing dna and protein sequence motifs. *Nucleic Acids Research*, 34(Web Server).

[Barzilai et al., 2012] Barzilai, N., Huffman, D. M., Muzumdar, R. H., and Bartke, A. (2012). The critical role of metabolic pathways in aging. *Diabetes*, 61(6):1315–1322.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

[Bhardwaj, 2005] Bhardwaj, N. (2005). Kernel-based machine learning protocol for predicting dna-binding proteins. *Nucleic Acids Research*, 33(20):6486–6493.

[Bjedov et al., 2010] Bjedov, I., Toivonen, J. M., Kerr, F., Slack, C., Jacobson, J., Foley, A., and Partridge, L. (2010). Mechanisms of life span extension by rapamycin in the fruit fly drosophila melanogaster. *Cell Metabolism*, 11(1):35–46.

[Blagosklonny, 2008] Blagosklonny, M. V. (2008). Aging: Ros or tor. *Cell Cycle*, 7(21):3344–3354.

[Blagosklonny, 2011] Blagosklonny, M. V. (2011). Hormesis does not make sense except in the light of tor-driven aging. *Aging*, 3(11):1051–1062.

[Blasco et al., 1997] Blasco, M. A., Lee, H.-W., Hande, M., Samper, E., Lansdorp, P. M., Depinho, R. A., and Greider, C. W. (1997). Telomere shortening and tumor formation by mouse cells lacking telomerase rna. *Cell*, 91(1):25–34.

[Bloomer and Lee, 2014] Bloomer, R. and Lee, S. (2014). Dietary and caloric restriction for human health. *Reference Module in Biomedical Sciences*.

[Boulias and Horvitz, 2012] Boulias, K. and Horvitz, H. R. (2012). The c. elegans microrna mir-71 acts in neurons to promote germline-mediated longevity through regulation of daf-16/foxo. *Cell Metabolism*, 15(4):439–450.

[Broer and Duijn, 2015] Broer, L. and Duijn, C. M. V. (2015). Gwas and meta-analysis in aging/longevity. *Longevity Genes Advances in Experimental Medicine and Biology*, page 107–125.

[Brooks-Wilson, 2013] Brooks-Wilson, A. R. (2013). Genetics of healthy aging and longevity. *Human Genetics*, 132(12):1323–1338.

[Brunet et al., 1999] Brunet, A., Bonni, A., Zigmond, M. J., Lin, M. Z., Juo, P., Hu, L. S., Anderson, M. J., Arden, K. C., Blenis, J., Greenberg, M. E., and et al. (1999). Akt promotes cell survival by phosphorylating and inhibiting a forkhead transcription factor. *Cell*, 96(6):857–868.

[Bussemaker et al., 2001] Bussemaker, H. J., Li, H., and Siggia, E. D. (2001). Regulatory element detection using correlation with expression. *Nature Genetics*, 27(2):167–171.

[Cao et al., 2011] Cao, K., Blair, C. D., Faddah, D. A., Kieckhaefer, J. E., Olive, M., Erdos, M. R., Nabel, E. G., and Collins, F. S. (2011). Progerin and telomere dysfunction collaborate to trigger cellular senescence in normal human fibroblasts. *Journal of Clinical Investigation*, 121(7):2833–2844.

[Chang et al., 2013] Chang, A. L. S., Bitter, P. H., Qu, K., Lin, M., Rapicavoli, N. A., and Chang, H. Y. (2013). Rejuvenation of gene expression pattern of aged human skin by broadband light treatment: A pilot study. *Journal of Investigative Dermatology*, 133(2):394–402.

[Claesson et al., 2012] Claesson, M. J., Jeffery, I. B., Conde, S., Power, S. E., O'Connor, E. M., Cusack, S., Harris, H. M. B., Coakley, M., Lakshminarayanan, B., O'Sullivan, O., and et al. (2012). Gut microbiota composition correlates with diet and health in the elderly. *Nature*, 488(7410):178–184.

[Colman et al., 2009] Colman, R. J., Anderson, R. M., Johnson, S. C., Kastman, E. K., Kosmatka, K. J., Beasley, T. M., Allison, D. B., Cruzen, C., Simmons, H. A., Kemnitz, J. W., and et al. (2009). Caloric restriction delays disease onset and mortality in rhesus monkeys. *Science*, 325(5937):201–204.

[Consortium, 2012] Consortium, T. E. P. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489:57 EP –.

[Crooks, 2004] Crooks, G. E. (2004). Weblogo: A sequence logo generator. *Genome Research*, 14(6):1188–1190.

[Cuellar-Partida et al., 2011] Cuellar-Partida, G., Buske, F. A., Mcleay, R. C., Whitington, T., Noble, W. S., and Bailey, T. L. (2011). Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics*, 28(1):56–62.

[Cusanovich et al., 2014] Cusanovich, D. A., Pavlovic, B., Pritchard, J. K., and Gilad, Y. (2014). The functional consequences of variation in transcription factor binding. *PLoS Genetics*, 10(3).

[de Hoon et al., 2004] de Hoon, M., Imoto, S., Nolan, J., and Miyano, S. (2004). Open source clustering software. *Bioinformatics*, 20(9):1453–1454.

[Dhaeseleer, 2006] Dhaeseleer, P. (2006). How does dna sequence motif discovery work? *Nature Biotechnology*, 24(8):959–961.

[D'haeseleer, 2006] D'haeseleer, P. (2006). What are dna sequence motifs? *Nature Biotechnology*, 24(4):423–425.

[Dong et al., 2008] Dong, X. C., Copps, K. D., Guo, S., Li, Y., Kollipara, R., Depinho, R. A., and White, M. F. (2008). Inactivation of hepatic foxo1 by insulin signaling is required for adaptive nutrient homeostasis and endocrine growth regulation. *Cell Metabolism*, 8(1):65–76.

[Doonan et al., 2008] Doonan, R., Mcelwee, J. J., Matthijssens, F., Walker, G. A., Houthoofd, K., Back, P., Matscheski, A., Vanfleteren, J. R., and Gems, D. (2008). Against the oxidative damage theory of aging: superoxide dismutases protect against oxidative stress but have little or no effect on life span in caenorhabditis elegans. *Genes and Development*, 22(23):3236–3241.

[Ducrest et al., 2002] Ducrest, A.-L., Szutorisz, H., Lingner, J., and Nabholz, M. (2002). Regulation of the human telomerase reverse transcriptase gene. *Oncogene*, 21(4):541–552.

[Eden et al., 2009a] Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009a). Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinformatics*, 10(1):48.

[Eden et al., 2009b] Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009b). Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinformatics*, 10(1):48.

[Fan et al.,] Fan, Y., Kon, M., and DeLisi, C. Transcription factor-dna binding via machine learning ensembles. *ARXIV*, page 33.

[Flachsbart et al., 2009] Flachsbart, F., Caliebe, A., Kleindorp, R., Blanché, H., Eller-Eberstein, H. V., Nikolaus, S., Schreiber, S., and Nebel, A. (2009). Association offoxo3avariation with human longevity confirmed in german centenarians. *Proceedings of the National Academy of Sciences*, 106(8):2700–2705.

[Florian and Geiger, 2010] Florian, M. C. and Geiger, H. (2010). Concise review: Polarity in stem cells, disease, and aging. *Stem Cells*, 28(9):1623–1629.

[Foat et al., 2006] Foat, B. C., Morozov, A. V., and Bussemaker, H. J. (2006). Statistical mechanical modeling of genome-wide transcription factor occupancy data by matrixreduce. *Bioinformatics*, 22(14).

[Fontana and Partridge, 2015] Fontana, L. and Partridge, L. (2015). Promoting health and longevity through diet: From model organisms to humans. *Cell*, 161(1):106–118.

[Gagnon-Bartsch and Speed, 2011] Gagnon-Bartsch, J. A. and Speed, T. P. (2011). Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552.

[Gorgoulis and Halazonetis, 2010] Gorgoulis, V. G. and Halazonetis, T. D. (2010). Oncogene-induced senescence: the bright and dark side of the response. *Current Opinion in Cell Biology*, 22(6):816–827.

[Grant et al., 2011] Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018.

[Gratacós and Brewer, 2010] Gratacós, F. M. and Brewer, G. (2010). The role of auf1 in regulated mrna decay. *Wiley Interdisciplinary Reviews: RNA*, 1(3):457–473.

[Green et al., 2011] Green, R. A., Kao, H.-L., Audhya, A., Arur, S., Mayers, J. R., Fridolfsson, H. N., Schulman, M., Schloissnig, S., Niessen, S., Laband, K., and et al. (2011). A high-resolution c. elegans essential gene network based on phenotypic profiling of a complex tissue. *Cell*, 145(3):470–482.

[Guarente, 2011] Guarente, L. (2011). Sirtuins, aging, and medicine. *New England Journal of Medicine*, 364(23):2235–2244.

[Gupta et al., 2007] Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., and Noble, W. (2007). Quantifying similarity between motifs. *Genome Biology*, 8(2).

[Han et al., 2015] Han, H., Shim, H., Shin, D., Shim, J. E., Ko, Y., Shin, J., Kim, H., Cho, A., Kim, E., and Lee, T. e. a. (2015). Trrust: a reference database of human transcriptional regulatory interactions. *Scientific Reports*, 5(1).

