AGE REVERSAL BY TRANSCRIPTOMIC REPROGRAMMING

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Division of Medical Sciences

in the subject of Biological and Biomedical Sciences

have examined a dissertation entitled

Age reversal by transcriptomic reprogramming

presented by Alex Plesa

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature: Amy Wagers
Typed Name: Dr. Amy Wagers

Signature: Vadim Gladyshev
Typed Name: Dr. Vadim Gladyshev

Signature:
Typed Name: Dr. Alexander Soukas

Signature:
Typed Name: Dr. Joao de Magalhaes

Date: June 28, 2022

AGE REVERSAL BY TRANSCRIPTOMIC REPROGRAMMING

A dissertation presented

by

Alex Pleșa

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Sciences

Harvard University
Cambridge, Massachusetts
June 2022

Dissertation Advisor: George M. Church Alex Pleșa

Age reversal by transcriptomic reprogramming

ABSTRACT

Aging is a progressive multifaceted functional decline of a biological system. Chronic age-related conditions such as cancer, cardiovascular, and neurodegenerative diseases are leading causes of death worldwide. As the elderly population is experiencing an exponential growth, aging is becoming a pressing problem for our society. To address this global challenge, there is a need for novel, safe, and effective rejuvenation therapies aimed at reversing age-related phenotypes and improving human health. In my dissertation work, I sought to advance the discovery of age reversal therapies by developing novel cell engineering tools and applying them to the aging problem. With gene expression being a key determinant of cell identity and function, I have focused my efforts on reprogramming the cell transcriptome to a youthful state. To this end, we first developed an integrated pipeline that enables fast and efficient mammalian cell engineering using algorithms for target identification, modular tools for library cloning, and next-generation sequencing-coupled readouts. Next, we built a functionally interpretable RNA clock as an integrative aging assay that can accurately predict the age of human fibroblasts, as well as the effect of several pro- and anti- aging interventions. Lastly, we used our cell engineering pipeline along with our aging clock to perform the first age reversal screen in primary human cells. Our work uncovered several targets for cellular rejuvenation that induced younger transcriptomes in middle-aged and old human fibroblasts. The most promising of these targets was serine and arginine rich splicing factor 1 (SRSF1), whose overexpression had a robust

rejuvenating effect on both the transcriptome and function of aged human fibroblasts. Throughout these studies I propose a new paradigm for the discovery of age reversal interventions by using transcriptomic reprogramming screens.

TABLE OF CONTENTS

AGE REVERSAL BY TRANSCRIPTOMIC REPROGRAMMING i

COPYRIGHT ii

ABSTRACT iii

TABLE OF CONTENTS v

ACKNOWLEDGEMENTS viii

TABLE OF FIGURES AND TABLES xi

CHAPTER I : INTRODUCTION 1

AUTHORS AND CONTRIBUTIONS 2

INTRODUCTION 2

The problem of aging 2

What is aging? 4

Aging clocks 7

Age reversal 9

Cell engineering 11

Transcriptomic reprogramming 12

SUMMARY AND GOALS FOR THIS DISSERTATION 13

CHAPTER II : AN INTEGRATED PIPELINE FOR MAMMALIAN GENETIC SCREENING 16

AUTHORS AND CONTRIBUTIONS 17

ABSTRACT 17

INTRODUCTION 18

RESULTS 20

In silico methods for gene target identification 20

DEG network analysis 21

GRN centrality analysis 22

Modular perturbation tools 24

CRISPRa performance in hiPSCs 25

CRISPRi performance in hiPSCs 25

cDNA overexpression performance in hiPSCs and primary human fibroblasts 28

Dual gene perturbation via cDNA overexpression and CRISPR 29

Library cloning 30

MegaGate—A unique toxin-less recombination cloning tool 30

Validation of MegaGate in single and pooled cloning 32

Screening and NGS-coupled readouts 33

DISCUSSION 37
Limitations of the study *38*
CHAPTER III : TRANSCRIPTOMIC REPROGRAMMING SCREEN USING FUNCTIONAL RNA CLOCK FOR CELLULAR REJUVENATION 40
AUTHORS AND CONTRIBUTIONS 41
ABSTRACT 41
INTRODUCTION 42
RESULTS 43
Process-based transcriptomic aging clock *43*
Age reversal screen identifies genes for cellular rejuvenation *46*
Transcriptomic variability in aging phenotypes *49*
SRSF1 induces robust cellular rejuvenation *52*
DISCUSSION 54
CHAPTER IV : DISCUSSION 56
AUTHORS AND CONTRIBUTIONS 57
SUMMARY 57
New tools for cell engineering screens *57*
Transcriptomic reprogramming screen using RNA clock *59*
FUTURE DIRECTIONS 62
The future of cell engineering *62*
Using transcriptomic reprograming screens to uncover age reversal targets *64*
CONCLUSION 66
APPENDIX I : METHODS FOR CHAPTERS II-III 67
CELL LINES UTILIZED IN CHAPTER II 68
DEG NETWORK TARGET PREDICTION 69
GRN INFERENCE AND PAGERANK TARGET IDENTIFICATION 70
sgRNA PREDICTION FOR CRISPRa AND CRISPRi 70
CRISPR TOOL EVALUATION IN hiPSCs 72
cDNA VECTOR COMPARISON 73
DUAL cDNA AND CRISPR SCREENING 74
MEGAGATE MOLECULAR CLONING VECTOR CONSTRUCTION 75
MEGAGATE REACTION CONDITIONS OPTIMIZATION 76
MEGAGATE VECTOR DNA BARCODING 78
SINGLE AND POOLED MEGAGATE CLONING AND NGS READOUT 79
PIGGYBAC COPY NUMBER TITRATION 80
gDNA BARCODE ENRICHMENT ANALYSIS (BAR-SEQ) 81

TARGETED RNA-SEQ (TAR-SEQ) 82
RNA-SEQ AND COUPLED BARCODE ENRICHMENT 83
QUANTIFICATION AND STATISTICAL ANALYSIS 84
RNA-SEQ DATA PROCESSING AND ANALYSIS 85
DIFFERENTIAL GENE EXPRESSION ANALYSIS 86
DIFFERENTIAL SPLICING ANALYSIS 86
DECOMPOSITION OF BULK RNA-SEQ 87
IDENTIFICATION OF CELLULAR PROCESSES PREDICTIVE OF AGE 87
TRAINING OF PROCESS-BASED TRANSCRIPTOMIC CLOCKS 89
PREDICTING THE AGE OF NON-TRAINING SAMPLES 89
COMPUTING PROCESS ACTIVITY SCORES 90
SCREENING LIBRARY CONSTRUCTION 91
TISSUE CULTURE 91
RNA SEQUENCING 92
FLOW CYTOMETRY 93
IN VITRO SCRATCH ASSAY 93
CODE AVAILABILITY 94
APPENDIX II : SUPPLEMENTAL FIGURES AND TABLES FOR CHAPTER II 95
APPENDIX III : SUPPLEMENTAL FIGURES AND TABLES FOR CHAPTER III . 107
REFERENCES 148

ACKNOWLEDGEMENTS

First, I would like to thank Prof. George M. Church, my thesis advisor, for his unwavering support and insight over the past five years. I am grateful for having been mentored by George, who has given me the freedom necessary to develop as an independent scientist, while also providing guidance and direction when needed. I also want to thank all lab members, past and present, for being knowledgeable, supportive, helpful, and great people to spend a great part of my time with, both inside and outside the lab. I would like to specifically thank Dr. Stepan Jerabek, who has played a pivotal role in my graduate studies through his continued mentorship and friendship over the years. Similarly, I want to thank Dr. Dima Ter-Ovanesyan and Dr. Xiaoge Guo for her their mentoring and invaluable technical expertise. Thank you, Dr. Noah Davidsohn for your mentorship and helpful career advice. Thank you, Dr. Anush Chiappino-Pepe for all the helpful coffee breaks and chats that gave me the motivation to try new experiments and continue to push through roadblocks. Thank you, Christian Kramme for being a great bay mate and showing me that there are more important things in life than scientific publications. Thank you, Dr. Jenny Tam and Dr. Richie Kohman for your mentorship and administrative support. Thank you, Emma Taddeo for being a fantastic and supportive lab manager, making sure I don't miss any deadlines and helping me navigate through all the administrative requirements. Thank you, Nicole D'Aleo, for your willingness to answer my questions about funding and approving all my (many) purchases.

I would like to specifically thank all the people that I had the pleasure to directly work with over the years. Thank you, Helen Wang for being a great and hard-working student, and for all your scientific and graphic contributions to my research. Thank you, Michael Shadpour for all the tissue culture help during late-nights and even early mornings, and for reminding me to take my

own advice and “be better”. Thank you, George Cai for your computational help with the many on-going projects and the last-minute data analysis. Thank you, Fawad Omar for your experimental help and feedback on my dissertation. Thank you, Dr. Sascha Jung for being a great collaborator.

I want to extend my gratitude to the members of my DAC, Dr. Amy Wagers, Dr. Will Mair, and Dr. Steve Elledge for their helpful suggestions and constructive criticism throughout my Ph.D. I would like to specifically thank Dr. Amy Wagers and Dr. Will Mair for taking extra time to mentor me outside of the DAC meetings, providing invaluable scientific and career advice. Thank you to all the BBS faculty and administrators I had the pleasure to talk to and work with, including Dr. Susan Dymecki, Dr. Tom Bernhardt, Dr. David Van Vactor, Dr. Monica Colaiácovo, Dr. Fred Winston, Dr. Max Heiman, Anne O’Shea, and Daniel Gonzalez. To my fellow DMS students, including (but definitely not limited to) Irene Wong, Vicky Chou, Nikita Sturrock, Adrija Navarro, Aparna Nathan, Lisa Stiu, Sam Sedor, and Ben Wolf, thank you for being there during the many ups and downs of a Ph.D., and for the many fun times we had. A special thanks to Patrick Loi for being a great roommate and friend throughout the years.

I am also very appreciative of the mentorship I received from my early scientific career mentor, Dr. Ron Peck. His support and guidance through my undergraduate studies have enabled me to attend this program and pursue a scientific career. I hope your future students will be able to keep up with your running requirements.

None of this would have been possible without the constant support and encouragement from my parents and my brother. The weekly Skype calls and yearly vacations in Greece were a source of fun and relaxation that kept me motivated and energized throughout the past five years.

Lastly, I want to thank my funding sources, the Aging and Longevity-Related Research Fund charitable gift, the Astera Fund at the Silicon Valley Community Foundation charitable gift, and the Wyss Institute Synthetic Biology Platform for making this work possible.

TABLE OF FIGURES AND TABLES

Figure 1.1: Mortality rate for leading causes of death in the U.S. 3
Figure 1.2: Age distribution of world population in 2020 and in 2100. 4
Figure 1.3: Schematic of transcriptomic reprogramming paradigm. 12
Figure 2.1: STAMPScreen schematic workflow. 20
Figure 2.2: *In silico* target ascertainment. 23
Figure 2.3: Gene perturbation tool evaluation. 26
Figure 2.4: MegaGate, a toxin-less cDNA cloning method. 31
Figure 2.5: NGS-coupled readouts. 34
Figure 3.1: Process-based transcriptomic clock reports on biological age of human fibroblasts. 44
Figure 3.2: Age reversal screen identifies genes for cellular rejuvenation. 46
Figure 3.3: Variability of the aging phenotype and its response to perturbations. 50
Figure 3.4: SRSF1 induces robust cellular rejuvenation. 52
Figure A2.1: Relative expression of gene targets, Related to Figure 2.3. 96
Figure A2.2: cDNA expression vector dynamics were compared using flow cytometry, Related to Figure 2.3. 98
Figure A2.3: Optimization reactions for MegaGate Cloning, Related to STAR Methods. 99
Figure A2.4: Copy number in single cell isolated and expanded clones for three different gene sets and input ratios, Related to Figure 2.4. 101
Figure A2.5: Validation of Targeted RNA-Sequencing (TAR-Seq) with custom gene panels, Related to STAR Methods. 102
Figure A3.1: Functional annotation of clock genes. 108
Figure A3.2: RNA clock performance across different fibroblast samples. 109
Figure A3.3: RNA clock response to age-modulating interventions. 110
Figure A3.4: Workflow and validation for network scoring approach applied to fibroblast aging. 111
Figure A3.5: Barcode validation of gene overexpression. 112
Figure A3.6: Correlations between different metrics for our candidate genes. 113
Figure A3.7: Flow cytometry assays for M55 line across all overexpressed genes. 114
Figure A3.8: Flow cytometry assays for M79 line across all overexpressed genes. 115
Figure A3.9: Process-specific scaled age effect for the strong perturbations in our screen. 116
Figure A3.10: Scratch assay in M79 line expressing SRSF1, OSKM, or BFP. 117
Figure A3.11: Splicing and cell identity analysis in SRSF1 overexpression lines. 118
Figure A3.12: Fibroblast identity analysis across all genes and lines in the screen. 119
Figure A3.13: Representative flow cytometry cell gating. 120
Table A2.1: Primers used in Chapter 2 103
Table A3.1: Genes included in the RNA clock. 121
Table A3.2: Candidate genes for age reversal screen. 126
Table A3.3: Plasmids and barcodes for the candidate genes. 129
Table A3.4: Predicted age of samples in the age-reversal screen. 134
Table A3.5: Public data accession IDs. 147

CHAPTER I : INTRODUCTION

AUTHORS AND CONTRIBUTIONS

Author contributions: I wrote this chapter with input from my advisor Prof. George M. Church.

INTRODUCTION

Aging remains the greatest unsolved biological problem. The main reason for the limited progress in the aging field is the high complexity of the process, which spans all areas of known and unknown biology. Attempting to reverse aging is thus similar to trying to fix a car before fully understanding how its parts work, both individually and together. Nonetheless, the discovery of several longevity interventions in the last few decades has shown that a full understanding of human biology is not necessary for uncovering interventions that modulate the aging process. Therefore, while difficult and complex, solving aging may be a feasible goal that I seek to contribute to through this dissertation.

The problem of aging

Aging is a major risk factor for multiple prevalent human diseases including cancer, cardiovascular disease, neurodegeneration, diabetes, and most recently COVID-19 (Figure 1.1)[1,2]. The mortality rate associated with all these diseases increases exponentially with age. Cancer, heart diseases, and diabetes have an early onset during the 20s, while COVID-19 affects people over 35, and neurodegenerative diseases start spiking in the early 50s. Although deaths are the most evident outcomes of age-related diseases, the suffering and drop in quality of life associated with these dysfunctions also have a significant negative impact on our society[3].

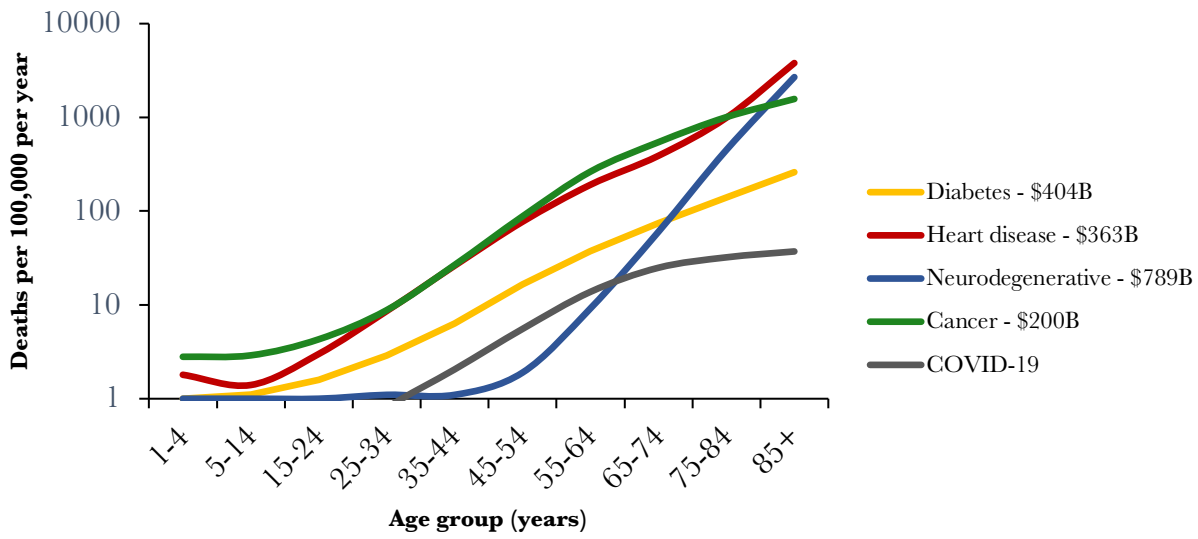


Figure 1.1: Mortality rate for leading causes of death in the U.S.

Death rate per 100,000 people per year across multiple age groups in the U.S. The data for diabetes, heart disease, neurodegenerative, and cancer is representative of 2019[1], while the COVID-19 deaths are compiled for 2020-2021[2]. The estimated associated healthcare costs for diabetes (2017), heart disease (2017), neurodegenerative diseases (2017), and cancer (2020) are also reported in the figure.

Another major aspect of aging that affects our entire society, not just the elderly, is the negative impact on the economy. In the U.S., the population over 55 years old accounts for more than half the total healthcare spending[4], and most of these costs are associated with age-related diseases. For example, diabetes, heart disease, and neurodegenerative diseases have an estimated associated healthcare cost of approximately $409B (2017)[5], $363B (2017)[6], $789B (2017)[7], and $200B (2020)[8], respectively. This doesn't include the impact of the recent COVID-19 pandemic on the U.S. health sector, which is currently estimated at $363B[9] for 2020. Besides the aforementioned direct costs, aging also incurs indirect costs on our society through the associated drop in productivity and GDP growth. A recent study has modeled the economic value of targeting aging to approximately $38 trillion for an increase in life expectancy by one year[10], thereby highlighting the hidden cost of aging.

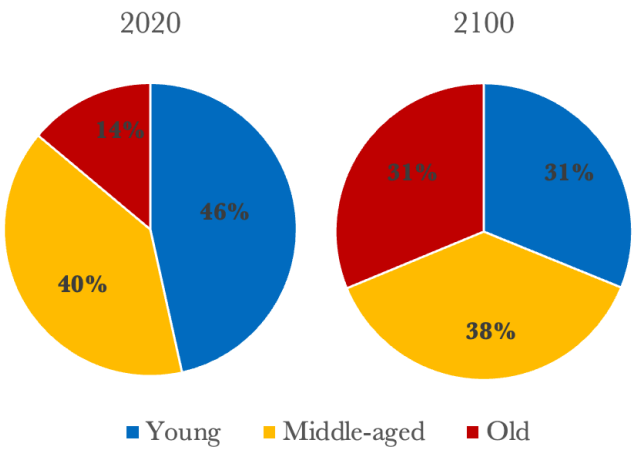


Figure 1.2: Age distribution of world population in 2020 and in 2100.

The proportion of young (<30), middle-aged (30-60), and old (>60) individuals in the world population as estimated in 2020 and as predicted in 2100[11].

While the present social and economic challenges associated with aging are already pronounced, they are predicted to only increase in scale in the next decades, if not addressed. The United Nations has projected the elderly population will nearly double by the end of the century[11], exacerbating the negative effects of aging on our society (Figure 1.2). It is thus important and urgent to tackle the growing complex problem of aging not only to avoid its effects on our economies, but to also increase the well-being of our global population.

What is aging?

There is currently no widely-accepted comprehensive definition of aging, but most evolutionary biologists agree the aging process is characterized by an age-dependent decline in physiological function that leads to an associated increase in the mortality[12]. For the purposes of this dissertation, I will use a similar but broader version of this definition and refer to aging as a progressive functional decline of a biological system.

Under this definition, it follows that aging manifests itself in biological systems at different levels of abstraction. The most apparent of these levels is the entire human organism, which follows a universal life trajectory. Newborns only attain peak fitness after development, when they reach early adulthood, a period characterized by stable organismal function and good health. This is followed by late adulthood, which marks the start of the progressive functional decline that ends in frailty and ultimately death. However, a similar trajectory is apparent at the tissue and cellular level, where the units of the systems slowly decay and lead to many functional impairments.

In an effort to better describe this functional decay, López-Otín et al.[13] have identified 9 hallmarks of aging as common cellular dysfunctions that develop throughout mammalian life and whose amelioration can retard the aging process. These hallmarks represent afflictions of several important cellular functions: genomic instability, telomere attrition, epigenetic alterations, loss of proteostasis, deregulated nutrient-sensing, mitochondrial dysfunction, cellular senescence, stem cell exhaustion, and altered intercellular communication. Similarly, Kennedy et al.[14] have defined 7 pillars of aging, as critical areas of aging research: epigenetics, proteostasis, macromolecular damage, metabolism, inflammation, stem cells and regeneration, and adaptation to stress. While the hallmarks and pillars have some differences, it is apparent that both lists have highly interconnected terms with hierarchical relationships, where some functions such as epigenetics are more causal than others like stem cell exhaustion[15].

Gene expression is a central function in the aging process under the direct regulation of epigenetics[16]. Through the coordinated regulation of transcription, gene expression is the main determinant of a cell's structure and function. While all cells of an organism share the same (unmutated) genome, their specification and function are dictated by the transcriptome. As such, since aging manifests through a decline in cellular function, the transcriptome is an evident layer that can both report on and modulate the state of a cell.

Indeed, several studies have thoroughly described the aging epigenome and transcriptome across species and cell types[17]. One of the central age-related epigenetic changes is a disruption of chromatin homeostasis characterized by histone loss[18], general genome-wide DNA hypomethylation[19], and chromatin remodeling due to altered histone modifications balance[20]. For example, genomic regions marked by broad H3K4me3 domains were linked to increased transcriptional consistency, and perturbing the balance of this histone modification was shown to increase transcriptional variability[21]. In line with these observations, Benayoun et al. reported common age-related chromatin and transcriptional changes in the mouse heart, liver, and cerebellum that upregulated the innate immune response, with similar pathways also being affected in African turquoise killifish, rat, and humans[22].

It is widely believed that the age-related loss of chromatin homeostasis leads directly to increased transcriptomic instability[17,19,20,23], and a multitude of studies have confirmed this assertion in different biological systems. One of the first such studies was that of Bahar et al., who showed increased transcriptional noise in the aging mouse heart, as determined by the cell-to-cell variability in gene expression[24]. Since then, with the advent of single-cell technologies, researchers have uncovered similar gene expression instability in mouse[25] and human[26] immune cells, human lung cells[27], and in the human pancreas[28]. Another related factor that increases the transcriptional noise with aging is the dysregulation of the splicing process, which is essential in maintaining the resilience and regulation of the transcriptome in response to different stimuli[29]. As such, splicing has been linked to aging and longevity, and due to its universal nature across species was proposed as a new hallmark of aging[30–32].

A promising mechanistic link between the age-related transcriptional instability and the associated functional decline is a loss of cell identity, where the coordinated transcriptional program required for youthful function is affected by a downregulation of cell-type specific genes[33].

In support of this paradigm, a decrease in the expression of cell identity genes was initially observed in the aging human brain[33–35] and pancreas[36], and then confirmed in human skin[37] and multiple mouse tissues[38–40]. These results, along with the previously discussed roles of gene expression aging, support a view of the transcriptome as a key cellular feature that can serve as a proxy for cell function in aging, making it an optimal target for assay development.

Aging clocks

Aging clocks are assays that use molecular features to predict the age of a biological sample. While most of these clocks use algorithms trained to predict the chronological age (days since birth), some of them manage to assess the biological age, which is a biomarker that attempts to quantify the aging phenotype of a given biological system.

The first epigenetic age predictor (average accuracy of 5.2 years) was developed by the Vilain lab in 2011 and used cells from saliva samples to assess the average methylation status of only 2 CpG sites associated with the Edaradd and NPTX2 genes[41]. This proof of concept paved the way for the development of several other first generation epigenetic clocks using methylation array data from different tissues[42–45]. Interestingly, while the clocks were directly trained on the chronological age of the sample donors, they managed to learn features of biological aging and were able to predict all-cause mortality [46–49]. To further improve this biological age prediction, second generation clocks[50,51] incorporated phenotypic age measurements such as plasma proteins and blood cell counts and achieved more accurate predictions of all-cause mortality and lifespan. However, while most DNA methylation clocks can accurately predict chronological and even biological age, their role in aging is still being investigated. Current theories posit that the CpG sites used in these clocks have no causal relationship with aging and their methylation levels represent molecular footprints of the aging processes[52]. Moreover, besides the lack of causality,

some CpG sites also have low interpretability due to limited information about the role of intergenic DNA methylation. As such, due to the unknown role and function of most of the predictive CpG sites, it is difficult to use epigenetic clocks as discovery tools for age-modulating interventions.

A more interpretable layer of gene expression that has also been successfully used as a biomarker of aging is the cellular transcriptome. Unlike DNA methylation levels, transcript counts can integrate multiple layers of aging biology such as histone abundance and modification status, as well as chromatin organization and homeostasis, and even proteome diversity through alternative splicing. One of the first transcriptomic aging clocks was developed by Peters et al. using microarray gene expression data from whole blood, and it achieved an average absolute error of 7.8 years[53]. Using similar data, Mamoshina et al. trained a human muscle tissue transcriptomic clock with an mean absolute error of 6.2 years[54]. With the advent of next generation sequencing, more recent transcriptomic clocks have used RNA-Seq data to build accurate age predictors for human T-cells[55] and skin[56], as well as mouse lung[57] and worms[58]. More importantly, these transcriptomic clocks can also report on features of biological aging, such as premature aging[56–58] and longevity interventions[58].

Nevertheless, while aging clocks can accurately predict the chronological and even biological age of multiple cell types, there are still several limitations that hinder their use in the discovery of novel age reversal interventions. On the one side, the widely-used DNA methylation clocks lack throughput, are cost ineffective, and have limited interpretability. On the other side, while RNA clocks have low cost and high throughput, they sometimes use abstract representations of the transcriptome, such as principal components or deep neural networks, which have similar limited interpretability. More importantly though, RNA clocks suffer from overtraining, whereby the algorithm learns to distinguish between the different chronological ages in the initial dataset,

but performs considerably worse on datasets from different sequencing platforms. This problem is further exacerbated by lack of standardization between different RNA-Seq platforms. Therefore, there is an unmet need in the aging field for accurate, high-throughput, interpretable clocks that can serve as integrative aging assays for the discovery of effective age reversal interventions.

Age reversal

In this dissertation, I refer to age reversal as the process through which a subset or all of the features of an aged biological system are changed, irrespective of the path or trajectory, to resemble their youthful counterparts, thereby reverting the entire system to a more functional state. Unlike slowing down the aging process, which can delay the age-related functional decline, the goal of age reversal is to recapitulate a previous biological state representative of improved functionality.

While aging is a highly complex unsolved problem that spans multiple areas of biology, decades of research have shown that a complete understanding of eukaryotic biology is not fully necessary for identifying interventions that can modulate the aging process. The first study to identify such a genetic perturbation is that of Kenyon et al., who showed that a single mutation in the *daf-2* gene can increase the lifespan of worms by more than 50%[59]. This study showed that aging was indeed malleable and under the control of certain genetic programs, thereby inspiring more people to research aging. Since then, there have been countless studies showing the key roles of the insulin/IGF-1, mTOR, AMPK, and SIR2 pathways in longevity and lifespan[60,61]. However, until recently, it has been unclear whether aging is just malleable, or actually reversible.

Conceptually, the act of reversing the functional decline brought by the aging process is possible and actually takes place with each new life cycle. Through reproduction, a biologically aged germ cell, upon fertilization, gives rise to a young organism, thereby achieving age reversal at the species level. Rando et al. used this observation to argue that given the right signals a similar

age reversal could be achieved at the cellular level, through a reprogramming of the epigenome to a youthful state[62]. For example, Gurdon et al. showed that, through somatic cell nuclear transfer (SCNT), adult differentiated nuclei can be transferred into enucleated oocytes and give rise to healthy frogs[63]. Similar to the fertilization process, the cytoplasm of the egg reprograms the aged nuclei to reverse any accumulated age-related changes, as was later shown in the first cloned animal, Dolly the sheep[64].

More recently, following a similar development reversal path, the induction of the four Yamanaka factors (Oct4, Sox2, Klf4, and c-Myc, or OSKM) was shown to not only reprogram an aged somatic cell to an induced pluripotent stem cell (iPSC) state[65], but also reverse several features of aging such as cellular senescence, telomere size, oxidative stress, mitochondrial metabolism, and transcriptomic and epigenetic profiles[45,66,67]. Moreover, several recent studies have shown beneficial rejuvenating effects through the partial induction of the Yamanaka factors *in vivo*[68–74].

To explain the relationship between aging and development, inherent to all discovered age reversal paradigms, Vadim Gladyshev proposed a "ground zero" of biological aging, whereby the early embryonic stages are associated with a universal rejuvenation process[75]. This theory was recently tested and supported by applying multiple epigenetic clocks to mouse and human prenatal development, which revealed a rejuvenation period during early embryogenesis followed by the start of the aging process[76].

It is thus evident that age reversal is possible and has been achieved to varying degrees at both the species and cellular level, but it still remains elusive *in vivo.* The problem with fertilization, SCNT, and cellular reprogramming as approaches to age-reversal, is that they pass through a developmental stage. This transition is either prohibitive at the organism level, or possesses a great teratogenic risk due to the loss of cell identity and acquisition of stem-like potential[69,77–79]. However, it has been proposed that it is possible to achieve age reversal without dedifferentiation[62,80,81] and

a recent study showed there are several sets of factors that can perform cellular rejuvenation with varying degrees of cell identity loss[82]. Therefore, a promising future direction for the aging field is the search for safe and effective interventions that can ultimately induce age reversal *in vivo*.

Cell engineering

The Yamanaka factors were discovered more than 16 years ago through a genetic screen aimed at engineering somatic cells to a pluripotent state[65]. Ever since, there haven't been any new factors discovered that can robustly reverse the age of cells *in vitro*. While genetic screens are powerful tools for the discovery of new biology, their success is predicated on the robustness of the experimental system, which is lacking in a complex problem such as mammalian aging.

Unbiased whole-genome screens are optimal for increasing the likelihood of identifying genes that can modulate the aging process. However, the limitations in culturing primary human cells, the high cost and lack of robustness of current aging assays, and the large search space for gene combinations reduce the feasibility of whole-genome aging screens and call for a targeted approach. With the increase in publicly available genomic and epigenomic resources[83,84], predictive algorithms have been developed for identifying targets for cell type conversion[85,86] and age-related diseases[87]. Nevertheless, these tools are limited to specific applications and sometimes require multiple types of data that are not always available, hindering their use for applications such as anti-aging screens.

The perturbation system is another aspect of genetic screens where innovation can further enable the usage of screening technologies in aging research. Over the last few years, the advent of genome-editing has provided tools for large-scale perturbation experiments such as CRISPR[88], CRISPRa[89], CRISPRi[90], and cDNA overexpression libraries[91]. Systematic comparisons of these

systems in relevant cell lines accompanied by improved scalable molecular cloning methods would greatly facilitate the use of these tools.

Lastly, another major hurdle to the implementation of genetic screens in aging research is a lack of integration of these diverse computational and experimental tools into one unified workflow for the generation of desired cellular phenotypes. The development of an integrated pipeline for identifying genetic factors that can reverse the age of mammalian cells would greatly accelerate the progress in the development of anti-aging therapies and likely provide mechanistic insight into the aging process.

Transcriptomic reprogramming

In this dissertation, I propose transcriptomic reprogramming, defined as the transition between two defined transcriptomic states, as a promising method for age reversal. The central hypothesis of this thesis is that there exist certain genetic perturbations that can reprogram a specific gene regulatory network to a more youthful state and thereby rescue the function of the associated aged biological system (Figure 1.3).

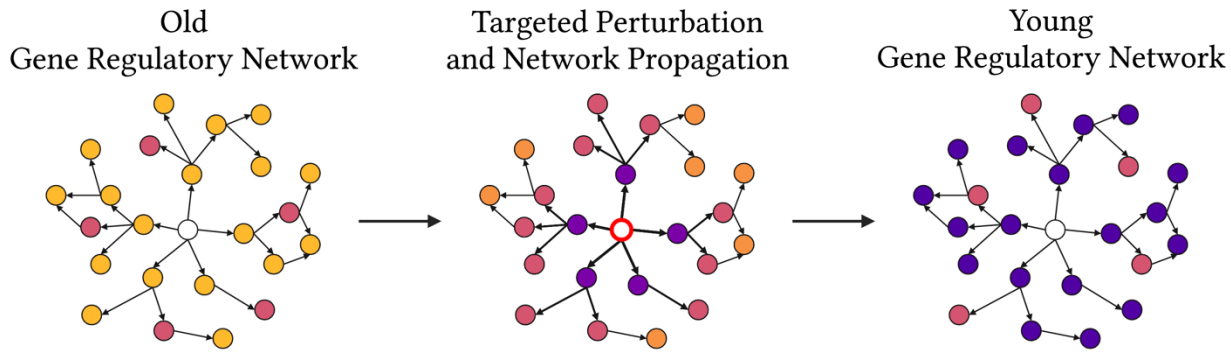


Figure 1.3: Schematic of transcriptomic reprogramming paradigm.

Figure 1.3 Continued

An old gene regulatory network, where the gene expression of different members is dysregulated (yellow color), can be reprogrammed to a younger state (purple color) through the targeted perturbation of specific nodes (white color) and the subsequent network propagation of these rejuvenating effects.

The permanence of a reprogrammed state is determined by the strength and duration of the initial perturbations, whose effect on key epigenetic regulators can in turn induce stable alterations in the chromatin landscape. While cellular reprogramming refers specifically to the dedifferentiation of somatic cells to a pluripotent state through the use of the Yamanaka factors, transcriptomic reprogramming is applicable to any somatic or stem cell whose gene regulatory network is shifted towards a new steady state.

To identify these specific genetic perturbations that can induce transcriptomic reprogramming to youthful states, I propose a three-step workflow inspired by the landmark transcriptomic aging screen performed by Janssens et al.[92]. First, transcriptomic data from differently aged primary cells are used to develop an aging assay and identify promising genetic targets for modulating the aging process. Second, these candidate genes are perturbed *in vitro* using high throughput gene modulating tools, and their effect on the aging phenotype is assayed using transcriptomic clocks. Lastly, the most promising hits from the screen are validated for their rejuvenation potential using cell type specific functional assays *in vitro* and *in vivo*.

SUMMARY AND GOALS FOR THIS DISSERTATION

Aging is an important societal and economic problem that is difficult to understand and tackle due to its complex nature, which involves multiple areas of eukaryotic biology[1–10]. While the definition of aging is still being debated in the field, there is a broader consensus that the process is

defined by a progressive functional decline, which is inextricably linked to the cellular transcriptome[12,16,17,19–28]. As such, accurate and robust aging assays have been developed that measure the fidelity of gene expression at the epigenetic[42,45,51] and transcriptomic level[53–58]. More importantly, it was shown that aging is a malleable process[59–62] and that the reversion of the epigenome and transcriptome to youthful states can improve the function of various biological systems[45,66–74]. However, none of these interventions have yet achieved age reversal at the organismal level, and some studies have even highlighted safety concerns with using teratogenic factors[77–79]. Informed by the recent aging screening studies[82,92], I believe a promising path towards the discovery of safe and effective age reversal interventions is through transcriptomic reprogramming screens.

Thus, here I aim to further enable the use of genetic screens in aging, provide proof of concept for transcriptomic reprogramming screens as a promising aging target discovery method, and identify new interventions that can rejuvenate aged human primary cells. Toward this goal, we first present an integrated pipeline for mammalian genetic screens by developing unique algorithms for target identification, a new efficient molecular cloning tool, and combining them with existing genetic perturbation methods and next generation sequencing (NGS) readouts to enable versatile cell engineering[93,94]. Next, we develop a robust and accurate NGS readout for human cell aging, by using the transcriptome to integrate several cellular processes into a proxy metric for cell function. We use the proposed screening pipeline to perform the first large-scale age reversal screen in human fibroblasts and provide proof of concept for the use of transcriptomic reprogramming in identifying new rejuvenating interventions. Lastly, we discover the overexpression of arginine rich splicing factor 1 (SRSF1) as a robust perturbation for transcriptomic reprogramming to a more youthful and functional state. We anticipate future genetic studies to improve on all aspects of our work and use transcriptomic reprogramming

screens for the discovery and development of other safe and effective perturbations that can reverse the age of cells both *in vitro* and *in vivo*.

CHAPTER II : AN INTEGRATED PIPELINE FOR MAMMALIAN GENETIC SCREENING

AUTHORS AND CONTRIBUTIONS

This chapter is published as a research article: Christian Kramme, Alexandru M. Plesa*, Helen H. Wang, Bennett Wolf, Merrick Pierson Smela, Xiaoge Guo, Richie E Kohman, Pranam Chatterjee, George M Church. An integrated pipeline for mammalian genetic screening. Cell Reports Methods 100082 (2021).*

Author contributions: C.K. designed experiments, invented and optimized MegaGate cloning, built vectors, designed and conducted CRISPRa/I screening and hiPSCs experiments, developed copy number assay, and conducted NGS library preparation and data analysis. A.M.P. designed experiments, invented and tested MegaGate cloning, developed the DEG network scoring method, conducted NGS library preparation and data analysis, NHDF culturing, and developed the copy number assay and TAR-seq method. H.H.W. developed the DEG network scoring method and conducted RNA-seq data analysis. B.W. assisted with MegaGate and transcriptome analysis. M.P.S. conducted copy number and FACS analysis. X.G. assisted with CRISPR screening protocols. P.C. conceived, designed, and implemented machine learning and bioinformatics protocols for target identification and sgRNA selection. C.K., A.M.P., and P.C. wrote the paper, with input from all authors. R.E.K. and G.M.C. acquired funding. P.C., R.E.K., and G.M.C. supervised the project.

ABSTRACT

With the recent advancements in genome editing, next-generation sequencing (NGS), and scalable cloning techniques, scientists can now conduct genetic screens at unprecedented levels of scale and precision. With such a multitude of technologies, there is a need for a simple yet comprehensive pipeline to enable systematic mammalian genetic screening. In this study, we develop unique algorithms for target identification and a toxin-less Gateway cloning tool, termed MegaGate, for library cloning which, when combined with existing genetic perturbation methods

and NGS-coupled readouts, enable versatile engineering of relevant mammalian cell lines. Our integrated pipeline for sequencing-based target ascertainment and modular perturbation screening (STAMPScreen) can thus be utilized for a host of cell state engineering applications.

INTRODUCTION

Recent technological advancements in large-scale cloning, genome editing, and next-generation sequencing (NGS) have enabled scientists to perform genetic screens at unprecedented levels of scale and precision. Previous studies have leveraged these tools for the interrogation of fundamental biological processes that underpin mammalian development, physiology, and pathology. Studies such as genome-wide CRISPR screens for identifying chemotherapeutic drug targets[95], single-cell profiling of the entire human embryo developmental trajectory[96,97], and characterization of the effect of transcription factor (TF) overexpression in the context of human induced pluripotent stem cell (hiPSC) differentiation for nearly all human TFs[98] demonstrate the value of pairing effective gene engineering tools with NGS-coupled readouts. However, despite these technological advances, there is not yet a straightforward, unified pipeline for gene target identification, perturbation tool selection, library cloning, and readout assessment, which can effectively serve as an end-to-end workflow for mammalian genetic screening.

As publicly available gene expression and chromatin accessibility datasets have drastically increased in quality and abundance, new in silico tools have become more effective at predicting screening targets[85,86,99]. However, many of these tools require multiple types of data that might not be available, and they are mostly limited to cell differentiation projects. To expand the scope and the feasibility of computational target prediction, versatile tools are required that can be applied to

any cell engineering screen and can effectively analyze different types of data to accurately infer gene targets.

With the advent of genome-editing tools and the major decrease in the cost of NGS, there now exist readily available tools for the large-scale perturbation of mammalian genomes, such as CRISPR single-guide RNA (sgRNA) libraries and the cDNA hORFeome, which enable researchers to target nearly any human gene for functional characterization[100,101]. However, the recent development of robust gene perturbation technologies has created a need for systematic studies that evaluate and compare the performance of these tools in relevant human cell lines, to allow researchers to select the appropriate tool for their specific application. In addition, common genetic tools, such as cDNA overexpression have so far been difficult to apply at scale due to hurdles in molecular cloning, thereby limiting their utility.

Considering the technological and scientific advancements in each of these areas of biology and their inherent limitations and applicability to different biological systems, there is a need for a salient, efficient pipeline that provides an integrated solution for systematic mammalian genetic screening studies. This article seeks to directly address this gap in the literature through the development of target identification and library cloning tools, which can be combined with NGS-coupled readouts for cellular engineering of relevant mammalian cells. Using these tools, we developed an all-in-one integrated pipeline for sequencing-based target ascertainment and modular perturbation screening (STAMPScreen) (Figure 2.1).

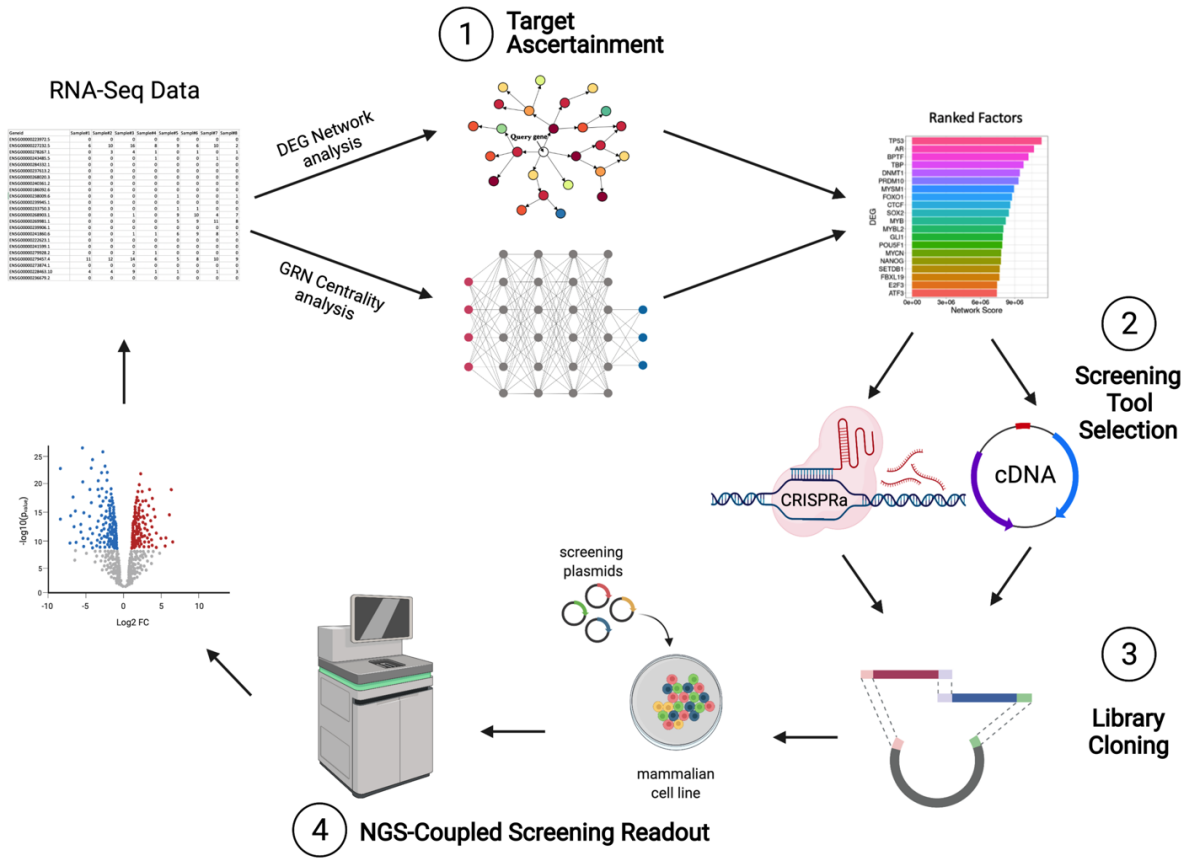


Figure 2.1: STAMPScreen schematic workflow.

Schematic representation of the STAMPScreen pipeline, highlighting *in silico* target ascertainment, screening tool selection, library cloning, and NGS-coupled screening readout. STAMPScreen generates data that feed into iterative cycles of the workflow.

RESULTS

In silico methods for gene target identification

The rapid improvement of NGS technologies has led to a decrease in the cost of sequencing and a subsequent increase in the amount of biological data generated from different cell types under various conditions. These data serve as a powerful resource for hypothesis-generating studies, especially for projects aimed at cell engineering. However, a major bottleneck in this area

of biology has been the lack of accurate tools that can analyze different types of large datasets and extract accurate information for cell engineering target identification. Here, we present two different methods for predicting phenotype conversion perturbations from bulk and single-cell RNA sequencing (RNA-seq) data.

Differential gene expression analysis (DGEA) has been the most widely used method for identifying statistically significant transcriptomic changes between treatment samples and controls[102,103]. While DGEA has helped identify causal genes for various transcriptomic changes and performs adequately in a well-controlled study with a defined intervention, the method prioritizes factors overexpressed at specific stages, thus inferring correlation rather than causality, and it loses its predictive power when applied to more drastic gene expression perturbations. More recently, gene regulatory network (GRN)-based approaches, such as CellOracle and IRENE, have been developed to identify factors that increase cell conversion efficiencies by integrating RNA-seq and chromatin accessibility single-cell data[86,104]. However, these methods require access to both transcriptomic and epigenomic single-cell datasets that may not be readily available for cell types of interest. In addition, network analyses of gene essentiality have been conducted, but to this point have been limited to CRISPR screening data[105]. To address the shortcomings of these frameworks, we present two complementary target ascertainment methods that use genetic and protein networks to accurately infer targets for cell engineering screens.

DEG network analysis

Using a similar approach to Rackham et al.[85], we developed a differentially expressed gene (DEG) network scoring method utilizing transcriptomic data from a starting cell state and a target cell state. In our method, we first performed a DGEA to determine significant gene expression changes, and then we generated a DEG score for each gene by combining the traditional DEG

metrics (fold-change, p value) with cell phenotype information (correlation with desired phenotype). To infer phenotype causality as well as identify DEGs with small changes but potential large effects, we added a layer of protein network connectivity to our DEG scoring. Using the STRING interaction database[106], we traversed each DEG's protein network across three layers and calculated a score that combines its DEG score with the degree of connectivity. The resulting list preferentially ranks DEGs with large significant changes between the two cell states that are also highly connected to other highly differentially expressed DEGs (Figure 2.2 A). We validated our method using RNA-seq data from differently aged human primary fibroblasts[56] to identify causal aging genes. Our analysis shows that the DEG network scoring method generates a ranked list of targets that has a significantly high enrichment of known experimentally validated aging genes[107], unlike plain DGEA or ranking based on common metrics (Figure 2.2 B).

GRN centrality analysis

While our DEG network analysis approach addresses the shortcomings of traditional DGEA, it is inherently dependent on the availability of protein interaction data and not highly sensitive to small intermediary transcriptomic changes across time-series data. To provide an alternative tool that overcomes these limitations, we sought to develop a unique algorithm by combining time-series transcriptomic data with graph theory-based centrality analysis. To do this, we utilized stochastic gradient boosting machines to train GRNs and calculated the PageRank of each genetic factor post network construction and graph pruning[108,109]. Our simple pipeline requires a normalized fold-change representation for each gene at different stages of the cell state conversion and generates a graphical representation of ranked factors with the highest global importance (Figure 2.2 C). We validated our pipeline using existing RNA-seq datasets of neuronal stem cell, myoblast, and melanocyte differentiation[110–112]. Our results demonstrate that the

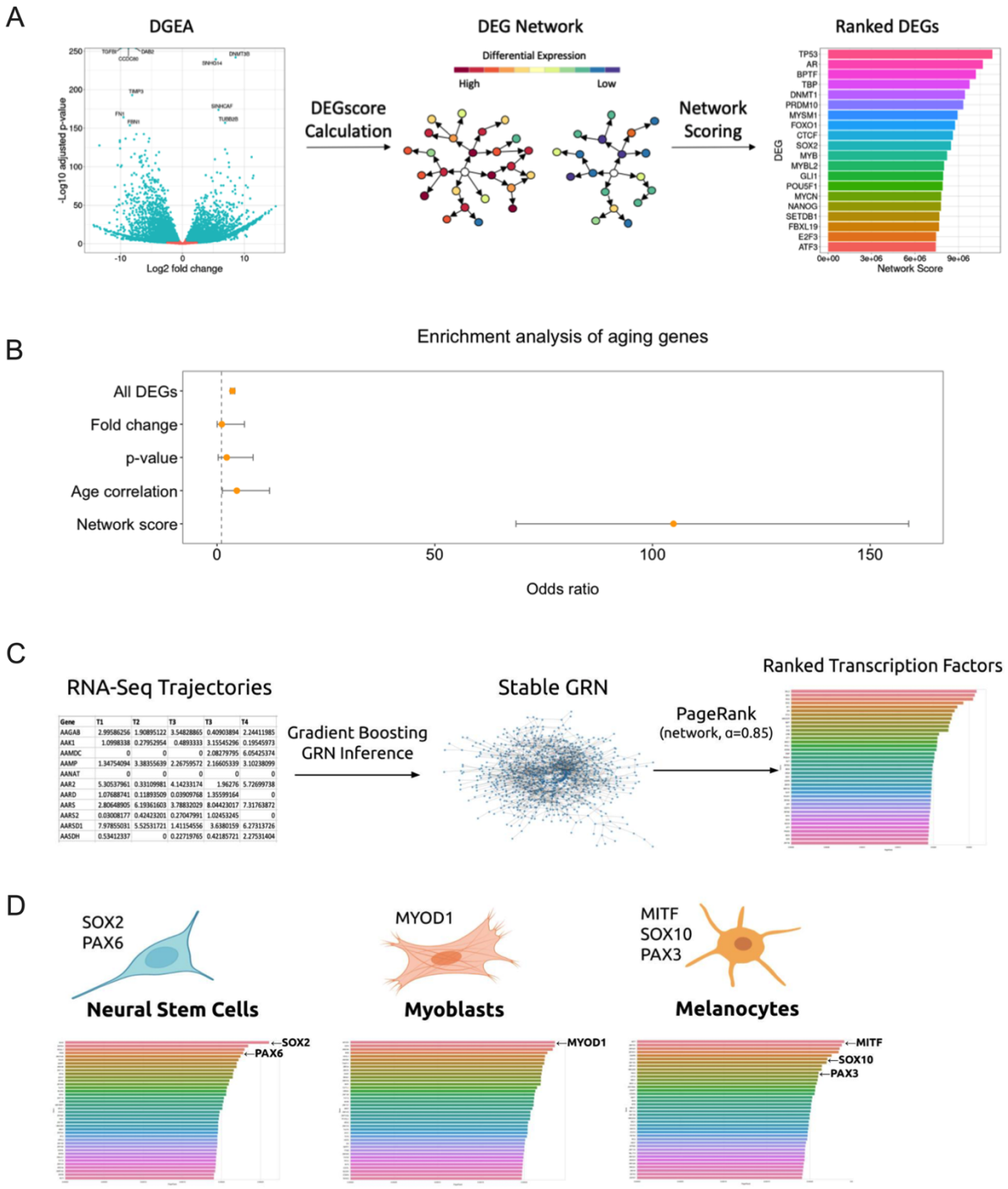


Figure 2.2: *In silico* target ascertainment.

(A) DEG network target identification pipeline. DGEA was used to determine significant changes between the starting and target cell state. Using publicly available protein interaction networks, DEGs are scored based on their connectivity and differential expression levels.

Figure 2.2 continued

(B) Validation of the DEG network analysis method applied to the fibroblast aging phenotype. The list of all DEGs and top 100 ranked lists based on common metrics and our network score were tested for enrichment of known aging genes from the GenAge database[107].
(C) Graph theory-based TF discovery pipeline. Gene regulatory networks are inferred from time-series RNA sequencing data. PageRank with a standard residual probability of 0.85 was utilized to rank TFs by centrality score.
(D) Prediction of central TFs in known differentiation protocols using graph theory-based TF discovery pipeline. Experimentally validated TFs are indicated, demonstrating predictive capability of the pipeline.

algorithm can effectively identify known experimentally validated causal regulators within the predicted top factors (Figure 2.2 D). Compared with traditional DEG approaches, these two methods generate candidate gene lists enriched for known genes related to the studied phenotype, showcasing the accuracy of our methods when applied to complex phenotypes.

Modular perturbation tools

After ascertaining the targets for a desired cell engineering screen using the previously presented methods, the STAMPScreen workflow requires the choice of a perturbation tool. Gene perturbation in mammalian cells can be performed either at the endogenous locus of a specific target via effector proteins that activate transcription, such as CRISPR-Cas[113], or exogenously through cDNA expression via inducible or constitutive promoters[114]. Both CRISPR and cDNA technologies have been leveraged in recent mammalian screens to effectively target genes of interest in an individual or combinatorial manner[89,90,112,113,115–124].

To generate an integrated pipeline for gene perturbation, we evaluated existing gene perturbation tools and developed powerful technologies for cDNA and CRISPR screening. We performed a rigorous evaluation in hiPSCs of the relative performance of three popular CRISPRa constructs, two CRISPRi tools, and five cDNA overexpression vectors. Finally, we generated

unique tools for dual cDNA and CRISPR screening, which enable researchers to overexpress genes via cDNA while simultaneously expressing sgRNAs to target native genes for suppression or induction.

CRISPRa performance in hiPSCs

Recent comprehensive comparisons of the three most popular CRISPRa tools[113,119,125] (synergistic activation mediator [SAM], VP64-based SunTag, and VP64, p65, and RtTa [VPR]) have been performed in a variety of cell types, but no study has compared the three activators in hiPSCs. We therefore evaluated the performance of the three CRISPRa tools for overexpressing 47 target genes in the hiPSC line PGP1. On average, dCas9-VPR induced stronger expression across the tested genes, when compared with dCas9-SAM and dCas9-SunTag, while SAM showed modest levels of improvement over SunTag (Figure 2.3 A). For all three activators, genes with high basal expression were more difficult to overexpress using any of the three CRISPRa technologies, in line with previous studies (Figure A2.1A)[89]. Based on these results, we conclude that, for CRISPRa screens in hiPSCs, dCas9-VPR is the optimal tool as compared with dCas9-SAM and dCas9-SunTag.

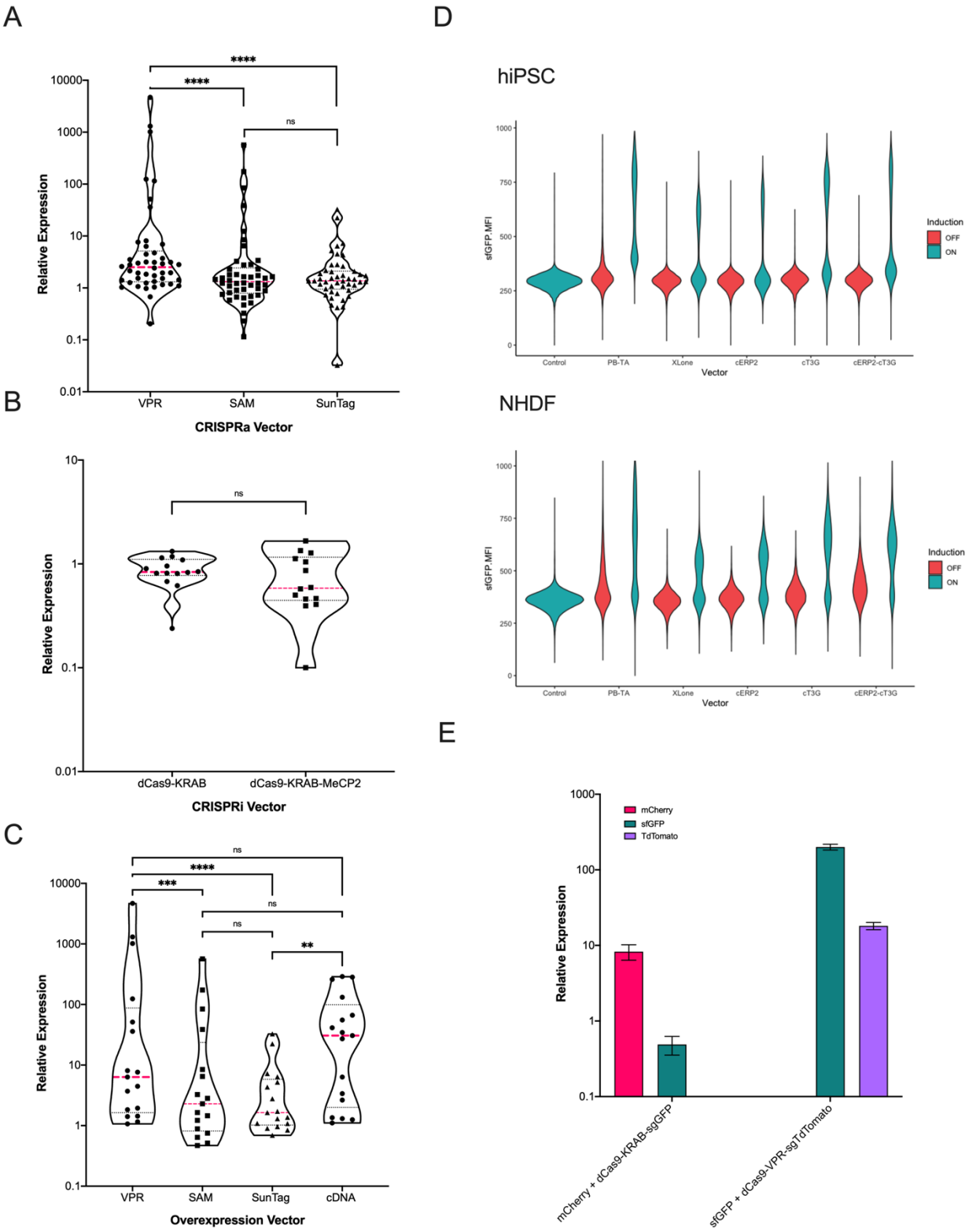
CRISPRi performance in hiPSCs

We additionally evaluated the performance of two highly utilized CRISPRi tools, dCas9-KRAB and dCas9-KRAB-MeCP2, against 12 gene targets in hiPSCs[90]. Our results demonstrate that dCas9-KRAB-MeCP2 induces higher gene knockdown compared with dCas9-KRAB, in line with previous findings (Figure 2.3 B). We also observe that the level of repression was locus dependent, with a knockdown of greater than 95% at certain loci while others showed minimal

Figure 2.3: Gene perturbation tool evaluation.

Systematic comparison of CRISPRa, CRISPRi, and cDNA function were performed in the hiPSC line PGP1. Analysis was performed using qRT-PCR with duplicates and the ΔΔCq method was utilized to determine relative expression to a no plasmid control duplicate. For CRISPRa and cDNA (A and C), cells were harvested 48 h post-transfection; for CRISPRi (B), cells were harvested at 72 h post-transfection. Induction using dCas9-VPR, SAM, and SunTag was performed on 47 gene targets in hiPSCs, cDNA induction was performed on 17 genes targets, and CRISPRi repression was performed on 12 gene targets. Significance by Mann-Whitney test. The pink line shows median values, dashed black lines are 95% CI. For systematic comparison of cDNA induction using the PB-TA-ERP2, XLone, cT3G, cERP2, and cERP2-cT3G vectors expressing sfGFP (D). Cells were harvested for flow cytometry after 24 h of doxycycline induction and screens were performed in duplicate with and without doxycycline. Geometric mean fluorescent intensity of all live singlet cells was plotted for each condition. For demonstration of dual expression vectors (E) a constitutive mCherry cDNA was expressed along with a sgRNA targeting a Tet promoter. This vector was co-nucleofected with dCas9-KRAB-MeCP2 into an hiPSC line harboring a Tet-sfGFP under induction. Cells were harvested 96 h later for flow cytometry. Relative expression was calculated as MFI compared with no plasmid control. The right panel shows a similar setup but with a sfGFP cDNA vector under Tet promoter and an sgRNA targeting an integrated tdTomato construct. The plasmid was co-nucleofected with dCas9-VPR and cells were harvested 48 h later for flow cytometry, relative expression was calculated as MFI compared with no plasmid control. Calculated p values are represented as follows: ***$p \leq 0.001$, ****$p \leq 0.0001$.

Figure 2.3 Continued



change in transcript abundance. The source of this resistance to suppression could be from poorly functioning sgRNAs, as our study screened only one sgRNA per target. In addition, local chromatin architecture, interference from native transcriptional regulators, or insufficient waiting period between knockdown and observation may also play a role[90]. In addition, it may be possible that genes that lie near the target are likewise affected in their expression by KRAB silencing. Previous studies have looked at this possible leakage with no definitive conclusions but, based on their data, it seems that this is uncommon but warrants further research[90]. Based on our data, we conclude that dCas9-KRAB-MeCP2 is the optimal CRISPRi tool for gene suppression in hiPSCs.

cDNA overexpression performance in hiPSCs and primary human fibroblasts

To compare the expression dynamics of CRISPRa and cDNA overexpression, we tested constitutive expression of 17 open reading frames (ORFs) in hiPSCs that we previously targeted using CRISPRa. Overall, cDNA overexpression demonstrated higher induction levels compared with CRISPRa, but similar to CRISPRa showed minimal overexpression for genes with higher basal expression (Figures 2.3 C and A2.1 B). In addition, we compared two common doxycycline-inducible plasmids, XLone[126] and PB-TA-ERP2[127], and found that XLone-based expression had low background (signal in the absence of doxycycline) and low inducibility (signal in the presence of doxycycline), while PB-TA-EPR2 showed high background and high inducibility in both hiPSCs and primary normal human fibroblasts (NHDFs) (Figures 2.3 D and A2.2 D). To combine the desirable features of both vectors, we used the XLone plasmid as a base vector and replaced the blasticidin selection with puromycin and additionally constructed three modified plasmids utilizing components of PB-TA-ERP2. We cloned a minimal cytomegalovirus (CMV) promoter downstream of the Tet promoter (cT3G), a CMV enhancer upstream of the rtTA-3G (cERP2), and both together (cERP2-cT3G). We then tested the performance of the three new vectors in

hiPSCs and NHDFs and show that the CMV promoter enhances induction in both fibroblasts and hiPSCs and that the CMV enhancer additionally increases induction in hiPSCs (Figures 2.3 D and A2.2 C). The cT3G and cERP2-cT3G vectors both exhibit robust inducibility to a broad range of doxycycline concentrations, and demonstrate stable, increasing expression for over a week and negligible silencing after a month (Figures A2.2 A and A2.2 B). We additionally cloned IRES2 fluorescent constructs mTagBFP2, mCherry, and mNeonGreen into cT3G and cERP2-cT3G and observed robust expression of both the primary ORF and fluorescent construct (Figure A2.2 C). For many genetic studies, high levels of overexpression of targets may not be beneficial to address the underlying question as this may lead to spurious, non-biologically relevant results. However, we believe that providing a tunable vector with a broad range of induction levels allows the user to best customize their expression to their desired outcome. Where extensive levels of induction are needed, such as in synthetic differentiation screens or polycistronic cassettes like IRES-coupled fluorescent expression, we believe these vectors perform optimally. In addition, the tunable nature of our vectors demonstrates strong doxycycline response curves that have linear ranges of response that can be utilized for precise dosage control in studies that require it. Based on these findings, we conclude that cDNA overexpression produces similar or higher transcript levels than CRISPRa in hiPSCs. We further develop robust doxycycline-inducible tools for tunable gene expression in hiPSCs and NHDFs with high induction efficiency and low background expression.

Dual gene perturbation via cDNA overexpression and CRISPR

Recent studies have shown that CRISPRa and CRISPRi can function as potent combinatorial gene perturbation tools, capable of target overexpression, repression, or both within a single cell[124,128,129]. To further expand the mammalian gene perturbation toolkit, we aimed to combine cDNA overexpression with the CRISPRa/i system in the same vector. We designed

plasmids that express both a cDNA construct from an inducible or constitutive promoter and an sgRNA from the human U6 promoter. We transfected the aforementioned plasmids into a hiPSC cell line harboring an integrated dCas9-VPR or dCas9-KRAB-MeCP2, and dual gene induction and dual overexpression and repression was achieved, showing 10- to 200-fold cDNA induction, 10-fold CRISPRa induction, and 50% CRISPRi repression (Figure 2.3 E). Based on our findings, our dual cDNA-CRISPR perturbation tool will provide a valuable resource for screening studies that aim to perform both exogenous and endogenous perturbations, such as cDNA screens paired with genome-wide CRISPRi libraries for pathway analysis.

Library cloning

Having decided on a gene target list and perturbation tool, the next step in our proposed STAMPScreen workflow consists of generating a viable screening library. For CRISPR studies, sgRNA cloning through Golden Gate reactions are efficient and scalable, having enabled many genome-wide CRISPR screens[130,131]. However, molecular cloning of cDNA, particularly at library scale into barcoded vectors, has proved to be cumbersome and difficult with existing technologies such as Golden Gate, Gibson Assembly, or Gateway cloning[132–134]. We therefore developed a powerful method for cDNA cloning, termed MegaGate, which combines the lack of sequence bias found in Gateway cloning with the combinatorial capacity of restriction cloning.

MegaGate—A unique toxin-less recombination cloning tool

MegaGate utilizes the lambda phage recombination technology of Gateway cloning to shuttle a DNA sequence of interest into a desired destination vector. Uniquely, our method eliminates the need for a ccdB toxin cassette by replacing it entirely with restriction enzyme recognition sites (Figure 2.4 A). Through recombinase insertion of the cDNA construct and

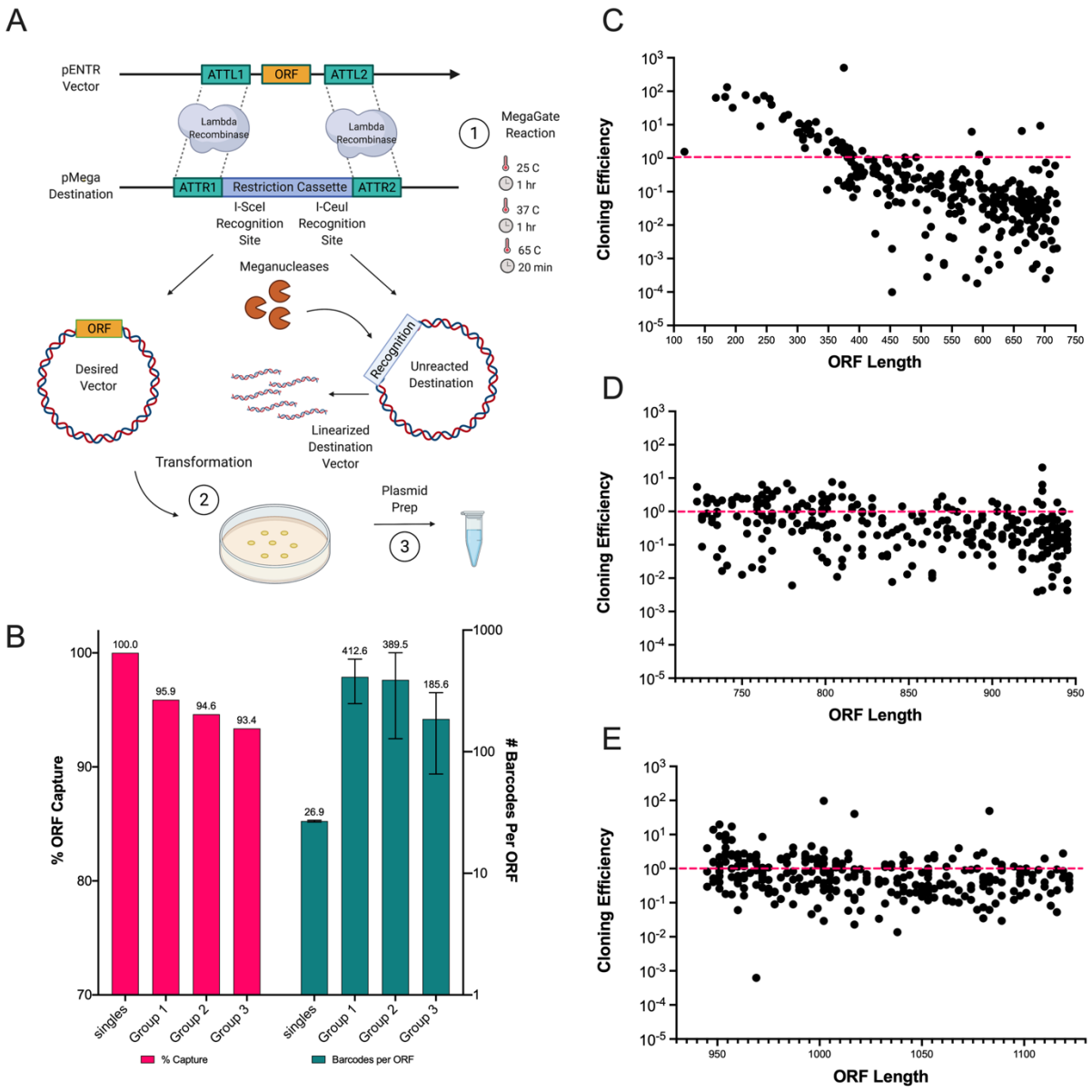


Figure 2.4: MegaGate, a toxin-less cDNA cloning method.

(A and B) (A) Schematic representation of the MegaGate cloning reaction. (B) Percent ORF capture (pink) measured as the ratio of input genes captured in expression vectors in a single MegaGate cloning reaction for single genes and pooled groups. Number of barcodes captured per gene (teal) for the single genes and pools was determined via NGS alignment of destination vector amplicons.

(C–E) Cloning efficiency as a function of ORF length. Cloning efficiency was measured as the relative abundance of the gene in the expression vector pool divided by the relative abundance of the gene in the pDONOR pool, as measured by NGS counts. Genes are arrayed by length on the x axis.

linearization of the unreacted destination vectors, only the desired vectors propagate in the bacterial transformant pool (Figure 2.4 A). We identified an optimized set of reaction conditions that regularly demonstrates 99.8% or greater cloning efficiencies and high positive colony plating (Figure A2.3). Moreover, the utilization of the Gateway att sites makes MegaGate compatible with all existing pENTR ORF libraries such as the hORFeome and TFome, enabling it to be readily usable in many labs.

The MegaGate reaction is highly modular, amenable to volume scaling, isothermal reactions, all-in-one or multi-step reaction mixtures, differential enzyme recognition sites, and highly variable insert sizes (Figure A2.3). In addition, the utilization of meganucleases, which have highly rare and specific recognition sequences, allows the use of a single MegaGate vector for cloning almost any known coding sequence without further plasmid or gene modification. Furthermore, we showed that other restriction enzymes, such as those in the type IIS family, can also be used in place of meganucleases, making the cloning strategy highly modular for use in a variety of cloning applications (Figure A2.3). In addition, we highlight MegaGate's compatibility with the BP Clonase enzyme mix to clone pENTR ORFs from cDNA PCR inserts (Figure A2.3). Based on our findings, MegaGate is a simple and efficient cDNA cloning tool for the generation of ORF libraries using pENTR libraries and non-barcoded or barcoded MegaGate destination vectors.

Validation of MegaGate in single and pooled cloning

To evaluate the performance of MegaGate single gene cloning, we cloned 200 pENTR genes individually into two different barcoded MegaGate destination vectors, yielding 100% ORF capture and 25–200 barcoded transformants per reaction (Figure A2.4). For testing pooled cloning, we pooled a total of 948 genes from the hORFeome into 3 pools of 316 pENTR vectors normalized

by size. We then utilized MegaGate to clone the pools into a pool of barcoded MegaGate destination vectors. Utilizing amplicon sequencing, we then characterized our initial pENTR pool and our resulting expression vector pools to determine the percentage of our plasmids that were converted into expression vectors and to assign barcodes to each ORF. On average, we captured 95% of input ORFs (900/948) from the 3 pools, each with 1–500 unique DNA barcodes per gene (Figure 2.4 B). In addition, our data show that MegaGate demonstrates a size dependence, with smaller ORFs cloning more efficiently relative to larger ORFs (Figure 2.4 C). When there is little variance in insert length, MegaGate cloning generally maintained the ORF relative abundance, whereby the distribution of ORFs in the expression pool mirrors that in the pENTR pool (Figure 2.4 C). Based on the results, MegaGate is an efficient cDNA cloning tool for generating barcoded ORF libraries, capable of cloning single or pools of 300+ ORFs at high throughput and low cost.

Screening and NGS-coupled readouts

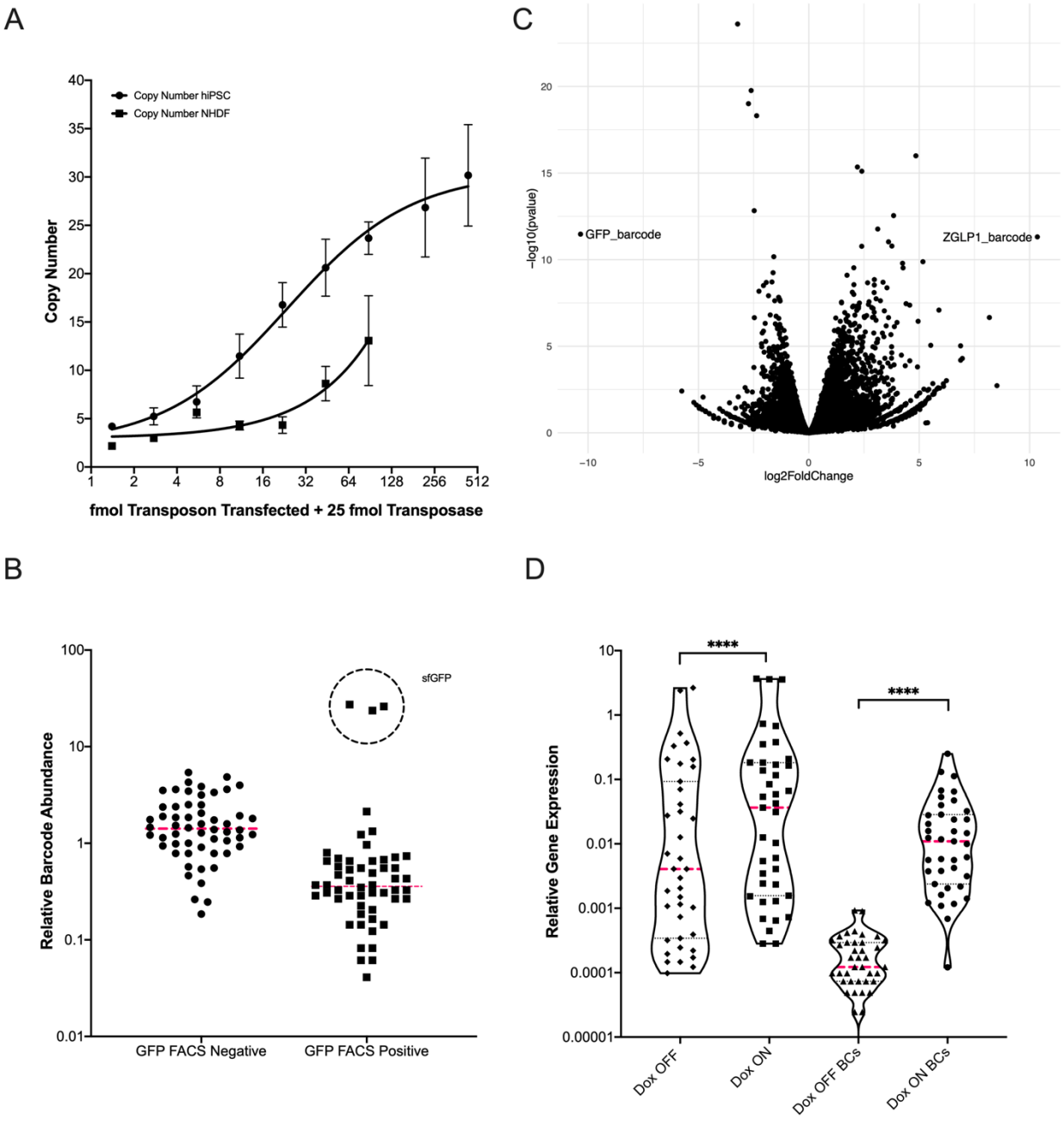
Having ascertained the genetic targets for the phenotype of interest, selected a gene perturbation tool, and generated the screening library for the relevant biological system, our cell conversion workflow next focuses on optimizing the screening pipeline, performing the screen, and analyzing the data. Here, we present current and relevant methods for these final steps in the STAMPScreen protocol.

The two most common integrative gene delivery systems are the lentiviral (LV) transduction and the PiggyBac (PB) transposon system[135,136]. Their respective advantages and disadvantages have been discussed previously[137], but one important consideration when using the PB system for screening purposes is the resulting copy number. While this variable is easily optimized in the LV system through multiplicity of infection titration, the PB system requires cell line-specific optimization, since the copy number will be correlated with the amount of DNA and

Figure 2.5: NGS-coupled readouts.

(A) PiggyBac integration copy number was measured using qRT-PCR on genomic DNA. Integrant-specific primers and a single copy gene RPP30 were used. A $2 \times \Delta Cq$ measurement was used to determine copy number for n = 2 biological replicates. Error bars are standard deviation.
(B) Barcode enrichment analysis (Bar-seq) was performed on hiPSC pools with 54 barcoded gene insertions and barcoded GFP insertions. RNA was harvested after FACS 72 h post-induction, and barcodes were amplified for NGS. Relative abundance was measured as the relative read count of each gene barcode compared with the total read count of all barcodes.
(C) Two barcoded genes, GFP and ZGLP1, were integrated into hiPSCs and induced for 72 h in duplicate. RNA was harvested and RNA-seq was performed. DEseq2 was utilized to calculate log2fc and p values and an exact match k-mer was used to identify the gene barcode.
(D) Forty barcoded genes were integrated into hiPSCs and induced for 3 days and RNA was harvested and converted to cDNA. A primer pool for all 40 genes was used to amplify the 40 targets for NGS as well as their barcodes in the DOX OFF and DOX ON pools. Relative gene expression was calculated as relative read counts normalized to GAPDH for each pool. Calculated p values are represented as follows: ****$p \leq 0.0001$.