[Harrison et al., 2009] Harrison, D. E., Strong, R., Sharp, Z. D., Nelson, J. F., Astle, C. M., Flurkey, K., Nadon, N. L., Wilkinson, J. E., Frenkel, K., Carter, C. S., and et al. (2009). Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. *Nature*, 460(7253):392–395.

[Hayflick and Moorhead, 1961] Hayflick, L. and Moorhead, P. (1961). The serial cultivation of human diploid cell strains. *Experimental Cell Research*, 25(3):585–621.

[Heinz et al., 2010] Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., Glass, C. K., and et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular Cell*, 38(4):576–589.

[Hoenicke and Zender, 2012] Hoenicke, L. and Zender, L. (2012). Immune surveillance of senescent cells–biological significance in cancer- and non-cancer pathologies. *Carcinogenesis*, 33(6):1123–1126.

[Holloway et al., 2006] Holloway, D. T., Kon, M., and Delisi, C. (2006). Machine learning for regulatory analysis and transcription factor target prediction in yeast. *Systems and Synthetic Biology*, 1(1):25–46.

[Honkela et al., 2010] Honkela, A., Girardot, C., Gustafson, E. H., Liu, Y.-H., Furlong, E. E. M., Lawrence, N. D., and Rattray, M. (2010). Model-based method for transcription factor target identification with limited data. *Proceedings of the National Academy of Sciences*, 107(17):7793–7798.

[Houtkooper et al., 2010] Houtkooper, R. H., Cantó, C., Wanders, R. J., and Auwerx, J. (2010). The secret life of nad : An old metabolite controlling new metabolic signaling pathways. *Endocrine Reviews*, 31(2):194–223.

[Infante et al., 2014] Infante, A., Gago, A., Eguino, G. R. D., Calvo-Fernández, T., Gómez-Vallejo, V., Llop, J., Schlangen, K., Fullaondo, A., Aransay, A. M., Martín, A., and et al. (2014). Prelamin a accumulation and stress conditions induce impaired oct-1 activity and autophagy in prematurely aged human mesenchymal stem cell. *Aging*, 6(4):264–280.

[Jacoby et al., 2003] Jacoby, J. J., Kalinowski, A., Liu, M.-G., Zhang, S. S.-M., Gao, Q., Chai, G.-X., Ji, L., Iwamoto, Y., Li, E., Schneider, M., and et al. (2003). Cardiomyocyte-restricted knockout of stat3 results in higher sensitivity to inflammation, cardiac fibrosis, and heart failure with advanced age. *Proceedings of the National Academy of Sciences*, 100(22):12929–12934.

[Jaskelioff et al., 2010] Jaskelioff, M., Muller, F. L., Paik, J.-H., Thomas, E., Jiang, S., Adams, A. C., Sahin, E., Kost-Alimova, M., Protopopov, A., Cadiñanos, J., and et al. (2010). Telomerase reactivation reverses tissue degeneration in aged telomerase-deficient mice. *Nature*, 469(7328):102–106.

[Jayaram et al., 2016] Jayaram, N., Usvyat, D., and Martin, A. C. R. (2016). Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics*.

[Jiang et al., 2007] Jiang, C., Xuan, Z., Zhao, F., and Zhang, M. Q. (2007). Tred: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Research*, 35(Database):D137–D140.

[Jin et al., 2011] Jin, C., Li, J., Green, C. D., Yu, X., Tang, X., Han, D., Xian, B., Wang, D., Huang, X., Cao, X., and et al. (2011). Histone demethylase utx-1 regulates c. elegans life span by targeting the insulin/igf-1 signaling pathway. *Cell Metabolism*, 14(2):161–172.

[Kang et al., 2011] Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A. M. M., Pletikos, M., Meyer, K. A., Sedmak, G., and et al. (2011). Spatio-temporal transcriptome of the human brain. *Nature*, 478(7370):483–489.

[Kang et al., 2009] Kang, J., Gemberling, M., Nakamura, M., Whitby, F. G., Handa, H., Fairbrother, W. G., and Tantin, D. (2009). A general mechanism for transcription regulation by oct1 and oct4 in response to genotoxic and oxidative stress. *Genes and Development*, 23(2):208–222.

[Kankainen and Löytynoja, 2007] Kankainen, M. and Löytynoja, A. (2007). Matlign: a motif clustering, comparison and matching tool. *BMC Bioinformatics*, 8(1):189.

[Karimzadeh and Hoffman, 2018] Karimzadeh, M. and Hoffman, M. M. (2018). Virtual chip-seq: Predicting transcription factor binding by learning from the transcriptome.

[Kenyon et al., 1993] Kenyon, C., Chang, J., Gensch, E., Rudner, A., and Tabtiang, R. (1993). A c. elegans mutant that lives twice as long as wild type. *Nature*, 366(6454):461–464.

[Kenyon, 2010] Kenyon, C. J. (2010). The genetics of ageing. *Nature*, 464(7288):504–512.

[Kim et al., 2014] Kim, D. H., Park, M. H., Chung, K. W., Kim, M. J., Jung, Y. R., Bae, H. R., Jang, E. J., Lee, J. S., Im, D. S., and Yu, B. P. e. a. (2014). The essential role of foxo6 phosphorylation in aging and calorie restriction. *AGE*, 36(4).

[Kisseleva et al., 2002] Kisseleva, T., Bhattacharya, S., Braunstein, J., and Schindler, C. (2002). Signaling through the jak/stat pathway, recent advances and future challenges. *Gene*, 285(1-2):1–24.

[Koga et al., 2011] Koga, H., Kaushik, S., and Cuervo, A. M. (2011). Protein homeostasis and aging: The importance of exquisite quality control. *Ageing Research Reviews*, 10(2):205–215.

[Kulakovskiy et al., 2010] Kulakovskiy, I. V., Boeva, V. A., Favorov, A. V., and Makeev, V. J. (2010). Deep and wide digging for binding motifs in chip-seq data. *Bioinformatics*, 26(20):2622–2623.

[Kulakovskiy et al., 2017] Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Sharipov, R. N., Fedorova, A. D., Rumynskiy, E. I., Medvedeva, Y. A., Magana-Mora, A., Bajic, V. B., and Papatsenko, D. A. e. a. (2017). Hocomoco: towards a complete collection of transcription factor binding models for human and mouse via large-scale chip-seq analysis. *Nucleic Acids Research*, 46(D1):D252–D259.

[Lambert et al., 2016] Lambert, S. A., Albu, M., Hughes, T. R., and Najafabadi, H. S. (2016). Motif comparison based on similarity of binding affinity profiles. *Bioinformatics*.

[Lambert et al., 2018] Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and Weirauch, M. T. (2018). The human transcription factors. *Cell*, 172(4):650–665.

[Larson et al., 2012] Larson, K., Yan, S.-J., Tsurumi, A., Liu, J., Zhou, J., Gaur, K., Guo, D., Eickbush, T. H., and Li, W. X. (2012). Heterochromatin formation promotes longevity and represses ribosomal rna synthesis. *PLoS Genetics*, 8(1).

[Latchman, 1999] Latchman, D. S. (1999). Pou family transcription factors in the nervous system. *Journal of Cellular Physiology*, 179(2):126–133.

[Lee et al., 2012] Lee, S. S., Vizcarra, I. A., Huberts, D. H. E. W., Lee, L. P., and Heinemann, M. (2012). Whole lifespan microscopic observation of budding yeast aging through a microfluidic dissection platform. *Proceedings of the National Academy of Sciences*, 109(13):4916–4920.

[Li et al., 2013] Li, X., Wang, K., Wang, F., Tao, Q., Xie, Y., and Cheng, Q. (2013). Aging of theory of mind: The influence of educational level and cognitive processing. *International Journal of Psychology*, 48(4):715–727.

[Liu et al., 2012] Liu, L., Cousens, S., Lawn, J. E., and Black, R. E. (2012). Global regional and national causes of child mortality – authors reply. *The Lancet*, 380(9853):1556–1557.

[Lord and Ashworth, 2012] Lord, C. J. and Ashworth, A. (2012). The dna damage response and cancer therapy. *Nature*, 481(7381):287–294.

[Lu et al., 2014] Lu, T., Aron, L., Zullo, J., Pan, Y., Kim, H., Chen, Y., Yang, T.-H., Kim, H.-M., Drake, D., Liu, X. S., Bennett, D. A., Colaiácovo, M. P., and Yankner, B. A. (2014). Rest and stress resistance in ageing and alzheimer's disease. *Nature*, 507:448 EP –.

[Lu et al., 2004] Lu, T., Pan, Y., Kao, S.-Y., Li, C., Kohane, I., Chan, J., and Yankner, B. A. (2004). Gene regulation and dna damage in the ageing human brain. *Nature*, 429(6994):883–891.

[López-Otín et al., 2013] López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. *Cell*, 153(6):1194–1217.

[Madsen et al., 2017a] Madsen, J. G. S., Rauch, A., Hauwaert, E. L. V., Schmidt, S. F., Winnefeld, M., and Mandrup, S. (2017a). Integrated analysis of motif activity and gene expression changes of transcription factors. *Genome Research*, 28(2):243–255.