Figure 2.5 Continued



its delivery efficiency. Our data suggest that the transgene copy number varies linearly with the ratio of transposon:transposase within a certain range where the transposon is not in saturation or ultra limiting, after which the number of integrations plateaus (Figure 2.5 A). We selected single-cell clones from hiPSC integrant pools and additionally showed the copy number varies within a population following a Poisson distribution (Figure A2.4). We suggest performing transposon titration at a fixed amount of transposase to determine the appropriate ratio for obtaining the desired copy number in a given biological system.

One of the main advantages of using the STAMPScreen pipeline in a given screen is its coupling of phenotype-specific readouts with NGS, enabling high-throughput characterization. If a given cellular phenotype does not have a well-established imaging-based high-throughput assay, targeted RNA sequencing (TAR-Seq) can offer accurate information about a subset of the cellular transcriptome in a scalable manner[138]. For example, a screen for the activation of certain cellular pathways, where the primary goal is the induction of specific transcripts, would be an obvious application. Our data show that, with proper design considerations and reaction optimizations, up to 70 RNA targets, including vector barcodes, can be quantified in a single reaction at precision levels comparable with qPCR (Figure A2.5). This method allows for the interrogation of a limited number of transcripts across hundreds of perturbations in a scalable and cost-effective manner.

When considering pooled screens of barcoded vectors, NGS readout coupling becomes necessary for matching the perturbation to the observed phenotype. Regardless of the screening readout (e.g., flow cytometry, targeted sequencing, scRNA sequencing, combination of flow and RNA sequencing), a barcode enrichment analysis is required for pooled screening approaches. Upon isolating a specific cell population and assigning it a signal value, the associated genetic perturbation can be identified by amplifying and sequencing barcodes from the gDNA or RNA of the cells (Bar-seq)[139]. To demonstrate Bar-seq utilizing our vectors, we FACS isolated barcoded

GFP-expressing cells from a pool of cells harboring 159 barcodes and showed that the GFP-associated barcodes were significantly enriched in the FACS-positive pool (Figure 2.5 B). In addition, we overexpressed two genes in separate hiPSC populations and performed bulk RNA-seq and demonstrated that we could capture both the whole transcriptome of the population as well as the gene-associated barcode for RNA-seq coupled to barcode enrichment analysis (Figure 2.5 C). To demonstrate TAR-seq coupled to barcode enrichment, we overexpressed 40 barcoded genes in hiPSC performed TAR-seq, which showed effective capture of the 40 gene panel and associated barcodes (Figure 2.5 D). We demonstrate that our vectors can be effectively integrated into hiPSCs and read out for coupled barcode enrichment and transcriptomic characterization via targeted RNA-seq, RNA-seq, or gDNA Barcode-seq, thus completing the STAMPScreen integrated workflow.

DISCUSSION

In this study, we present STAMPScreen, a sequencing-based target ascertainment and modular perturbation screening method for enabling cellular engineering. Our proposed workflow consists of four main parts: target identification, perturbation tool selection, library cloning, and cell screening. We showcase two unique computational approaches for identifying candidate genes for a specific phenotypic conversion starting with transcriptomic data of the initial and target cells. Next, we present data systematically comparing different genetic perturbation methods and develop vectors for genetic screening. We then introduce a powerful method for large-scale recombination-based cloning for library creation. Last, we highlight screening optimization considerations and NGS-based high-throughput readout tools for performing library screens in human cells.

STAMPScreen provides an integrated workflow for cell state engineering, which can be used for a multitude of biological applications. We envision that the tools and methods presented in this article will enable researchers to perform a varied array of large-scale screens for target overexpression (using cDNA or CRISPRa), knockdown (using CRISPRi), or dual overexpression and knockdown (using cDNA and CRISPRi). The modularity of these systems will enable pathway and GRN studies, differentiation factor screening, drug and complex pathway characterizations, and mutation modeling. Therefore, we envision that STAMPScreen will serve as an efficient and high-throughput workflow for the study and engineering of cellular phenotypes.

Limitations of the study

We present STAMPScreen as a widely applicable method that can be utilized in its entirety or in sections, based on the user's needs. While the method is broadly applicable, there are still limitations that will require further research. Currently, the computational target identification methods solely utilize RNA-seq data, which can be limiting when epigenome datasets are available to help better define the underlying genetic regulation architecture. However, we believe that our methods can be easily adapted and integrated into additional pipelines for utilization of these novel data types. In addition, our CRISPR-based experiments only compare three common CRISPRa tools and two common CRISPRi tools, thus more research on novel enzymes or gene perturbation platforms is needed. Furthermore, we believe MegaGate will be a widely utilized tool for cDNA cloning, with near universal applicability to ORF cloning. Except in extremely rare instances of an ORF containing one of the described meganuclease sites, MegaGate can be utilized without customization. However, this method of cloning is easily customized and can be modified to fit nearly any ORF constraint. Finally, we demonstrate that gene barcodes can be captured alongside transcripts using TAR-seq and RNA-seq. For lower read depth methods, such as scRNA-seq,

barcodes can be difficult to capture effectively. In these instances, it will be necessary to further PCR-enrich the barcodes separately to best capture the genes in each cell. Users may also need to optimize these conditions for their specific experimental design.

CHAPTER III : TRANSCRIPTOMIC REPROGRAMMING SCREEN USING FUNCTIONAL RNA CLOCK FOR CELLULAR REJUVENATION

AUTHORS AND CONTRIBUTIONS

This chapter is a submitted research article: Alexandru M. Plesa, Sascha Jung*, Michael Shadpour, Helen H. Wang, Fawad Omar, Antonio del Sol, George M. Church. Transcriptomic reprogramming screen using functional RNA clock for cellular rejuvenation. (2022).*

Author contributions: A.M.P. and G.M.C. conceived the idea of the study. A.M.P. wrote the manuscript with input from all co-authors and was involved in all experiments and analyses. S.J. developed the transcriptomic clock with input from A.M.P, and performed the computational analyses. M.S. contributed to the experiments and analysis for the RNA-Seq and flow cytometry. H.H.W. contributed to prediction and generation of the candidate list, and assisted with experiments and analysis for the RNA-Seq and flow cytometry. G.M.C and A.d.S supervised the work.

ABSTRACT

Aging is a highly complex process that manifests itself through a progressive multifaceted functional decline. Unstable transcriptomic profiles have emerged as a hallmark of aging organisms[16,20], and their modulation through interventions such as reprogramming has been shown to reverse aging signatures and improve function[74,80]. Furthermore, aging clocks use gene expression information at the epigenetic or transcriptomic level to predict the biological age of a cell[140,141] but they currently lack the interpretability necessary for assessing cellular function and identifying specific rejuvenating targets. To address this, here we integrate gene-cellular process associations[142] with RNA-Sequencing datasets to develop a functionally interpretable transcriptomic age predictor. Our RNA clock showed high predictive accuracy (r=0.93) when applied to multiple datasets, as well as robust response to known age-modulating interventions such

as reprogramming and cell stressors. Moreover, we used our clock as an integrative aging assay and performed a transcriptomic reprogramming screen for rejuvenation of primary human fibroblasts. Using our predictor's functional interpretability, we uncovered four distinct aging phenotypes based on process dysfunctions, which influenced the response to most of our perturbations. Nevertheless, we discovered that SRSF1 overexpression led to robust transcriptomic age reversal and functional improvement. These findings propose a new paradigm for aging as a transcriptomic state and SRSF1 as a promising target for cell rejuvenation.

INTRODUCTION

The transcriptome is a key determinant of the cell phenotype and regulates its identity and function. As such, there is a growing body of evidence that supports a systems-level view of cell engineering, whereby exogenous signals in the form of gene perturbations can induce desired cell states[98,143]. For example, overexpressing the four Yamanaka factors was shown to not only dedifferentiate a somatic cell to a pluripotent state, but also reverse the age-related functional decline in old cells, thereby supporting a view of aging as a transcriptomic state[65,66].

Throughout aging, there is an increased instability in the gene expression program[21,38], which serves as the basis for numerous aging clocks trained at either the epigenetic or transcriptomic level[52,141]. However, there is little known about the functional relevance of the DNA methylation sites or the thousands of transcripts driving currently used aging predictors[56,140]. This lack of interpretability along with the complexity of inducing discrete epigenetic perturbations precludes the use of these clocks as proxies for cellular function in the study of aging.

In this study, we set out to explore transcriptomic reprogramming as an approach to cellular rejuvenations. We first developed a functionally interpretable transcriptomic predictor that

can accurately measure the age of human fibroblasts and respond to known biological interventions. Then, we used our transcriptomic assay to perform a cDNA overexpression screen for rejuvenating perturbations in aged human cells. Leveraging the functional interpretability of our approach, we describe transcriptomic differences in the aging phenotype of cells from different old donors and analyze the effect of our perturbations on key cellular processes in aging. Lastly, we discovered SRSF1 as a novel age-modulating gene, whose overexpression reprogrammed transcriptomes to younger states through the differential splicing of genes involved in histone methylation and translation initiation.

RESULTS

Process-based transcriptomic aging clock

Using a published RNA-Sequencing (RNA-Seq) dataset[56] from primary normal human dermal fibroblasts (NHDF) aged 1-96 years old, along with our own sequencing data from similar fibroblasts, we trained a machine learning predictor on the chronological age of the sample donor. Our approach included a layer of functional interpretability by sub-setting the transcriptome into cellular processes using the Molecular Biology of the Cell Ontology[142]. To develop our clock, we first trained process-specific weak age predictors using a generalized linear model (glm), and identified the cellular processes most predictive of age (Figure 3.1 a). This approach stably selected 8 processes (WNT signaling, Histone methylation, Junction organization, Translation initiation, Actin polymerization, Lipid transport, ER quality control (ERQC), Sodium transport) with varying relative contributions to the age prediction (Figure A3.1 a). Furthermore, we used the gene expression levels for the genes in these 8 processes to train the ensemble predictor, which achieved accurate age predictions using only 146 genes (Figures A3.1 b, A3.2 a, Table A3.1). Finally, we

analyzed process-specific aging across the human lifespan and identified three distinct trajectories (Figure 3.1 b): early logarithmic increase (Histone methylation), middle-age inflection (Junction organization, Lipid transport, Sodium transport), and late-life exponential increase (Translation initiation, Actin polymerization, ERQC, WNT signaling).

We applied our predictor to 12 other different datasets and observed a similar performance to that of the widely-used epigenetic clocks (Pearson's r=0.93, p-value < 2.2e-16; MAE=6.66)[140], showcasing the accuracy and robustness of our RNA clock (Figure 3.1 c). Moreover, our transcriptomic assay showed good agreement with the epigenetic clock[144] in predicting the *in vitro*

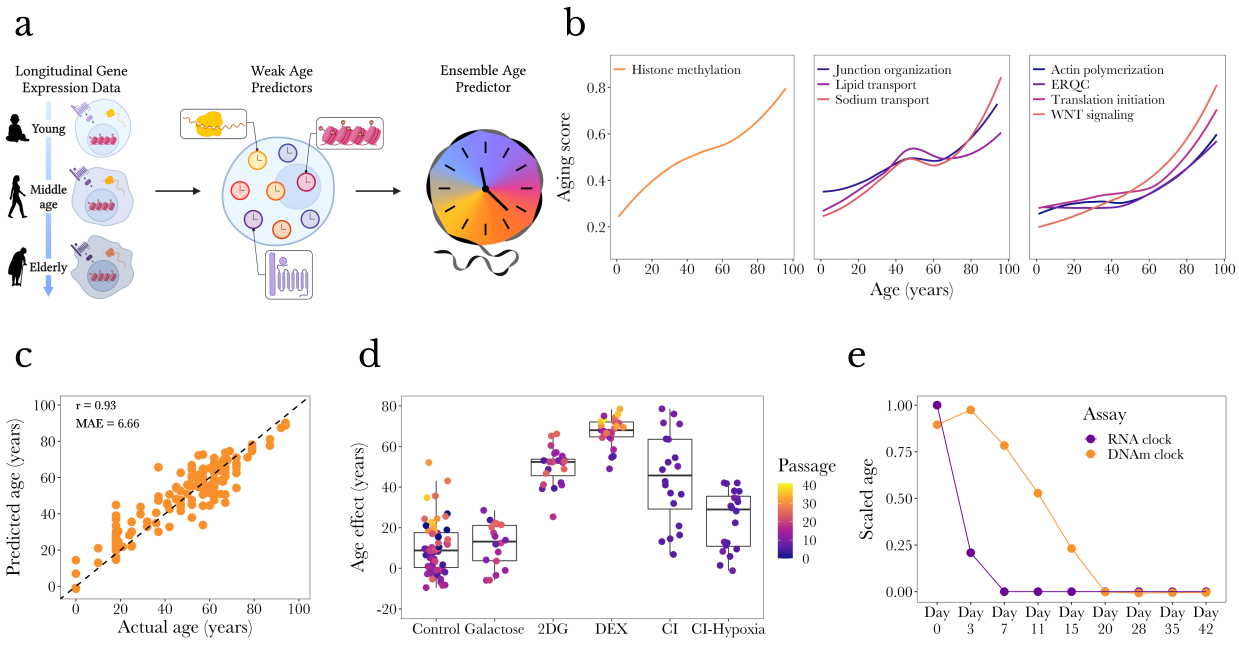


Figure 3.1: Process-based transcriptomic clock reports on biological age of human fibroblasts.

(a) Schematic for transcriptomic clock training. Longitudinal transcriptomic data from human fibroblasts is used to train process-specific weak age predictors, which are then combined into an ensemble predictor. (b) Aging score for different clock processes across the human lifespan. Score is calculated as a weighted average of process specific clock genes and normalized across all ages (n=133), such as the highest score seen in the oldest samples. The processes are split between three clusters based on their trajectories. Histone methylation experiences an early logarithmic increase; junction organization, lipid transport and sodium transport have a middle-age inflection point; actin polymerization, ERQC, translation initiation, and WNT signaling exhibit an exponential

Figure 3.1 Continued

with age. (c) Validation of our transcriptomic clock on human dermal fibroblasts from independent datasets (n=12). The RNA clock maintains high accuracy (Pearson's r=0.93, p-value < 2.2e-16; MAE=6.66) across different sequencing platforms. (d) Response of RNA clock to different age-modulating in vitro interventions. Galactose has no effect on the age of the cells, unlike 2-deoxy-D-glucose (2DG), dexamethasone (DEX) treatment and contact inhibition (CI), which showed significant increases in the predicted age. Growth in hypoxia (3% O^2) showed partial reversal of the detrimental effect of contact inhibition. (e) Response of RNA clock and DNA methylation (DNAm) clock (Horvath Skin and Blood) to cellular reprogramming time course. The predicted age is scaled across the entire time course and exhibits a steeper and faster decrease in the RNA compared to the DNAm clock. The two clocks reach a consensus at the iPS stage suggesting that the rejuvenation effect of reprogramming peaks after 7 days of reprogramming, in line with previous studies[80].

aging of skin fibroblasts (1 year ≃ 0.7 PD), but was not as accurate when applied to lung fibroblast, implying a tissue specificity (Figures A3.2 c, A3.2 d). When using our predictor on cells exposed to age-modulating treatments we observed expected age-related changes in the transcriptomes of progeroid samples, as well as cells exposed to Rapamycin, hypoxia, and various stresses[145] (Figure 3.1 d, Figure A3.3).

We also used our predictor on a time course dataset of fibroblast reprogramming, and we observed a drastic decrease in the predicted age after only 3 days (Figure 3.1 e). In comparison, the widely used DNA methylation (DNAm) clock reached similar age predictions only at day 15 of reprogramming, after which the two assays reached a consensus at the iPSC stage. This suggests a shorter response time for the transcriptional assay, which is in line with the measured effects of Rapamycin in young and middle-aged cells (Figure A3.3 f). We believe the observed accelerated transcriptional rejuvenation is due to the dynamic nature of gene expression networks as compared to the more stable and slower-changing DNAm patterns. When analyzing the specific process activity across the reprogramming timeline, we discovered that early rejuvenation is driven by changes in genes part of the WNT signaling, histone methylation, and lipid transport processes, which acquire a stable young state after 7 days (Figure A3.2 e).

Age reversal screen identifies genes for cellular rejuvenation

We next applied our RNA clock to identify novel rejuvenating interventions by performing a cDNA overexpression screen in human cells. Using our previously published target identification method[93], we analyzed a publicly available NHDF RNA-Seq dataset[56] and generated a ranked list of genes predicted to be highly influential in the aging process (Figure A3.4, Table A3.2). We chose 89 genes to test for their effect on the aging phenotype and included KAT7 (a gene implicated in fibroblast senescence[146]) as well as the Yamanaka factors in a polycistronic cassette (OSKM), as positive controls.

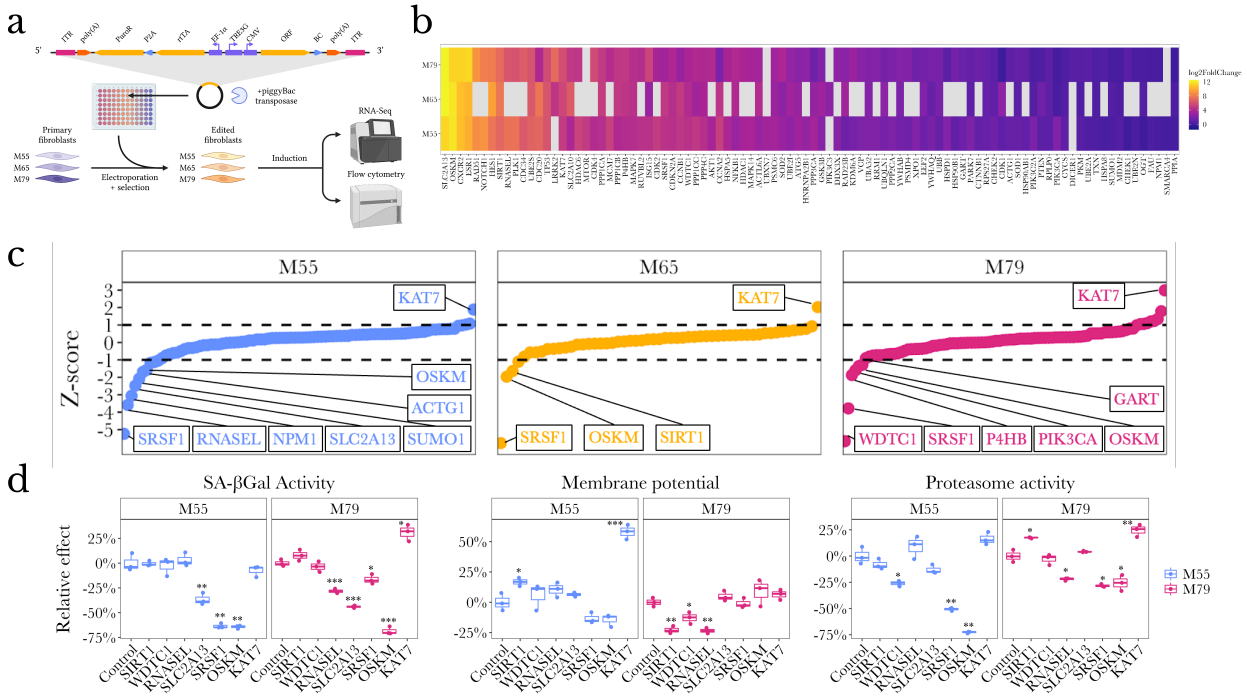


Figure 3.2: Age reversal screen identifies genes for cellular rejuvenation.

(a) Workflow of cDNA overexpression screen for rejuvenating interventions. We electroporated 95 different plasmids containing our overexpression cassette along with the gene of interest into 3 different NHDF lines. The edited fibroblasts were induced using Dox (1 ug/mL) for 3 days and assayed for age-related changes using RNA-Seq and flow cytometry assays. (b) Heatmap of gene induction over all our overexpression lines in the three different NHDF backgrounds (n=2). There is a consensus across the log2FoldChange levels across the 3 different lines, and we observe a range of induction values across all tested genes. (c) Z-score for the age effect, as measured by our RNA clock, in the overexpression screen across the three NHDF lines (n=2). Genes with a z-score lower than -1 (their age lowering effect is more than 1 standard deviation away from the mean) and

Figure 3.2 Continued

KAT7 (positive control for pro-aging effect) are labeled. (d) Staining data for representative gene hits and controls across the three assayed cellular processes in the M55 and M79 lines (n=3). Data is normalized with respect to the gene specific No Doxycycline control and the line specific gene control, and significance is tested using the two-tailed Student's t-test.

To test the potential age-modulating effect of our target genes, we generated individual overexpression lines of our candidates in 3 different primary NHDF lines from 55-, 65-, and 79-year-old donors. After selecting for integrants, we induced the expression of our transgenes and assayed age-related changes using RNA-Seq and flow cytometry-based assays (Figure 3.2 a). We observed a wide range of gene induction levels among our 95 genes, ranging from no overexpression to 4000-fold overexpression (Figure 3.2 b). However, there was good agreement in the induction levels between the 3 NHDF lines, indicating that induction differences are gene specific. We confirmed the expression of most of our transgenes and the identity of our lines by performing a barcode search on the raw sequencing reads (Figure A3.5, Table A3.3). The results agreed with our fold change data, and we discovered that the basal endogenous gene expression level is negatively correlated with the log2FoldChange values (Figure A3.6 a), confirming our previous findings that there are gene-specific induction limits[93].

We then analyzed the effect of our genetic perturbations on the aging phenotype of the three NHDF lines using our transcriptomic age predictor. The RNA clock correctly estimated the ages of our WT NHDF lines (Table A3.4) and predicted the ages of our controls to be significantly different than that of the BFP lines (Figure 3.2 c). OSKM, as expected, had a high rejuvenating effect across all three lines, while KAT7 had the highest pro-aging effect. Most notably, we discovered several genes that induced significant changes towards a young transcriptome across the 3 NHDF lines (Figure 3.2 c), ranging from -20 to -50 years, including SIRT1, SLC2A13, SRSF1, RNASEL, and WDTC1. While the role of SIRT1 in aging has been thoroughly studied[147–

[150], the other four hits have limited information available about their role in aging. SLC2A13 was found to decrease during aging in the human dorsolateral prefrontal cortex[151] and was recently identified as a risk gene for Parkinson's disease[152]. The expression of SRSF1 was significantly negatively associated with age[153] and with parental longevity in humans[154]. RNASEL levels in human serum were shown to inversely correlate with metabolic syndrome and age[155], and WDTC1 has not been previously linked to aging, but it was associated with lower fat mass and heightened insulin sensitivity in humans[156].

For an orthogonal readout of our transcriptomic clock, we also performed three staining assays that report on important cellular features: senescence associated β-galactosidase activity (SA-βGal activity), mitochondrial membrane potential, and proteasomal activity. Our data showed that OSKM, RNASEL, and SLC2A13, and SRSF1 reduced the senescence phenotype in both our tested lines (Figure 3.2 d). In the mitochondrial assay, we observed opposing effects in our two lines, with SIRT1 increasing the mitochondrial potential of M55 and decreasing that of M79 (Figure 3.2 d). SRSF1 and OSKM had similar opposing effects between the two lines, with a decrease of the potential in M55 and an increase in M79, but not reaching the significance threshold. This, along with our prior observations of large variability in the mitochondrial potential with age, suggests a potential homeostasis of the membrane potential, whose increase or decrease can lead to dysfunction. Lastly, the proteasome staining data showed a decrease in proteasomal activity in the OSKM and SRSF1 overexpression lines across both M55 and M79 (Figure 3.2 d). This reduction of the 26S proteasome activity was unexpected, but when compared to the known pro-aging KAT7 overexpression, we noticed there was an opposite effect, whereby KAT7 increased the staining of our assay, but our hits decreased it. We hypothesized that rejuvenating interventions would lead to increased proteostasis, thereby lowering the burden on the proteasome and reducing its expression and measured activity. Future studies aimed at measuring proteotoxicity would be

necessary to better explain this result. Finally, the observed differential responses between the NHDF lines in both our transcriptomic and functional assays (Figures A3.7, A3.8) suggest that cell line-intrinsic differences in the starting transcriptome might influence the outcome of our perturbations.

Transcriptomic variability in aging phenotypes

To further study the differences in transcriptomic aging between our three initial cell lines, we performed a Uniform Manifold Approximation and Projection (UMAP)[157] analysis and observed distinct clusters corresponding to the three NHDF donors (Figure 3.3 a). This suggests that following the stress of our nucleofection and selection process, the three cell lines developed considerable transcriptomic differences, which are more pronounced than gene expression changes induced by our perturbations. We then set out to analyze differences in the WT aging phenotypes through the perspective of our clock processes by focusing on isogenic samples from individuals at different ages across their lifespan (2-19 years age difference). When quantifying the process activity for our eight processes, we observed the NHDFs fall into four distinct groups defined by different dysfunctions (Figure 3.3 d): group I = ERQC and lipid transport, group II = WNT signaling and sodium transport, group III = histone methylation, and group IV = translation initiation and sodium transport. These results contextualize the different responses of the same perturbation in multiple cell lines and showcase the variability of human aging, in line with a recent study that described four distinct human aging patterns based on the types of molecular patterns that change over time[158].

Interestingly, while most of our gene inductions showed distinct age effects across the three lines (Table A3.4), we discovered that some of the perturbations clustered together in UMAP space (Figure 3.3 a), overcoming line-specific differences. We hypothesized that these genes (KDM6A,

NOTCH1, OSKM, and SRSF1) induced robust transcriptomic reprogramming due to their central role in regulating gene expression, and overcame differences in the starting gene regulatory networks. In agreement with this hypothesis, we observed a strong negative correlation (Pearson's r = -0.64, p-value = 1.678e-11) between the number of differentially expressed clock genes and the mean distance in UMAP space between overexpression lines for the same gene (Figure 3.3 b).

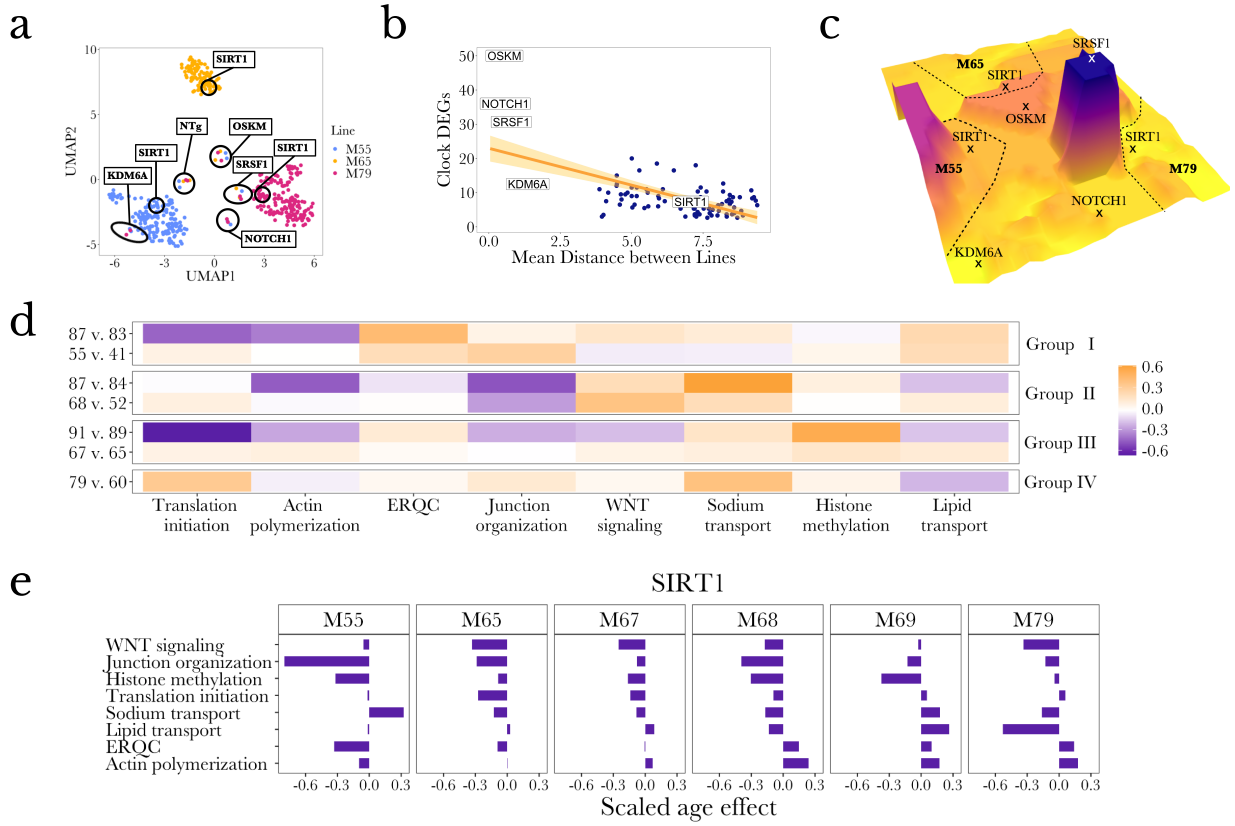


Figure 3.3: Variability of the aging phenotype and its response to perturbations.

(a) Total transcriptome UMAP on overexpression lines from age reversal screen. There is a strong clustering based on donor cell lines with only a few perturbations overcoming line-specific differences. (b) Relationship between transcriptomic variability of interventions across cell lines and their induced differentially expressed clock genes. We observed a strong inverse correlation between the mean umap distance of biological replicates of the same gene and the associated number of clock DEGs, with OSKM, NOTCH1, SRSF1, and KDM6A having the largest effects. (c) 3D landscape of all 480 NHDF transcriptomes from our large-scale screen, with the x-y axis representing the UMAP coordinates from a, while the Z-axis represents the predicted age, with the youngest samples having the largest values. (d) Heatmap of differential (older age - younger age) process activity in 7 pairs of isogenic cell lines. Larger activity is associated with an older phenotype, while lower activity is representative of the young cells. We observed four distinct

Figure 3.3 Continued

groups based on the nature of the process dysfunction: group I = ERQC and lipid transport, group II = WNT signaling and sodium transport, group III = histone methylation, group IV = translation initiation and sodium transport. (e) Scaled age effect of SIRT1 overexpression on the clock processes across six lines (n=2). Negative age effects represent younger gene expression profiles. We observed no coordinated effects across the cell lines, but we noticed that the responder lines (M65, M67, M68, M79) all had moderate rejuvenation in WNT signaling.

Moreover, similar to the Waddington landscape analogy[159], we generated a transcriptomic landscape of fibroblast aging by embedding the predicted age of the samples to the UMAP plot and noticed that SRSF1 and OSKM had strong robust rejuvenating effects, while KDM6A and NOTCH1 had significant pro-aging effects (Figure 3.3 c). This suggests that perturbations with strong effects on the aging phenotype, as measured by our RNA clock, push the cells into a common transcriptomic space, while the effect of weaker perturbations is dependent on the initial GRN state (Figure A3.9).

One example of a weak perturbation that has cell-line dependent effects is SIRT1 overexpression. There have been reports linking SIRT1 levels to lifespan extension in yeast, but not in worms, flies, or mice, thereby questioning its role in the aging process[147–150]. In our initial screen, SIRT1 showed transcriptomic rejuvenation effects in two of the three lines, with only one being statistically significant. Similarly, when we repeated the gene overexpression in 6 different lines, we observed strong age reversal effects in only half the lines (Figure 3.4 a), further confirming SIRT1 overexpression as a weak perturbation with cell-line specific effects. When looking at the scaled age effect on the clock processes, we observed no common pattern between all lines, which suggests that SIRT1 overexpression has line-specific aging effects (Figure 3.3 e). Nevertheless, in the responder lines (M65, M67, M68), we observed a slight improvement in WNT signaling, suggesting SIRT1 might act on the WNT pathway in these cells.

SRSF1 induces robust cellular rejuvenation

To validate our initial findings and further study the effect of our hits on the aging phenotype, we repeated the experiment by generating new overexpression lines in NHDFs from 6 different donors (M55, M65, M67, M68, M69, M79). In agreement with our initial results, we observed that most of the genes had cell-line specific effects, but some of them didn't show the

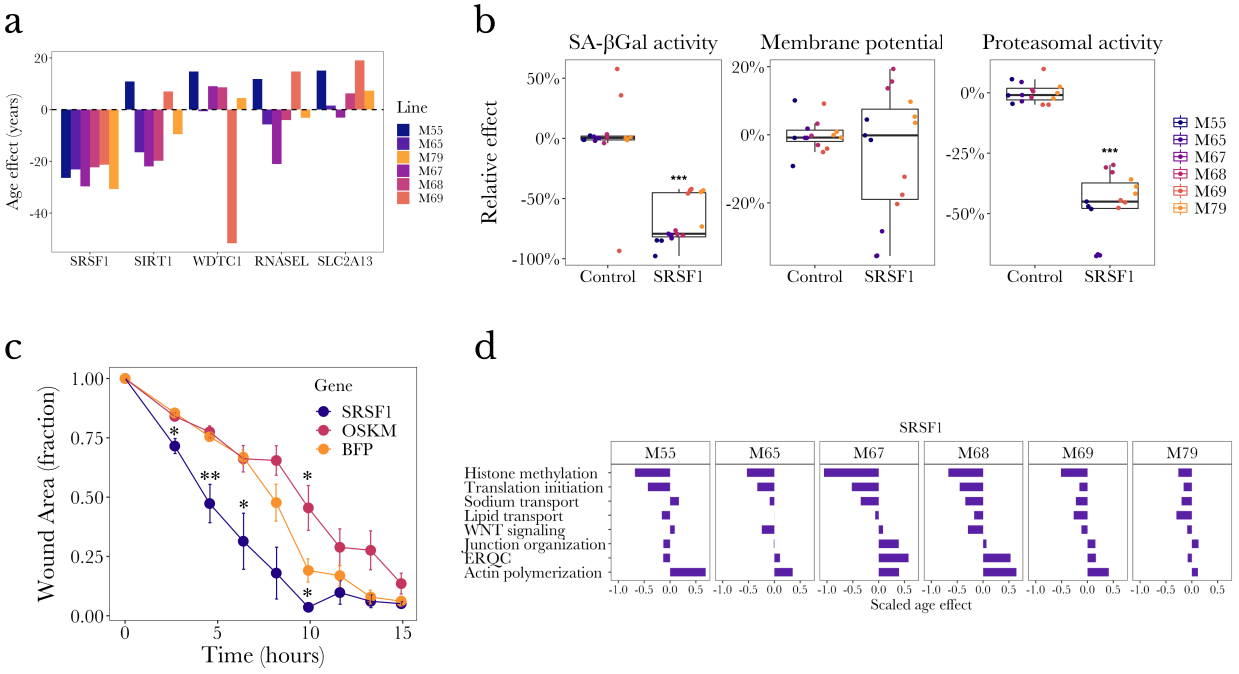


Figure 3.4: SRSF1 induces robust cellular rejuvenation.

(a) Predicted age effect of hit overexpression in validation experiments using 6 NHDF lines (n=2). We observed robust age reduction in the SRSF1 samples, and a beneficial effect of SIRT1 on the transcriptome in 4 out of 6 lines. The other 3 genes had weak cell-line specific effects with the exception of WDTC1, which showed marked rejuvenation in the M69 line. (b) Staining data for SRSF1 across three assayed cellular processes in the six cell lines (n=3). Data is normalized with respect to the uninduced samples and is reported as a relative effect to the induced control, and significance is tested using the two-tailed Student's t-test. We observed a robust significant reduction in the SA-βGal and proteasomal activity, but no coordinated effect on the mitochondrial membrane potential. (c) Scratch assay time course, expressed as fraction of wound area, for the M79 line after the induction of OKSM, SRSF1, or BFP. Compared to the BFP control, we observed a faster wound area closure in the SRSF1 treated cells and a slower response in those that were induced with OSKM. Data is averaged over replicates (n=3-5) and significance with respect to the BFP control is tested using the two-tailed Student's t-test. (d) Scaled age effect of SRSF1

Figure 3.4 Continued

overexpression on the clock processes across six lines (n=2). Negative age effects represent younger gene expression profiles. We observed high correlation between the process effects across all 6 lines, with a significant improvement in histone methylation and translation initiation.

same beneficial changes in our transcriptomic assay (Figure 3.4 a, Table A3.4). However, SRSF1 displayed a robust transcriptomic age reversal effect in all 6 lines, with an average rejuvenation of approximately 26 years, which was expected considering its strong effect on the clock genes (Figure 3.3 b).

Our flow cytometry assays showed that SRSF1 induction reduced cellular senescence by more than 50% across all 6 lines, suggesting functional rejuvenation of these cells (Figure 3.4 b). However, we did not observe any significant effect on the mitochondrial potential, in agreement with our initial screening data. Similarly, the proteasome activity was reduced in the SRSF1 samples, matching the observed OSKM induction phenotype and opposite to the pro-aging effect of KAT7 and CDKN2A (Figure 3.2 d, Figures A3.7, A3.8). To further assess rejuvenation upon SRSF1 induction, we tested wound healing efficacy, a key function of dermal fibroblasts, using an *in vitro* scratch assay. Our data showed that SRSF1 expressing cells reached half wound closure 50% faster than the BFP control, while OSKM seemed to slightly delay full wound closure (Figure 3.4c, Figure A3.10), in agreement with previous mouse wound healing results[160].

We then studied the effect of SRSF1 overexpression on the individual processes of our clock, and observed a coordinated rejuvenation of gene expression associated with histone methylation and translation initiation (Figure 3.4 c). Since histone methylation has a central role in regulating gene expression, it is expected that a reversal of the aging associated transcriptional changes would have a large and robust beneficial effect on cells. Conversely, we see that the overexpression of KDM6A has a similar but detrimental coordinated effect on the histone

methylation and WNT signaling processes (Figure A3.9 a). Considering the strong effect of OSKM and SRSF1 on the transcriptome, and the known dedifferentiation potential of the Yamanaka factors, we then asked whether SRSF1 also induces a loss of cell identity. By performing an enrichment analysis for fibroblast, myofibroblast, and mesenchymal stem cell (MSC) identity genes, we discovered that both SRSF1 and OSKM lead to a reduction of fibroblast cell identity, when compared to the WT and BFP controls (Figure A3.11c, A3.12).

Since SRSF1 is a known splicing factor, we performed a differential splicing analysis in our overexpression lines and observed great agreement in the distribution of splicing events between the 6 lines, with an alternative first exon accounting for more than 50% of all differential isoforms (Figure A3.11 a). Interestingly, when we looked at the parent processes for the alternatively spliced genes, we discovered that histone methylation and large ribosomal subunit organization were the only processes common for all 6 lines (Figure A3.11 b). Moreover, the alternatively spliced genes have been previously confirmed to directly interact with SRSF1[161], providing further evidence for the link between SRSF1 and gene expression changes in histone methylation and translation initiation. Therefore, we propose a mechanism by which SRSF1 overexpression leads to differential splicing of important regulators of histone methylation and protein translation that induce young gene expression profiles in these processes (Figure A3.11 d).

DISCUSSION

In this study, we took advantage of the transcriptome's multidimensional integration and functional interpretability to build an accurate aging clock with higher responsiveness than the epigenetic clocks. The increased sensitivity of our assay is likely due to the more dynamic nature of the transcriptome, compared to the epigenome, which would also imply a higher susceptibility

to biological and technical noise. Nevertheless, the functional interpretability of our RNA clock revealed histone methylation as the earliest process to show aging-related changes, suggesting a potential causal role in aging, which aligns with recent reports of reprogramming and its associated epigenome remodeling as a general age reversal strategy among different cell types.

Our study also provided a proof of concept for using the transcriptome in combination with functional assays as *in vitro* aging readouts for large scale rejuvenation screens. We believe that such approaches will be instrumental in discovering novel perturbations for transcriptomic reprogramming in the context of aging and cell engineering. Moreover, we showed that the outcome of such gene expression modulation experiments is influenced by both the initial transcriptomic state of the cell and the strength of the perturbation, with transcription factors and chromatin modifiers having the most robust effects across different backgrounds.

This observation proposes an interesting problem for the development of universal age reversal therapies, which would require strong perturbations without unintended transcriptional shifts that could alter cell identity and function. For example, the use of strong exogenous perturbations such as OSKM is known to have robust rejuvenation effects at the expense of loss of cell identity, thereby requiring careful dosage for therapeutic usage. Similarly, we find that SRSF1 also induces a loss of cell identity, but we hypothesize that this effect could be temporary, since SRSF1 is endogenously expressed in these cells and it was shown to be transiently activated upon regeneration signals[162]. Future experiments are needed to determine the persistence of this effect and whether the gene overexpression possesses similar teratogenic risks as OSKM.

Overall, we propose a novel paradigm for studying aging and rejuvenation as transcriptomic cell states using process-based prediction algorithms, and we identify SRSF1 overexpression as a potential age reversal intervention.

CHAPTER IV : DISCUSSION

AUTHORS AND CONTRIBUTIONS

Author contributions: I wrote this chapter with input from my advisor Prof. George M. Church.

SUMMARY

In this dissertation, I present a collaborative effort to develop a new method for the discovery of age-reversal interventions through transcriptomic reprogramming. We first developed new tools as part of an integrated pipeline to enable large-scale screens for mammalian cell engineering. Next, we built a robust and accurate transcriptomic aging assay, which we used to perform a transcriptomic reprogramming screen aimed at identifying human age reversal perturbations. Finally, we provided a proof of concept for the proposed discovery pipeline by identifying SRSF1 overexpression as a robust rejuvenating intervention.

New tools for cell engineering screens

In Chapter 2, we present an integrated pipeline for mammalian cell engineering as well as newly-developed tools to aid in the discovery of genetic interventions for cell state transitions. Our pipeline starts with RNA-Seq data for establishing the initial and final transcriptomic states of a given cell type of interest. We then propose leveraging network-based target identification algorithms, which use publicly available data or predicted gene regulatory network interactions, to identify important nodes that can have large effects in shifting the transcriptome towards a desired state. To inform the screening of these candidate genes, we compare the latest genome editing tools and provide new vectors for a modular control of the perturbation platform. Lastly, we show how NGS-coupled readouts can connect cell phenotypes to genetic perturbations and provide useful data for refining target prediction and achieving higher fidelity cell conversion.

With the increased accessibility of large gene expression datasets in different cell types[96,163], new algorithms have been developed for uncovering the underlying gene regulatory network of specific cell types and predicting ways to influence it in desired directions[85,86,164]. Similarly, we provided two new tools that are cell-type agnostic and only require transcriptomic data from the starting and target cell states. Such datasets are slowly becoming available for the aging phenotype, with consortia like tabula muris[165], tabula sapiens[166], and GTEx[167] providing data for multiple tissues of varied ages. Furthermore, both our tools leverage network information, either through publicly available data[168] or by identifying gene-gene interactions from time course transcriptomics[108], to further refine their predictions and infer causality. Since our current understanding of cellular gene regulatory networks is incomplete, future research in this area will undoubtedly improve the performance of these algorithms.

Another important aspect of genetic screens that we address in this work is the perturbation system. Here we evaluated the performance of multiple gene modulation technologies, and showed that cDNA overexpression is, on average, more potent than CRISPRa activation, while the two tested CRISPRi tools have similar efficiencies. Moreover, cDNA overexpression has the advantage of requiring less genetic material for achieving the desired effect, whereas the CRISPR systems usually require two separate plasmids, thereby increasing cell toxicity. However, the choice between cDNA overexpression and CRISPRa should take into account the efficiency and toxicity aspects as well as any potential post-transcriptional regulation systems. In some instances where the dosage of a gene and the relative abundance of its isoforms is important, or where the identity of the active isoform is unknown, CRISPRa applications are preferred.

We decided to further focus on gene induction as opposed to gene knockdown tools, since most cell engineering studies seek gain-of-function phenotypes, where the starting cell acquires new or different functions. As such, based on previous research[169], it is more likely that such cell state

transitions occur due to the expression of new factors rather than the downregulation of specific transcripts. Between overexpression and activation, we chose the former due to the improved precision and induction range offered by cDNA over CRISPRa, and also since ORF induction offers a defined perturbation (known coding sequence) that can be more tightly controlled. To further improve the precise control of the overexpression, we generated new Tet-inducible vectors with indetectable levels of induction in the absence of Doxycycline. Furthermore, we also developed a new molecular cloning tool for library scale cloning of ORFs, which offers efficient and fast (<3 days) cloning of any desired genes from expanding resources such as the human ORFeome[91] and TFome[98].

To address the applicability of this pipeline to complex cell engineering, such as age reversal, we also developed TAR-Seq, a target sequencing NGS readout that can measure high-dimensionality phenotypes using a subset of the transcriptome. TAR-Seq can provide accurate quantification of specific transcripts relevant to the cell state of interest at a fraction of the cost of RNA-Seq. This tool together with the previously mentioned ones are integrated in our pipeline to enable large-scale screens for cell engineering studies with complex phenotypes.

Transcriptomic reprogramming screen using RNA clock

In Chapter 3 we applied the pipeline developed in Chapter 2 to the aging problem, aiming to engineer old cells towards younger states by reprogramming their transcriptome. First, we developed an accurate high-throughput aging assay that can be used as a primary readout for our screen. Then we used our algorithm and cloning tools from Chapter 2 to predict and build a screening library for fibroblast rejuvenation. In performing our screen, we identified great heterogeneity in the aging phenotype and in the perturbation responses of our different donor

lines. Nevertheless, we discovered SRSF1 as a promising target, whose overexpression reprogrammed the transcriptome of 6 different fibroblast lines to younger states.

Aging clocks have revolutionized the field by providing accurate biomarkers that can be used to better understand and modulate the aging process. In Chapter 3 we present a functionally interpretable clock that uses transcripts from 8 different cellular processes to assess cellular age. Conceptually, the algorithm is measuring process efficacy by comparing the transcript levels of specific genes with those in young cells. An interesting observation is that some of the chosen processes relate to cell type agnostic functions (histone methylation, lipid transport, WNT signaling, translation initiation, ERQC), while others are more specific to fibroblasts (junction organization, actin polymerization, sodium transport). This distinction is also apparent in the process trajectories (Figures 3.1, A3.2), where histone methylation seems to be most responsive, suggesting it may have a more causal role in aging, which is in line with the large body of literature linking epigenome instability to aging. Furthermore, this observation indicates that there are core shared features of aging, as well as cell type specific ones. If true, such a distinction implies that multi-tissue clocks are either incomplete in their depiction of cell aging, or they require considerably more features to accurately predict the age of different cell types.

Using this newly developed transcriptomic clock and our target genes predicted in Chapter 2, we performed an overexpression screen in primary fibroblasts from middle-aged and old donors. Interestingly, we observed a wide range of overexpression levels in our candidate genes, which was inversely correlated with the basal expression levels of these transcripts. This is in agreement with our data from Chapter 2 that shows lower overexpression of highly expressed genes, but it also poses a problem for the therapeutic induction of endogenously expressed genes, which likely have negative feedback mechanisms. Possible ways to overcome this limitation is to use stronger promoters[170] or prevent mRNA degradation by optimizing the secondary structure[171]. It is also

worth noting that the highest age reversal effects were seen in the youngest cell line (M55), suggesting that it may be easier to rejuvenate younger cells. While more systematic studies are required to determine whether this effect replicates at the population level, it is somewhat expected to more easily rejuvenate a middle-aged GRN than that of an old cell, simply due to the higher similarity between the young and middle-aged transcriptomic states. However, this effect size difference was not apparent in our flow cytometry assays, which, despite showing mostly opposing effects for the anti-aging and pro-aging therapies, were not always consistent with the transcriptomic clock results. One potential confounding factor is that the transcriptomic changes that serve as the basis of our clock can represent beneficial cellular responses to the aging phenotype. Since there is currently no method to determine the effect of such changes *a priori*, it's important to use integrative functional readouts both *in vitro* and *in vivo*, such as scratch or wound healing assays for fibroblasts, to assess the age reversal of the entire biological systems. As such, future studies should select readily available primary cells with defined functions and established assays that show large and robust changes with aging.

Our results also highlight the variability of aging across different donors. We identified different characteristic dysfunction profiles by applying our clock to isogenic samples from different donors at two timepoints in their life. This observation agrees with previous studies[158] and raises the possibility that the genotype or the different insults throughout one's life can influence the cellular aging phenotype and, in turn, the response to different interventions. Such a varied response is indeed present in our screening data, where most of the genetic perturbations have donor cell line specific effects, likely due to different initial transcriptomic states. Inspired by Waddington's epigenetic landscape[159], we generated a transcriptomic landscape of fibroblast aging and showed that strong perturbations that induce large gene expression changes overcome initial steady state differences and shift all cell lines towards a common transcriptomic space. While such

interventions are desirable for their robust effect, they could potentially lead to a disproportionate reprogramming of the transcriptomic network away from the initial normal state, thereby inducing a change in the cellular identity.

An example of such an intervention is the overexpression of the SRSF1 splicing factor, which induces a similar age reversal effect in 6 different initial transcriptomic states. With splicing dysfunctions being recently linked to aging and longevity[29,30,32,154,172–174], and considering the key role of alternative splicing in regulating gene expression, it is not unexpected that SRSF1 modulation can induce robust transcriptomic reprogramming towards a younger state. However, along with cellular rejuvenation, we also observed a downregulation of cell identity genes comparable to that caused by OSKM induction. Moreover, SRSF1 has also been implicated in senescence[175,176] and has been identified as an oncogenic downstream target of Myc[177]. Nevertheless, unlike OSKM in somatic cells, SRSF1 is an endogenously expressed gene under negative self-regulation[178], whose expression declines with age in humans[153] and is required for regeneration and adult cell homeostasis in mice[162]. Therefore, while our current data show SRSF1 is a promising age-reversal target, future experiments are required to determine the safety and efficacy of its overexpression in aged cells.

FUTURE DIRECTIONS

The future of cell engineering

While current algorithms have moderate success in predicting cell differentiation targets, future studies aimed at complex phenotype conversions will likely require the use of large-scale combinatorial perturbations and deep learning representations of the underlying gene regulatory networks. Since the space of all possible combinations of a given set of genetic factors is too large

for an exhaustive search, computationally-guided experimental design will be necessary for choosing the perturbations most informative of the unknown parts of a cell's GRN. Nonetheless, with the progress of gene modulation and NGS technologies[179], as well as that of deep neural networks[180], I am confident that in the future we will have a better representations of how complex gene interactions give rise to cellular phenotypes, and we'll be able to design specific genetic perturbations for reprogramming aged and diseased cell states to youthful and healthy ones.

An important aspect of such therapeutic perturbations will be their durability, which was recently addressed by the development of heritable gene modulating systems such as CRISPR-OFF[181]. However, there is also a recent interest in the development of transient mRNA-based therapies[182]. Despite being temporary, given the right dosage and duration, such mRNA perturbations can induce stable changes in a cell's GRN, by leveraging the cell's internal epigenetic machinery. Moreover, while the CRISPR tools will undoubtedly be useful in a research setting, the safety profile of transient perturbations will make mRNA interventions more attractive for therapeutic usage. As such, advances in RNA biology will be important for informing the future design of therapeutic perturbations.

Similarly, progress in the development of new inducible vectors under the control of different small molecules will offer a more granular control over the simultaneous induction and/or repression of multiple genes of interest, thereby expanding the space of possible cellular perturbations. Although most differentiation studies use gene activation for cell engineering, there is evidence showing that the downregulation of specific genetic modules can also increase cell conversion efficiencies[183]. Lastly, while our MegaGate cloning method will undoubtedly facilitate library scale cloning of ORFs for specific screening applications, the field is slowly moving towards a synthesis-based approach, similar to that used in cloning gRNAs. Once the cost per base pair is substantially reduced and the complexity problems (high GC content, repetitive regions) are solved,

gene synthesis will become the preferred method for building screening libraries of both natural and synthetic DNA sequences.

Furthermore, since the innovation of multiplex DNA sequencing[184] and the subsequent improvement of NGS technologies, the cost of sequencing has dropped precipitously, surpassing the predictions of Moore's law[185]. For this reason, there has been a shift towards coupling every layer of cell biology to an NGS readout, starting with the genome and transcriptome, and more recently focusing on the epigenome, metabolome, and proteome. As such, future studies will most likely integrate multiple such NGS assays to enable comprehensive measurements of complex phenotypes induced by specific perturbations.

Using transcriptomic reprograming screens to uncover age reversal targets

The main innovation of our transcriptomic clock lies in its interpretability due to the process specific representations. A similar shift towards linking predictor features to cell biology is happening in the epigenetic clock field[186], and it's likely that future studies will also adopt this methodology due to its mechanistic insight. However, since clock processes differ in their predictive power and causality, cell type specific clocks will probably be more accurate than their multi-tissue counter-parts and will be preferred if the necessary data is available. When such data is sparse, multi-tissue clocks will likely be applied to difficult to obtain cell types such as neurons and cardiomyocytes.

Another future direction of innovation for aging biomarkers is the adoption of single cell technologies, which is starting to be used for RNA-Seq[187–189] and methylome[190–193] studies of a variety of cells across different ages. With this data, multiple groups have already developed transcriptomic[189] and epigenetic[194] single cell clocks, which offer an unprecedented resolution of the aging process. Such tools will be crucial for studying the heterogeneity of aging and for enabling

pooled screening studies. Future work on single cell proteomics[195] and metabolomics[196], as well as the incorporation of spatial dimensionality[197] will also expand the depth of phenotypic profiling available for aging studies and provide a more comprehensive picture of aging and rejuvenation.

Better perturbations methods will allow for more complex screening workflows, such as pooled and *in vivo* screens. Pooled screening will expand the number of perturbations tested in an experiment and enable whole genome screens, while *in vivo* experiments will recapitulate the complex environment of cells in an organism, thereby providing more biologically relevant results. By combining these two technologies, *in vivo* pooled screens will present a scalable method for identifying systemic mediators of organismal aging[198].

Future human aging studies will need to consider both safety and generalizability in their development of age reversal therapies. Currently the field is focused on cellular reprogramming, since it has been the most promising and robust intervention in the last few decades[199]. However, several groups have observed serious risks for *in vivo* OSKM induction[69,77–79,200], underlining the need for safer targets for human interventions. An alternative to strong and robust perturbations like OSKM induction, are more specific and safe therapies informed by the heterogeneity of human aging phenotypes. We and several other groups have observed this variability using different cellular or molecular biomarkers[158,201,202], suggesting that the future of personalized medicine might also include patient-specific anti-aging interventions.

Lastly, we expect future studies to further pursue SRSF1 induction as a promising strategy for cellular age reversal. Our efforts will focus next on assessing the stability of the reprogrammed transcriptome and duration of the rejuvenating effect, as well as determining the mechanism of action (cell autonomous vs cell non-autonomous) and validating the effect in *vivo* through worm lifespan and mouse wound healing assays. Furthermore, since there have been reports linking SRSF1 to thymocyte development[203] and satellite cell proliferation[204], it will be interesting to test

SRSF1 induction in other cell types such as hematopoietic stem cells or muscle stem cells. If safety and efficacy studies show positive results, we can envision using *in vivo* or *ex vivo* mRNA-based therapies for transient induction of SRSF1.

CONCLUSION

In conclusion, through the work presented in this dissertation, I propose a new paradigm for identifying age-reversal genetic targets through transcriptomic reprogramming screens. Towards this goal, we first developed an integrated screening pipeline for mammalian cell engineering that offers a streamlined workflow for the discovery of cell state transition interventions. To further enable such studies, we provided two unique algorithms for target identification, new modular cloning tools for high efficiency library cloning, and NGS-based assays for cost-efficient and high-dimensional readouts. We then developed a functionally interpretable transcriptomic clock that can serve as a scalable, robust, and accurate cellular aging assay. This enabled us to apply our proposed pipeline to the aging problem and perform the first large-scale age reversal screen in primary human cells. The most promising hit from our screen was SRSF1, whose overexpression showed robust transcriptomic and functional rejuvenation in multiple fibroblast lines. The safety, long-term efficacy, mode of action, and limitations of SRSF1 induction will be the focus of our future research. Overall, these experiments provide a proof of concept for using transcriptomic reprogramming screens to discover novel age-reversal targets, and we expect future improvements and applications of this pipeline to enable the discovery of safe and effective interventions that will serve as the basis for human age reversal therapies.

APPENDIX I : METHODS FOR CHAPTERS II-III

CELL LINES UTILIZED IN CHAPTER II

PGP1 hiPSC line

- PGP1 were obtained from the Personal Genome Project. This cell line is male and reprogrammed from primary fibroblasts using Sendai virus-based reprogramming.
- PGP1 cells were utilized between passages 27 to 46 for all experiments.
- Cells were grown at 37°C at 5% CO2 and standard oxygen.
- Cells were maintained on Corning hESC qualified Matrigel, feeder-free. Cells were cultured in Stem Cell Technologies mTeSR Plus.
- Cells were passaged every 3-4 days using Accutase and cultured for 24 hours post-seeding in the presence of Y-27632.

F3 hiPSC line (ATCC BXS0116)

- F3 hiPSCs were obtained from ATCC, catalog: ATCC BXS0116. This cell line is female and reprogrammed from LCLs using Sendai virus-based reprogramming.
- F3 cells were utilized between passages 42 to 56 for all experiments.
- Cells were grown at 37°C at 5% CO_2 and standard oxygen.
- Cells were maintained on Corning hESC qualified Matrigel, feeder-free. Cells were cultured in Stem Cell Technologies mTeSR Plus.
- Cells were passaged every 3-4 days using Accutase and cultured for 24 hours post-seeding in the presence of Y-27632.

M1 - HDFn M1 line (PCS-201-010)

- M1 line was obtained from ATCC Institute, catalog: PCS-201-010. This primary cell line is male and generated from neonatal foreskin.
- M1 cells were utilized between passages 10 to 20 for all experiments.
- Cells were grown at 37°C under hypoxic conditions with 5% CO_2 and 5% O_2
- Cells were maintained in low-glucose DMEM with 15% FBS and 1% pen-strep
- Cells were passaged every 3-4 days using TrypLE-express

M2 – HDFa line (PCS-201-012)

- M2 line was obtained from ATCC Institute, catalog: PCS-201-012. This primary cell line is male and generated from adult skin biopsy.
- M2 cells were utilized between passages 10 to 20 for all experiments.
- Cells were grown at 37°C under hypoxic conditions with 5% CO_2 and 5% O_2
- Cells were maintained in low-glucose DMEM with 15% FBS and 1% pen-strep
- Cells were passaged every 3-4 days using TrypLE-express

For the PGP1 and F3 hiPSC lines, STR analysis was performed for cell line authentication using bioSYNTHESIS to generate STR profiles for both cell lines to ensure authenticity. In addition, all cell lines were regularly tested for mycoplasma contamination.

DEG NETWORK TARGET PREDICTION

The network scoring method was adapted from the Mogrify algorithm[85] with modifications for application to complex phenotypes like aging. RNA-seq datasets were obtained from the Gene

Expression Omnibus (GEO) database, aligned to the hg38 build using STAR aligner[205], and the raw counts were generated using featureCounts[206]. DGEA was performed using DESeq2[102], and the resulting log-transformed fold change, p value, as well as the Pearson correlation of the gene expression with the sample age were combined into a DEGscore. Lastly, to generate a ranked list of candidate genes, we calculated a network score by performing a weighted sum of gene scores over a local gene network constructed from STRING[168], centered on the query gene. The enrichment analysis was performed by performing a hypergeometric test using the top 100 candidates from each ranked list and the GenAge database of experimentally-validated aging genes[84]. The full code and corresponding Juptyer notebooks can be found at: https://github.com/pranam16/STAMPScreen/.

GRN INFERENCE AND PAGERANK TARGET IDENTIFICATION

RNA-seq datasets were obtained from the Gene Expression Omnibus (GEO) database, and log2fc values for each aligned gene for each sample were calculated using the DESeq2 package[102]. Gene regulatory networks were inferred utilizing the GRNBoost2 algorithm in the Arboreto computational framework[108]. PageRank was calculated for each transcription factor in the resulting network via the NetworkX package[207], and ranked factors were visualized using Seaborn. The full code and corresponding Juptyer notebooks can be found at: https://github.com/pranam16/STAMPScreen/.

sgRNA PREDICTION FOR CRISPRa AND CRISPRi

In order to develop highly predictive CRISPRa and CRISPRi tools deep learning models for sgRNA selection, three separate deep learning-based architectures were trained for each task:

a model with only fully connected layers (a fully connected neural network - FCNN), a model with convolutional layers (a convolutional neural network - CNN), and one with recurrent long-short term memory layers (an LSTM model). The suite of neural networks was implemented using Keras, a minimalist, highly modular neural networks library, written in Python (https://keras.io). The Theano library was used as its backend and a Titan GPU was utilized for fast neural network training. The pandas library[208] was utilized to first load and preprocess sgRNA sequences, which were each encoded as a vector of 20 one-hot vectors: A(1,0,0,0), C(0,1,0,0), G(0,0,1,0), and T(0,0,0,1). Hyperas (https://github.com/maxpumperla/hyperas) was utilized to optimize the hyperparameters of each model architecture, and mean squared error was used to select the best model. For the CRISPRa model, activity score dataset of 2898 sgRNAs from 9 CRISPRa screens were collected[121]. A two-layered fully connected layer, with the number of units in the first layer equivalent to the number of sequences in the training set and 810 units in the second layer, followed by a final output layer, proved to have the lowest loss value of all architecture combinations tested. To train an independent CRISPRi model, data from 30 CRISPRi screens consisting of activity scores for 18,380 sgRNAs was obtained[121]. A CNN with a single convolutional layer, filter size of 4, employing l2 regularization, dropout, and a sigmoid activation function, demonstrated the most optimal performance for the held-out test set. To predict highly-active sgRNAs using our models for a given transcription factor gene, the hsEPDnew database for the corresponding promoter sequence was queried[209]. Potential sgRNA sequences, possessing a 5′–NGG–3′PAM within −550 and −25 bp (for CRISPRa) and −25 and 500 bp (for CRISPRi) upstream of the TSS, were used as inputs into the optimized activity models for ranking and downstream experimental testing.

CRISPR TOOL EVALUATION IN hiPSCs

The PGP1 cell line was utilized for all CRISPRa comparisons. dCas9-VPR, dCas9-SAM, and dCas9-SunTag were procured from Addgene (Addgene 63,798, 102,559, 60,903 and 60,904) as gifts from the original labs in which they were generated. sgRNAs were cloned into a common hU6 promoter plasmid. sgRNAs for use with SAM were cloned into a modified sgRNA plasmid that contained the SAM-specific guide RNA scaffold. Experimental methods for assessing CRISPRa performance were as follows:

- Briefly 2 x 10^5 cells hiPSCs were harvested for transfection using the Lonza 4D Nucleofector on setting CM 113 with P3 reagent.
- 100 fmol of each CRISPRa plasmid were combined with 800 fmol of sgRNA plasmid and transfected in duplicate.
- RNA was harvested from the cells at 48 hr post-transfection using the Qiagen RNAeasy kit.
- cDNA synthesis was then performed using the First Strand Synthesis IV kit.
- RT-qPCR was then performed on the duplicate control cells which had only the CRISPRa plasmid transfected and not a sgRNA versus the duplicate wells with a CRISPRa tool and sgRNA targeting one of 47 targets.
- Cq values for the target gene and the housekeeping gene GAPDH were obtained in triplicate and averaged and fold change was calculated using the ΔΔCq method.

Two sgRNAs were first tested for each of the 47 genes in combination with dCas9-VPR and 16 low performing genes were additionally tested with a third sgRNA in all three CRISPRa tools. The best performing sgRNA was then chosen and subsequently used for testing with all three

CRISPRa tools. Nucleofections, RNA harvesting, cDNA synthesis and qPCR were performed for all targets for each of the three CRISPRa tools at the same time to minimize batch effects.

Similarly, PGP1 cells were utilized for all CRISPRi comparisons. dCas9-KRAB and dCas9-KRAB-MeCP2 were obtained from Addgene (Plasmids 110,820 and 110,821) as gifts from the original labs in which they were generated. Screening of the CRISPRi tools was performed in the same manner as the CRISPRa comparison, detailed above, with the only change being that RNA was harvested from cells at 72 hr post-transfection, based on the methods of previous studies. A single sgRNA for each of the 12 targets was utilized for comparing the two tools.

cDNA VECTOR COMPARISON

17 genes targeted in the CRISPRa screen were cloned into a PB-CA vector (Addgene, 20,960) using Gateway cloning. Screening was performed in PGP1 hiPSCs and compared to previously generated CRISPRa data. To perform the cDNA expression comparison to CRISPRa a similar procedure to what is detailed above was performed.

- Briefly, 2 µg of either cDNA vector was nucleofected in duplicate.
- RNA was harvested at 48 hr post nucleofection.
- RT-qPCR was utilized to determine differences in expression of target genes relative to a duplicate control (pMAX-GFP transfection).

For the 5 vector comparison, the following procedure was performed:

- 1 µg of each cDNA vector harboring a sfGFP construct was nucleofected into hiPSCs or HDFs in duplicate.
- Cells were seeded into 1 µg/mL dox media or in no doxycycline media.

- Flow cytometry was performed at 24 hr post-nucleofection and expression was compared to no plasmid control cells.
- Data analysis was performed in FlowJo and mean fluorescent intensity of GFP was graphed using ggplot2.

DUAL cDNA AND CRISPR SCREENING

A superfolder GFP (sfGFP) construct was cloned into the PB-cT3G-MG-cERP2-hU6 plasmid using MegaGate cloning and a sgRNA targeting a DDX4-tdTomato reporter was inserted using Golden Gate cloning for dual cDNA-CRISPRa testing. A mCherry cDNA was cloned into PB-EF1α-MG-U6 and a sgRNA targeting the Tet promoter (sgTET) was cloned into the EF1a vector for dual cDNA-CRISPRi testing. Dual cDNA-CRISPRa screening was performed as follows:

- A F3 hiPSC line harboring a DDX4-TdTomato reporter was utilized.
- 2 x 10^5/well hiPSCs were harvested and transfected as previously described.
- 1 µg of the PB-cT3G-GFP-hU6-sgDDX4 and 1 µg of PB-cT3G-dCas9-VPR were transfected for the cDNA-CRISPRa test.
- 1000 ng/mL of doxycycline was added to the media each day for 48 hr for cDNA-CRISPRa testing.
- Cells for cDNA-CRISPRa were harvested at 48 hr post-transfection for flow cytometry on the CytoFlex LX for sfGFP and tdTomato expression.
- Fold induction was determined by division of the Mean Fluorescent Intensity in the experimental samples versus an uninduced sample and a no plasmid control sample.

Dual cDNA-CRISPRi screening was performed as follows:

- A F3 hiPSC line harboring a stably integrated PB-EF1α-dCas9-KRAB-MeCP2 and a Tet-sfGFP was used for screening
- 1 μg of PB-EF1α-mCherry-sgTET was transfected into an hiPSC line.
- For cDNA-CRISPRi, cells were induced in 1000 ng/mL doxycycline for 108 hr.
- For cDNA-CRISPRi, RNA was harvested at 108 hr post-induction.
- Flow cytometry was performed on the CytoFlex LX for sfGFP and mCherry expression.
- Fold induction of mCherry was determined by division of the Mean Fluorescent Intensity in the experimental samples versus an uninduced sample and fold suppression of sfGFP was determined by division of the Mean Fluorescent Intensity in the experimental samples versus an a dox induced sample with no plasmid.

MEGAGATE MOLECULAR CLONING VECTOR CONSTRUCTION

A common MegaGate backbone plasmid was created for doxycycline induction based on the XLone-GFP plasmid (Addgene 96,930), a gift from the original lab in which it was generated. Modifications to this plasmid include switching blasticidin selection for puromycin, the addition of a CMV enhancer driving the EF1α promoter, replacing the minimal RNA Poll II Tet promoter with a minimal CMV promoter, replacing GFP with a megagate cassette, and an addition of a DNA barcode region all using Gibson Assembly. This core vector is termed PB-cT3G-MG-cERP2. Additionally, an hU6 promoter and sgRNA scaffold were added to the PB-cT3G-MG-cERP2 plasmid to create the plasmid PB-cT3G-MG-cEPR2-hU6. Additionally, an EF1α MegaGate vector was created by modification of Addgene 104,543. BsaI and SapI recognition sites were removed and the Tet-On-3G was replaced with a Megagate cassette, in addition a DNA

barcode region was added to make the plasmid we term PB-EF1α-MG. An hU6 promoter and sgRNA scaffold were added to the PB-EF1α-MG plasmid to make the plasmid PB-EF1α-MG-hU6. sfGFP versions of all plasmids were tested in hiPSCs and primary human fibroblasts to ensure proper function. The Megagate cassette in each vector is created by synthesis of a DNA construct containing an attR1 sequence, 15 bp of intervening sequence, an I-SceI recognition site, 45 bp of intervening sequence, an I-CeuI recognition site, 15 bp of intervening sequence and an attR2 sequence. Additional MegaGate cassettes were tested that contained other enzyme recognition sites such as BsaI, as well as multiple meganuclease restriction sites.

MEGAGATE REACTION CONDITIONS OPTIMIZATION

All MegaGate reaction optimization conditions were performed on the above listed plasmids. The pENTR insert used in the optimization reactions was a pENTR-sfGFP. For negative control reactions (no insert) the pENTR was replaced with water. The 99.8% efficient MegaGate reaction is performed as follows:

For MegaGate LR reaction assemble on ice:

- 24 fmol or 50 ng of pENTR insert
- 16 fmol or 75ng of MegaDestination vector
- 1 µL of the Gateway LR Clonase II Enzyme mix (Invitrogen 11,791,020)
- 1 µL (5U) of the MegaNuclease I-SceI (NEB R0694)
- 1 µL (5U) of the MegaNuclease I-CeuI (NEB R0699)
- 5 µL of 10X CutSmart Buffer (NEB B7204).
- Nuclease free water was then utilized to bring the reaction to the desired volume of 50 µL.

For MegaGate BP reaction assemble on ice:

- 16 fmol or 10 ng of AttB flanked PCR insert
- 16 fmol or 75ng of MegaDONOR221 vector
- 2 µL of the Gateway BP Clonase II Enzyme Mix (Invitrogen 11,789,100)
- 1 µL (5U) of the MegaNuclease I-SceI (NEB R0694)
- 1 µL (5U) of the MegaNuclease I-CeuI (NEB R0699)
- 5 µL of 10X CutSmart Buffer (NEB B7204)
- Nuclease free water was then utilized to bring the reaction to the desired volume of 50 µL.

For both the MegaGate LR and MegaGate BP reactions, place MegaGate reaction into thermocycler with the following settings:

- 1 hr at 25°C
- 1 hr at 37°C
- 20 min at 65°C.

2 µL of MegaGate reaction was then transformed into 10 µL of NEB 5-alpha cells and colonies were picked and counted the following day for sequencing.

In practice, pENTR amounts as low as 10ng and as high as 150ng yielded successful MegaGate LR products. Similarly, PCR inserts as low as 2ng and as high as 150 ng yielded successful MegaGate BP products. Additional meganuclease was found to lead to negligible increase in efficiency, but decreased meganuclease resulted in lower efficiency. Additionally, lower

reaction volumes of 20 µL were found to likewise decrease reaction efficiency. Splitting the reaction into two separate steps, one with only Gateway components incubated at 25°C for 1 hr then adding MegaNucleases and CutSmart and incubating at 37°C for 1 hr and 20 min at 65°C was was shown to slightly increase colony number with the same cloning efficiency. However, for ease of use, we found the all-in-one reaction to be simpler and sufficient for all cloning reactions. Additionally, isothermal reactions at 25°C and 37°C, single meganuclease reactions cutting for 1, 2, and 3 hr and scaled down 25 µL were tested and show varying albeit lower efficiencies. Multiple MegaGate destination vectors and different inserts of varying length were additionally tested to demonstrate that the reaction is efficient across ORF sizes and destination vectors, and that the colony number scales linearly down with increase in ORF size. Reaction efficiency was measured as the total number of positive plate colonies minus the number of colonies on no-insert plates divided by the total number of positive plate colonies. Sanger sequencing was performed on 100 colonies per plate to validate reaction efficiency.

MEGAGATE VECTOR DNA BARCODING

For DNA barcoding, a 40,000 + member 20 bp barcode library was used as a PCR template. Primers were utilized that added Golden Gate enzyme compatibility with the DNA barcode region of the MegaGate destination vectors. A Golden Gate reaction was then performed between the UMI PCR products and the MegaGate destination vectors. The resulting plasmids were then transformed in mass onto multiple selection plates. Colonies were then scrapped and midiprepped to obtain the barcoded MegaGate Destination pool. The resulting barcode distribution in the MegaGate pool was analyzed via NGS using the Illumina MiSeq platform. Briefly, barcoded destination plasmids were used as PCR templates and primers amplified the

barcode region and added Illumina adaptors. Resulting reads were then aligned and barcodes identified using the Geneious alignment software package.

SINGLE AND POOLED MEGAGATE CLONING AND NGS READOUT

For single gene MegaGate cloning, 50 ng of pENTR plasmid was utilized in all 185 MegaGate cloning reactions along with 75 ng of barcoded MegaGate destination vectors. Colonies were subsequently picked and Sanger sequencing was used to link a specific DNA barcode with the inserted ORF.

For pooled MegaGate cloning, 300 ORFs were selected from the ORFeome and pooled for transformation. Transformants were then scraped and midiprepped to yield a pENTR pool. 50 ng of the pENTR pool was then utilized for the MegaGate reaction along with 75 ng of barcode MegaGate destination vector. Transformants were then scraped and midiprepped the following day to yield the barcoded ORF pool in the expression vector. To determine the resulting cloning efficiency and to assign barcodes to the ORFs, primers were utilized that bound upstream of the ORF and downstream of the ORF in the pENTR vector. Additional primers were used that bound a constant region upstream of the inserted gene in the expression vector and the 3′end of the DNA barcode and Illumina adaptors were added. The pENTR and expression plasmid amplicons were then run on an Illumina MiSeq micro kit. STAR aligner was then utilized to identify ORFs in the pENTR and expression plasmid pools to determine the percentage of ORFs in the pENTR library that were converted into the expression vector. Biopython's fastQ parser was then utilized to identify barcodes to match to the corresponding ORF in the expression pool (https://github.com/biopython/biopython). Cloning efficiency is calculated as the TPM

normalized abundance of the ORF in the expression pool divided by the TPM normalized abundance of the ORF in the donor pool, calculated from the NGS counts.

PIGGYBAC COPY NUMBER TITRATION

For targeted copy number integration in hiPSCs and NHDFs, we developed a high throughput RT-qPCR readout for average copy number in a given cell pool. Utilizing qPCR primers amplifying a constant region on the integrated transposon (Table A2.1) in reference to the RNASEP gene (RPP30) (Table A2.1) which is known to have two autosomal copies per human genome, copy number was assessed in human cells. To determine piggyBAC-based integration dynamics the following procedure was performed:

- A super PiggyBac transposase vector vector driven by a CMV promoter (CMV-SPB) was nucleofected at 100 ng per reaction (24 fmol) using the Lonza 4D Nucleofector into F3 hiPSCs on setting CM 113 in P3 solution and HDFs on setting DS 150 in P2 solution.
- A sfGFP vector (PB-cT3G-sfGFP-cERP2) was serially titrated from 400 fmol to 1.2 fmol and co-nucleofected with CMV-SPB.
- Cells were then drug-selected for 13 days on puromycin (400 ng/mL in hiPSCs and 500 ng/mL in NHDFs) using a 3 day ON and 3 day OFF drug selection regimen to ensure pool purity and loss of any unintegrated plasmid.
- RT-qPCR using SYBR Green master mix was then performed after gDNA extraction using the DNAeasy kit.
- 10 ng of input gDNA was used per reaction based on the standard curve, with an anneal temperature of 60°.
- To calculate copy number, the $2^{\Delta Cq+1}$ method was used, with RNASEP as a reference.

- The resultant value was multiplied by two to account for the two autosomal copies of RPP30.

gDNA BARCODE ENRICHMENT ANALYSIS (BAR-SEQ)

BAR-Seq was performed using the F3 hiPSC line and an equimolar pooled plasmid library of 40 TFs with three unique barcodes each and a GFP vector with 9 unique barcodes (129 unique total plasmids) as follows:

- The TF library was co-nucleofected at 4fmol with the super PiggyBac transposase vector at 24 fmol using the Lonza 4D Nucleofector into F3 hiPSCs on setting CM 113 in P3 solution
- Copy number was estimated to average 5 copies per cell using the methods described above.
- The drug selected pure population was then induced with doxycycline at 500 ng/mL for two days and an uninduced sample was used as control.
- The cells were then sorted using a Sony cell sorter (SH800S) based on GFP expression.
- The FACS positive and negative pools were then gDNA extracted using the DNAeasy kit.
- 5 ng of input gDNA was then used to amplify the integrated barcodes using two rounds of PCR. The first round added illumina R1 and R2 adaptors using primers 1.25 µL of 10 µM BC-NGS-Fwd and BC-NGS-Rev primers. Cycling was done in a qPCR machine using KAPA SYBR mix and samples were removed when they reached their Cq to prevent over amplification.