[Madsen et al., 2017b] Madsen, J. G. S., Rauch, A., Van Hauwaert, E. L., Schmidt, S. F., Winnefeld, M., and Mandrup, S. (2017b). Integrated analysis of motif activity and gene expression changes of transcription factors. *Genome Research*, 28(2):243–255.

[Magalhães et al., 2009] Magalhães, J. P. D., Curado, J., and Church, G. M. (2009). Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics*, 25(7):875–881.

[Mahony and Benos, 2007] Mahony, S. and Benos, P. V. (2007). Stamp: a web tool for exploring dna-binding motif similarities. *Nucleic Acids Research*, 35(Web Server).

[Maleszewska et al., 2016] Maleszewska, M., Mawer, J. S. P., and Tessarz, P. (2016). Histone modifications in ageing and lifespan regulation. *Current Molecular Biology Reports*, 2(1):26–35.

[Marbach et al., 2016] Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., and Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature Methods*, 13(4):366–370.

[Martins et al., 2015] Martins, R., Lithgow, G. J., and Link, W. (2015). Long live foxo: unraveling the role of foxo proteins in aging and longevity. *Aging Cell*, 15(2):196–207.

[Matheu et al., 2007] Matheu, A., Maraver, A., Klatt, P., Flores, I., Garcia-Cao, I., Borras, C., Flores, J. M., Viña, J., Blasco, M. A., Serrano, M., and et al. (2007). Delayed ageing through damage protection by the arf/p53 pathway. *Nature*, 448(7151):375–379.

[Matys et al., 2006] Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, 34(suppl_1):D108–D110.

[Mercier et al., 2011] Mercier, E., Droit, A., Li, L., Robertson, G., Zhang, X., and Gottardo, R. (2011). An integrated pipeline for the genome-wide analysis of transcription factor binding sites from chip-seq. *PLoS ONE*, 6(2).

[Mesquita et al., 2010] Mesquita, A., Weinberger, M., Silva, A., Sampaio-Marques, B., Almeida, B., Leao, C., Costa, V., Rodrigues, F., Burhans, W. C., Ludovico, P., and et al. (2010). Caloric restriction or catalase inactivation extends yeast chronological lifespan by inducing h2o2 and superoxide dismutase activity. *Proceedings of the National Academy of Sciences*, 107(34):15123–15128.

[Naito et al., 1998] Naito, T., Matsuura, A., and Ishikawa, F. (1998). Circular chromosome formation in a fission yeast mutant defective in two atm homologues. *Nature Genetics*, 20(2):203–206.

[Neph et al., 2012] Neph, S., Stergachis, A., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J. (2012). Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, 150(6):1274–1286.

[Nogueira et al., 2004] Nogueira, M. L., Wang, V. E. H., Tantin, D., Sharp, P. A., and Kristie, T. M. (2004). Herpes simplex virus infections are arrested in oct-1-deficient cells. *Proceedings of the National Academy of Sciences*, 101(6):1473–1478.

[Novack, 2017] Novack, J. (2017). Autophagy and proteostasis: A unifying theory of neurodegenerative disease. 6:3–18.

[O'Brown et al., 2015] O'Brown, Z. K., Van Nostrand, E. L., Higgins, J. P., and Kim, S. K. (2015). The inflammatory transcription factors nfÎºb, stat1 and stat3 drive age-associated transcriptional changes in the human kidney. *PLOS Genetics*, 11(12):1–28.

[Ottaviani et al., 2011] Ottaviani, S., Jean-Luc, B., Thomas, B., and Pascal, R. (2011). Effect of music on anxiety and pain during joint lavage for knee osteoarthritis. *Clinical Rheumatology*, 31(3):531–534.

[Overington et al., 2006] Overington, J. P., Al-Lazikani, B., and Hopkins, A. L. (2006). How many drug targets are there? *Nature Reviews Drug Discovery*, 5(12):993–996.

[Page, 1996] Page, R. D. (1996). Tree view: An application to display phylogenetic trees on personal computers. *Bioinformatics*, 12(4):357–358.

[Pont et al., 2012] Pont, A. R., Sadri, N., Hsiao, S. J., Smith, S., and Schneider, R. J. (2012). mrna decay factor auf1 maintains normal aging, telomere maintenance, and suppression of senescence by activation of telomerase transcription. *Molecular Cell*, 47(1):5–15.

[Portales-Casamar et al., 2007] Portales-Casamar, E., Kirov, S., Lim, J., Lithwick, S., Swanson, M. I., Ticoll, A., Snoddy, J., and Wasserman, W. W. (2007). Pazar: a framework for collection and dissemination of cis-regulatory sequence annotation. *Genome Biology*, 8(10):R207.

[Powers et al., 2009] Powers, E. T., Morimoto, R. I., Dillin, A., Kelly, J. W., and Balch, W. E. (2009). Biological and chemical approaches to diseases of proteostasis deficiency. *Annual Review of Biochemistry*, 78(1):959–991.

[Qin and Feng, 2017] Qin, Q. and Feng, J. (2017). Imputation for transcription factor binding predictions based on deep learning. *PLOS Computational Biology*, 13(2).

[Raamsdonk and Hekimi, 2009] Raamsdonk, J. M. V. and Hekimi, S. (2009). Deletion of the mitochondrial superoxide dismutase sod-2 extends lifespan in caenorhabditis elegans. *PLoS Genetics*, 5(2).

[Ragnauth et al., 2010] Ragnauth, C. D., Warren, D. T., Liu, Y., Mcnair, R., Tajsic, T., Figg, N., Shroff, R., Skepper, J., and Shanahan, C. M. (2010). Prelamin a acts to accelerate smooth muscle cell senescence and is a novel biomarker of human vascular aging. *Circulation*, 121(20):2200–2210.

[Rawlings, 2004] Rawlings, J. S. (2004). The jak/stat signaling pathway. *Journal of Cell Science*, 117(8):1281–1283.

[Readhead et al., 2018] Readhead, B., Haure-Mirande, J.-V., Funk, C. C., Richards, M. A., Shannon, P., Haroutunian, V., Sano, M., Liang, W. S., Beckmann, N. D., Price, N. D., Reiman, E. M., Schadt, E. E., Ehrlich, M. E., Gandy, S., and Dudley, J. T. (2018). Multiscale analysis of independent alzheimer's cohorts finds disruption of molecular, genetic, and clinical networks by human herpesvirus. *Neuron*, 99(1):64 – 82.e7.

[Ristow and Schmeisser, 2011] Ristow, M. and Schmeisser, S. (2011). Extending life span by increasing oxidative stress. *Free Radical Biology and Medicine*, 51(2):327–336.

[Rothwell et al., 2011] Rothwell, P. M., Fowkes, F. G. R., Belch, J. F., Ogawa, H., Warlow, C. P., and Meade, T. W. (2011). Effect of daily aspirin on long-term risk of death due to cancer: analysis of individual patient data from randomised trials. *The Lancet*, 377(9759):31–41.

[Rubinsztein et al., 2011] Rubinsztein, D. C., Mariño, G., and Kroemer, G. (2011). Autophagy and aging. *Cell*, 146(5):682–695.

[Rudolph et al., 1999] Rudolph, K. L., Chang, S., Lee, H.-W., Blasco, M., Gottlieb, G. J., Greider, C., and Depinho, R. A. (1999). Longevity, stress response, and cancer in aging telomerase-deficient mice. *Cell*, 96(5):701–712.

[Salekin et al., 2017] Salekin, S., Zhang, J. M., and Huang, Y. (2017). A deep learning model for predicting transcription factor binding location at single nucleotide resolution. *2017 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*.

[Salminen et al., 2012] Salminen, A., Kaarniranta, K., and Kauppinen, A. (2012). Inflammaging: disturbed interplay between autophagy and inflammasomes. *Aging*, 4(3):166–175.

[Sandelin, 2004] Sandelin, A. (2004). Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(90001):91D–94.

[Schmeier et al., 2016] Schmeier, S., Alam, T., Essack, M., and Bajic, V. B. (2016). Tcof-db v2: update of the database of human and mouse transcription co-factors and transcription factor interactions. *Nucleic Acids Research*, 45(D1):D145–D150.

[Schumacher et al., 2008] Schumacher, B., Pluijm, I. V. D., Moorhouse, M. J., Kosteas, T., Robinson, A. R., Suh, Y., Breit, T. M., Steeg, H. V., Niedernhofer, L. J., Ijcken, W. V., and et al. (2008). Delayed and accelerated aging share common longevity assurance mechanisms. *PLoS Genetics*, 4(8).

[Segil et al., 1991] Segil, N., Roberts, S., and Heintz, N. (1991). Mitotic phosphorylation of the oct-1 homeodomain and regulation of oct-1 dna binding activity. *Science*, 254(5039):1814–1816.

[Sen et al., 2016] Sen, P., Shah, P. P., Nativio, R., and Berger, S. L. (2016). Epigenetic mechanisms of longevity and aging. *Cell*, 166(4):822–839.

[Sena and Chandel, 2012] Sena, L. A. and Chandel, N. S. (2012). Physiological roles of mitochondrial reactive oxygen species. *Molecular Cell*, 48(2):158–167.