- The cycling conditions were as follows: 95°C-3′, Cq-determined x[98°C-10”; 58°C-20”; 72°C-30”].The samples were then bead purified using Promega Size-Selective beads at 1.75X for a 185 bp product and eluted in 20 µL.
- 1ul of PCR1 input was then used for a second round of PCR amplification with 2X Q5 master mix and 1.25 µL of 10 µM Illumina index primers.
- Cycling was done with the following conditions: 98°C-30″, 8x[98°C-5”; 61°C-20”; 72°C-5”], 72°C-2′and the 255bp product was again bead purified using 1.3X chemistry.
- The indexed samples were then run on NGS using an Illumina MiSeq machine at 20 pM with 10% PhiX spike in.
- The counts of each barcode were then generated using Geneious RNA-Aligner against a reference list of barcodes and normalized by the total counts in the sample to measure relative abundance.
- Data was then plotted using the Prism software.

TARGETED RNA-SEQ (TAR-SEQ)

TAR-Seq was performed using the F3 hiPSC line and an equimolar pooled plasmid library of 40 TFs with three unique barcodes each and a GFP vector with 9 unique barcodes (129 unique total plasmids) as follows:

- RNA was extracted from doxycycline induced and uninduced pools from the same experiment as the above detailed Bar-Seq.
- For each sample, cDNA was generated from 1 µg total RNA using the NEB Luna RT SuperMix kit with random hexamers.

- The targets of interest were amplified in a 25 µL reaction using the NEB 2x Q5 mastermix with 2 µL 1:10 diluted cDNA, 0.1 µM final concentration for each primer pair (Table A2.1),
- the following cycling conditions were used: 98°C-5′, 26x[98°C-1'; 60°C-30"; 68°C-15"], 72°C-3'.
- The PCR product was gel purified, prepared for NGS using the same workflow as previously detailed in BAR-Seq, and sequenced on a MiSeq machine at 20 pM with with a 10% PhiX spike in.
- Counts were then generated in Geneious using RNA-Aligner against a reference list of full length CDS sequences and barcodes.
- Counts were then normalized per sample and further normalized to GAPDH in the sample for TAR-Seq analysis to calculate relative expression.

For the TAR-Seq validation experiments, the total cell RNA from four different primary NHDF lines was converted to cDNA using the Thermo SuperScript IV First Strand Synthesis system with random hexamers. The TAR-Seq protocol was then performed as described above.

RNA-SEQ AND COUPLED BARCODE ENRICHMENT

The F3 hiPSC line was utilized for the RNA-Seq and barcode capture experiment as follows:

- Cells were nucleofected as described above with 100fmol of a barcoded cDNA vector harboring sfGFP or the gene ZGLP1 and 24 fmol of super piggyBac transposase in duplicate.
- Integrants were purified through puromycin drug selection as detailed above

- The cells were then induced under 1 μg/mL doxycycline for 3 days and RNA was harvested using the Qiagen RNAeasy plus kit.
- RIN scores were determined using a Bioanalyzer and only samples with RIN score greater than 8 were utilized.
- An RNA-Seq cDNA library was constructed using the KAPA Hyper Kit with RiboErase
- RNA-Sequencing was performed on the Next-Seq 500 Illumina platform.
- Reads were processed and aligned to the hg19 build using STAR Aligner
- DEGs were determined using DESeq2.
- Gene barcodes were identified using BBMap.
- DESeq2 results for gene expression log2fc and their matching p values were plotted utilizing the R package ggplot2.

QUANTIFICATION AND STATISTICAL ANALYSIS

Specific quantification and statistical analysis details for each experiment can be found in the figure legends. Statistical analysis was performed using the software GraphPad. For Figure 2.1, p values for DEGs are determined by the DESeq2 software package. For Figure 2.2 perturbation tool comparisons, n = 2 biological replicates were utilized in which the same hiPSC line was independently nucleofected and screened for each condition. Statistical analysis is performed using the Mann Whitney Test with significance determined at p value less than 0.5 threshold. Indicated values are given as median of the population for each perturbation tool. For dual cDNA-CRISPR comparisons, n = 3 biological replicates, where replicates are independently nucleofected hiPSCs. Values are presented as mean with standard deviations of replicates. For Figure 2.3 MegaGate cloning outcomes, values in Figure 2.3 B are given as the mean with standard deviations of cloning

efficiency and barcodes captured for the gene. For single gene cloning n = 185 and for pooled cloning reactions n = 300 where n represents a single cloned gene. For Figure 2.4 NGS Coupled readouts, copy number was determined in n = 2 biological replicates for hiPSCs and HDFs. Biological replicates are independent hiPSC and HDF lines that are independently nucleofected in duplicate, drug selected and screened. Data is plotted as the mean copy number with standard deviation of the technical replicates across the two biological replicates. For RNA-Seq coupled to barcode enrichment, p value is determined by the DESeq2 software package. For TAR-Seq, significance is determined by Mann Whitney test and significance threshold is set at p value less than 0.05. n = 40, where n represents the expression of a single gene in the sample.

For Chapter 3 statistical analyses were performed with R version 4.0.3, using two-tailed Student's t-tests, or a Z-score. All of the statistical tests performed are indicated in the figure legends. The data are presented as mean or individual points with box plots that show median and quartiles.

RNA-SEQ DATA PROCESSING AND ANALYSIS

Raw RNA-Seq reads were aligned to the GRCh38 human genome using STAR v2.5.2b[205] and checked for quality control using FastQC v0.11.5[210]. The alignment files were then indexed using SAMtools v1.3.1[211] and mapped reads were counted using featureCounts from the Subread v2.0.1 package[206]. To reduce the variation between our samples caused by disproportionate sequencing depth, we downsampled the fastq files to 20M reads using Seqtk-1.3[212]. For the overexpression analysis, to address the high number of discarded multi-mapped reads due to the sequence similarity between the endogenous genes and our barcoded genes (Table A3.3), we aligned our reads to a genome index that contains our transgene sequences (including their

barcodes) using kallisto v.0.46.2[213], and we used the tximport[214] package to generate TPM from estimated counts. Barcode search was performed on the downsampled fastq files by using the agrep[215] tool for approximate string matching with 3 mismatches. Subsequent differential expression analysis was performed using DESeq2 v1.32.0[102] unless otherwise noted.

DIFFERENTIAL GENE EXPRESSION ANALYSIS

Differential gene expression analysis was performed using the DESeq2 R package v1.30.1[102]. In particular, DESeq data objects were constructed from raw, untransformed read counts and a design formula reflecting the attempted comparisons. In particular, to detect differentially expressed genes in response to perturbation effects within the screening data, samples of each line were separated and individually processed. Differentially expressed genes were obtained by first running the "DESeq" function with standard parameters followed by the "contrast" function to obtain differentially expressed genes between overexpressed genes and wildtype or blue fluorescent protein (BFP)-transduced controls. For each gene, computed p-values were corrected by applying Bonferroni correction and significance was determined at the 1% level.

DIFFERENTIAL SPLICING ANALYSIS

Differential splicing analysis was performed using SUPPA2[216]. In particular, we first invoked the 'generateEvents' subcommand of SUPPA on the GENCODE V29[217] annotation and parameter setting '-e SE SS MX RI FL -f ioe' to generate all potential splicing events. Next, we next generated the proportion spliced-in (PSI) for all potential events by invoking the 'psiPerEvent' subcommand on our transcript-aligned, large-scale screening assay (including 6 NHDF lines with SRSF1 over-expression) RNA-seq data in TPM format. Finally, we invoked the 'diffSplice'

command with an empirical model ('-m empirical') to compare changes between wildtype and SRSF1 over-expressing lines. Automatic gene correction of p-values is applied by supplying the '-gc' parameter.

DECOMPOSITION OF BULK RNA-SEQ

RNA-seq samples are deconvoluted into individual cell types using CibersortX[218]. First, droplet-based single-cell reference datasets of stromal cells have been collected from Tabula Sapiens[166]. Subsequently, for each population, 100 cells were randomly sampled and combined into a single matrix. A signature matrix containing 300 to 500 genes per cell type was computed using CibersortX. RNA-seq counts of the initial perturbation screen served as an input for deconvolution. CibersortX was run in "absolute mode" to allow for comparison between samples.

IDENTIFICATION OF CELLULAR PROCESSES PREDICTIVE OF AGE

The identification of processes that are predictive of chronological age is based on the Molecular Biology of the Cell (MBotC) Ontology, a previously curated resource relating genes to cellular processes[142]. The MBotC Ontology is organized in four different layers where broader categories are iteratively split into more and more specialized functions. Therefore, we first empirically selected the third layer and created a set of processes that are potentially informative about aging. In order to quantify the predictability of chronological age from these processes, we built "weak" age predictors for each process as follows: (1) Subset the gene expression matrix to genes in the process under consideration; (2) Perform principal component analysis (PCA) on the subset matrix and replace it with a matrix containing all principal components. PCA is performed using the implementation in the h2o v3.36.0.2 R package[219] and imbalanced contribution of

individual genes is alleviated by standardizing the data through the 'transform' parameter; (3) Train a generalized linear model (GLM) with 5-fold cross-validation using the 'h2o.glm' function in the h2o R package. In particular, the model is based on Gaussian distributions with an identity link function, which standardizes the input before training and implements automatic lambda search with ridge regression; (4) Train a "strong" age predictor based on the in-bag predictions of all weak predictors. Like weak predictors, the strong age predictor is a GLM with 10-fold cross-validation that is based on Gaussian distributions, an identity link function and standardization of the input. However, in contrast to the weak models, the strong predictor employs Lasso regression to select only a minimal set of processes that are most informative of chronological age. The combination of processes with non-zero coefficients is considered to be predictive of age.

Importantly, the identification of cellular processes that are predictive of age requires assembling a training dataset in which the chronological age of each sample is *a priori* known. In this regard, both the weak and strong age predictors are not trained based on ages expressed in years. Instead, the age is transformed by a piecewise, approximately linear function defined as:

$$f(x) = \begin{cases} \left(\left(\log\left(\frac{1}{71} - \frac{1}{101}\right) \cdot \log 25\right) - \left(\log\left(\frac{1}{70} - \frac{1}{101}\right) \cdot \log 25\right)\right) \cdot (x - 70), & x \leq 70 \\ 3 \cdot \left(\left(\log\left(\frac{1}{71} - \frac{1}{101}\right) \cdot \log 25\right) - \left(\log\left(\frac{1}{70} - \frac{1}{101}\right) \cdot \log 25\right)\right) \cdot (x - 70), & x > 70 \end{cases}$$

The function is a suitable transformation, since its inverse has the following closed form that can be used to map the predicted ages back to actual ages:

$$f^{-1}(x) = \begin{cases} \dfrac{-x + 70 \cdot \log\left(\frac{7070}{31}\right) - 70 \cdot \log\left(\frac{7171}{30}\right)}{\log\left(\frac{7070}{31}\right) - \log\left(\frac{7171}{30}\right)}, & x \leq 0 \\ \dfrac{-x + 210 \cdot \log\left(\frac{7070}{31}\right) - 210 \cdot \log\left(\frac{7171}{30}\right)}{3 \cdot \left(\log\left(\frac{7070}{31}\right) - \log\left(\frac{7171}{30}\right)\right)}, & x > 0 \end{cases}$$

TRAINING OF PROCESS-BASED TRANSCRIPTOMIC CLOCKS

The transcriptomic clock introduced in this study resides on a set of cellular processes that are predictive of age, as described in the previous subsection. In particular, training of the clock follows a two-step process. First, for each process, the most predictive set of genes is selected by starting from a model with only an intercept, iteratively adding or removing a gene and scoring each iteration using Akaike's Information Criterion[220] by performing linear regression between the current gene set and sample age. In total, 1000 iterations are performed for each process. Second, a GLM, implemented in the h2o v3.36.0.2 R package[219], is trained on all selected genes of all processes that were selected in the first step. More specifically, the GLM employs Gaussian distributions with an identity link function, lambda search and ridge regression on the standardized gene expression data. Since the number of genes selected in the first step may be larger than the number of training samples, an upper bound on the number of active predictors in the GLM was set to the number of training samples.

PREDICTING THE AGE OF NON-TRAINING SAMPLES

Due to the variability in raw RNA-seq data, which is related to sequencing depth, library preparation and other experimental confounding factors, it is necessary to pre-process the training and non-training samples together to detect and correct batch effects. Therefore, for a set of new non-training samples (Table A3.5), we employed a common pre-processing pipeline. First, training and non-training samples are merged into one matrix and TMM[221] normalized using the 'tmm'-function of the NOIseq v2.34.0 R package[222]. Secondly, in case of significant batch effects visible in the first two principal components, batch correction is applied between the training and non-training samples. For the majority of datasets in this work, we employed the naïve Removal of

Unwanted Variance (RUV)[223] algorithm from the RUVnormalize v1.24.0 R package. RUV requires the selection of an appropriate set of control genes that are expected to be uncorrelated to the variable of interest, i.e. age. Thus, we selected appropriate controls as having a low correlation with age in the training data (Pearson's $r < 0.01$). Nevertheless, depending on the observed batch effects, other dataset-specific control genes should be included (see individual scripts for the batch correction methods and parameters employed for each dataset).

After normalization and batch correction, the transcriptional clock was trained on the training data as described in the previous section. The age of non-training samples was then predicted using the 'h2o.predict' function in the H2O R package[219] on the trained model.

COMPUTING PROCESS ACTIVITY SCORES

Using the estimated coefficients of our transcriptomic clock, we quantified the activity of all eight age-associated processes in a sample by computing the scalar product of the model coefficients and the expression values of the corresponding genes. In case of multiple replicates, the activity scores of the same processes in different samples are aggregated into their arithmetic mean. Since RNA-seq data has been shown to be sensitive to the preprocessing pipeline employed to transform raw read counts, we computed a reference process activity for all samples of our training data. These reference activities define a range of values for each process that correspond to physiological aging and are employed to uniformly scale the activity scores of new samples. Due to the generalized linear model underlying our clock, lower process activity scores correspond to lower whereas higher values correspond to higher transcriptional age.

SCREENING LIBRARY CONSTRUCTION

Q5 high-fidelity 2X master mix (NEB M0492S) was used to amplify all of the ORFs from their original vectors (Addgene or ORFeome) in order to add attB sites, the Kozak consensus sequence "GCCACC", and the WT STOP codon. The amplified fragments were gel purified (QIAGEN 28506) and shuttled into pDONR221 (ThermoFisher 12536017) using the BP Clonase II enzyme mix (ThermoFisher 11789020). The reactions were transformed into 5-alpha competent E. coli (NEB C2987H), clones were picked and sequence confirmed using Sanger sequencing. The resulting plasmids were miniprepped (NEB T1010L) and reacted with a barcoded pool of destination vectors (PB-CT3G-ERP2-MG-BC) in a MegaGate reaction. After transforming into 5-alpha cells, clones were picked, sequence confirmed, and barcodes were assigned to specific ORFs (Table A3.3). Final plasmids were miniprepped and used for nucleofection.

TISSUE CULTURE

NHDF lines (Table A3.4) were cultured at 37 °C, 5% CO_2, 5% O_2 in fibroblast media (FM): low glucose DMEM (ThermoFisher 11885-084) supplemented with 15% FBS (GenClone 25-550) and 1% Penicillin-Streptomycin (ThermoFisher 15140122). When induced with Doxycycline, cells were switched to media made with Tet System Approved FBS (Takara 631367) to reduce background induction. Media was changed every other day.

For generating the overexpression lines, 200,000 cells were nucleofected with 50 fmol transposon and 50 fmol transposase (Super piggyBac Transposase - SystemBio PB210PA-1) or with 300 ng pmaxGFP using the P2 Primary Cell 4D Nucleofector kit (Lonza V4SP-2096) and the Lonza 4D-Nucleofector with the DS-150 program. Cells were recovered at room temperature for 45 min and then plated to a 24 well plate in 500 uL FM. The following day after nucleofection,

dead cells were washed with PBS and media was replenished. 3-4 days post-nucleofection (depending on cell confluency) selection was started using 400 ng/mL Puromycin (ThermoFisher A1113802) in FM (PFM). Cells were selected and expanded at the same time, switching between FM and PFM every 4 days, until the pmaxGFP control reached 0% viability. At the 10 cm dish stage, the cells were frozen in FM with 5% DMSO and stored in liquid nitrogen until thaw. Due to cell line specific sensitivity to the nucleofection and selection process, as well as the increased toxicity of plasmids harboring larger genes, we couldn't generate overexpression lines for all our 95 genes in all three primary NHDF lines.

When assayed, cells were thawed into 2 wells of a 6 well plate in FM, and switched to the Tet-free media the next day. 2 days after thaw, cells were harvested, counted, and seeded for RNA-Seq and flow cytometry. For RNA-Seq, each cell line was seeded into 2 wells of a 6 well plate at a density of 60,000 cells/well in FM supplemented with Doxycycline (1 ug/mL for the initial screen and 2ug/mL for subsequent experiments). For flow cytometry, each line was seeded into 18 wells of a 24 well plate at a density of 10,000 cells/well in FM (the +Dox wells received Doxycycline at 1 ug/mL). 72 h post seeding, the cells were washed with PBS and either stained and harvested for flow cytometry or lysed using the Monarch DNA/RNA Protection Reagent (NEB T2011L) and stored at -80 °C.

RNA SEQUENCING

RNA was extracted from the cell lysates using the Monarch Total RNA Miniprep kit (NEB T2010S) and its quality was spot-checked for random samples using the Bioanalyzer High Sensitivity RNA Kit (Agilent 5067-1513). 100 ng of total RNA was quantified using the Qubit RNA High Sensitivity Assay (ThermoFisher Q32852) and used for library preparation with the

NEBNext Ultra II Directional RNA Library Prep Kit for Illumina and the polyA mRNA workflow. The library quality was spot-checked for random samples using the Bioanalyzer High Sensitivity DNA Kit (5067-4626), and then all libraries were quantified using the Qubit dsDNA High Sensitivity Assay (ThermoFisher Q33230) and pooled together. RNA extraction, as well as library preparation, was performed on the same day for all samples that were sequenced together. Sequencing was performed by the Harvard Biopolymers Facility on an Illumina NextSeq or NovaSeq instrument.

FLOW CYTOMETRY

Cells were stained for cellular senescence, mitochondrial membrane potential, and proteasome activity. For cellular senescence, cells were incubated with 100 nM Bafilomycin A1 (VWR 102513) for 2 h and with 33 nM C_{12}FDG (5-Dodecanoylaminofluorescein Di-β-D-Galactopyranoside; ThermoFisher D2893) for 1 h. Mitochondrial membrane potential and proteasome activity were multiplexed by incubating cells with 20 nM or 40 nM (M65: [46, 55, 56, 57, 60, 64, 72, 77, 82, 94, 95]) TMRM (ThermoFisher M20036), 0.125x proteasome LLVY-R110 substrate, and 0.0625x assay buffer (Millipore Sigma MAK172) for 2 h. Samples were analyzed using either a Cytoflex LX or BD LSRfortessa instrument. Analysis was done using FlowJo (Version 10.8.1).

IN VITRO SCRATCH ASSAY

Cells from the M79 line harboring overexpression cassettes for SRSF1, OSKM, or mTagBFP2 were grown in 10 cm dishes and treated with doxycycline (1 ug/mL) for 3 days. Cells were subsequently harvested and seeded into 24 well plates at a density of 120,000 cells per well.

The next morning, plates were scratched with a p200 pipette tip followed by a PBS wash. Plates were imaged using the Cellcyte live-cell imaging system (Cytena) with a 10x objective at an interval of 1.5 hours.

Scratch images were stitched, processed, and analyzed as virtual stacks for each time course using Fiji[224,225]. After cropping to the scratch area, the image backgrounds were subtracted using a 10-pixel rolling ball radius and contrast enhanced to 10% saturated pixels, normalized for all images in the stack. The Python package Bowhead v1.1.3[226] was used to identify the largest contiguous wound area in the images using a threshold of 0.5.

CODE AVAILABILITY

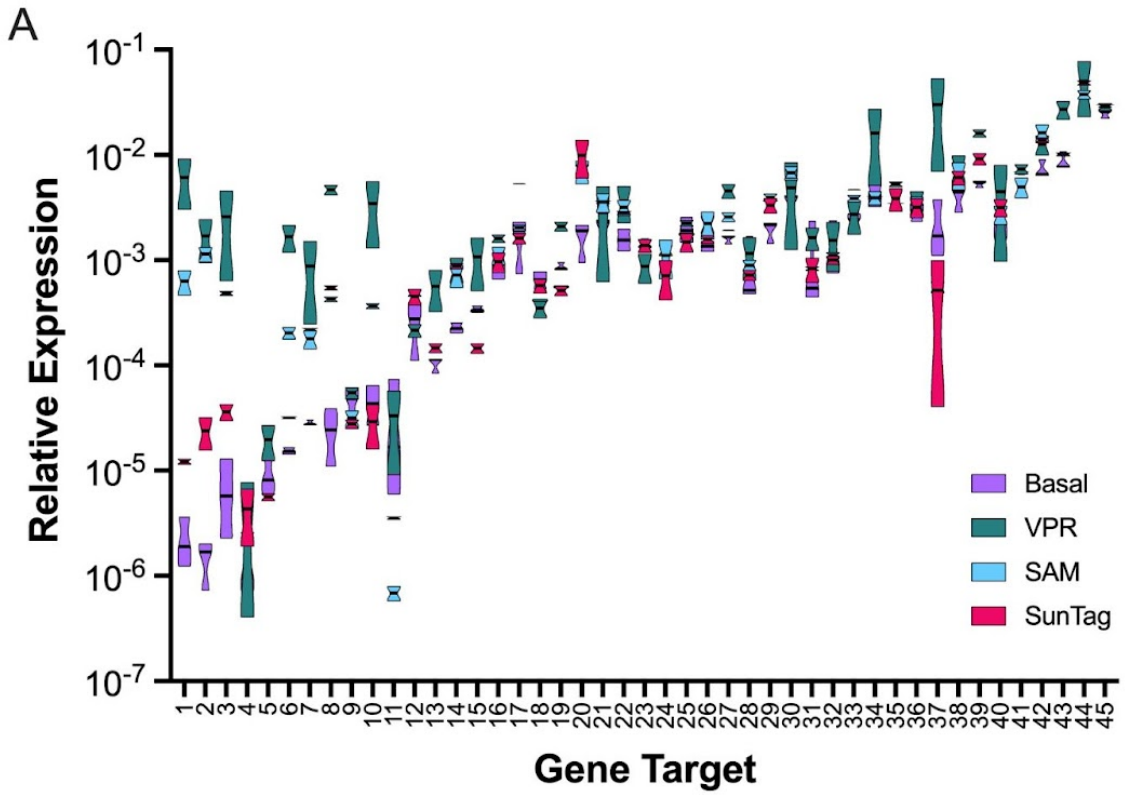
All original code for computational methods for target identification has been deposited on a Github repository (https://github.com/pranam16/STAMPScreen/), while the source code for transcriptomic clock analyses is available at https://github.com/saschajung/Clock.

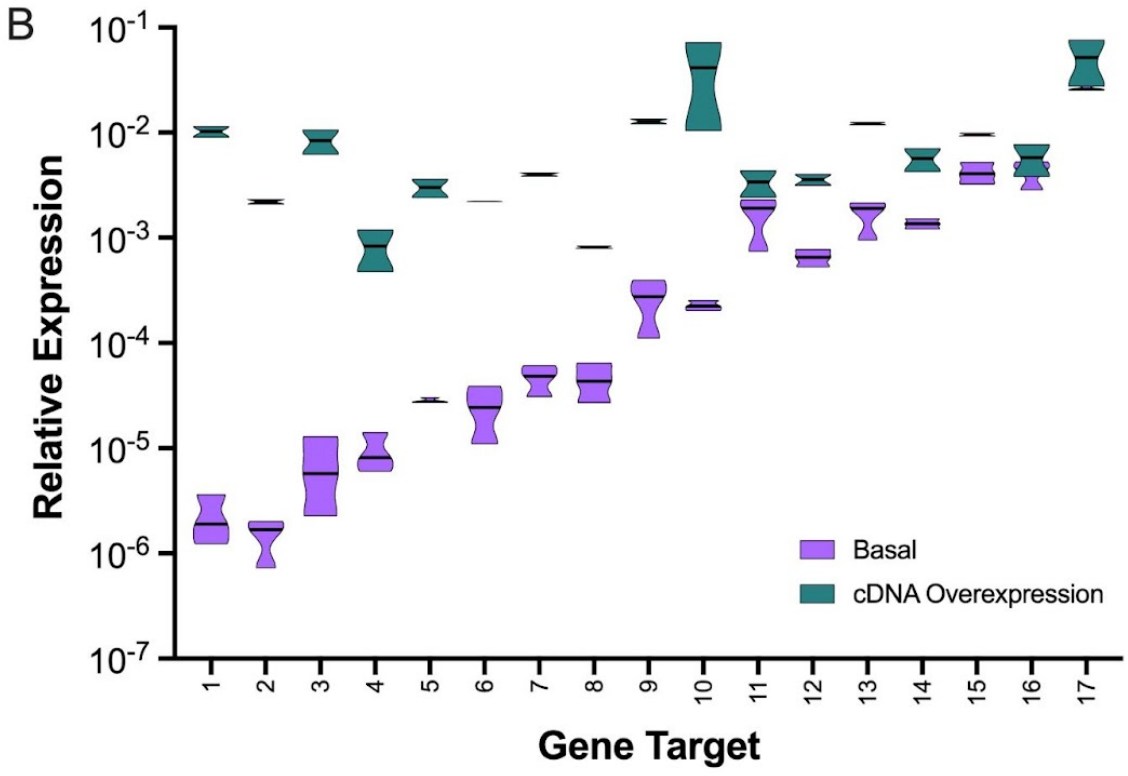
APPENDIX II : SUPPLEMENTAL FIGURES AND TABLES FOR CHAPTER II

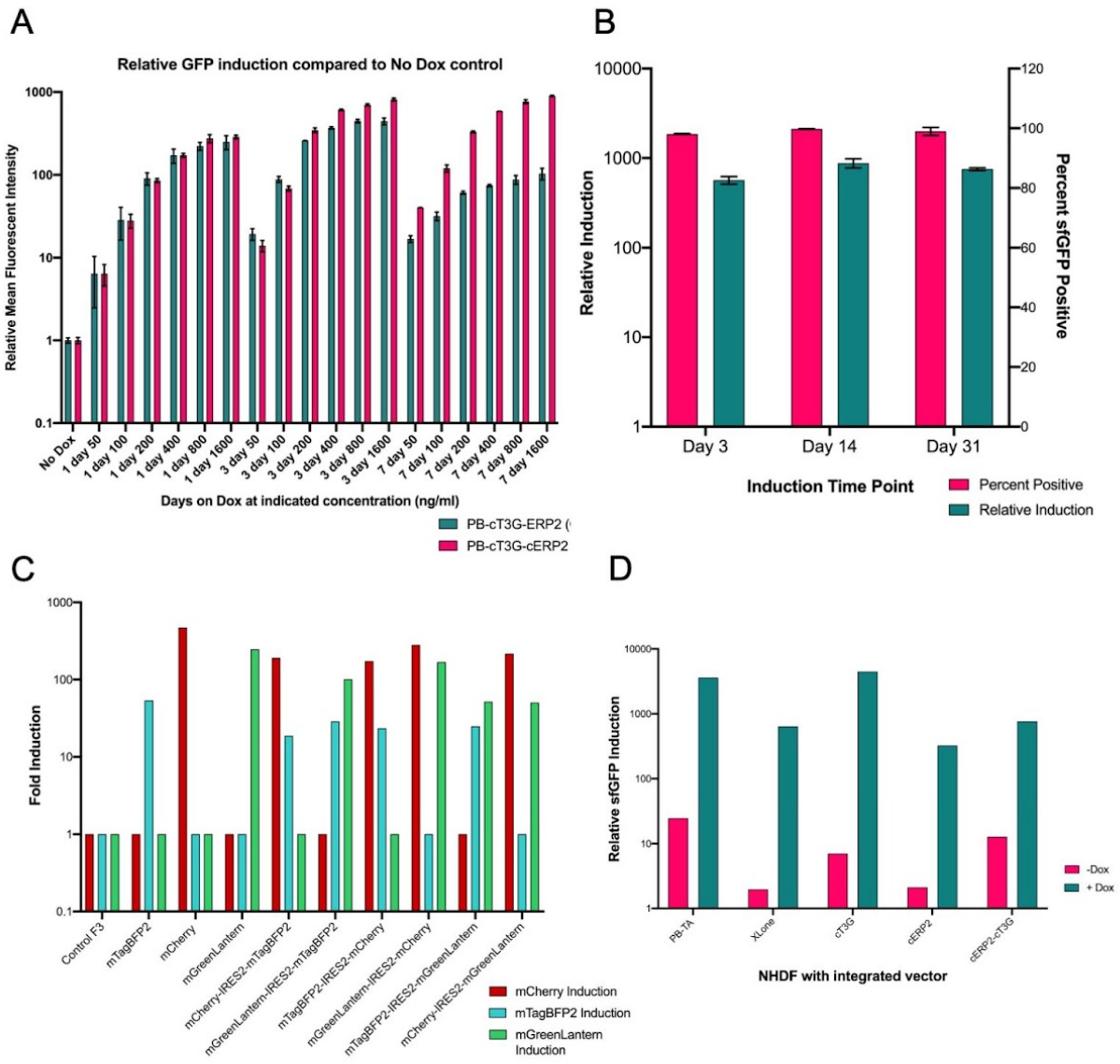
Figure A2.1: Relative expression of gene targets, Related to Figure 2.3.

RT-qPCR was performed to determine the expression for 45 genes (A) or 17 genes (B) after 48-hour induction via CRISPRa (A) or cDNA (B) in biological duplicates. Relative expression is calculated using ΔΔCq method relative to GAPDH. The basal gene expression level is calculated from a no induction vector control condition.

Figure A2.1 Continued







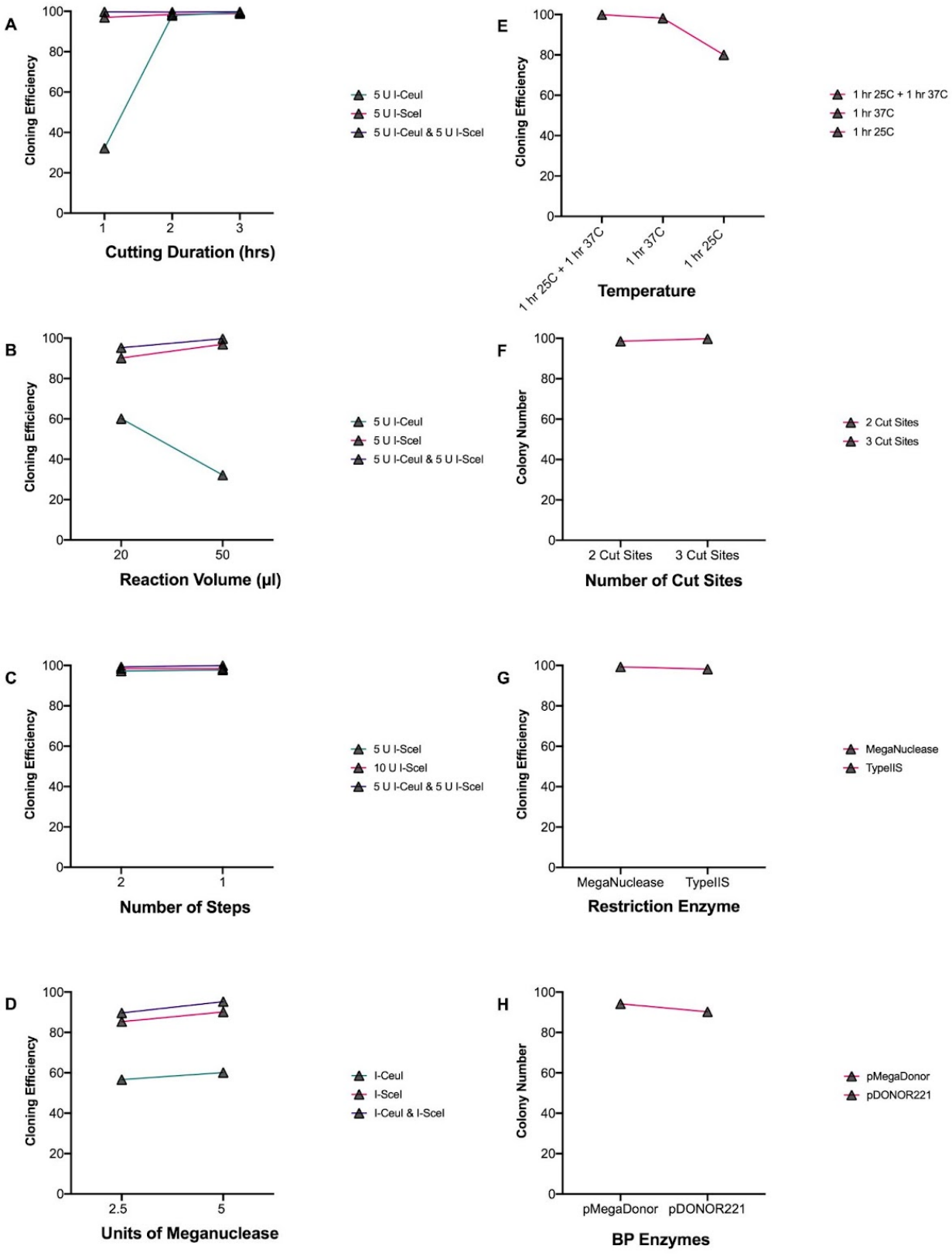
Supplementary Figure A2.2: cDNA expression vector dynamics were compared using flow cytometry, Related to Figure 2.3.

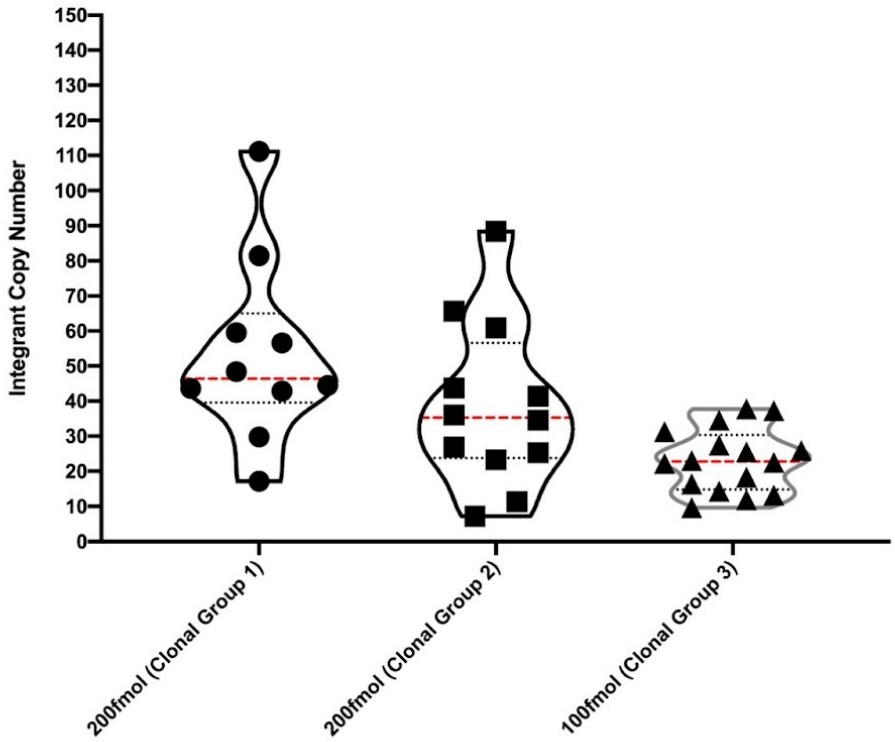
Relative expression was calculated as the relative mean fluorescent intensity of the fluorescent construct compared to a basal no induction condition. A) Expression dynamics under dox titration at three time points B) Stable vector expression over one month in hiPSCs. C) dual expression of IRES2-Fluorescent proteins D) Expression comparison of five vectors in NHDFs after integration and selection.

Supplementary Figure A2.3: Optimization reactions for MegaGate Cloning, Related to STAR Methods.

Cloning efficiency is determined as the number of colonies on the plus insert plate compared to the total number of colonies.

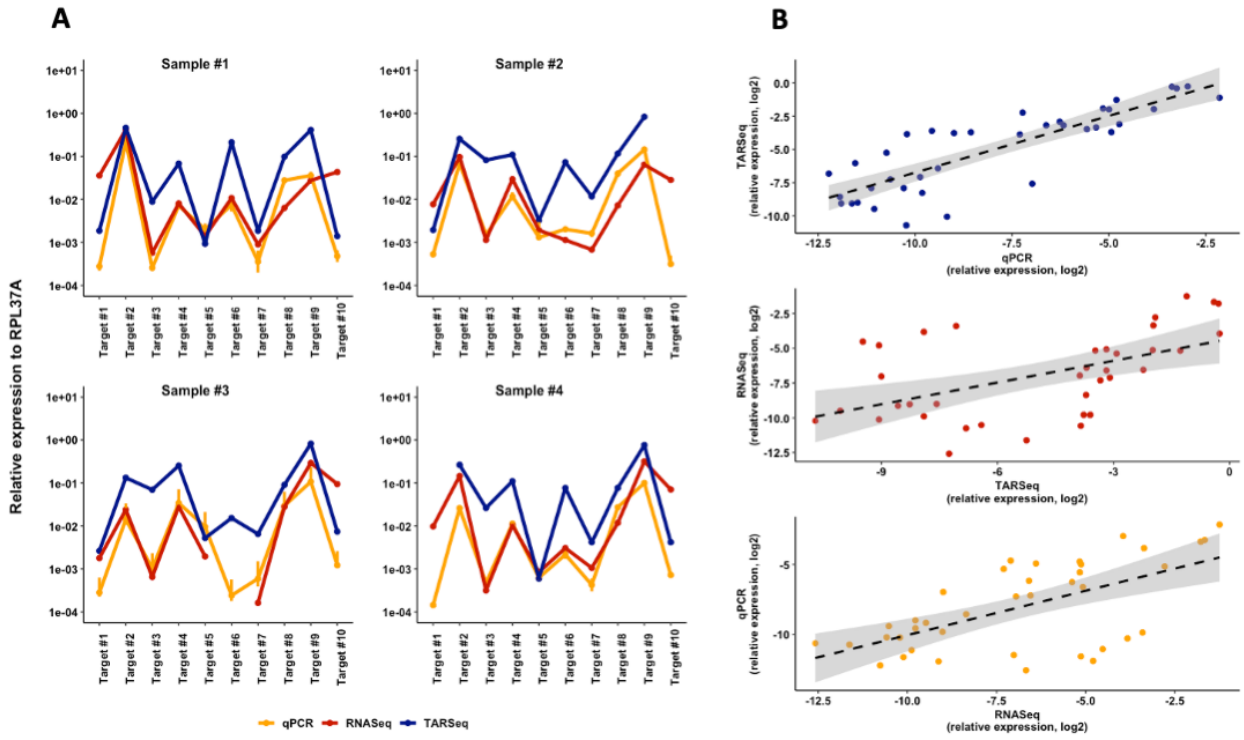
Figure A2.3 Continued





Supplementary Figure A2.4: Copy number in single cell isolated and expanded clones for three different gene sets and input ratios, Related to Figure 2.4.