[Serrano et al., 1997] Serrano, M., Lin, A. W., Mccurrach, M. E., Beach, D., and Lowe, S. W. (1997). Oncogenic ras provokes premature cell senescence associated with accumulation of p53 and p16ink4a. *Cell*, 88(5):593–602.

[Sherry L. Murphy et al., 2017] Sherry L. Murphy et al., D. o. V. S. (2017). Deaths: Final data for 2015. *National Vital Statistics Reports*, 66(6):75.

[Slack et al., 2011] Slack, C., Giannakou, M. E., Foley, A., Goss, M., and Partridge, L. (2011). dfoxo-independent effects of reduced insulin-like signaling indrosophila. *Aging Cell*, 10(5):735–748.

[Slowikowski, 2018] Slowikowski, K. (2018).

[Smith-Vikos and Slack, 2012] Smith-Vikos, T. and Slack, F. J. (2012). Micrornas and their roles in aging. *Journal of Cell Science*, 125(1):7–17.

[Soares et al., 2013] Soares, H., Marinho, H. S., Real, C., and Antunes, F. (2013). Cellular polarity in aging: role of redox regulation and nutrition. *Genes and Nutrition*, 9(1).

[Sokal and Michener, 1958] Sokal and Michener (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*.

[Strong et al., 2008] Strong, R., Miller, R. A., Astle, C. M., Floyd, R. A., Flurkey, K., Hensley, K. L., Javors, M. A., Leeuwenburgh, C., Nelson, J. F., Ongini, E., and et al. (2008). Nordihydroguaiaretic acid and aspirin increase lifespan of genetically heterogeneous male mice. *Aging Cell*, 7(5):641–650.

[Szczepańska, 2011] Szczepańska, A. (2011). Research design and statistical analysis, third edition by jerome l. myers, arnold d. well, robert f. lorch, jr. *International Statistical Review*, 79(3):491–492.

[Tabas, 2009] Tabas, I. (2009). Macrophage death and defective inflammation resolution in atherosclerosis. *Nature Reviews Immunology*, 10(1):36–46.

[Tan and Lenhard, 2016] Tan, G. and Lenhard, B. (2016). Tfbstools: an r/bioconductor package for transcription factor binding site analysis. *Bioinformatics*, 32(10):1555–1556.

[Tanaka et al., 2011] Tanaka, E., Bailey, T., Grant, C. E., Noble, W. S., and Keich, U. (2011). Improved similarity scores for comparing motifs. *Bioinformatics*, 27(12):1603–1609.

[Thomas-Chollier et al., 2008] Thomas-Chollier, M., Sand, O., Turatsinze, J.-V., Janky, R., Defrance, M., Vervisch, E., Brohee, S., and Helden, J. V. (2008). Rsat: regulatory sequence analysis tools. *Nucleic Acids Research*, 36(Web Server).

[Tilstra et al., 2012] Tilstra, J. S., Robinson, A. R., Wang, J., Gregg, S. Q., Clauson, C. L., Reay, D. P., Nasto, L. A., Croix, C. M. S., Usas, A., Vo, N., and et al. (2012). Nf-κb inhibition delays dna damage–induced senescence and aging in mice. *Journal of Clinical Investigation*, 122(7):2601–2612.

[Toledano et al., 2012] Toledano, H., D'Alterio, C., Czech, B., Levine, E., and Jones, D. L. (2012). The let-7–imp axis regulates ageing of the drosophila testis stem-cell niche. *Nature*, 485(7400):605–610.

[Tomás-Loba et al., 2008] Tomás-Loba, A., Flores, I., Fernández-Marcos, P. J., Cayuela, M. L., Maraver, A., Tejera, A., Borrás, C., Matheu, A., Klatt, P., Flores, J. M., and et al. (2008). Telomerase reverse transcriptase delays aging in cancer-resistant mice. *Cell*, 135(4):609–622.

[Ugalde et al., 2011] Ugalde, A. P., Ramsay, A. J., Rosa, J. D. L., Varela, I., Mariño, G., Cadiñanos, J., Lu, J., Freije, J. M., and López-Otín, C. (2011). Aging and chronic dna damage response activate a regulatory pathway involving mir-29 and p53. *The EMBO Journal*, 30(11):2219–2232.

[Wang et al., 2009] Wang, C., Jurk, D., Maddick, M., Nelson, G., Martin-Ruiz, C., and Zglinicki, T. V. (2009). Dna damage response and cellular senescence in tissues of aging mice. *Aging Cell*, 8(3):311–323.

[Wasserman and Sandelin, 2004] Wasserman, W. W. and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4):276–287.

[Weirauch et al., 2015] Weirauch, M. T., Yang, A., Albu, M., Cote, A., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J. S., Govindarajan, S., Shaulsky, G., Walhout, A. J., Bouget, F.-Y., Ratsch, G., Larrondo, L. F., Ecker, J. R., and Hughes, T. R. (2015). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*.

[Wilkinson et al., 2012] Wilkinson, J. E., Burmeister, L., Brooks, S. V., Chan, C.-C., Friedline, S., Harrison, D. E., Hejtmancik, J. F., Nadon, N., Strong, R., Wood, L. K., and et al. (2012). Rapamycin slows aging in mice. *Aging Cell*, 11(4):675–682.

[Willcox et al., 2006] Willcox, D. C., Willcox, B. J., Todoriki, H., Curb, J. D., and Suzuki, M. (2006). Caloric restriction and human longevity: what can we learn from the okinawans? *Biogerontology*, 7(3):173–177.

[Wingender, 1996] Wingender, E. (1996). Transfac: a database on transcription factors and their dna binding sites. *Nucleic Acids Research*, 24(1):238–241.

[Wingender et al., 2012] Wingender, E., Schoeps, T., and Dönitz, J. (2012). Tfclass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Research*, 41(D1):D165–D170.

[Wu et al., 2014] Wu, K.-C., Lu, Y.-H., Peng, Y.-H., Tsai, T.-F., Kao, Y.-H., Yang, H.-T., and Lin, C.-J. (2014). Decreased expression of organic cation transporters, oct1 and oct2, in brain microvessels and its implication to mptp-induced dopaminergic toxicity in aged mice. *Journal of Cerebral Blood Flow and Metabolism*, 35(1):37–47.

[Xia, 2012] Xia, X. (2012). Position weight matrix, gibbs sampler, and the associated significance tests in motif characterization and prediction. *Scientifica*, 2012:1–15.

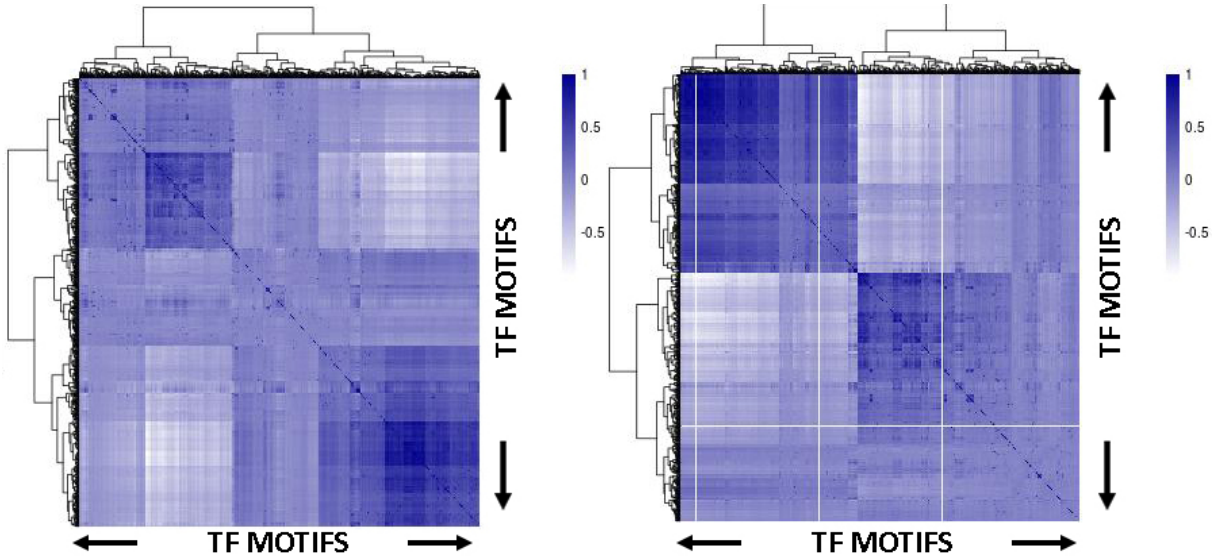
[Xue et al., 2007] Xue, W., Zender, L., Miething, C., Dickins, R. A., Hernando, E., Krizhanovsky, V., Cordon-Cardo, C., and Lowe, S. W. (2007). Senescence and tumour clearance is triggered by p53 restoration in murine liver carcinomas. *Nature*, 445(7128):656–660.

[Zhang et al., 2014] Zhang, H.-M., Liu, T., Liu, C.-J., Song, S., Zhang, X., Liu, W., Jia, H., Xue, Y., and Guo, A.-Y. (2014). Animaltfdb 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Research*, 43(D1):D76–D81.

[Zhang and Manning, 2015] Zhang, Y. and Manning, B. D. (2015). mtorc1 signaling activates nrf1 to increase cellular proteasome levels. *Cell Cycle*, 14(13):2011–2017.

[Zheng et al., 2008] Zheng, G., Tu, K., Yang, Q., Xiong, Y., Wei, C., Xie, L., Zhu, Y., and Li, Y. (2008). Itfp: an integrated platform of mammalian transcription factors. *Bioinformatics*, 24(20):2416–2417.

A Appendix



(a) upstream **(b)** upstream + downstream

Figure 32: Clusters of z-score-transformed TF-motif genome scanning score vectors. Using the Li lab implementation, the vectors containing the scores for all promoter regions per TF-motif were clustered symmetrically. The pairwise similarity values used to derive the clusters were generated by computing Pearson's correlation coefficient between z-score transformed vectors. In the left cluster (a), the score vectors were derived based on the 5 kb region upstream of the transcription start site (TSS). In the right cluster (b), the score vectors were derived based on the full 10 kb region surrounding the TSS. Darker regions in the figure represent transcription factor motifs with higher similarity in gene scores. This is an indication that either the motifs are similar, or that the corresponding group of TF-motifs target a similar subset of genes. The hierarchical, agglomerative clustering was done using the *hclust* function for R.

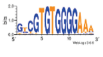
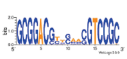
Table 8: The top 20 most effective transcription factor DNA-binding motifs in terms of target gene prediction based on 5 kb upstream promoter regions. For every motif with known targets, the derived scores for potential target genes were ranked. Subsequently, a Wilcoxon rank-sum test is applied to test if the database-verified targets are enriched towards the top of the ranked target gene list. This table presents the top 20 most effective transcription factor DNA-binding motifs in terms of target gene prediction. For this analysis, only the motifs hits for the 5 kb region upstream of the TSS were used.

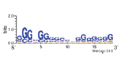
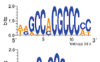
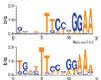
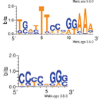
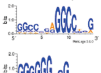
index	names	p-value	test statistic	effective	corrected effective
0	KLF14motif1	0.000000e+00	-45.116459	True	True
1	NRF1motif1	0.000000e+00	-45.091346	True	True
2	SP4motif2	0.000000e+00	-42.014388	True	True
3	SP4motif1	0.000000e+00	-39.769473	True	True
4	KLF16motif1	0.000000e+00	-38.640737	True	True
5	EGR1motif2	0.000000e+00	-37.726864	True	True
6	SP1motif1	2.978978e-291	-36.472937	True	True
7	SP3motif1	2.272655e-273	-35.326861	True	True
8	SP8motif1	3.081557e-260	-34.460999	True	True
9	SP1motif2	6.196400e-236	-32.798445	True	True
10	TFAP2Amotif4	1.208114e-230	-32.425294	True	True
11	KLF4motif1	1.757964e-221	-31.768597	True	True
12	ZFXmotif1	8.369305e-219	-31.574117	True	True
13	TFAP2Amotif1	2.863536e-218	-31.535174	True	True
14	ZNF219motif1	1.425017e-212	-31.116853	True	True
15	KLF7motif1	1.124920e-209	-30.901939	True	True
16	GABPAmotif1	1.322111e-201	-30.295297	True	True
17	TFAP2Cmotif4	6.384028e-199	-30.090846	True	True
18	HINFPmotif5	4.152383e-191	-29.487596	True	True
19	HINFPmotif2	4.155835e-191	-29.487568	True	True

Table 9: The top 20 most effective transcription factor DNA-binding motifs in terms of target gene prediction based on the full 10 kb region surrounding the TSS. For every motif with known targets, the derived scores for potential target genes were ranked. Subsequently, a Wilcoxon rank-sum test is applied to test if the database-verified targets are enriched towards the top of the ranked target gene list. This table presents the top 20 most effective transcription factor DNA-binding motifs in terms of target gene prediction. For this analysis, the motifs hits for the full 10 kb region surrounding the TSS (upstream + downstream) was used.

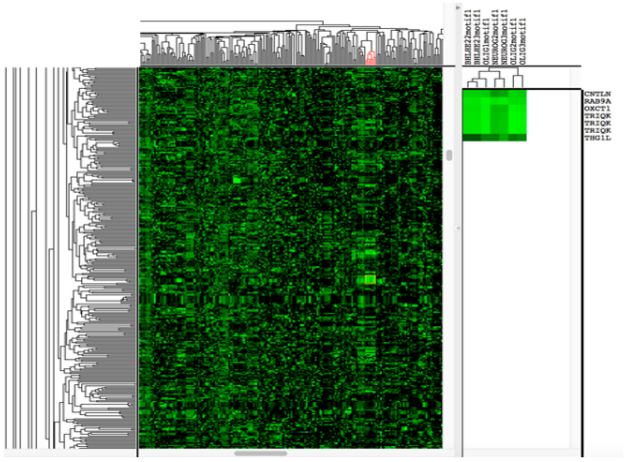
index	names	p-value	statistic	effective	corrected effective
0	NRF1motif1	0.000000e+00	-56.163250	True	True
1	KLF14motif1	0.000000e+00	-47.089085	True	True
2	SP4motif2	0.000000e+00	-44.457499	True	True
3	GABPAmotif1	0.000000e+00	-43.039728	True	True
4	SP4motif1	0.000000e+00	-40.743876	True	True
5	KLF16motif1	0.000000e+00	-39.705562	True	True
6	EGR1motif2	0.000000e+00	-39.269258	True	True
7	SP1motif1	0.000000e+00	-39.049749	True	True
8	ETV3motif1	4.708306e-298	-36.899493	True	True
9	SP3motif1	1.732832e-276	-35.529334	True	True
10	SP8motif1	1.226814e-265	-34.819645	True	True
11	TFAP2Amotif4	8.910784e-248	-33.618807	True	True
12	ZNF219motif1	2.504080e-244	-33.381975	True	True
13	ELK1motif1	1.662160e-241	-33.186926	True	True
14	ZFXmotif1	1.931399e-239	-33.043458	True	True
15	HINFPmotif4	7.175126e-236	-32.793978	True	True
16	HINFPmotif5	5.482240e-226	-32.093218	True	True
17	HINFPmotif2	5.482948e-226	-32.093214	True	True
18	ELF4motif1	1.411336e-225	-32.063768	True	True
19	SP1motif2	5.884898e-224	-31.947326	True	True

Table 10: Details of the most commonly mentioned motifs in the genome scanning section of this manuscript. This sorted table represents a list of the commonly mentioned motifs in the genome scanning section of this document. In the first column, the HGNC-symbol of the trancsription factor can be found. In the second column, the corresponding motif number is listed. The third column represents the visual representation (WebLogo) of the motif, whereas the motif matrix on which the logo is based can be found in the fourth column.