Copy number was determined via RT-qPCR via the $2^{\Delta Cq}$ method compared to the single copy number gene RPP30.



Supplementary Figure A2.5: Validation of Targeted RNA-Sequencing (TAR-Seq) with custom gene panels, Related to STAR Methods.

A) Relative expression with respect to RPL37A of a target gene panel in 4 different NHDF samples across 10 different targets as quantified by TAR-Seq, RNA-Seq and RT-qPCR. Target 10 in Sample #2 was not identified due to low abundance. B) Gene expression correlation of 10 genes in 4 different samples between TAR-Seq, RT-qPCR and RNA-Seq.

Table A2.1: Primers used in Chapter2.

Primers listed were utilized for qPCR amplification of targets for CRISPRa/i and cDNA expression characterization in Figures 2.2A-C, Figures A2.1, A2.4, and Figures 2.4A, B, and D.

Primer	Sequence
GAPDH_Fwd	GGAGCGAGATCCCTCCAAAAT
GAPDH_Rev	GGCTGTTGTCATACTTCTCATGG
OTX2_Fwd	CATGCAGAGGTCCTATCCCAT
OTX2_Rev	AAGCTGGGGACTGATTGAGAT
ZNF155_Fwd	GGCATCAACCGTTCCACCAA
ZNF155_Rev	TCTTCTGAGTATGACTGCTGAGG
SOX13_Fwd	GATCTCGCTGGAAGTCCATGA
SOX13_Rev	GGATCACGTAGCTCTGGCG
DLX5_Fwd	GTCTTCAGCTACCGATTCTGAC
DLX5_Rev	CTTTGCCATAGGAAGCCGAG
ETV5_Fwd	TCAGCAAGTCCCTTTTATGGTC
ETV5_Rev	GCTCTTCAGAATCGTGAGCCA
ZNF502_Fwd	AGGAAGGAGGTTTTGGGAGAA
ZNF502_Rev	AGTGAGATGTGAGTGATTGCG
SATB1_Fwd	TCGGGCCATCTGATGAAAACC
SATB1_Rev	CCAACCTGGATTAGCCCTTTG
LHX8_Fwd	AAGGGGAATGTCTATCACTTGGC
LHX8_Rev	TTCCACCAAAGCAAACTCCTC
ZBTB39_Fwd	GAGCGTCTGGGTATGGAGGA
ZBTB39_Rev	CCGAAGGACAACTCAGGGT
KLF2_Fwd	CTACACCAAGAGTTCGCATCTG
KLF2_Rev	CCGTGTGCTTTCGGTAGTG
HHEX_Fwd	GGCAAACCTCTACTCTGGAGC
HHEX_Rev	GTCGTTGGAGAATCTCACCTG
SOHLH2_Fwd	TGGCAATGGGCTTGAATTAAATG
SOHLH2_Rev	GTTTTTCTCGAACTCTGACAACG
SOX17_Fwd	CTCCGGTGTGAATCTCCCC
SOX17_Rev	CACGTCAGGATAGTTGCAGTAAT
TFAP2C_Fwd	CTGTTGCTGCACGATCAGACA
TFAP2C_Rev	CTCAGTGGGGTTCATTACGGC
PRDM1_Fwd	AAGCAACTGGATGCGCTATGT
PRDM1_Rev	GGGATGGGCTTAATGGTGTAGAA
NANOS3_Fwd	CTTTGACCTGTGGACAGATTACC
NANOS3_Rev	GCCTGGTTTCAGGACCCTC

LYAR_Fwd	GCCTTTCTTGCATTGACTGCG
LYAR_Rev	TGTGGGTTTTACCTTCATAGCCT
MTF1_Fwd	CACAGTCCAGACAACAACATCA
MTF1_Rev	GCACCAGTCCGTTTTTATCCAC
KMT2C_Fwd	ACTCAGCACCACGAAAACAAA
KMT2C_Rev	GGCATACTCCTAGTGACCACTC
NR2C1_Fwd	ACGGCTCTACTCCAAGCAAAG
NR2C1_Rev	TTGTCTCCACATACTACGCAAAG
CARHSP1_Fwd	CCCCGTCTACAAAGGAGTCTG
CARHSP1_Rev	GGGTGGGATGGAGCACATTT
NFATC2_Fwd	GAGCCGAATGCACATAAGGTC
NFATC2_Rev	CCAGAGAGACTAGCAAGGGG
ZNF281_Fwd	AGGACCTCAGTATTCTCCACC
ZNF281_Rev	CCATCTCCAACCAAAGAAGGTTT
NR1D2_Fwd	TTTAGTGGCATGGTTCTACTGTG
NR1D2_Rev	AGCCTTCGCAAGCATGAACT
NFYC_Fwd	CCAAGAGATGAACTGAAACCTCC
NFYC_Rev	GCTGAGCCAGCGTGAAATAGT
DMTF1_Fwd	GTCTGAACCGGCCTTTGTTTG
DMTF1_Rev	GCCCAGTCATTGCCATGCT
ZNF419_Fwd	GTAGGACTGCTCAGTTCAAACAT
ZNF419_Rev	ACAGTTACTACACCCGTAAGGC
CENPA_Fwd	TTCCTCCCATCAACACAGTCG
CENPA_Rev	CACACCACGAGTGAATTTAACAC
NOBOX_Fwd	GAGACCCTCAAATCACCCCAA
NOBOX_Rev	GCCCCTTGTGAGTTCCCTTTT
TCF4_Fwd	CAAGCACTGCCGACTACAATA
TCF4_Rev	CCAGGCTGATTCATCCCACTG
ZGLP1_Fwd	ATCATCCCCACGTACAGCCT
ZGLP1_Rev	CAGCGAGTGCCGTATTTCTTG
TOX_Fwd	TATGAGCATGACAGAGCCGAG
TOX_Rev	GGAAGGAGGAGTAATTGGTGGA
ZNF57_Fwd	ACAGTCTTCACGCATCTTTCTTC
ZNF57_Rev	CCACATGCTTGACACTTATACGG
NR1H2_Fwd	AGAAGATTCGGAAACAACAGCA
NR1H2_Rev	GCTGGATCATTAGTTCTTGAGCC
ZNF134_Fwd	AAGCCGTTTGAGTGCATTGAA
ZNF134_Rev	CTTTGGTGTCGAACAAGGTGG
ZNF589_Fwd	GCAGAAGATCAACGAGTGGAA

ZNF589_Rev	CCAAACGTGACTGCTCCAAAC
CRX_Fwd	GCCCCACTATTCTGTCAACG
CRX_Rev	GTCTGGGTACTGGGTCTTGG
PLAGL2_Fwd	GTTTCACCGCAAGGACCATCT
PLAGL2_Rev	GGTAGCCCAGCTTCGTATTGTA
ZNF574_Fwd	ACATTGAGCACCGCTATGTCT
ZNF574_Rev	CCTGCACAAGGGTCTGATAGA
CLOCK_Fwd	CAACACCAACCAAGATCCCGA
CLOCK_Rev	AATGATGACCTTCTTTGCACCA
HSF2_Fwd	AGAATGAGTCCCTTTGGAAGGA
HSF2_Rev	TTCTTTTGGGCTCCATTAGTGTT
ZNF227_Fwd	CTTATAGATGCGACAGTTGCGG
ZNF227_Rev	CTCCAACCGAAGCCCTTACC
TCF3_Fwd	CCGACTCCTACAGTGGGCTA
TCF3_Rev	CGCTGACGTGTTCTCCTCG
CCDC88A_Fwd	ATGCCTCACTTAGAATGCACAA
CCDC88A_Rev	AGACATTTGGCAACGACATCA
SP1_Fwd	TGGCAGCAGTACCAATGGC
SP1_Rev	CCAGGTAGTCCTGTCAGAACTT
KLF16_Fwd	CAAGTCCTCGCACCTAAAGTC
KLF16_Rev	AGCGGGCGAACTTCTTGTC
ZBTB26_Fwd	CCATCTCCACAATTATGCCCTT
ZBTB26_Rev	GGCCTTTGTCTACACCTCGAAT
RnaseP_Fwd (RPP30)	AGATTTGGACCTGCGAGCG
RnaseP_Rev (RPP30)	GAGCGGCTGTCTCCACAAGT
BC-NGS-Fwd	ATACTCAGAAGATGTCACTCACC
BC-NGS-Rev	CATTCATAGTTCTTGCTCAGTGG
Copy_Number_Fwd	GAGCGTACAGTGCGTTCTCC
Copy_Number_Rev	CGGGGCTAAAGTGCATCTCG
i501	AATGATACGGCGACCACCGAGATCTACACTATAGCCTAC ACTCTTTCCCTACACGACGCTCTTCCGATCT
i502	AATGATACGGCGACCACCGAGATCTACACATAGAGGCAC ACTCTTTCCCTACACGACGCTCTTCCGATCT
i503	AATGATACGGCGACCACCGAGATCTACACCCTATCCTAC ACTCTTTCCCTACACGACGCTCTTCCGATCT
i504	AATGATACGGCGACCACCGAGATCTACACGGCTCTGAAC ACTCTTTCCCTACACGACGCTCTTCCGATCT
i505	AATGATACGGCGACCACCGAGATCTACACAGGCGAAGAC ACTCTTTCCCTACACGACGCTCTTCCGATCT
i506	AATGATACGGCGACCACCGAGATCTACACTAATCTTAACA CTCTTTCCCTACACGACGCTCTTCCGATCT

i507	AATGATACGGCGACCACCGAGATCTACACCAGGACGTAC ACTCTTTCCCTACACGACGCTCTTCCGATCT
i508	AATGATACGGCGACCACCGAGATCTACACGTACTGACAC ACTCTTTCCCTACACGACGCTCTTCCGATCT
i701	CAAGCAGAAGACGGCATACGAGATCCGTGAGAGTGACTG GAGTTCAGACGTGTGCTCTTC
i702	CAAGCAGAAGACGGCATACGAGATCATCAAGTGTGACTG GAGTTCAGACGTGTGCTCTTC
i703	CAAGCAGAAGACGGCATACGAGATAAGACGGAGTGACTG GAGTTCAGACGTGTGCTCTTC
i704	CAAGCAGAAGACGGCATACGAGATCGAACTTAGTGACTG GAGTTCAGACGTGTGCTCTTC
i705	CAAGCAGAAGACGGCATACGAGATGATAGACAGTGACTG GAGTTCAGACGTGTGCTCTTC
i706	CAAGCAGAAGACGGCATACGAGATGGTGCGAAGTGACTG GAGTTCAGACGTGTGCTCTTC
i707	CAAGCAGAAGACGGCATACGAGATCACCTTACGTGACTG GAGTTCAGACGTGTGCTCTTC
i708	CAAGCAGAAGACGGCATACGAGATCCTAATCCGTGACTG GAGTTCAGACGTGTGCTCTTC

APPENDIX III : SUPPLEMENTAL FIGURES AND TABLES FOR CHAPTER III

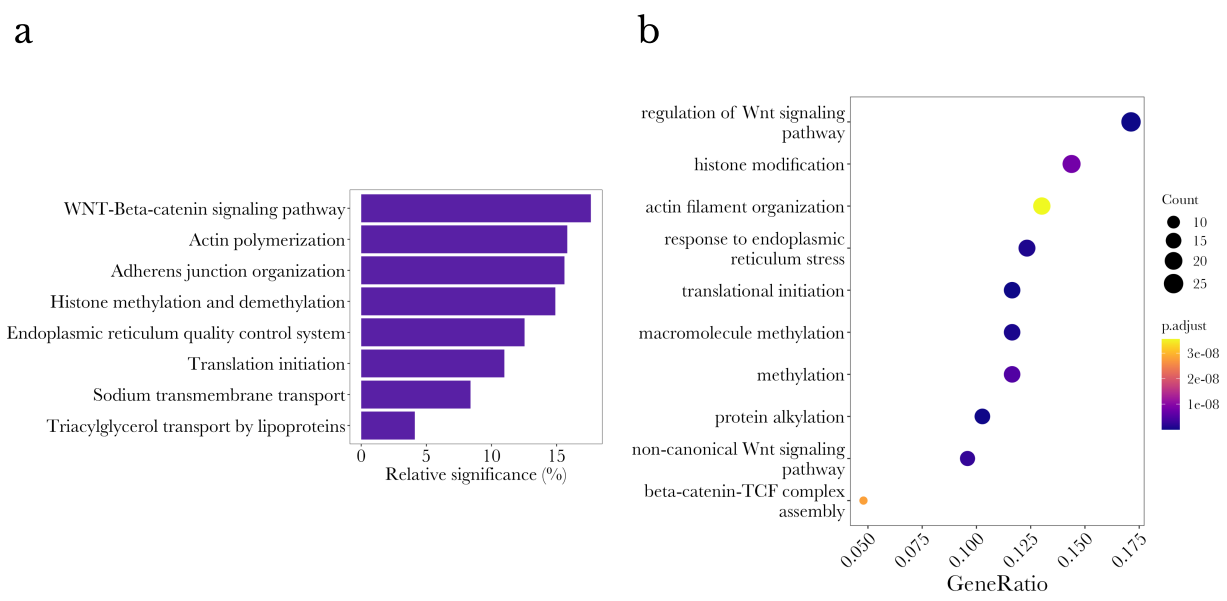


Figure A3.1: Functional annotation of clock genes.

(a) RNA clock processes and their relative significance in the overall age prediction. The relative significance exhibits a right skewed distribution across all processes, with WNT signaling accounting for more than 15% of the signal, while lipid transport has a contribution of less than 5%. (b) GO term enrichment analysis for clock genes agrees with process annotation and reveals significant enrichment for WNT signaling, histone modification, actin organization, ER stress, and translation initiation.

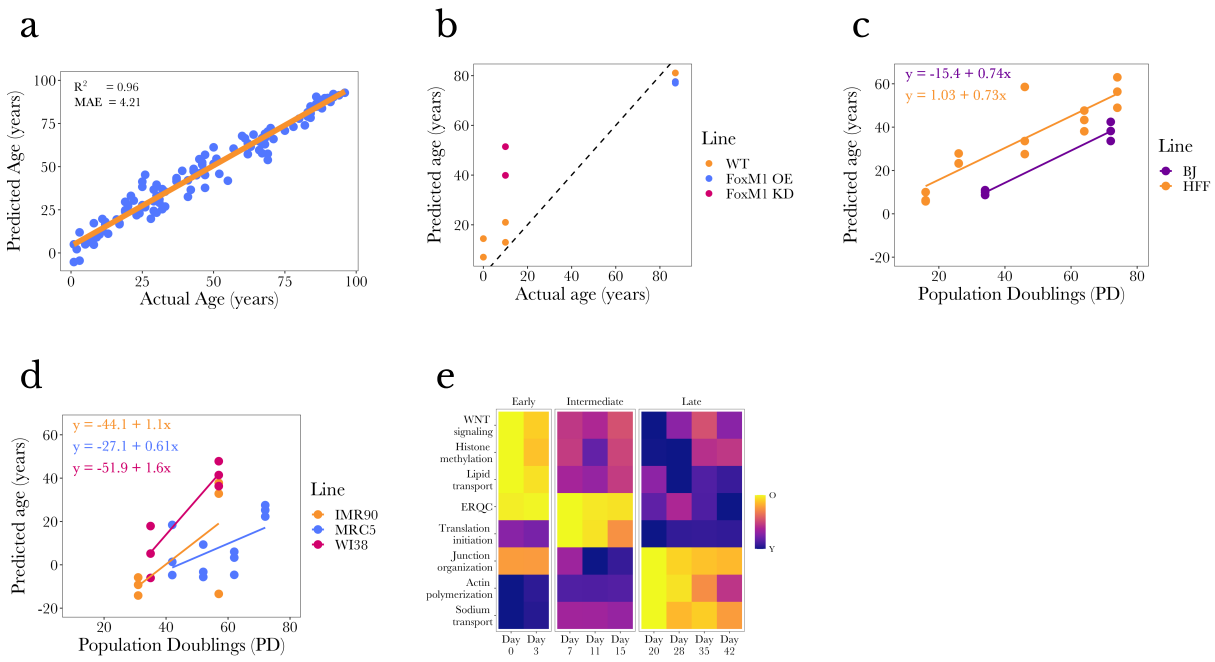


Figure A3.2: RNA clock performance across different fibroblast samples.

(a) Performance of RNA clock on training data shows high correlation (R^2=0.96) and low absolute error (MAE=4.2). (b) Predicted ages showing the expected biological effect on NHDFs treated with siRNA for FoxM1 (shown to induce senescence) or a constitutively active truncated form of FoxM1 (dNdK), which was shown to partially rescue aging phenotypes in old cells. (c) RNA clock predictions for established skin fibroblast lines (BJ, HFF) across multiple population doublings (PD) show a linear relationship between biological aging and passaging, with a slope of approximately 0.7 years/PD. (d) RNA clock predictions for established lung fibroblast lines (IMR90, MRC5, WI38) across multiple population doublings (PD) show a disagreement in the aging rates of the different lines suggesting either distinct aging phenotypes or cell type specificity of our transcriptomic predictor. (e) Process activity across cellular reprogramming timelines. We observed three distinct phases of reprogramming (early, intermediate, and late), characterized by the activity of specific clock processes. The early age reversal is driven by a rejuvenation of the WNT signaling, histone methylation, and lipid transport processes, which reach steady young state in the intermediate phase. ERQC and translation initiation resemble young states only late in reprogramming, while junction organization, actin polymerization, and sodium transport seem to become older in the late stage. We hypothesize that these last three processes are more specific to fibroblasts, while WNT signaling, histone methylation and lipid transport represent a more core signature of cellular aging.

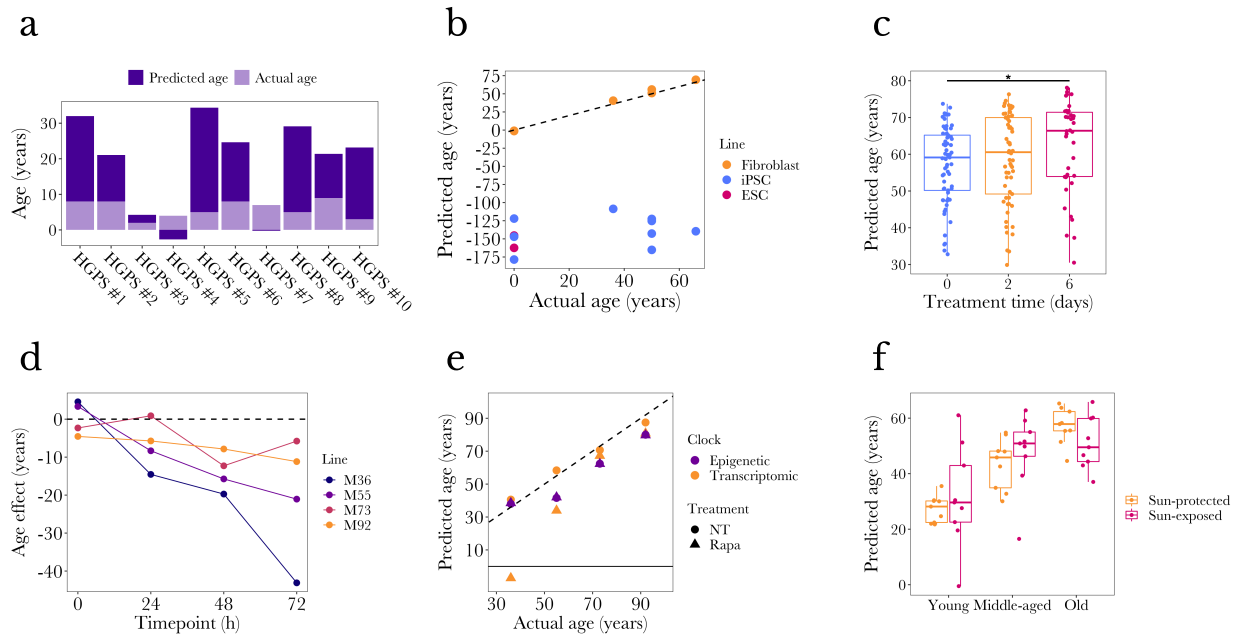


Figure A3.3: RNA clock response to age-modulating interventions.

(a) Predicted ages of fibroblast samples from Hutchinson-Gilford progeria syndrome (HGPS) donors. We observe the expected age acceleration in all but two samples. (b) Predicted ages of NHDFs, age-matched fibroblast derived iPSCs, and embryonic stem cells (ESCs). There is an expected significant reduction in the biological age of ESCs and iPSCs, as previously reported. (c) Predicted ages of NHDFs treated with poly(I:C), a known immunostimulant in human cells, for 0, 2 or 6 days. We observe a significant average increase in the biological age of approximately 7 years after 6 days of treatment with poly(I:C). Significance is tested using the two-tailed Student's t-test. (d) Age effect (actual age subtracted from predicted age) of Rapamycin (Rapa) (100 nM) as measured by our RNA clock at different treatment timepoints in four different NHDF lines. All lines exhibit an age reduction across the treatment timespan, with the younger NHDFs showing the highest response to Rapa. € Predicted age of Rapa treated samples (100nM for 72h) as measured by our RNA clock and the Horvath Skin&Blood DNAm clock[140]. Our RNA clock shows a higher accuracy in the no treatment (NT) samples as well as a higher sensitivity to the Rapa treatment, with the difference between Rapa and NT being more apparent in our assay than in the epigenetic clock. (f) Predicted ages for young, middle-aged and old NHDFs isolated from a sun protected or sun-exposed area of the body. While there is no significant age difference induced by the sun exposure, we observe a slight increase in the median predicted age in sun-exposed middle-aged samples but a decrease in the old donors. An interesting explanation could be that sun exposed old skin has a higher neoplastic potential which might appear younger in our transcriptomic assay.

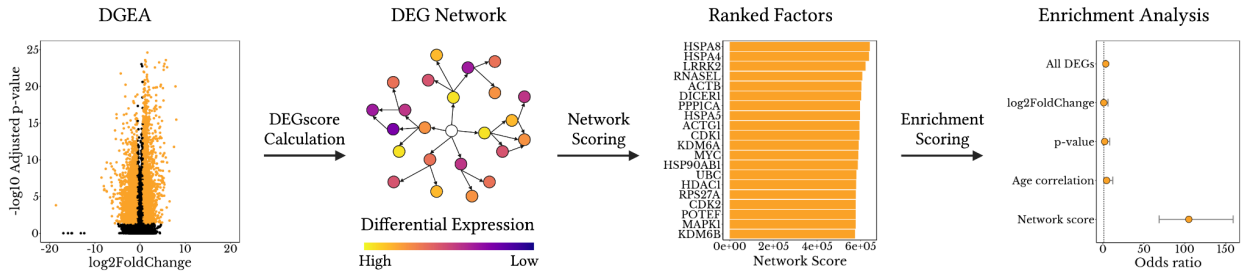


Figure A3.4: Workflow and validation for network scoring approach applied to fibroblast aging.

Differential gene expression analysis (DGEA) was first performed on a recent human dermal fibroblast RNA-Seq dataset generated from 133 individuals with ages 1-94 years old[15]. A gene score (DEGscore) was calculated for each differentially expressed gene (DEG). Then, each candidate gene was assigned a network score based on its local DEG network, derived from the STRING database. Large network scores indicate high connectivity and differential expression levels. Candidate genes were then ranked by decreasing network score. For a preliminary validation of this workflow, Fisher's exact test was used to compare the statistical enrichment for known aging genes from the GenAge database[84]. The list of all DEGs and top 100 ranked lists based on common candidate selection metrics and our network scoring approach were compared.

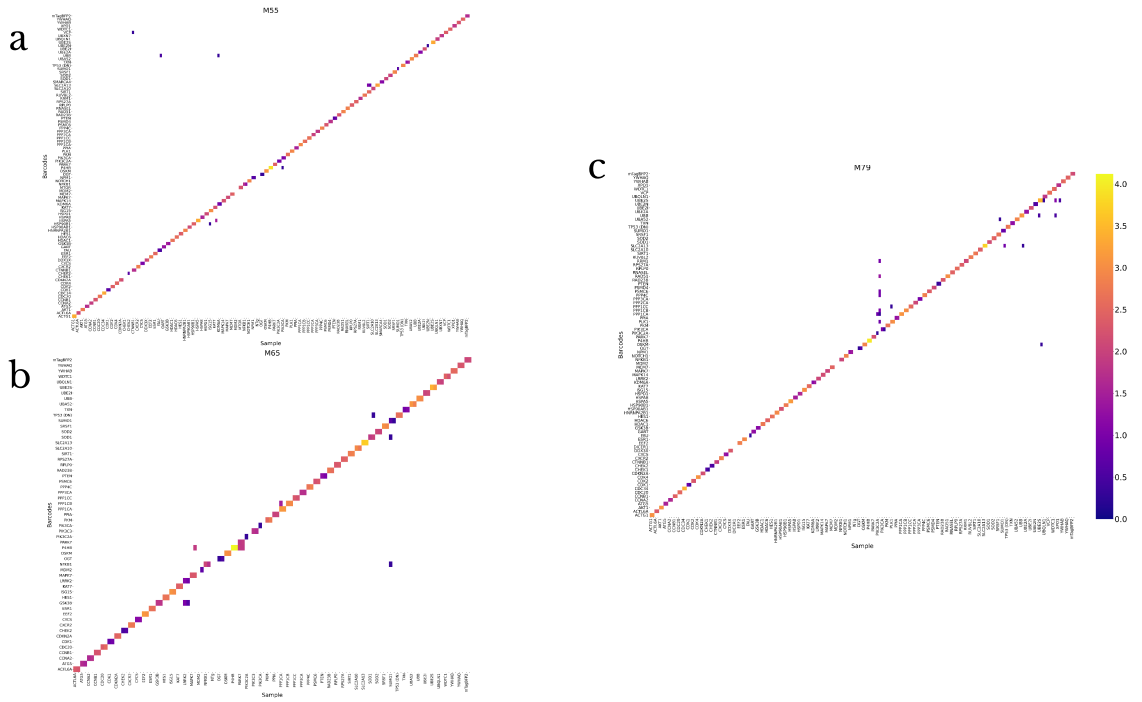


Figure A3.5: Barcode validation of gene overexpression.

Barcode counts for all sequencing samples in our screen. Raw FASTQ files for each sample were searched for the specific barcode sequence of each gene. Reads were assigned to any barcode within a hamming distance of 3 and normalized by total number of reads per sample. (a) M55. (b) M65. (c) M79.

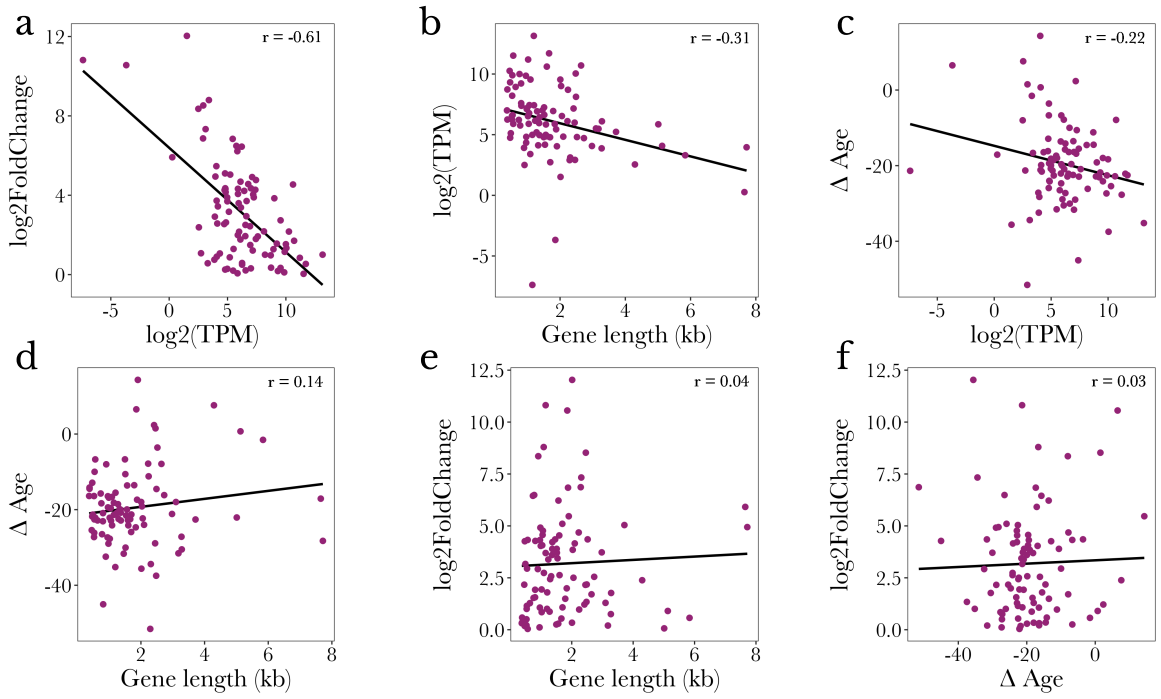


Figure A3.6: Correlations between different metrics for our candidate genes.

(a) Transgene induction (log2FoldChange with respect to mTagBFP2) vs endogenous expression (log2(TPM)) shows a moderate negative correlation of Pearson's r=-0.61 (p-value = 1.112e-10), with the lowest overexpressed genes having the highest levels of endogenous expression. (b) Endogenous expression vs gene length shows a weak negative correlation of Pearson's r=-0.31 (p-value = 0.002226), with the largest genes having the lowest endogenous expression. (c) Predicted age effect (Age_{gene} - $Age_{mTagBFP2}$) vs endogenous expression shows a weak correlation of Pearson's r=-0.22 (p-value = 0.03254). (d) Predicted age effect vs gene lengths shows no correlation (Pearson's r=0.1397267, p-value = 0.184). (e) Transgene induction vs gene length shows no correlation (Pearson's r=0.0425474, p-value = 0.6872). (f) Transgene induction vs predicted age effect shows no correlation (Pearson's r=0.03259279, p-value = 0.7578). All metrics represent data for the candidate genes averaged across the three NHDF lines used in the screen.

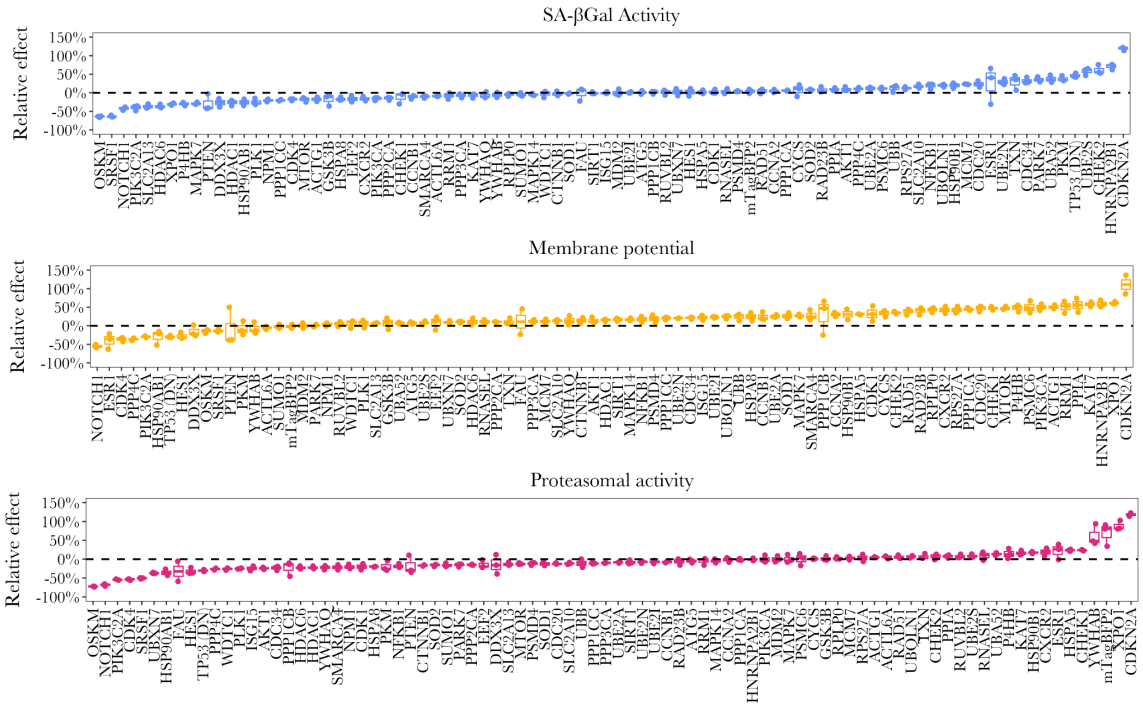


Figure A3.7: Flow cytometry assays for M55 line across all overexpressed genes.

Staining data for all overexpression lines across the three assayed cellular processes in M55 line (n=3). Top. SA-βGal activity measured for cells expressing gene of interest using C_{12}FDG fluorescent substrate. OSKM and SRSF1 reduced cellular senescence, while CDKN2A, HNRNPA2B1, and CHEK2 greatly increased it. Middle. Mitochondrial membrane potential measured for cells expressing gene of interest using TMRM dye. NOTCH1 and ESR1 greatly reduced the membrane potential, while CDKN2A and XPO1 significantly increased it. Bottom. Proteasomal activity measured for cells expressing gene of interest using LLVY fluorescent substrate. OSKM and NOTCH1 reduced proteasomal activity while CDKN2A, and XPO1 greatly increased it. Data is normalized with respect to the gene specific No Doxycycline control and the line specific BFP control.

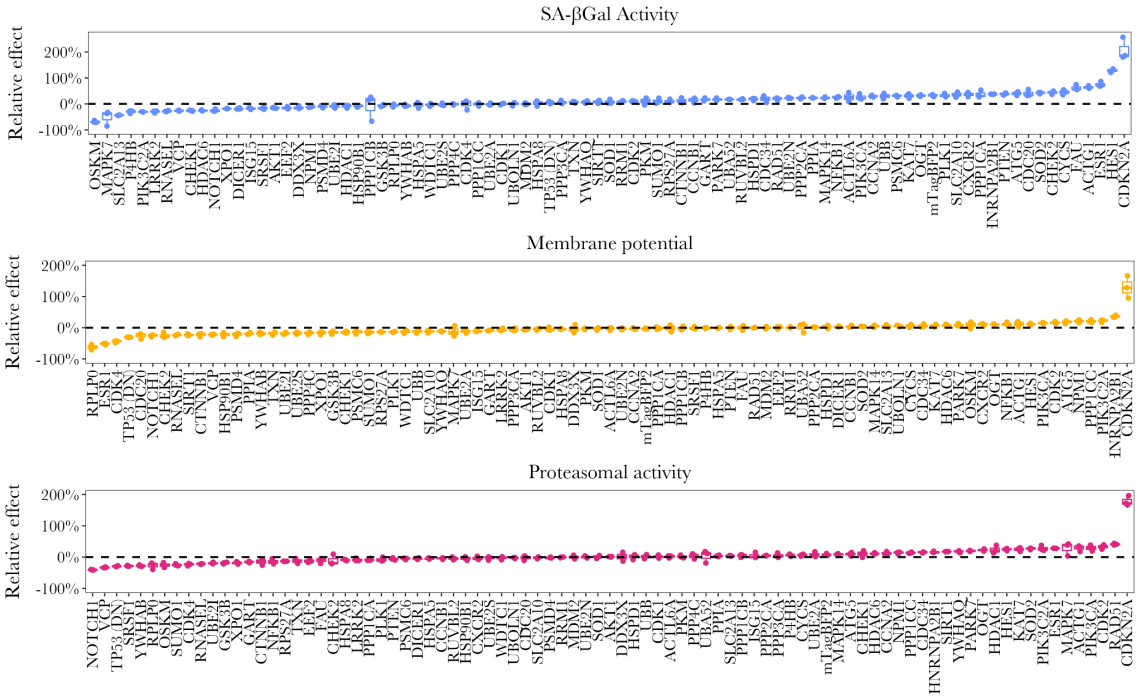


Figure A3.8: Flow cytometry assays for M79 line across all overexpressed genes.

Staining data for all overexpression lines across the three assayed cellular processes in M79 line (n=3). **Top**. SA-βGal activity measured for cells expressing gene of interest using C_{12}FDG fluorescent substrate. OSKM and MAPK7 reduced cellular senescence, while CDKN2A, HES1, and ESR1 greatly increased it. **Middle**. Mitochondrial membrane potential measured for cells expressing gene of interest using TMRM dye. RPLP0 and ESR1 greatly reduced the membrane potential, while CDKN2A and HNRNPA2B1 significantly increased it. **Bottom**. Proteasomal activity measured for cells expressing gene of interest using LLVY fluorescent substrate. NOTCH1 and VCP reduced proteasomal activity while CDKN2A, and RAD51 greatly increased it. Data is normalized with respect to the gene specific No Doxycycline control and the line specific BFP control.

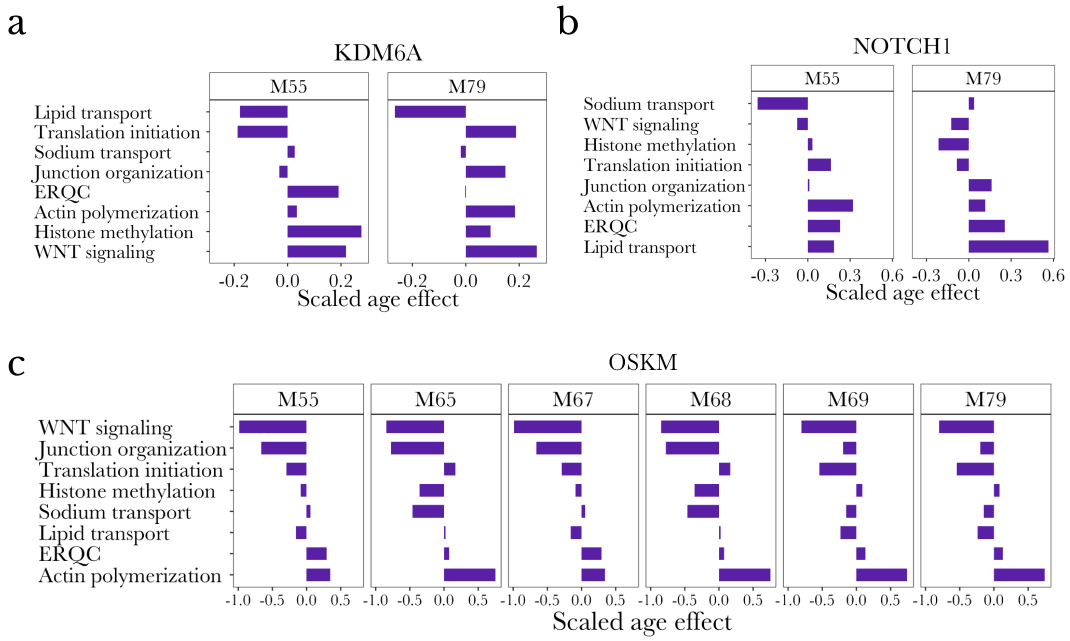


Figure A3.9: Process-specific scaled age effect for the strong perturbations in our screen.

(a) Scaled age effect of KDM6A overexpression on the clock processes across M55 and M79 (n=2). We observed a coordinated pro-aging effect in WNT signaling, actin polymerization, and histone methylation, with a rejuvenation effect in lipid transport. (b) Scaled age effect of NOTCH1 overexpression on the clock processes across M55 and M79 (n=2). We observed a coordinated pro-aging effect in lipid transport, ERQC, actin polymerization, and WNT signaling. (c) Scaled age effect of OSKM overexpression on the clock processes across all 6 lines (n=2). We observed a coordinated rejuvenating effect in WNT signaling, junction organization, and translation initiation (except for M65 and M68).

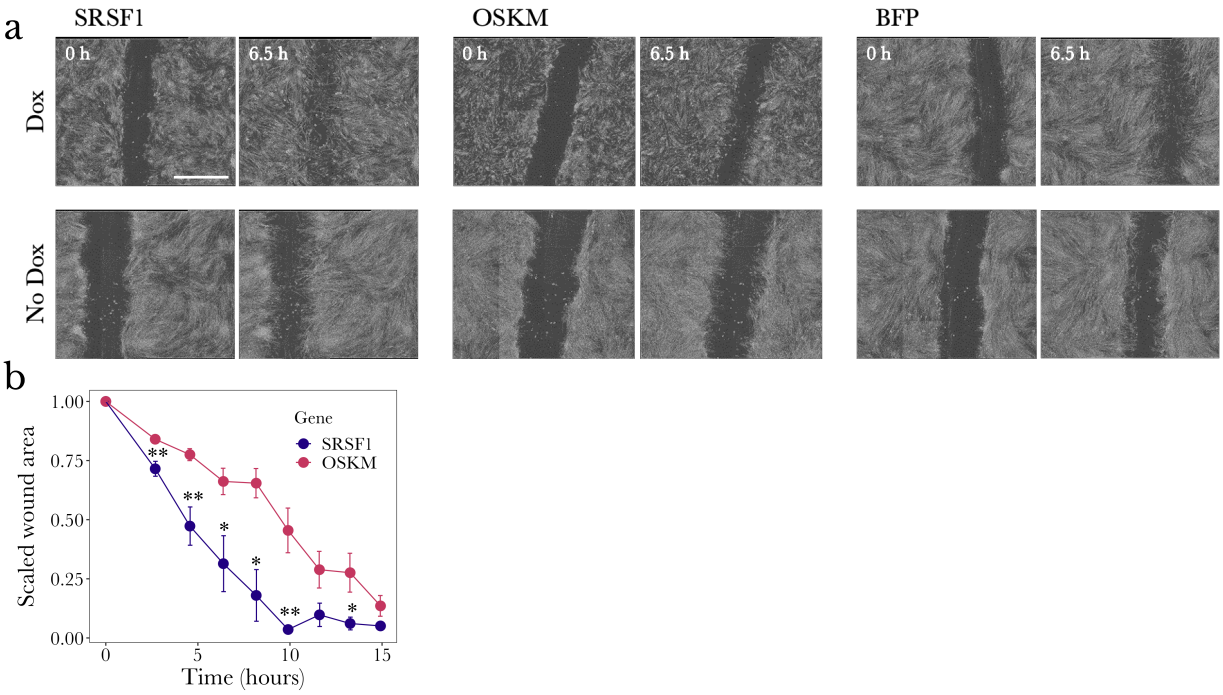


Figure A3.10: Scratch assay in M79 line expressing SRSF1, OSKM, or BFP.

(a) Representative bright-field images of SRSF1-, OSKM-, or BFP-induced cells at 0 and 6.5h after scratch. Scale bar, 1 mm. Experiments were performed independently at least three times with similar results. **(b)** Scratch assay time course, expressed as wound area scaled to the 0 h time point, for the M79 line after the induction of OKSM or SRSF1. We observed a faster wound area closure in the SRSF1 treated cells compared to those induced with OSKM. Data is averaged over replicates (n=3-5), error bars represent standard error of the mean, and significance with respect to the BFP control is tested using the two-tailed Student's t-test.

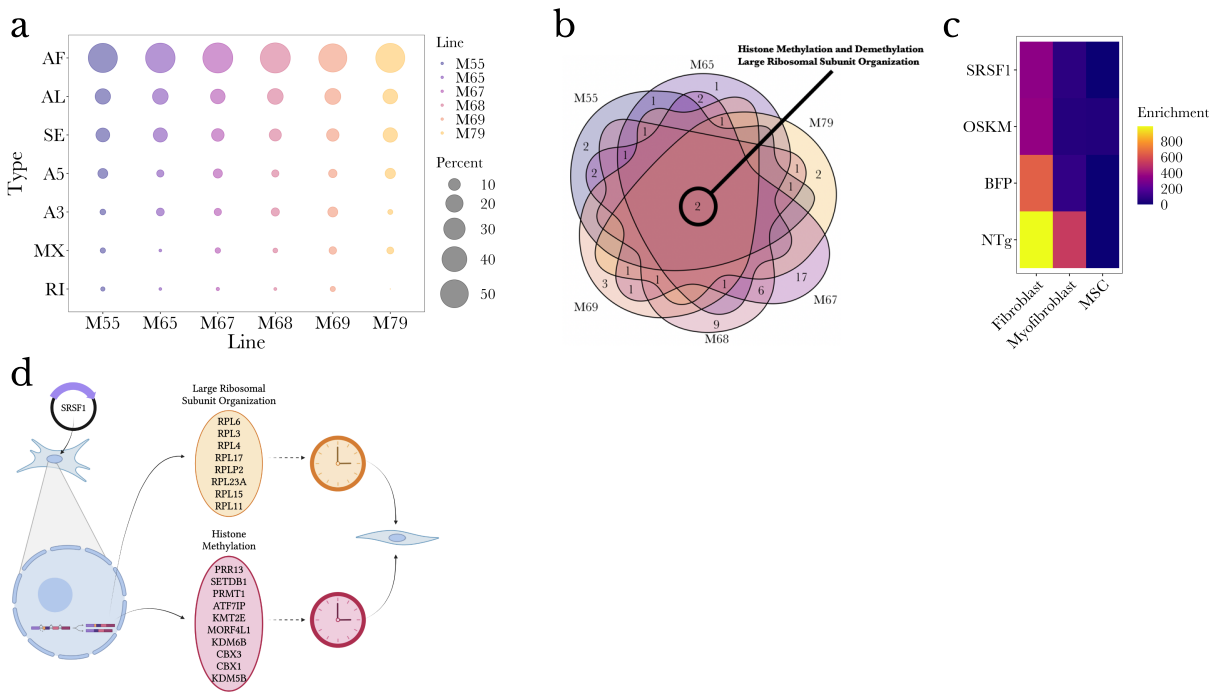


Figure A3.11: Splicing and cell identity analysis in SRSF1 overexpression lines.

(a) Splicing analysis of SRSF1 overexpression in 6 different NHDF lines (n=2) (AF = alternative first; AL= alternative last; SE = skipped exon, A5 = alternative 5' site, A3 = alternative 3' site, MX = mutually exclusive exons, RI = retained intron). We saw a large agreement in the distribution of splicing events with AF, AL, and SE representing most of the events, while RI had the lowest abundance. (b) Overlap between cellular processes that alternatively spliced genes are part of, across all 6 lines. We observed only two processes common for all 6 NHDF lines: histone methylation and large ribosomal subunit organization. (c) Enrichment analysis for cell identity genes related to fibroblast, myofibroblast, and mesenchymal stem cells. Data is averaged across the 3 NHDF lines (M55, M65, M79) and shows that both SRSF1 and OSKM lead to a loss of the fibroblast gene expression profile, compared to the NTg and BFP samples, which have a large enrichment for fibroblast (and myofibroblast to a lesser extent) genes. (d) Proposed mechanism for SRSF1 rejuvenation. The alternative splicing of several genes part of the large ribosomal subunit organization and the histone methylation pathways leads to a beneficial effect on the expression of clock genes part of the same processes, which ultimately induces a younger transcriptome in the treated cells.

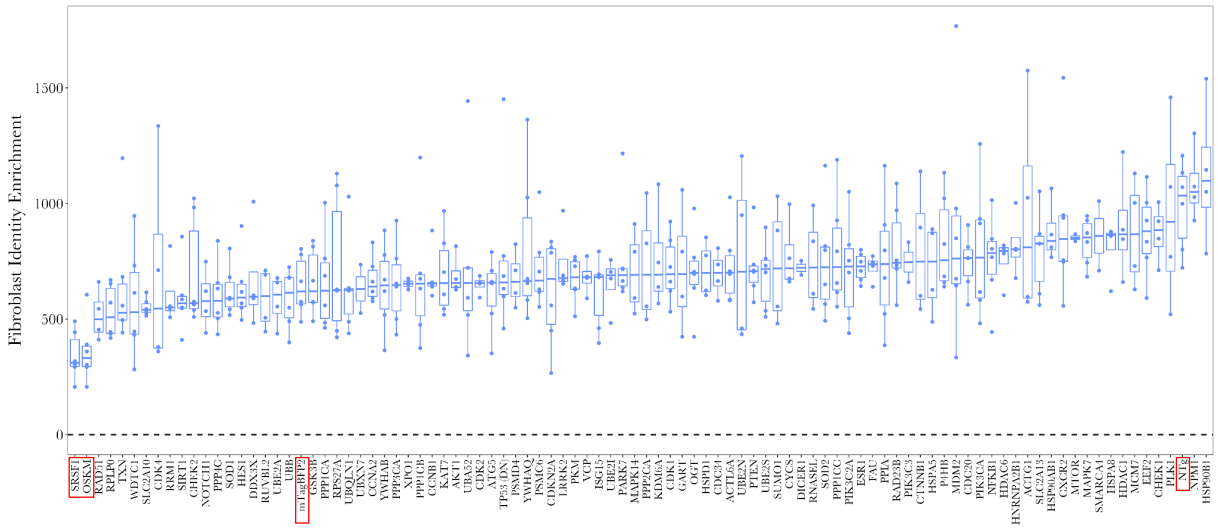


Figure A3.12: Fibroblast identity analysis across all genes and lines in the screen.

Enrichment of fibroblast identity gene expression for all overexpression lines across the three screened cell lines (n=6). Each dot represents a single replicate of an overexpression line. Gene overexpression lines are sorted by their median enrichment. SRSF1, OSKM, mTagBFP2, and NTg are highlighted in red boxes.

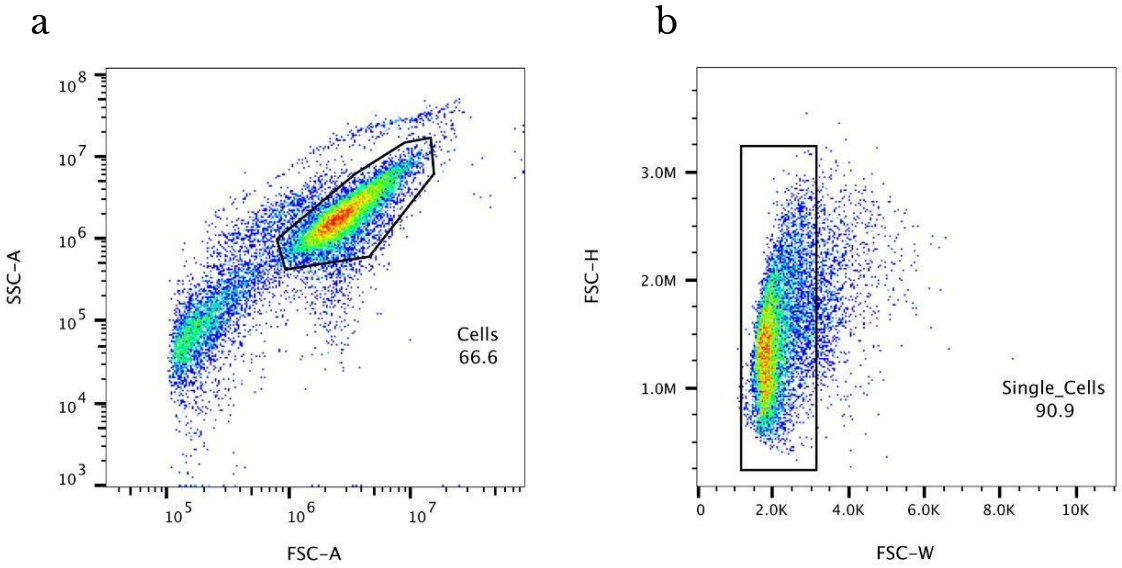


Figure A3.13: Representative flow cytometry cell gating.

Representative gating strategy for flow cytometry. (a) Fibroblast cell isolation using forward scatter area and side scatter area, all 18,434 events shown. (b) Single cell isolation using forward scatter width and height, all 12,277 events shown.

Table A3.1: Genes included in the RNA clock.

Genes included in the RNA clock are listed as Ensembl-IDs ("Predictor" column) and gene symbols. The unnormalized and standardized contribution of each gene to the age prediction is provided in the "Coefficient" and "Standardized Coefficient" column, respectively. In addition, the processes each gene is associated with in the Molecular Biology of the Cell Ontology is shown ("Process" column).

Predictor	Gene Symbol	Coefficient	Standardized Coefficient	Process
Intercept	-	3.676114809	-0.547921048	-
ENSG00000148737	TCF7L2	0.203423062	0.145205231	WNT signaling
ENSG00000165617	DACT1	-0.104577043	-0.11158401	WNT signaling
ENSG00000168036	CTNNB1	1.155505293	0.241076129	Junction organization WNT signaling
ENSG00000134245	WNT2B	-0.080958237	-0.09063118	WNT signaling
ENSG00000162552	WNT4	0.065692601	0.146128577	WNT signaling
ENSG00000184675	AMER1	-0.340425536	-0.11510598	WNT signaling
ENSG00000228630	HOTAIR	-0.02660401	-0.070668056	Histone methylation WNT signaling
ENSG00000106483	SFRP4	0.054716313	0.1254804	WNT signaling
ENSG00000131650	KREMEN2	-0.374212687	-0.111098616	WNT signaling
ENSG00000082701	GSK3B	0.165730251	0.048646235	WNT signaling
ENSG00000163348	PYGO2	0.602236784	0.12687599	WNT signaling
ENSG00000164736	SOX17	0.17694256	0.099658795	WNT signaling
ENSG00000070018	LRP6	-0.400894674	-0.145217517	WNT signaling
ENSG00000004975	DVL2	-0.137932933	-0.035690819	WNT signaling
ENSG00000174804	FZD4	0.04796222	0.05108055	WNT signaling
ENSG00000134569	LRP4	0.024747349	0.020981309	WNT signaling
ENSG00000138795	LEF1	0.084010082	0.109813402	WNT signaling
ENSG00000108375	RNF43	0.057468549	0.034638674	WNT signaling
ENSG00000162337	LRP5	-0.064028332	-0.061243317	WNT signaling
ENSG00000154856	APCDD1	0.013634361	0.027832253	WNT signaling
ENSG00000183762	KREMEN1	-0.192852626	-0.112828097	WNT signaling
ENSG00000165879	FRAT1	0.099806158	0.067259579	WNT signaling
ENSG00000135925	WNT10A	0.138072373	0.037215726	WNT signaling
ENSG00000188763	FZD9	-0.05298353	-0.056740297	WNT signaling
ENSG00000155011	DKK2	0.038131278	0.084632705	WNT signaling
ENSG00000114251	WNT5A	-0.057537968	-0.061460551	WNT signaling
ENSG00000075290	WNT8B	0.389034126	0.140364369	WNT signaling

ENSG00000157240	FZD1	-0.078188613	-0.033630775	WNT signaling
ENSG00000177283	FZD8	0.050571533	0.054071337	WNT signaling
ENSG00000007372	PAX6	-0.043215988	-0.079755094	WNT signaling
ENSG00000125249	RAP2A	0.16908097	0.060736505	WNT signaling
ENSG00000122515	ZMIZ2	-0.038528623	-0.01385182	WNT signaling
ENSG00000126457	PRMT1	-0.227667683	-0.087422616	Histone methylation
ENSG00000115548	KDM3A	-0.297243909	-0.109263437	Histone methylation
ENSG00000164256	PRDM9	302.2036775	0.016097875	Histone methylation
ENSG00000138336	TET1	-0.024934291	-0.021778778	Histone methylation
ENSG00000187605	TET3	0.069005463	0.052496081	Histone methylation
ENSG00000142453	CARM1	-0.302278901	-0.089161122	Histone methylation
ENSG00000136504	KAT7	0.919204701	0.168482405	Histone methylation
ENSG00000143499	SMYD2	0.309824718	0.075025188	Histone methylation
ENSG00000133895	MEN1	-0.1676316	-0.045443562	Histone methylation
ENSG00000078403	MLLT10	0.087182019	0.024400756	Histone methylation
ENSG00000127663	KDM4B	0.226919391	0.091026956	Histone methylation
ENSG00000055609	KMT2C	-0.081517358	-0.027199734	Histone methylation
ENSG00000145391	SETD7	0.361775259	0.117057071	Histone methylation
ENSG00000107077	KDM4C	0.102464121	0.031609414	Histone methylation
ENSG00000185787	MORF4L1	-0.150952912	-0.03924061	Histone methylation
ENSG00000116455	WDR77	0.082572644	0.025394591	Histone methylation
ENSG00000171681	ATF7IP	0.333836327	0.107711525	Histone methylation
ENSG00000171988	JMJD1C	-0.166670894	-0.060685489	Histone methylation
ENSG00000143379	SETDB1	-0.385548129	-0.075841229	Histone methylation
ENSG00000129691	ASH2L	-0.364779419	-0.075588272	Histone methylation
ENSG00000176476	SGF29	0.173076479	0.079762393	Histone methylation
ENSG00000176658	MYO1D	0.118977805	0.115935697	Junction organization
ENSG00000138771	SHROOM3	-0.03467626	-0.025399217	Junction organization
ENSG00000196405	EVL	-0.321164897	-0.142415479	Actin polymerization Junction organization
ENSG00000123572	NRK	-0.005980471	-0.009225339	Junction organization
ENSG00000123094	RASSF8	0.119457967	0.054678427	Junction organization
ENSG00000105245	NUMBL	0.216588615	0.075611974	Junction organization
ENSG00000079112	CDH17	0.009452929	0.004021506	Junction organization

ENSG00000028203	VEZT	-0.30110926	-0.072680461	Junction organization
ENSG00000107263	RAPGEF1	0.24958251	0.05802994	Junction organization
ENSG00000142279	WTIP	-0.075224123	-0.031969176	Junction organization
ENSG00000066032	CTNNA2	-0.191004371	-0.164717544	Junction organization
ENSG00000128849	CGNL1	0.133127066	0.160244025	Junction organization
ENSG00000148498	PARD3	-0.68872521	-0.168172831	Junction organization
ENSG00000111052	LIN7A	-0.055094728	-0.064272703	Junction organization
ENSG00000170558	CDH2	-0.081070801	-0.084281445	Junction organization
ENSG00000092820	EZR	-0.0692457	-0.060596153	Junction organization
ENSG00000196581	AJAP1	0.014087075	0.036484885	Junction organization
ENSG00000105879	CBLL1	0.24540869	0.07905453	Junction organization
ENSG00000173801	JUP	0.03880525	0.03794143	Junction organization
ENSG00000198910	L1CAM	-0.021540397	-0.046639908	Junction organization
ENSG00000163531	NFASC	0.148456805	0.086367754	Junction organization
ENSG00000117155	SSX2IP	-0.058873474	-0.064761919	Junction organization
ENSG00000156299	TIAM1	0.018751852	0.026172225	Junction organization
ENSG00000163412	EIF4E3	0.05776898	0.053575212	Translation initiation
ENSG00000148730	EIF4EBP2	0.04712675	0.014552469	Translation initiation
ENSG00000063046	EIF4B	-0.2280096	-0.086328	Translation initiation
ENSG00000197111	PCBP2	0.106929224	0.033134444	Translation initiation
ENSG00000134030	CTIF	0.047308078	0.017456987	Translation initiation
ENSG00000125977	EIF2S2	-0.22274827	-0.065375001	Translation initiation
ENSG00000158417	EIF5B	0.40563521	0.094520568	Translation initiation
ENSG00000106682	EIF4H	-0.372446843	-0.093474236	Translation initiation
ENSG00000132507	EIF5A	0.076915329	0.042019631	Translation initiation
ENSG00000067248	DHX29	-0.252563915	-0.062864723	Translation initiation
ENSG00000178982	EIF3K	0.032302086	0.006318761	Translation initiation
ENSG00000100353	EIF3D	-0.042319294	-0.010052054	Translation initiation

ENSG00000110321	EIF4G2	-0.42109715	-0.084644846	Translation initiation
ENSG00000134001	EIF2S1	0.601898328	0.125769833	Translation initiation
ENSG00000130741	EIF2S3	-0.043741951	-0.013654019	Translation initiation
ENSG00000086232	EIF2AK1	0.049218399	0.008409422	Translation initiation
ENSG00000243056	EIF4EBP3	0.140509067	0.133625323	Translation initiation
ENSG00000034510	TMSB10	-0.258687809	-0.090344164	Actin polymerization
ENSG00000136754	ABI1	0.703312755	0.161875393	Actin polymerization
ENSG00000187239	FNBP1	-0.190728503	-0.081926167	Actin polymerization
ENSG00000090924	PLEKHG2	-0.330897417	-0.120198563	Actin polymerization
ENSG00000136732	GYPC	0.001727358	0.001148747	Actin polymerization
ENSG00000106211	HSPB1	0.043295352	0.037462604	Actin polymerization
ENSG00000134215	VAV3	-0.039381555	-0.058022034	Actin polymerization
ENSG00000007237	GAS7	-0.02781765	-0.032878578	Actin polymerization
ENSG00000254999	BRK1	0.302218194	0.067602702	Actin polymerization
ENSG00000125753	VASP	0.015685437	0.007135417	Actin polymerization
ENSG00000070087	PFN2	-0.144624282	-0.071408446	Actin polymerization
ENSG00000132970	WASF3	-0.225682275	-0.145618229	Actin polymerization
ENSG00000137478	FCHSD2	-0.057108071	-0.020695706	Actin polymerization
ENSG00000105443	CYTH2	-0.318138825	-0.092354343	Actin polymerization
ENSG00000197969	VPS13A	0.055511854	0.024584931	Actin polymerization
ENSG00000100592	DAAM1	0.057196414	0.030164331	Actin polymerization
ENSG00000131504	DIAPH1	0.090678132	0.02788895	Actin polymerization
ENSG00000055163	CYFIP2	0.099728834	0.119196271	Actin polymerization
ENSG00000147676	MAL2	0.093367114	0.054684027	Actin polymerization
ENSG00000198848	CES1	0.08904118	0.213807278	Lipid transport
ENSG00000130208	APOC1	-0.069950992	-0.073675416	Lipid transport
ENSG00000087237	CETP	-0.040538899	-0.028076057	Lipid transport
ENSG00000102595	UGGT2	-0.07181183	-0.027930465	ERQC
ENSG00000185825	BCAP31	0.436033394	0.133785732	ERQC
ENSG00000204308	RNF5	-0.303471855	-0.08985336	ERQC
ENSG00000116661	FBXO2	-0.129076237	-0.137552948	ERQC
ENSG00000239857	GET4	0.174054233	0.11960688	ERQC
ENSG00000155313	USP25	0.089393503	0.037796398	ERQC
ENSG00000147475	ERLIN2	-0.201119054	-0.053173061	ERQC
ENSG00000133872	SARAF	0.210988556	0.092044456	ERQC
ENSG00000135506	OS9	-0.09068598	-0.026793206	ERQC
ENSG00000100350	FOXRED2	0.252944926	0.122488326	ERQC
ENSG00000184787	UBE2G2	0.15662805	0.032885851	ERQC
ENSG00000170881	RNF139	0.140211116	0.061521906	ERQC

ENSG00000113194	FAF2	0.425512085	0.070863665	ERQC
ENSG00000136731	UGGT1	-0.583578738	-0.136072435	ERQC
ENSG00000172071	EIF2AK3	0.06871913	0.022151185	ERQC Translation initiation
ENSG00000115307	AUP1	-0.036892608	-0.010667756	ERQC
ENSG00000116406	EDEM3	0.044310763	0.015588211	ERQC
ENSG00000144285	SCN1A	0.056868128	0.127184712	Sodium transport
ENSG00000110881	ASIC1	-0.039031864	-0.021913201	Sodium transport
ENSG00000136531	SCN2A	0.043863329	0.088218121	Sodium transport
ENSG00000090020	SLC9A1	0.260266387	0.07173017	Sodium transport
ENSG00000138074	SLC5A6	-0.129596731	-0.044213498	Sodium transport
ENSG00000166828	SCNN1G	55.58189494	0.016097875	Sodium transport
ENSG00000108684	ASIC2	0.109669358	0.093913509	Sodium transport
ENSG00000153253	SCN3A	-0.094709183	-0.121966637	Sodium transport
ENSG00000010379	SLC6A13	-0.229526716	-0.138485092	Sodium transport
ENSG00000197818	SLC9A8	0.077551666	0.019261143	Sodium transport
ENSG00000072182	ASIC4	-0.0368954	-0.030325705	Sodium transport
ENSG00000111319	SCNN1A	0.121327298	0.117784901	Sodium transport
ENSG00000158865	SLC5A11	-0.108365157	-0.076438251	Sodium transport
ENSG00000089327	FXYD5	0.014910649	0.007213908	Sodium transport

Table A3.2. Candidate genes for age reversal screen.

Library of candidate genes listed by HGNC gene symbol. DEG denotes differential expression in at least one age model used in DGEA. DEGscores calculated from age groups- (ag) and decades- (d) based DGEAs were rounded to 3 decimal places. NA in either DEGscore column indicates the gene was not differentially expressed using that age model. Ranks in network-scored lists for each age model given by ag_rank and d_rank. Network scores in binary gene networks given by ag_ns and d_ns. Summing ag_rank and d_rank yields a combined rank score in Borda.

Gene symbol	DEG	ag_ DEG score	d_ DEG score	ag_ rank	d_ rank	ag_ns	d_ns	Borda
ACTG1	DEG	0.456	3.927	14	9	1.1E+05	5.9E+05	23
ACTL6A	DEG	0.529	0.934	19	43	1.1E+05	5.5E+05	62
AKT1	DEG	0.985	3.813	90	49	9.4E+04	5.4E+05	139
ATG5	DEG	0.056	NA	1026	NA	5.7E+04	NA	NA
CCNA2	DEG	0.541	2.625	36	48	1.0E+05	5.4E+05	84
CCNB1	DEG	1.296	4.689	21	30	1.1E+05	5.6E+05	51
CDC20	DEG	0.697	5.481	57	132	9.8E+04	4.9E+05	189
CDC34	DEG	0.115	2.258	65	130	9.7E+04	4.9E+05	195
CDK1	DEG	0.362	1.902	5	10	1.1E+05	5.9E+05	15
CDK2	DEG	0.541	1.753	13	17	1.1E+05	5.8E+05	30
CDK4	DEG	0.149	0.457	24	26	1.1E+05	5.7E+05	50
CDKN2A	nonDEG	NA	NA	17	8	1.1E+05	5.8E+05	25
CHEK1	DEG	0.288	0.608	58	102	9.8E+04	5.0E+05	160
CHEK2	DEG	0.529	0.847	75	89	9.6E+04	5.1E+05	164
CTNNB1	DEG	0.743	0.886	49	37	9.9E+04	5.5E+05	86
CXCR2	nonDEG	NA	NA	3030	2078	3.1E+04	1.8E+05	5108
CYCS	nonDEG	NA	NA	19	9	1.1E+05	5.8E+05	28
DDX3X	DEG	0.568	1.487	83	134	9.5E+04	4.9E+05	217
DICER1	DEG	0.358	0.94	7	6	1.1E+05	6.0E+05	13
EEF2	DEG	0.125	1.362	37	50	1.0E+05	5.4E+05	87
ESR1	DEG	4.343	19.16	128	72	9.1E+04	5.2E+05	200
FAU	DEG	0.339	4.264	62	73	9.7E+04	5.2E+05	135
GART	nonDEG	NA	NA	18	26	1.1E+05	5.5E+05	44
GSK3B	DEG	0.609	6.045	39	29	1.0E+05	5.6E+05	68
HDAC1	DEG	0.272	0.674	11	15	1.1E+05	5.8E+05	26
HDAC6	DEG	0.459	1.075	72	81	9.6E+04	5.1E+05	153
HES1	nonDEG	NA	NA	3271	2401	2.9E+04	1.6E+05	5672
HNRNPA2B1	DEG	0.399	0.516	85	156	9.5E+04	4.8E+05	241
HSP90AB1	DEG	0.143	0.453	9	13	1.1E+05	5.9E+05	22

HSP90B1	DEG	0.769	0.388	43	53	1.0E+05	5.4E+05	96
HSPA5	DEG	0.118	0.449	8	8	1.1E+05	6.0E+05	16
HSPA8	DEG	0.199	0.656	1	1	1.2E+05	6.4E+05	2
HSPD1	DEG	0.244	1.162	52	105	9.9E+04	5.0E+05	157
ISG15	DEG	0.161	5.821	84	86	9.5E+04	5.1E+05	170
KAT7	nonDEG	NA	NA	1477	1245	4.7E+04	2.4E+05	2722
KDM6A	DEG	0.059	0.329	4	11	1.1E+05	5.9E+05	15
LRRK2	DEG	5.109	19.83	3	3	1.1E+05	6.2E+05	6
MAPK14	DEG	0.235	1.305	34	22	1.0E+05	5.7E+05	56
MAPK7	DEG	0.403	3.262	115	88	9.2E+04	5.1E+05	203
MCM7	DEG	0.674	3.38	61	151	9.7E+04	4.8E+05	212
MDM2	DEG	3.946	21.69	94	97	9.4E+04	5.0E+05	191
MTOR	nonDEG	NA	NA	9	2	1.1E+05	6.1E+05	11
NFKB1	DEG	0.379	1.988	42	38	1.0E+05	5.5E+05	80
NOTCH1	DEG	0.212	3.723	125	74	9.1E+04	5.2E+05	199
NPM1	DEG	0.9	7.332	70	122	9.6E+04	4.9E+05	192
OGT	DEG	0.334	2.829	40	45	1.0E+05	5.5E+05	85
P4HB	DEG	0.259	0.618	116	112	9.2E+04	4.9E+05	228
PARK7	DEG	0.132	0.469	80	92	9.5E+04	5.1E+05	172
PIK3C2A	DEG	0.187	0.606	118	119	9.2E+04	4.9E+05	237
PIK3C3	DEG	0.281	0.349	126	110	9.1E+04	5.0E+05	236
PIK3CA	DEG	1.285	4.451	82	68	9.5E+04	5.2E+05	150
PKM	DEG	0.37	0.483	67	78	9.7E+04	5.2E+05	145
PLK1	DEG	0.759	6.376	35	64	1.0E+05	5.3E+05	99
PPIA	DEG	0.78	2.225	117	109	9.2E+04	5.0E+05	226
PPP1CA	DEG	0.693	2.348	2	7	1.1E+05	6.0E+05	9
PPP1CB	DEG	2.996	14.67	17	25	1.1E+05	5.7E+05	42
PPP1CC	nonDEG	NA	NA	4	3	1.1E+05	6.0E+05	7
PPP2CA	nonDEG	NA	NA	5	10	1.1E+05	5.8E+05	15
PPP3CA	DEG	1.133	4.202	93	61	9.4E+04	5.3E+05	154
PPP4C	DEG	0.195	1.236	30	52	1.0E+05	5.4E+05	82
PSMC6	DEG	0.141	0.353	51	87	9.9E+04	5.1E+05	138
PSMD4	DEG	1.503	7.353	55	111	9.8E+04	4.9E+05	166
PTEN	DEG	0.956	4.87	45	34	9.9E+04	5.6E+05	79
RAD23B	DEG	1.066	9.218	44	82	9.9E+04	5.1E+05	126
RAD51	DEG	0.679	0.984	18	47	1.1E+05	5.5E+05	65
RNASEL	DEG	0.268	0.692	6	4	1.1E+05	6.1E+05	10
RPLP0	DEG	0.641	1.676	86	118	9.5E+04	4.9E+05	204
RPS27A	DEG	0.513	3.82	10	16	1.1E+05	5.8E+05	26

RRM1	DEG	0.527	2.258	64	171	9.7E+04	4.7E+05	235
RUVBL2	DEG	0.786	5.295	68	159	9.6E+04	4.8E+05	227
SIRT1	DEG	0.137	0.171	54	40	9.8E+04	5.5E+05	94
SLC2A10	DEG	0.176	0.306	108	121	9.3E+04	4.9E+05	229
SLC2A13	DEG	0.574	2.635	26	46	1.0E+05	5.5E+05	72
SMARCA4	DEG	0.29	3.459	46	57	9.9E+04	5.3E+05	103
SOD1	DEG	0.024	0.087	66	51	9.7E+04	5.4E+05	117
SOD2	DEG	4.002	5.696	97	75	9.3E+04	5.2E+05	172
SRSF1	DEG	0.375	1.484	73	153	9.6E+04	4.8E+05	226
SUMO1	nonDEG	NA	NA	10	17	1.1E+05	5.6E+05	27
TP53	nonDEG	NA	NA	22	16	1.0E+05	5.6E+05	38
TXN	DEG	0.203	0.526	87	80	9.4E+04	5.1E+05	167
UBA52	DEG	0.324	0.417	47	79	9.9E+04	5.1E+05	126
UBB	DEG	0.426	1.037	25	23	1.1E+05	5.7E+05	48
UBE2A	DEG	0.047	0.107	60	115	9.7E+04	4.9E+05	175
UBE2I	DEG	0.363	1.359	16	36	1.1E+05	5.5E+05	52
UBE2N	DEG	0.077	0.157	22	55	1.1E+05	5.4E+05	77
UBE2S	DEG	0.334	8.326	69	138	9.6E+04	4.9E+05	207
UBQLN1	DEG	0.031	0.201	81	93	9.5E+04	5.1E+05	174
UBXN7	DEG	0.418	1.623	48	96	9.9E+04	5.1E+05	144
VCP	DEG	0.131	0.725	23	35	1.1E+05	5.5E+05	58
WDTC1	nonDEG	NA	NA	14	4	1.1E+05	6.0E+05	18
XPO1	nonDEG	NA	NA	3	5	1.2E+05	6.0E+05	8
YWHAB	DEG	0.534	2.066	20	24	1.1E+05	5.7E+05	44
YWHAQ	nonDEG	NA	NA	24	15	1.0E+05	5.7E+05	39

Table A3.3: Plasmids and barcodes for the candidate genes.