Transcription Factor	Motif Number	Logo	Motif
ELF4	1		[0.75 0.15 0.02 0.05 0. 0. 0.99 0.99 0.03 0.] [0.08 0.71 0.89 0.94 0. 0. 0. 0.01 0.01 0.05] [0.02 0.06 0.09 0.01 1. 1. 0. 0. 0.95 0.03] [0.16 0.08 0. 0. 0. 0. 0. 0.01 0. 0.92]
ELK1	1		[0.12 0.08 0. 0.01 0.98 0.88 0.22 0.04] [0.78 0.88 0. 0. 0. 0.04 0.05 0.17] [0.06 0.02 0.96 0.98 0.01 0. 0.71 0.06] [0.04 0.02 0.04 0.01 0.01 0.08 0.02 0.73]
ETV3	1		[0.71 0.06 0.04 0. 0. 0.95 0.75 0.2 0.] [0.09 0.77 0.91 0. 0. 0.04 0. 0.05 0.13] [0.12 0.05 0. 0.94 0.93 0. 0. 0.75 0.03] [0.08 0.11 0.05 0.06 0.07 0.02 0.25 0. 0.85]
FOXO6	1		[0.19 0.04 0.82 0.96 1. 0. 1. 0.05 0.03 0.01 0.03 0.04 0.91 0.01] [0.01 0.02 0.17 0.03 0. 0.95 0. 0.01 0. 0.03 0.01 0.01 0.02 0.81] [0.8 0. 0. 0. 0. 0. 0. 0.01 0.96 0. 0.05 0.15 0.03 0.] [0. 0.94 0. 0.01 0. 0.04 0. 0.93 0. 0.96 0.91 0.81 0.04 0.17]
FOXO6	2		[0.07 0.11 0.04 0.02 0. 0. 0. 0.02 0. 0.01 0.01 0.94 0.83 0.7] [0.11 0.07 0.79 0.01 0.01 0. 0. 0. 0. 0. 0. 0.02 0.03 0.12] [0.74 0.2 0.04 0.92 0. 1. 0. 0.97 0.99 0.99 0.98 0.02 0.04 0.08] [0.08 0.62 0.13 0.05 0.98 0. 1. 0.01 0. 0. 0. 0.02 0.1 0.1]
GABPA	1		[0.07 0.09 0. 0. 1. 0.97 0.09 0.09] [0.77 0.9 0. 0. 0. 0. 0.03 0.21] [0.16 0.01 1. 1. 0. 0. 0.88 0.06] [0. 0. 0. 0. 0. 0.03 0. 0.65]
HINFP	2		[0.11 0.05 0. 0.02 0.98 0.03 0.01 0.02] [0.04 0.93 0.05 0.02 0. 0.95 0.12 0.2] [0.79 0.03 0.95 0.96 0.02 0.01 0.82 0.04] [0.05 0. 0. 0. 0. 0. 0.05 0.75]
HINFP	4		[[0. 0. 0. 0. 0.99 0. 0. 0.02 0.09 0.03 0.58 0.55 0. 0. 0. 0.01 0. 0. 0.] [0.01 1. 0. 0. 0.01 1. 0.2 0.34 0.17 0.47 0.23 0.2 0.72 0. 0. 0.99 1. 0. 0.95] [0.98 0. 1. 1. 0. 0. 0.8 0.1 0.19 0.49 0.14 0.22 0.28 1. 0.01 0. 0. 1. 0.01] [0. 0. 0. 0. 0. 0. 0. 0.54 0.55 0.01 0.04 0.03 0. 0. 0.99 0. 0. 0. 0.04]]
HINFP	5		[0.75 0.05 0. 0. 0. 0. 0. 0.05] [0.04 0.82 0.01 0.02 0.96 0.95 0.03 0.79] [0.2 0.12 0.95 0. 0.02 0.05 0.93 0.04] [0.02 0.01 0.03 0.98 0.02 0. 0.05 0.11]
KLF4	1		[0.05 0.47 0.01 0.43 0.01 0.03 0.01] [0.53 0.18 0.79 0.01 0.64 0.64 0.64] [0.37 0.34 0.01 0.51 0.15 0.3 0.29] [0.06 0.02 0.19 0.05 0.2 0.03 0.05]
KLF7	1		[0.17 0. 0. 0.1 0.02] [0. 0. 0. 0.66 0.] [0.82 1. 1. 0.01 0.97] [0. 0. 0. 0.23 0.01]
KLF14	1		[0.33 0.22 0. 0.9 0. 0.02 0. 0. 0. 0.42 0.02] [0.02 0.57 0.86 0.07 1. 0.01 1. 0.99 0.96 0.56 0.71] [0.63 0.04 0.01 0.03 0. 0.96 0. 0. 0. 0. 0.01] [0.02 0.17 0.13 0. 0. 0.01 0. 0. 0.04 0.02 0.27]
KLF16	1		[0.18 0.02 0.06 0. 0. 0.15 0.01 0. 0.03 0.04] [0.01 0.01 0. 0. 0. 0.6 0.01 0.02 0. 0.03] [0.79 0.7 0.93 0.99 1. 0.06 0.94 0.28 0.93 0.6] [0.02 0.26 0.01 0. 0. 0.19 0.04 0.7 0.04 0.33]

Transcription Factor	Motif Number	Logo	Motif
NRF1	1		[0.1 0.11 0.08 0.03 0.03 0.57 0.01 0.01 0.01 0.01 0.02] [0.14 0.05 0.9 0.02 0.91 0.25 0.06 0.02 0.92 0.06 0.83] [0.1 0.83 0.02 0.94 0.05 0.12 0.09 0.97 0.01 0.88 0.01] [0.66 0.01 0. 0.01 0.02 0.06 0.83 0. 0.06 0.05 0.14]
SP1	1		[[0.13 0. 0.02 0.17 0.01 0.01 0.12 0.11 0.1 0.08 0.21 0.15 0.12 0.16 0.09 0.12 0.12 0.12] [0.21 0.25 0.23 0.52 0.23 0.25 0.2 0.23 0.59 0.46 0.14 0.2 0.17 0.2 0.27 0.23 0.15 0.12] [0.63 0.74 0.74 0.13 0.72 0.65 0.59 0.57 0.12 0.25 0.42 0.59 0.68 0.57 0.58 0.58 0.62 0.7] [0.04 0. 0.02 0.19 0.03 0.1 0.09 0.09 0.19 0.21 0.24 0.07 0.04 0.07 0.06 0.07 0.11 0.06]]
SP1	2		[0. 0. 0. 0.26 0. 0.05 0.11 0.26] [0. 0. 0.68 0. 0.11 0. 0.15 0.09] [1. 1. 0.22 0.52 0.63 0.74 0.69 0.] [0. 0. 0.11 0.22 0.26 0.22 0.05 0.65]
SP4	1		[0.35 0.23 0.34 0.07 0.04 0.13 0.06 0.06 0.35 0.21] [0.03 0.02 0.02 0.01 0.01 0.76 0.02 0.02 0.03 0.01] [0.55 0.63 0.61 0.89 0.93 0.01 0.85 0.85 0.56 0.74] [0.08 0.12 0.02 0.02 0.03 0.1 0.07 0.07 0.05 0.03]
SP4	2		[0.52 0.57 0.16 0.06 0. 0.75 0. 0. 0. 0. 0. 0.42 0.02] [0.04 0.03 0. 0.94 0.97 0.25 1. 0. 1. 1. 0.98 0.56 0.78] [0.02 0.36 0.83 0. 0. 0. 0. 0. 0.99 0. 0. 0. 0.] [0.43 0.04 0. 0. 0.03 0. 0. 0.01 0. 0. 0.02 0.02 0.19]
SP8	1		[0.32 0.12 0.79 0.04 0.09 0.03 0.01 0. 0.51 0.05] [0.66 0.81 0.2 0.93 0.14 0.93 0.97 0.9 0.44 0.7] [0. 0. 0.01 0.01 0.62 0.03 0.02 0.02 0.03 0.09] [0.02 0.06 0.01 0.03 0.15 0.01 0. 0.08 0.02 0.16]
STAT3	1		[0.21 0.04 0.86 0.88] [0.02 0.02 0.1 0.06] [0.75 0.86 0. 0.04] [0.02 0.08 0.04 0.02]
STAT3	2		[[0.02 0.1 0.17 0.49 0.16 0.03 0.03 0.14 0.03 0.18 0.09 0.05 0.96 0.98] [0.18 0.18 0.37 0.13 0.3 0.02 0.01 0.81 0.62 0.06 0. 0. 0.03 0.01] [0.59 0.39 0.28 0.2 0.16 0.02 0.19 0.01 0.01 0.49 0.91 0.89 0.01 0.] [0.2 0.33 0.18 0.18 0.38 0.93 0.77 0.04 0.34 0.27 0. 0.06 0. 0.01]
STAT3	3		[[0.05 0. 0.56 0. 0. 0.02 0.16 0.05 0.37 0.28 0. 0.88 0.98 0.7] [0. 0.35 0.09 0.3 0. 0. 0.67 0.63 0.09 0. 0. 0.12 0.02 0.02] [0.19 0.63 0.21 0.14 0. 0.09 0.05 0.02 0.42 0.72 0.65 0. 0. 0.19] [0.77 0.02 0.14 0.56 1. 0.88 0.12 0.3 0.12 0. 0.35 0. 0. 0.09]
TFAP2A	1		[0.07 0. 0.02 0.06 0.47 0.06 0.19 0.04] [0.77 0.75 0.47 0.67 0.13 0.14 0.02 0.4] [0.05 0.19 0.06 0.22 0.22 0.77 0.78 0.54] [0.11 0.06 0.45 0.05 0.18 0.02 0.01 0.01]
TFAP2A	4		[0.1 0.05 0.04 0.03 0.08 0.3 0.3 0.01 0. 0.01] [0.2 0.68 0.86 0.56 0.36 0.31 0.17 0. 0.06 0.89] [0.69 0.25 0.04 0.11 0.24 0.31 0.47 0.98 0.94 0.07] [0.02 0.03 0.06 0.3 0.33 0.08 0.06 0. 0. 0.04]
TFAP2C	4		[0.1 0.01 0.01 0.01 0.09 0.42 0.37 0.04 0.02 0.03] [0.22 0.9 0.97 0.61 0.37 0.25 0.06 0.04 0.15 0.76] [0.65 0.08 0.01 0.07 0.22 0.28 0.57 0.89 0.83 0.14] [0.03 0.01 0.02 0.31 0.32 0.04 0. 0.03 0.01 0.07]
ZFX	1		[0.09 0.06 0.06 0.08 0.07 0.14 0.45 0. 0. 0. 0. 0.02 0.23 0.2] [0.16 0.15 0.68 0.72 0.37 0.58 0.03 0. 0. 0.99 0.89 0.33 0.41 0.01] [0.71 0.74 0.17 0.12 0.33 0.18 0.29 0.99 1. 0.01 0.09 0.27 0.3 0.78] [0.04 0.05 0.09 0.08 0.22 0.11 0.03 0. 0. 0. 0.02 0.39 0.06 0.01]
ZNF219	1		[0.03 0. 0. 0.02 0. 0.01 0.03 0.04 0.] [0.01 0.09 0.01 0.03 0.01 0. 0.43 0.18 0.04] [0.88 0.88 0.79 0.92 0.97 0.99 0.15 0.75 0.91] [0.07 0.03 0.19 0.03 0.01 0. 0.39 0.03 0.04]

1. Cluster overview for upstream regions



2. Example of a STAT3 cluster

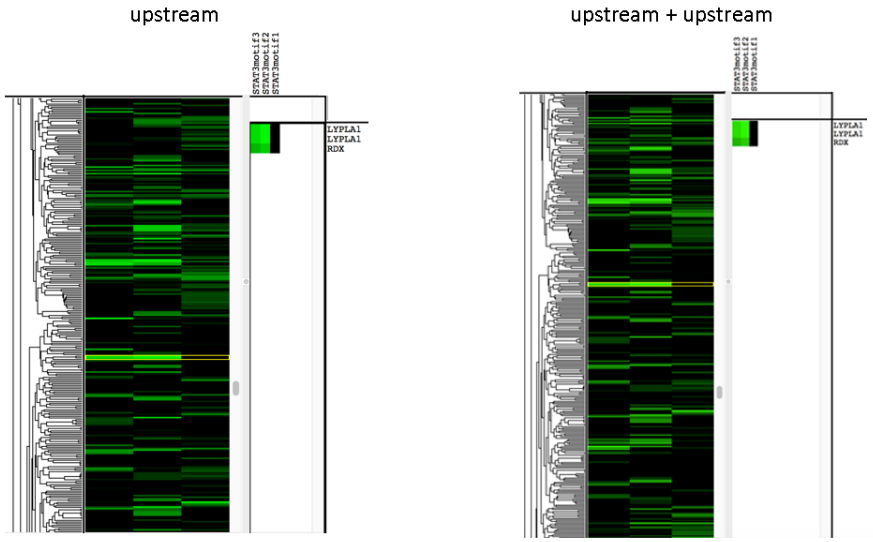


Figure 33: TreeView interface for the clustered z-score-transformed motif vectors. In this image, the CLUSTER 3.0 generated cluster is visualized using TreeView 3.0. Every TF-motif has a z-score-transformed score vector with values for every potential target gene. These vectors were hierarchically clustered using Pearson's correlation coefficient as pairwise distance measure with the complete-linkage approach to measure distance between clusters. The TF-motifs are shown horizontally, whereas the target genes are shown vertically. Target genes clustering together are likely to be similarly structured as they have similar scores for the majority of TF-motifs. TF-motifs clustering together likely have similar motif structures or similarly rank the majority of target genes. On the bottom, a selection for the STAT3 motifs is made. On the bottom left, the cluster for the values based on only the 5 kb upstream region is presented, whereas on the bottom right the cluster for the values based on the full 10 kb region surrounding the TSS is given. Motif 2 and 3 of STAT3 are considered effective and therefore show high similarity with each other, in contrary to STAT3 - motif 1. Further manually investigating the resulting clusters is certainly interesting, but beyond the scope of this thesis.

Table 11: Ranking of the 20 most relevant TFs of the middle-age old vs. young contrasts in the other aging contrasts. In this series of tables, the ranking of the 20 most relevant transcription factors in the middle-age vs. young contrasts are evaluated in the other contrasts (normal age stage and extreme age stage). A similar ranking in a different contrast is considered to indicate similar regulatory TF patterns as observed in the contrast under study. The samples of male (top) and the samples of female (bottom) origin are separated. Similar comparisons for normal age and extreme age can be found in table 12 and 13, respectively.

	mid age vs. young (male)	norm age vs. young (male)	extreme age vs. young (male)	mid age vs. young (female)	norm age vs. young (female)	extreme age vs. young (female)
1	STAT5A_motif2	126	180	116	203	185
2	PRDM6_motif1	83	99	108	170	266
3	STAT4_motif2	8	51	71	21	71
4	STAT6_motif2	193	244	185	250	234
5	NFATC1_motif4	60	39	107	99	124
6	TBP_motif3	108	86	117	103	206
7	TCF7_motif1	246	176	226	289	373
8	CEBPA_motif1	129	164	97	139	144
9	TBP_motif1	81	83	124	87	192
10	POU2F1_motif8	78	132	114	86	98
11	NR3C2_motif3	157	166	70	184	200
12	SRY_motif4	52	100	8	89	128
13	PAX2_motif3	4	61	59	25	123
14	HOXB13_motif1	40	60	64	46	92
15	NFATC2_motif1	111	118	176	145	308
16	SOX11_motif1	124	125	77	153	132
17	SOX12_motif1	118	120	68	138	127
18	SOX30_motif1	110	89	58	118	107
19	LEF1_motif1	219	226	230	346	306
20	FOXS1_motif1	101	122	61	130	137

	mid age vs. young (female)	mid age vs. young (male)	norm age vs. young (male)	extreme age vs. young (male)	norm age vs. young (female)	extreme age vs. young (female)
1	LHX8_motif1	41	58	79	80	75
2	NOBOX_motif1	32	12	19	15	12
3	ZFHX2_motif1	94	28	25	40	22
4	HOXB5_motif1	59	18	24	38	6
5	HOXA2_motif1	68	24	32	50	29
6	POU2F1_motif2	48	6	2	10	4
7	NKX6-2_motif1	39	9	10	27	5
8	SRY_motif4	12	52	100	89	128
9	ESX1_motif1	77	19	17	34	18
10	GSX1_motif1	78	10	13	14	3
11	DLX4_motif1	42	20	14	20	9
12	LHX4_motif1	47	90	81	92	78
13	VSX1_motif1	63	21	11	32	10
14	HOXD4_motif1	33	3	22	3	2
15	EN2_motif1	25	32	36	70	47
16	EVX1_motif1	100	25	12	36	26
17	HOXB4_motif1	70	55	57	64	58
18	HOXC5_motif1	65	57	59	68	62
19	HOXD3_motif1	86	56	56	59	53
20	LHX1_motif1	44	62	65	76	64

Table 12: Ranking of the 20 most relevant TFs of the normal-age old vs. young contrasts in the other aging contrasts. In this series of tables, the ranking of the 20 most relevant transcription factors in the normal-age vs. young contrasts are evaluated in the other contrasts (middle age stage and extreme age stage). A similar ranking in a different contrast is considered to indicate similar regulatory TF patterns as observed in the contrast under study. The samples of male (top) and the samples of female (bottom) origin are separated. Similar comparisons for mid age and extreme age can be found in table 11 and 13, respectively.

	norm age vs. young (male)	mid age vs. young (male)	extreme age vs. young (male)	mid age vs. young (female)	norm age vs. young (female)	extreme age vs. young (female)
1	BARHL2_motif1	72	6	38	1	43
2	HOXD11_motif1	115	1	66	2	42
3	HOXD4_motif1	33	22	14	3	2
4	PAX2_motif3	13	61	59	25	123
5	HOXC10_motif1	51	9	42	7	49
6	POU2F1_motif2	48	2	6	10	4
7	NFIL3_motif1	135	78	60	4	93
8	STAT4_motif2	3	51	71	21	71
9	NKX6-2_motif1	39	10	7	27	5
10	GSX1_motif1	78	13	10	14	3
11	HLF_motif1	129	64	65	6	84
12	NOBOX_motif1	32	19	2	15	12
13	VAX1_motif1	53	4	24	42	8
14	ARID3B_motif1	96	47	73	17	21
15	VAX2_motif1	38	3	36	58	11
16	MNX1_motif1	66	16	35	54	34
17	CEBPA_motif4	124	121	89	11	68
18	HOXB5_motif1	59	24	4	38	6
19	ESX1_motif1	77	17	9	34	18
20	DLX4_motif1	42	14	11	20	9

	norm age vs. young (female)	mid age vs. young (male)	norm age vs. young (male)	extreme age vs. young (male)	mid age vs. young (female)	extreme age vs. young (female)
1	BARHL2_motif1	72	1	6	38	43
2	HOXD11_motif1	115	2	1	66	42
3	HOXD4_motif1	33	3	22	14	2
4	NFIL3_motif1	135	7	78	60	93
5	HMX2_motif1	139	31	55	87	106
6	HLF_motif1	129	11	64	65	84
7	HOXC10_motif1	51	5	9	42	49
8	HOXB3_motif1	83	29	15	28	7
9	BSX_motif1	55	36	5	43	15
10	POU2F1_motif2	48	6	2	6	4
11	CEBPA_motif4	124	17	121	89	68
12	ALX1_motif2	183	86	58	102	60
13	HOXB2_motif1	90	38	29	31	14
14	GSX1_motif1	78	10	13	10	3
15	NOBOX_motif1	32	12	19	2	12
16	PHOX2A_motif1	226	161	101	154	59
17	ARID3B_motif1	96	14	47	73	21
18	PHOX2B_motif1	241	159	102	159	72
19	HMX3_motif1	88	51	7	41	54
20	DLX4_motif1	42	20	14	11	9

Table 13: Ranking of the 20 most relevant TFs of the extreme-age old vs. young contrasts in the other aging contrasts. In this series of tables, the ranking of the 20 most relevant transcription factors in the extreme-age vs. young contrasts are evaluated in the other contrasts (mid age stage and normal age stage). A similar ranking in a different contrast is considered to indicate similar regulatory TF patterns as observed in the contrast under study. The samples of male (top) and the samples of female (bottom) origin are separated. Similar comparisons for mid age and normal age can be found in table 11 and 12, respectively.

	extreme age vs. young (male)	mid age vs. young (male)	norm age vs. young (male)	mid age vs. young (female)	norm age vs. young (female)	extreme age vs. young (female)
1	HOXD11_motif1	115	2	66	2	42
2	POU2F1_motif2	48	6	6	10	4
3	VAX2_motif1	38	15	36	58	11
4	VAX1_motif1	53	13	24	42	8
5	BSX_motif1	55	36	43	9	15
6	BARHL2_motif1	72	1	38	1	43
7	HMX3_motif1	88	51	41	19	54
8	EVX2_motif2	107	30	32	35	27
9	HOXC10_motif1	51	5	42	7	49
10	NKX6-2_motif1	39	9	7	27	5
11	VSX1_motif1	63	21	13	32	10
12	EVX1_motif1	100	25	16	36	26
13	GSX1_motif1	78	10	10	14	3
14	DLX4_motif1	42	20	11	20	9
15	HOXB3_motif1	83	29	28	8	7
16	MNX1_motif1	66	16	35	54	34
17	ESX1_motif1	77	19	9	34	18
18	DLX6_motif1	71	33	29	30	16
19	NOBOX_motif1	32	12	2	15	12
20	DLX2_motif1	80	26	27	29	19

	extreme age vs. young (female)	mid age vs. young (male)	norm age vs. young (male)	extreme age vs. young (male)	mid age vs. young (female)	norm age vs. young (female)
1	POU3F2_motif4	275	84	33	150	65
2	HOXD4_motif1	33	3	22	14	3
3	GSX1_motif1	78	10	13	10	14
4	POU2F1_motif2	48	6	2	6	10
5	NKX6-2_motif1	39	9	10	7	27
6	HOXB5_motif1	59	18	24	4	38
7	HOXB3_motif1	83	29	15	28	8
8	VAX1_motif1	53	13	4	24	42
9	DLX4_motif1	42	20	14	11	20
10	VSX1_motif1	63	21	11	13	32
11	VAX2_motif1	38	15	3	36	58
12	NOBOX_motif1	32	12	19	2	15
13	DLX5_motif1	82	23	21	23	28
14	HOXB2_motif1	90	38	29	31	13
15	BSX_motif1	55	36	5	43	9
16	DLX6_motif1	71	33	18	29	30
17	ARID5A_motif1	26	71	67	119	123
18	ESX1_motif1	77	19	17	9	34
19	DLX2_motif1	80	26	20	27	29
20	SHOX2_motif1	76	44	27	50	43

Table 14: Gene ontology enrichment analysis for the top TFs versus all tested TFs in the context of the second aging study by Lu et al. In the second aging study by Tao Lu et al. (2014), three different stages of aging (middle age, normal age, and extreme age) are compared to young samples in the frontal cortex of the human brain. The top 20 most relevant TFs in every contrast (split by gender) are tested for a significant enrichment for any GO category by comparing with a background list of all tested transcription factors. The GO analysis was performed using the GOrilla webtool. In the first column, the contrast is specified. In the second column, the unique GO category identifier is listed. The description of the category follows in the third column. Adjusted p-values are presented in the next column. Finally, the TFs on which the enrichment is based are shown in the last column (right).

Contrast	GO term	Description	P-value	Genes
mid age (male)	GO:0043603	cellular amide metabolic process	1.37e-04	SOX11 CEBPA STAT5A
	GO:0051151	negative regulation of smooth muscle cell differentiation	4.63e-04	NFATC1 PRDM6 NFATC2
	GO:0002568	somatic diversification of T cell receptor genes	6.17e-04	TCF7 LEF1
	GO:0002681	somatic recombination of T cell receptor gene segments	6.17e-04	TCF7 LEF1
	GO:0033153	T cell receptor V(D)J recombination	6.17e-04	TCF7 LEF1
extreme age (female)	GO:0048562	embryonic organ morphogenesis	3.77E-04	HOXB3 HOXB2 HOXD4 DLX2 HOXB5 VAX2 SHOX2
	GO:0048704	embryonic skeletal system morphogenesis	5.49E-04	HOXB3 HOXB2 HOXD4 DLX2 HOXB5 SHOX2
norm age (male)	No GO enrichment under the significance threshold.			
extreme age (male)	No GO enrichment under the significance threshold.			
mid age (female)	No GO enrichment under the significance threshold.			
norm age (female)	No GO enrichment under the significance threshold.			

Table 15: WebLogo representation of the top 10 TF-motifs for all six contrasts presented in the second study by Lu et al. In the second aging study by Tao Lu et al. (2014), three different stages of aging (middle age, normal age, and extreme age) are compared to young samples in the frontal cortex of the human brain. In this table, the motifs of these highly relevant TFs are shown. The 6 different columns represent all contrasts (3 stages of aging separated by gender).

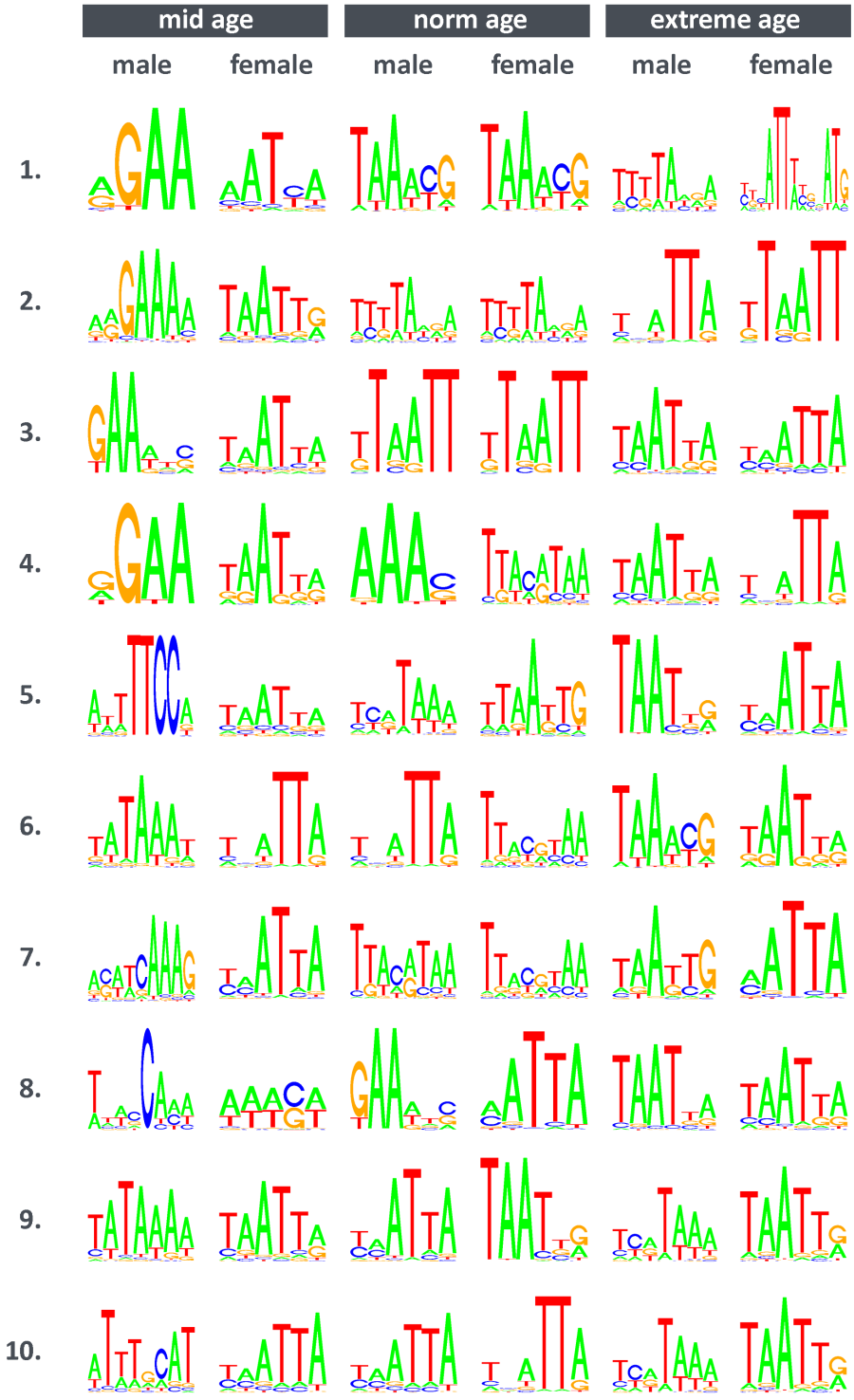


Table 16: WebLogo representation of the top 10 TF-motifs for all three contrasts presented in the rejuvenation experiment. In the rejuvenation experiment, three different aging-related contrasts (BBL treated old vs. young, untreated old vs. young, and untreated old vs. BBL treated old) are studied in human skin tissue. BBL treatment (broadband light treatment) is assumed to rejuvenate human skin tissue. In this table, the motifs of these highly relevant TFs are shown. The 3 different columns represent all three contrasts.