Library of plasmids used in the screen listed by internal gene_ID, internal plasmid id (pAMP#), HGNC gene symbol, and barcode sequence for associated gene. For most of the lines BC_A was used, unless otherwise specified.

gene_ ID	pAMP#	gene_symbol	BC_A	BC_B
1	450	ACTG1	CACAAACTTATCGTC TATTC	TTTCCTAACCCTGAA GAAAA
2	451	ACTL6A	CTGTTATGTGGTGC TTCTAT	AGTTAAGGTTTGTG CTGTGA
3	452	AKT1	CATAACATCCCTATT TGACA	GATTCAGAATCTTAT AATTT
4	453	ATG5	AAGAGGCACTTGCA CTGTCA	CATTCCCACGAAACC CATCT
5	455	CCNA2	GAGGTGTGCTTAGA GAAGAA	ATTTCTCAAGCATTA GTTTA
6	456	CCNB1	TATTATGCCTTTTCA ATATT	GCTGTGCCACGTGC GATTTT
7	457	CDC20	AAAGACGAAGTGAA GTATTC	TAGTCATTCAGGTCA GGTAA
8	458	CDC34	CTTATGAAAGAGCA TTTTTA	AAGTAAGAAACCGTC TTCCT
9	459	CDK1	CATATTTAGTTCTCC TTTCTT	TCGTCCGTGTTTGT AGTCAA
10	460	CDK2	AAAGTCTTCATTTCT CCATG	GAACGTCAAAGTGAC AAATT
11	461	CDK4	TTTTCAAAAAAGATT CGTTA	GCGTGGAACCATTC TTGTGA
12	462	CDKN2A	TTATTGGAGACAAC AACGCA	AATCAACAGTGAAAC ACCAA
13	463	CHEK1	CCGCAACCGTTACG AACCAA	GAAGTTAGAAACAAC AAACA
14	464	CHEK2	TAACATGGCATGAA TTCTAG	CACACCAAATCAGTT TGAAA
15	465	CTNNB1	AGAAGGTATATGCA AATTGA	TAGCGGCAAAATAGA GTTAA
16	466	CXCR2	GCCGGACCACCTCC AGAAAG	TTCGAAAAATCCAAA TAAAA
17	467	CYCS	CTCGGAATTCCACT AATTGA	GAATTAGAGGCTGG AGTACA
18	468	DDX3X	ATATACTAGTTACCG TATCA	GGTACCAGCCCCCA TTGCAA
19	469	DICER1	AATAGCCTCCAACTC AACTA	AGCCATCTTCCATCG TAATA

20	471	EEF2	ACGCAATTGTATCAAATCAA	AACAAAACACCTTGCTTCAT
21	472	ESR1	TGCCCACAGATGAGCTTTAT	TACAACAACTGCTGTTGTTG
22	473	FAU	TTACTGCAGATTTTAATAGT	TTAATCTTTAAAAGCAGTTT
23	474	GART	TAATTCCAGACTGAAACAAG	CTTGTACCATTCATATCAGA
24	475	GSK3B	TAGACCCTACGGTCAAACTA	GAACGCTAAGAAAAGATTTG
25	476	HDAC1	AGGCTAACGGATTAATCTCC	ACCAGATATGCCCCAGGTGT
26	477	HDAC6	TATAAAAAATCGGTTCGATT	GCCTTCCCTTCCAGATCCTA
27	478	HES1	GCGTGTGATGCAGTGGTGTA	GAATTGTAAAAATTTACAGA
28	479	HNRNPA2B1	TAGCCTTTCTCAAGAAATCA	GGCCATATTGCGCATATTGA
29	481	HSP90AB1	TTGAAGAACAGCAACTTAGA	CGGTGCATCCACTTGGGCGT
30	482	HSP90B1	CGACGTGCGCCTTCCTGATT	AGGTCAATAGACTTAAACTG
31	483	HSPA5	TTTTTAGTAATCTTGCCAGC	GGCAAAAAGTGGAAGCCTGA
32	484	HSPA8	AAATAATGTAAATAACAATT	TTTGACCCAGCTTTGCGTAG
33	485	HSPD1	GTAAGTTTCAGTTGGATGAG	AAATACTTGCAGTGTCCTCT
34	486	ISG15	ATTGCAGGATTCTCATTGTC	AGTTCTTGATCAGGAATATG
35	487	KAT7	GTTAATCTCATCCACCTCCT	ACTTCTACACGTAATTGCCA
36	488	KDM6A	TCTTTGGTTCAAATTGCTA	TTGACCAGAACAACACATCA
37	490	LRRK2	ATCCAAATCTACAGATGCTG	CCAACAATAACGTTCAGTTG
38	491	MAPK14	CAAAGTACAATTGAGTTCTT	TAAAGATTTACTAGATTTAA
39	492	MAPK7	GTTTGAATCAATATTGTTCT	TTGAATTTGAAGTGAGCGAA
40	493	MCM7	TGCTCAACAACTGTTGCCTT	TTAGGATGGTCTTGATCAAT
41	494	MDM2	AATGCGAGTTCCAACAGATT	ATGCACTGCATGTGGTTTCTG
42	495	MTOR	TGCATTTCCTTTCCCGGTTT	GTATTCATATAGTAAGTCA

43	496	NFKB1	CTACAAATAACAAGGCTAAA	ACTACTCTGAATTATCTTAT
44	497	NOTCH1	TAGTGTTGACGATCCTGCTG	AAGTGTACGTAGGTACATAT
45	498	NPM1	GAATACGAAGATTGAAATTA	AAGTAAGTTATTATCCAAAA
46	499	OGT	TATTCCGAGTTTATAAAAAT	TTGCGGTGGTTATGACGAGG
48	501	P4HB	ACACGTGACTGCTTGAAT	TGCTAGAGCGATTATTCCTA
49	502	PARK7	TGATGAGACTCTCATCATCA	GCTCATTTCATGTGAAGCTA
50	503	PIK3C2A	TTATTGGCGAGGTTAACAAC	CTCCGGCCTCCATCACCACG
51	504	PIK3C3	GGCAACGCACTTGACGCCA	CTGTTAGACACAATATAAAT
52	505	PIK3CA	CAGCGTCCCCATTAAATACA	TGTACATTAAGACATCCGCC
53	506	PKM	ATAATAGAGTCTTGTAGACT	AAACTTTCACCAATAACTCA
54	507	PLK1	TTGAAATACAAAGAGAAGAA	GAAGCCAGTAATGCTTCAGA
55	508	PPIA	AGAAGACCTTGATAAGCGCT	GGCAAGGCTTACGACAGTGA
56	509	PPP1CA	GGTAAGAAGACCGTCAAGAG	CACTTATAGTATCCGTTAAA
57	510	PPP1CB	GCCTTGTTCTAATCCATCTG	CCTCTTTGCAGTTTTTCAGT
58	511	PPP1CC	TAACAGTGAATTCACTATT	ACCGAACTTTTCCAAAAGCT
59	512	PPP2CA	TCACTGTTCAAATCAAAATA	GCGGTAGCGCCACGCCTACC
60	513	PPP3CA	AGCTGTAAGTGTAGATTCGC	CCATCTCTTTTCGAAATTAA
61	514	PPP4C	CCCCAATCATGAAGTTGCAA	TCTTCCTTGATCAATCAAAA
62	515	PSMC6	ACCCAGGGCTAAAGAACTTC	GTGATGGTCAGTGATATCCT
63	516	PSMD4	AACGCCCATTATGCAAAAGT	GGAAAAATCTTTATAGAGGG
64	517	PTEN	AACAGGAGCAGTATCCCCGA	TTGTCCATTGAAACCACCTC
65	518	RAD23B	TACCGAATGCAATGCCGCAC	TTCTTCCGACCAAACTGAAC
66	519	RAD51	TTCGTGAATCGGTTCCTTCT	AAGACCATCTGACTCAAACA

67	520	RNASEL	AGCTGTCTTGAAATCCTTCT	CATATTTAGTTCTCCTTCTT
68	521	RPLP0	CAACTTAATACCGAATTCGT	ACCAACGATTCAGAAAGATT
69	522	RPS27A	TGTGCGAAAATATCTCCTGA	TAAATTTGAAGCTAATGAAA
70	523	RRM1	GCTGCCTCTCAGATTTCTGA	TCCGCAGTTGGTGCTTATCC
71	524	RUVBL2	GACAAAACCGAGATTGCTGA	TACTGCTGCTGCTGCTGTT
72	525	SIRT1	ATCAGTTTCTGTGTTATTAT	TGTAAGGATTTGCTTATTAG
73	526	SLC2A10	TGTTGTTGCATCCTAATAAT	TGTGGGCATGATCTACCAAG
74	527	SLC2A13	TGGATACTAACCTAAGAGTT	GATACATATATCCACAGAGT
75	528	SMARCA4	AGGACCATCTGACTCAAACGA	TTTTCTGGGGTGGCTCCCAA
76	529	SOD1	CTGTAAATTGTGTTGTGCCG	ATTGGCGGTGCGCTCGCCGA
77	530	SOD2	GAGTCCGCACCAGGGCCGGC	AAGGAAAGACCAAGTTGTAT
78	531	SRSF1	GCTGCGTTTGATGAGTTTGT	TTCCAGCCGCTGTGTCCATT
79	532	SUMO1	ACAAGCCATCAACGTAGACG	GATTATTGTAATAAAAATAT
80	533	TP53	ATGCGCTGATTTCCAACTGT	TTCATATCGATCTTTAGCGC
81	534	TXN	CTCCATTGTACAAGACAACT	GAAATAAAACCGTTAACTAG
82	535	UBA52	GGGTTCTGCCCTTTGCGCTT	GACTCCGTCCTTCACACCGT
83	536	UBB	AAACGGTAGATCACTGTAAT	TGAATCACAGAAGAATACTC
84	537	UBE2A	CCGGCCTTGTGGTGCGAGA	GATATATTTAGTTGCTTCGC
85	538	UBE2I	GGCTTCATTATCATTGTCAC	CCATATTATGAATTTGAAGA
86	539	UBE2N	GCTTCGGGCTCTGTCATTCG	GGGTGAGCTAAATTTCATCG
87	540	UBE2S	TTTATCGTTTGTACAATGAA	AGCAGGTTTATCTGATGCAG
88	541	UBQLN1	TTTAGACGCTTGTGATTCAC	GTGGTATAACGTATGTATTA
89	542	UBXN7	GACTGTCTTCTTGTTAAAAT	TGACCATATCTGATATATTT

90	543	VCP	AATCGGTGACGTAAAATCGG	AGGTACAATTAATGAGTTCT
91	544	WDTC1	TCTCAAAAGCGTCAAAATCA	TGTGCGATATCTACGAGTAT
92	545	XPO1	TATGCCCAAAGTCGCAAGTA	CCCTTTTGATTGCCTGTTTA
93	546	YWHAB	CTTGATACCCTTGAATTCTT	GTGCTAGATACAAAGAAATC
94	547	YWHAQ	TGATGGTCTTTACACTTTGT	CACCATACGAATGGTGCATT
47	500	OSKM	CTTTTTAGGGAGTCTTTTCA	GAATACTTGGACACCGTTAT
95	548	mTagBFP	CTTTAGGCCTAATACCATTT	TGTTCCAAGACCTCAATCTT

Table A3.4: Predicted age of samples in the age-reversal screen.

Predicted ages of over-expression lines of the initial perturbation screen ("Predicted Age (years)" column). The over-expressed gene ("Gene" column), the number of passages before over-expression ("Passage" column) and the chronological age of the donor ("Line Age (years)" column) is shown. Wild-type samples are denoted as "NTg" in the "Gene" column. Replicate identifiers ("Replicate" column) were randomly assigned before conducting the experiments.

Line	Gene	Replicate	Passage	Line Age (years)	Predicted Age (years)
AG12657	mTagBFP2	1	12	55	66.35
AG12657	mTagBFP2	2	12	55	65.56
AG12657	NTg	1	9	55	48.21
AG12657	NTg	2	9	55	55.65
AG12657	ACTG1	1	12	55	41.07
AG12657	ACTG1	2	12	55	60.23
AG12657	ACTL6A	1	12	55	72.29
AG12657	ACTL6A	2	12	55	76.65
AG12657	AKT1	1	12	55	73.91
AG12657	AKT1	2	12	55	74.24
AG12657	ATG5	1	12	55	53.55
AG12657	ATG5	2	12	55	71.83
AG12657	CCNA2	2	12	55	73.69
AG12657	CCNA2	1	12	55	72.37
AG12657	CCNB1	1	12	55	76.48
AG12657	CCNB1	2	12	55	73.07
AG12657	CDC20	1	12	55	72.94
AG12657	CDC20	2	12	55	74.38
AG12657	CDC34	1	12	55	71.39
AG12657	CDC34	2	12	55	73.12
AG12657	CDK1	1	12	55	78.95
AG12657	CDK1	2	12	55	76.12
AG12657	CDK2	1	13	55	70.47
AG12657	CDK2	2	13	55	71.29
AG12657	CDK4	1	12	55	72.02
AG12657	CDK4	2	12	55	75.76
AG12657	CDKN2A	1	12	55	65.42
AG12657	CDKN2A	2	12	55	71.88
AG12657	CHEK1	1	12	55	75.66
AG12657	CHEK1	2	12	55	74.8
AG12657	CHEK2	1	12	55	71.62

AG12657	CHEK2	2	12	55	66.59
AG12657	CTNNB1	1	12	55	81.29
AG12657	CTNNB1	2	12	55	77.92
AG12657	CXCR2	1	12	55	75.74
AG12657	CXCR2	2	12	55	79.24
AG12657	CYCS	1	12	55	66.71
AG12657	CYCS	2	12	55	70.94
AG12657	DDX3X	1	12	55	77.94
AG12657	DDX3X	2	12	55	76.38
AG12657	EEF2	1	12	55	77.65
AG12657	EEF2	2	12	55	82.54
AG12657	ESR1	1	12	55	78.48
AG12657	ESR1	2	12	55	80.08
AG12657	FAU	1	12	55	77.16
AG12657	FAU	2	12	55	77.39
AG12657	GART	1	13	55	74.56
AG12657	GART	2	13	55	73.66
AG12657	GSK3B	1	12	55	67.29
AG12657	GSK3B	2	12	55	78.88
AG12657	HDAC1	1	12	55	80.09
AG12657	HDAC1	2	12	55	72.38
AG12657	HDAC6	1	12	55	71.62
AG12657	HDAC6	2	12	55	73.1
AG12657	HES1	1	12	55	75.04
AG12657	HES1	2	12	55	70.61
AG12657	HNRNPA2B1	1	12	55	74.54
AG12657	HNRNPA2B1	2	12	55	70.55
AG12657	HSP90AB1	1	12	55	64.62
AG12657	HSP90AB1	2	12	55	71.92
AG12657	HSP90B1	1	12	55	62.84
AG12657	HSP90B1	2	12	55	72.72
AG12657	HSPA5	1	12	55	73.38
AG12657	HSPA5	2	12	55	70.64
AG12657	HSPA8	1	12	55	75.96
AG12657	HSPA8	2	12	55	72.34
AG12657	HSPD1	1	12	55	71.33
AG12657	HSPD1	2	12	55	74.7
AG12657	ISG15	1	12	55	72.27
AG12657	ISG15	2	12	55	77.97

AG12657	KAT7	1	12	55	89.85
AG12657	KAT7	2	12	55	89.01
AG12657	KDM6A	1	12	55	74.05
AG12657	KDM6A	2	12	55	74.63
AG12657	MAPK14	1	12	55	74.11
AG12657	MAPK14	2	12	55	70.9
AG12657	MAPK7	1	12	55	74.54
AG12657	MAPK7	2	12	55	76.4
AG12657	MCM7	1	12	55	72.2
AG12657	MCM7	2	12	55	71.26
AG12657	MDM2	1	12	55	74.67
AG12657	MDM2	2	12	55	73.64
AG12657	MTOR	1	12	55	62.37
AG12657	MTOR	2	12	55	68.71
AG12657	NFKB1	1	12	55	69.85
AG12657	NFKB1	2	12	55	67.94
AG12657	NOTCH1	1	12	55	75.89
AG12657	NOTCH1	2	12	55	58.58
AG12657	NPM1	1	12	55	37.38
AG12657	NPM1	2	12	55	34.27
AG12657	OGT	1	13	55	73.33
AG12657	OGT	2	13	55	75.75
AG12657	OSKM	1	12	55	43.4
AG12657	OSKM	2	12	55	71.87
AG12657	P4HB	1	12	55	63.68
AG12657	P4HB	2	12	55	71.8
AG12657	PARK7	1	12	55	71.89
AG12657	PARK7	2	12	55	71.76
AG12657	PIK3C2A	1	12	55	82.16
AG12657	PIK3C2A	2	12	55	79.18
AG12657	PIK3CA	1	12	55	71.93
AG12657	PIK3CA	2	12	55	72.09
AG12657	PKM	2	12	55	71.01
AG12657	PKM	1	12	55	67.77
AG12657	PLK1	1	12	55	74.31
AG12657	PLK1	2	12	55	73.27
AG12657	PPIA	1	12	55	67.11
AG12657	PPIA	2	12	55	60.18
AG12657	PPP1CA	1	12	55	79.52

AG12657	PPP1CA	2	12	55	74.4
AG12657	PPP1CB	1	12	55	65.76
AG12657	PPP1CB	2	12	55	69.76
AG12657	PPP1CC	1	12	55	70.19
AG12657	PPP1CC	2	12	55	68.9
AG12657	PPP2CA	1	12	55	71.19
AG12657	PPP2CA	2	12	55	59.7
AG12657	PPP3CA	1	12	55	69.82
AG12657	PPP3CA	2	12	55	74.42
AG12657	PPP4C	1	12	55	73.6
AG12657	PPP4C	2	12	55	70.28
AG12657	PSMC6	1	12	55	71.28
AG12657	PSMC6	2	12	55	72.35
AG12657	PSMD4	1	10	55	66.51
AG12657	PSMD4	2	10	55	55.01
AG12657	PTEN	1	12	55	75.51
AG12657	PTEN	2	12	55	71.83
AG12657	RAD23B	1	12	55	73.58
AG12657	RAD23B	2	12	55	72.15
AG12657	RAD51	2	12	55	61.62
AG12657	RAD51	1	12	55	56.71
AG12657	RNASEL	1	12	55	44.44
AG12657	RNASEL	2	12	55	15.8
AG12657	RPLP0	1	12	55	70.5
AG12657	RPLP0	2	12	55	72.99
AG12657	RPS27A	1	12	55	72.4
AG12657	RPS27A	2	12	55	70.34
AG12657	RRM1	1	12	55	71.43
AG12657	RRM1	2	12	55	61.96
AG12657	RUVBL2	1	12	55	70.05
AG12657	RUVBL2	2	12	55	71.77
AG12657	SIRT1	1	12	55	67.15
AG12657	SIRT1	2	12	55	73.28
AG12657	SLC2A10	1	12	55	55.88
AG12657	SLC2A10	2	12	55	57.73
AG12657	SLC2A13	1	12	55	36.84
AG12657	SLC2A13	2	12	55	47.4
AG12657	SMARCA4	1	11	55	58.69
AG12657	SMARCA4	2	11	55	66.09

AG12657	SOD1	1	12	55	77.53
AG12657	SOD1	2	12	55	70.9
AG12657	SOD2	1	12	55	69.27
AG12657	SOD2	2	12	55	71.16
AG12657	SRSF1	1	12	55	7.98
AG12657	SRSF1	2	12	55	16.31
AG12657	SUMO1	1	12	55	43.17
AG12657	SUMO1	2	12	55	49.68
AG12657	TP53 (DN)	1	12	55	50.67
AG12657	TP53 (DN)	2	12	55	72.61
AG12657	TXN	1	12	55	71.81
AG12657	TXN	2	12	55	74.33
AG12657	UBA52	1	12	55	68.35
AG12657	UBA52	2	12	55	66.23
AG12657	UBB	1	12	55	75.7
AG12657	UBB	2	12	55	75.49
AG12657	UBE2A	1	12	55	74.66
AG12657	UBE2A	2	12	55	70.57
AG12657	UBE2I	1	12	55	72.95
AG12657	UBE2I	2	12	55	74.84
AG12657	UBE2N	1	12	55	79.9
AG12657	UBE2N	2	12	55	74.61
AG12657	UBE2S	1	12	55	72.16
AG12657	UBE2S	2	12	55	71.68
AG12657	UBQLN1	1	12	55	38.8
AG12657	UBQLN1	2	12	55	73.69
AG12657	UBXN7	1	13	55	72
AG12657	UBXN7	2	13	55	76.34
AG12657	VCP	1	13	55	76.77
AG12657	VCP	2	13	55	77.09
AG12657	WDTC1	1	12	55	76.74
AG12657	WDTC1	2	12	55	72.1
AG12657	XPO1	1	12	55	62.79
AG12657	XPO1	2	12	55	66.92
AG12657	YWHAB	1	12	55	67.07
AG12657	YWHAB	2	12	55	74.69
AG12657	YWHAQ	1	12	55	71.58
AG12657	YWHAQ	2	12	55	71.78
AG06294A	mTagBFP2	1	10	65	64.72

AG06294A	mTagBFP2	2	10	65	64.67
AG06294A	NTg	1	5	65	58.61
AG06294A	NTg	2	5	65	65.53
AG06294A	ACTL6A	1	10	65	74.83
AG06294A	ACTL6A	2	10	65	76.28
AG06294A	ATG5	1	10	65	73.44
AG06294A	ATG5	2	10	65	72.63
AG06294A	CCNA2	1	10	65	75.74
AG06294A	CCNA2	2	10	65	76.54
AG06294A	CCNB1	1	13	65	74.12
AG06294A	CCNB1	2	13	65	65.73
AG06294A	CDC20	2	10	65	74.84
AG06294A	CDC20	1	10	65	74.42
AG06294A	CDK1	1	10	65	72.3
AG06294A	CDK1	2	10	65	70.97
AG06294A	CDKN2A	1	10	65	78.64
AG06294A	CDKN2A	2	10	65	75.35
AG06294A	CHEK2	1	10	65	73.65
AG06294A	CHEK2	2	10	65	61.23
AG06294A	CXCR2	1	10	65	78.92
AG06294A	CXCR2	2	10	65	75.83
AG06294A	EEF2	1	10	65	70.75
AG06294A	EEF2	2	10	65	80.24
AG06294A	ESR1	1	10	65	77.45
AG06294A	ESR1	2	10	65	83.83
AG06294A	GSK3B	1	12	65	74.02
AG06294A	GSK3B	2	12	65	72.37
AG06294A	HES1	1	10	65	73.06
AG06294A	HES1	2	10	65	74.4
AG06294A	ISG15	1	10	65	77.47
AG06294A	ISG15	2	10	65	74.08
AG06294A	KAT7	1	10	65	91.9
AG06294A	KAT7	2	10	65	89.45
AG06294A	LRRK2	1	9	65	74.85
AG06294A	LRRK2	2	9	65	71.18
AG06294A	MAPK7	1	9	65	77.95
AG06294A	MAPK7	2	9	65	79.14
AG06294A	MDM2	1	10	65	75.79
AG06294A	MDM2	2	10	65	76.42

AG06294A	NFKB1	1	9	65	66.02
AG06294A	NFKB1	2	9	65	76.28
AG06294A	OGT	1	10	65	72.87
AG06294A	OGT	2	10	65	61.04
AG06294A	OSKM	1	9	65	52.08
AG06294A	OSKM	2	9	65	56.33
AG06294A	P4HB	1	10	65	62.65
AG06294A	P4HB	2	10	65	74.74
AG06294A	PARK7	1	10	65	73.19
AG06294A	PARK7	2	10	65	78.56
AG06294A	PIK3C2A	1	9	65	73.96
AG06294A	PIK3C2A	2	9	65	77.87
AG06294A	PIK3C3	1	9	65	69.67
AG06294A	PIK3C3	2	9	65	75.43
AG06294A	PIK3CA	1	9	65	64.56
AG06294A	PIK3CA	2	9	65	73.61
AG06294A	PKM	1	10	65	70.47
AG06294A	PKM	2	10	65	78.47
AG06294A	PPIA	1	10	65	71.86
AG06294A	PPIA	2	10	65	79.41
AG06294A	PPP1CA	1	10	65	78.36
AG06294A	PPP1CA	2	10	65	77.31
AG06294A	PPP1CB	1	10	65	78.51
AG06294A	PPP1CB	2	10	65	74.23
AG06294A	PPP1CC	1	10	65	74.1
AG06294A	PPP1CC	2	10	65	78.25
AG06294A	PPP3CA	1	9	65	72.45
AG06294A	PPP3CA	2	9	65	71.34
AG06294A	PPP4C	1	10	65	75.45
AG06294A	PPP4C	2	10	65	74.73
AG06294A	PSMC6	1	10	65	76
AG06294A	PSMC6	2	10	65	81.77
AG06294A	PTEN	1	10	65	78.3
AG06294A	PTEN	2	10	65	75.76
AG06294A	RAD23B	1	10	65	72.42
AG06294A	RAD23B	2	10	65	70.43
AG06294A	RPLP0	1	10	65	70.61
AG06294A	RPLP0	2	10	65	68.27
AG06294A	RPS27A	1	10	65	72.4

AG06294A	RPS27A	2	10	65	70.6
AG06294A	SIRT1	1	9	65	52.62
AG06294A	SIRT1	2	9	65	61.63
AG06294A	SLC2A10	1	10	65	75.26
AG06294A	SLC2A10	2	10	65	70.02
AG06294A	SLC2A13	1	10	65	63.84
AG06294A	SLC2A13	2	10	65	70.05
AG06294A	SOD1	1	10	65	65.53
AG06294A	SOD1	2	10	65	76.49
AG06294A	SOD2	1	10	65	76.12
AG06294A	SOD2	2	10	65	80.12
AG06294A	SRSF1	1	10	65	24.41
AG06294A	SRSF1	2	10	65	14.32
AG06294A	TP53 (DN)	1	10	65	72.22
AG06294A	TP53 (DN)	2	10	65	70.88
AG06294A	TXN	1	10	65	75.92
AG06294A	TXN	2	10	65	73.03
AG06294A	UBA52	1	10	65	73.89
AG06294A	UBA52	2	10	65	75.13
AG06294A	UBB	1	10	65	73
AG06294A	UBB	2	10	65	64.67
AG06294A	UBE2S	1	10	65	78.94
AG06294A	UBE2S	2	10	65	79.64
AG06294A	UBQLN1	2	10	65	71.22
AG06294A	UBQLN1	1	10	65	78.88
AG06294A	WDTC1	1	9	65	74.15
AG06294A	WDTC1	2	9	65	76.55
AG06294A	YWHAB	1	10	65	75.55
AG06294A	YWHAB	2	10	65	72.2
AG06294A	YWHAQ	1	10	65	72.6
AG06294A	YWHAQ	2	10	65	75.06
AG16359	mTagBFP2	1	13	79	78.78
AG16359	mTagBFP2	2	13	79	72.05
AG16359	NTg	1	10	79	82.83
AG16359	NTg	2	10	79	75.12
AG16359	ACTG1	1	13	79	77.26
AG16359	ACTG1	2	13	79	75.2
AG16359	ACTL6A	1	13	79	69.35
AG16359	ACTL6A	2	13	79	70.19

AG16359	AKT1	1	13	79	81.72
AG16359	AKT1	2	13	79	82.24
AG16359	ATG5	1	13	79	72.4
AG16359	ATG5	2	13	79	74.04
AG16359	CCNA2	2	13	79	74.27
AG16359	CCNA2	1	13	79	76.33
AG16359	CCNB1	1	13	79	73.65
AG16359	CCNB1	2	13	79	77.41
AG16359	CDC20	1	13	79	74.37
AG16359	CDC20	2	13	79	74.46
AG16359	CDC34	1	13	79	74.7
AG16359	CDC34	2	13	79	71.33
AG16359	CDK1	1	13	79	76.23
AG16359	CDK1	2	13	79	77.82
AG16359	CDK2	1	13	79	76.28
AG16359	CDK2	2	13	79	78.73
AG16359	CDK4	1	13	79	73.28
AG16359	CDK4	2	13	79	70.68
AG16359	CDKN2A	1	13	79	78.45
AG16359	CDKN2A	2	13	79	76.17
AG16359	CHEK1	1	13	79	77.32
AG16359	CHEK1	2	13	79	80.61
AG16359	CHEK2	1	13	79	73.01
AG16359	CHEK2	2	13	79	73.23
AG16359	CTNNB1	1	13	79	79.58
AG16359	CTNNB1	2	13	79	75.8
AG16359	CXCR2	1	13	79	62.12
AG16359	CXCR2	2	13	79	79.64
AG16359	CYCS	1	13	79	73.73
AG16359	CYCS	2	13	79	74.81
AG16359	DDX3X	1	13	79	72.69
AG16359	DDX3X	2	13	79	78.34
AG16359	DICER1	1	13	79	80.16
AG16359	DICER1	2	13	79	80.31
AG16359	EEF2	1	13	79	82.91
AG16359	EEF2	2	13	79	83.95
AG16359	ESR1	1	13	79	88.87
AG16359	ESR1	2	13	79	85.62
AG16359	FAU	1	13	79	70.02

AG16359	FAU	2	13	79	73.05
AG16359	GART	1	13	79	75.79
AG16359	GART	2	13	79	58.86
AG16359	GSK3B	1	13	79	74.42
AG16359	GSK3B	2	13	79	80.31
AG16359	HDAC1	1	13	79	71
AG16359	HDAC1	2	13	79	76.55
AG16359	HDAC6	1	13	79	71.33
AG16359	HDAC6	2	13	79	70.99
AG16359	HES1	1	13	79	71.35
AG16359	HES1	2	13	79	79.19
AG16359	HNRNPA2B1	1	13	79	74.68
AG16359	HNRNPA2B1	2	13	79	74.62
AG16359	HSP90AB1	1	13	79	78.67
AG16359	HSP90AB1	2	13	79	73.23
AG16359	HSP90B1	1	13	79	73.42
AG16359	HSP90B1	2	13	79	73.43
AG16359	HSPA5	1	13	79	83.18
AG16359	HSPA5	2	13	79	85.55
AG16359	HSPA8	1	13	79	77.09
AG16359	HSPA8	2	13	79	77.59
AG16359	HSPD1	1	13	79	79.53
AG16359	HSPD1	2	13	79	75.55
AG16359	ISG15	1	13	79	76.13
AG16359	ISG15	2	13	79	78.42
AG16359	KAT7	1	13	79	92.21
AG16359	KAT7	2	13	79	97.86
AG16359	KDM6A	1	13	79	82.23
AG16359	KDM6A	2	13	79	82.68
AG16359	LRRK2	1	13	79	73.83
AG16359	LRRK2	2	13	79	66.91
AG16359	MAPK14	1	13	79	78.32
AG16359	MAPK14	2	13	79	76.91
AG16359	MAPK7	1	13	79	83.06
AG16359	MAPK7	2	13	79	81.97
AG16359	MCM7	1	13	79	74.19
AG16359	MCM7	2	13	79	76.95
AG16359	MDM2	1	13	79	65.69
AG16359	MDM2	2	13	79	75.71

AG16359	NFKB1	1	13	79	67.34
AG16359	NFKB1	2	13	79	77.6
AG16359	NOTCH1	1	13	79	81.92
AG16359	NOTCH1	2	13	79	76.95
AG16359	NPM1	1	13	79	76.98
AG16359	NPM1	2	13	79	84.5
AG16359	OGT	1	13	79	76.36
AG16359	OGT	2	13	79	70.21
AG16359	OSKM	1	13	79	70.74
AG16359	OSKM	2	13	79	62.85
AG16359	P4HB	1	13	79	66.13
AG16359	P4HB	2	13	79	60.8
AG16359	PARK7	1	13	79	72.95
AG16359	PARK7	2	13	79	78.13
AG16359	PIK3C2A	1	13	79	78.2
AG16359	PIK3C2A	2	13	79	74.94
AG16359	PIK3CA	2	13	79	54.84
AG16359	PIK3CA	1	13	79	75.58
AG16359	PKM	1	13	79	79.55
AG16359	PKM	2	13	79	74.18
AG16359	PLK1	1	13	79	79.96
AG16359	PLK1	2	13	79	72.99
AG16359	PPIA	1	13	79	80.56
AG16359	PPIA	2	13	79	75.32
AG16359	PPP1CA	1	13	79	72.07
AG16359	PPP1CA	2	13	79	77.87
AG16359	PPP1CB	1	13	79	74.53
AG16359	PPP1CB	2	13	79	67.21
AG16359	PPP1CC	1	13	79	74.71
AG16359	PPP1CC	2	13	79	76.6
AG16359	PPP2CA	1	13	79	81.91
AG16359	PPP2CA	2	13	79	71.45
AG16359	PPP3CA	1	13	79	77.71
AG16359	PPP3CA	2	13	79	72.9
AG16359	PPP4C	1	13	79	74.91
AG16359	PPP4C	2	13	79	79.22
AG16359	PSMC6	1	13	79	77.26
AG16359	PSMC6	2	13	79	72.26
AG16359	PSMD4	1	13	79	81.51

AG16359	PSMD4	2	13	79	84.09
AG16359	PTEN	1	13	79	78.22
AG16359	PTEN	2	13	79	76.25
AG16359	RAD23B	1	13	79	69.73
AG16359	RAD23B	2	13	79	72.11
AG16359	RAD51	1	13	79	76.79
AG16359	RAD51	2	13	79	74.77
AG16359	RNASEL	1	13	79	77.37
AG16359	RNASEL	2	13	79	77.84
AG16359	RPLP0	1	13	79	72.7
AG16359	RPLP0	2	13	79	81.43
AG16359	RPS27A	1	13	79	77.6
AG16359	RPS27A	2	13	79	78.6
AG16359	RRM1	1	13	79	79.62
AG16359	RRM1	2	13	79	75.56
AG16359	RUVBL2	1	13	79	74.87
AG16359	RUVBL2	2	13	79	76.36
AG16359	SIRT1	1	13	79	71.7
AG16359	SIRT1	2	13	79	70.96
AG16359	SLC2A10	1	13	79	72.49
AG16359	SLC2A10	2	13	79	82.28
AG16359	SLC2A13	1	13	79	75.18
AG16359	SLC2A13	2	13	79	80.31
AG16359	SOD1	1	13	79	81.12
AG16359	SOD1	2	13	79	78.21
AG16359	SOD2	1	13	79	76.56
AG16359	SOD2	2	13	79	74.34
AG16359	SRSF1	1	13	79	51
AG16359	SRSF1	2	13	79	51.06
AG16359	SUMO1	1	13	79	81.35
AG16359	SUMO1	2	13	79	76.68
AG16359	TP53 (DN)	1	13	79	79.57
AG16359	TP53 (DN)	2	13	79	79.43
AG16359	TXN	1	13	79	82.81
AG16359	TXN	2	13	79	84.44
AG16359	UBA52	1	13	79	80.68
AG16359	UBA52	2	13	79	72.71
AG16359	UBB	1	13	79	70.71
AG16359	UBB	2	13	79	72

AG16359	UBE2A	1	13	79	76.5
AG16359	UBE2A	2	13	79	69.85
AG16359	UBE2I	1	13	79	76.65
AG16359	UBE2I	2	13	79	82.14
AG16359	UBE2N	1	13	79	78.13
AG16359	UBE2N	2	13	79	83.47
AG16359	UBE2S	1	13	79	81.37
AG16359	UBE2S	2	13	79	78.38
AG16359	UBQLN1	1	13	79	73.62
AG16359	UBQLN1	2	13	79	78.77
AG16359	VCP	1	13	79	78.45
AG16359	VCP	2	13	79	79.41
AG16359	WDTC1	1	13	79	1.78
AG16359	WDTC1	2	13	79	75.8
AG16359	XPO1	1	13	79	78.6
AG16359	XPO1	2	13	79	80.13
AG16359	YWHAB	1	13	79	74.55
AG16359	YWHAB	2	13	79	80.58
AG16359	YWHAQ	1	13	79	74.95
AG16359	YWHAQ	2	13	79	79.1

Table A3.5: Public data accession IDs.

List of datasets used to assess the response of the transcriptomic predictor to perturbations and independent wild type samples. For each study, the RNA-seq accessions are provided. In case DNA methylation was also profiled in a study, the respective accessions are provided as well.

Study	RNA-Seq accession	DNA methylation accession
A Multi-Omics and Bioenergetics Longitudinal Aging Dataset in Primary Human Fibroblasts with Mitochondrial Perturbations	PRJNA745334	GSE179847
Age, gender and UV-exposition related effects on gene expression in in vivo aged short term cultivated human dermal fibroblasts	PRJEB13731	N/A
High-efficiency RNA-based reprogramming of human primary fibroblasts	PRJNA381150	N/A
FoxM1 repression during human aging leads to mitotic decline and aneuploidy-driven full senescence	PRJEB27047	N/A
Reprogramming and differentiation-dependent transcriptional alteration of DNA damage response and apoptosis genes in human induced pluripotent stem cells	PRJNA450083	N/A
Gene expression variability across cells and species shapes innate immunity	PRJEB22019	N/A
p38MAPK plays a crucial role in stromal-mediated tumorigenesis	PRJNA243087	N/A
Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins	PRJNA275305	N/A
Rapamycin reduces fibroblast proliferation without causing quiescence and induces STAT5A/B-mediated cytokine production	PRJNA273277	N/A
Conserved Senescence Associated Genes and Pathways in Primary Human Fibroblasts Detected by RNA-Seq	PRJNA268292	N/A
	PRJNA311203	N/A
Hormetic effect of rotenone in primary human fibroblasts	PRJNA271291	N/A
Strong Components of Epigenetic Memory in Cultured Human Fibroblasts Related to Site of Origin and Donor Age	PRJNA286856	GSE77136
Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential	PRJNA243926	GSE54848

REFERENCES

1. Xu, J., Murphy, S. L., Kochanek, K. D. & Arias, E. Deaths: Final data for 2019. *Natl. Vital Stat. Rep.* **70**, (2021).
2. COVID-19 Mortality Overview. Available at: https://www.cdc.gov/nchs/nvss/vsrr/covid_weekly/index.htm. (Accessed: 6th September 2022)
3. Netuveli, G., Wiggins, R. D., Hildon, Z., Montgomery, S. M. & Blane, D. Quality of life at older ages: Evidence from the English longitudinal study of aging (wave 1). *J. Epidemiol. Community Health* (2006). doi:10.1136/jech.2005.040071
4. Jared, O., Matthew, M., Emma, W., Gary, C. & Krutika, A. How do health expenditures vary across the population? - Peterson-KFF Health System Tracker. *Peterson-KFF Health System Tracker* (2021).
5. O'Connell, J. M. & Manson, S. M. Understanding the economic costs of diabetes and prediabetes and what we may learn about reducing the health and economic burden of these conditions. *Diabetes Care* (2019). doi:10.2337/dci19-0017
6. Heart Disease and Stroke Statistics-2021 Update A Report from the American Heart Association. *Circulation* (2021). doi:10.1161/CIR.0000000000000950
7. Gooch, C. L., Pracht, E. & Borenstein, A. R. The burden of neurological disease in the United States: A summary report and call to action. *Annals of Neurology* (2017). doi:10.1002/ana.24897
8. Mariotto, A. B., Enewold, L., Zhao, J., Zeruto, C. A. & Robin Yabroff, K. Medical care costs associated with cancer survivorship in the United States. *Cancer Epidemiology Biomarkers and Prevention* (2020). doi:10.1158/1055-9965.EPI-19-1534
9. Actuary, O. of the. National Health Spending in 2020 Increases due to Impact of COVID-

19 Pandemic. (2021).

10. Scott, A. J., Ellison, M. & Sinclair, D. A. The economic value of targeting aging. *Nat. Aging* (2021). doi:10.1038/s43587-021-00080-0

11. United Nations. Department of Economic and Social Affairs, Population Division: World Population Prospects 2019, Volume II: Demographic Profiles. *World Popul. Prospect.* (2019).

12. Flatt, T. A new definition of aging? *Frontiers in Genetics* (2012). doi:10.3389/fgene.2012.00148

13. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The Hallmarks of Aging. *Cell* **153**, 1194–1217 (2013).

14. Kennedy, B. K. *et al.* Aging: a common driver of chronic diseases and a target for novel interventions. *Cell* (2014).

15. Zhang, R., Chen, H. Z. & Liu, D. P. The Four Layers of Aging. *Cell Systems* (2015). doi:10.1016/j.cels.2015.09.002

16. Booth, L. N. & Brunet, A. The Aging Epigenome. *Mol. Cell* **62**, 728–744 (2016).

17. Sen, P., Shah, P. P., Nativio, R. & Berger, S. L. Epigenetic mechanisms regulating longevity and aging. *Cell* (2016).

18. O'Sullivan, R. J., Kubicek, S., Schreiber, S. L. & Karlseder, J. Reduced histone biosynthesis and chromatin changes arising from a damage signal at telomeres. *Nat. Struct. Mol. Biol.* (2010). doi:10.1038/nsmb.1897

19. Pal, S. & Tyler, J. K. Epigenetics and aging. *Science Advances* (2016). doi:10.1126/sciadv.1600584

20. Benayoun, B. A., Pollina, E. A. & Brunet, A. Epigenetic regulation of ageing: linking environmental inputs to genomic stability. *Nat. Rev. Mol. Cell Biol.* **16**, 593–610 (2015).

21. Benayoun, B. A. *et al.* H3K4me3 Breadth Is Linked to Cell Identity and Transcriptional Consistency. *Cell* **158**, 673–688 (2014).

22. Benayoun, B. A. *et al.* Remodeling of epigenome and transcriptome landscapes with aging in mice reveals widespread induction of inflammatory responses. *Genome Res.* (2019). doi:10.1101/gr.240093.118

23. Perez-Gomez, A., Buxbaum, J. N. & Petrascheck, M. The aging transcriptome: read between the lines. *Current Opinion in Neurobiology* (2020). doi:10.1016/j.conb.2020.05.001

24. Bahar, R. *et al.* Increased cell-to-cell variation in gene expression in ageing mouse heart. *Nature* (2006). doi:10.1038/nature04844

25. Martinez-Jimenez, C. P. *et al.* Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science (80-.).* (2017). doi:10.1126/science.aah4115

26. Cheung, P. *et al.* Single-Cell Chromatin Modification Profiling Reveals Increased Epigenetic Variations with Aging. *Cell* (2018). doi:10.1016/j.cell.2018.03.079

27. Angelidis, I. *et al.* An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nat. Commun.* (2019). doi:10.1038/s41467-019-08831-9

28. Gupta, K., Yadav, P., Maryam, S., Ahuja, G. & Sengupta, D. Quantification of Age-Related Decline in Transcriptional Homeostasis. *J. Mol. Biol.* (2021). doi:10.1016/j.jmb.2021.167179

29. Ferrucci, L., Wilson, D. M., Donegà, S. & Gorospe, M. The energy–splicing resilience axis hypothesis of aging. *Nat. Aging* (2022). doi:10.1038/s43587-022-00189-w

30. Bhadra, M., Howell, P., Dutta, S., Heintz, C. & Mair, W. B. Alternative splicing in aging and longevity. *Human Genetics* (2020). doi:10.1007/s00439-019-02094-6

31. Li, H., Wang, Z., Ma, T., Wei, G. & Ni, T. Alternative splicing in aging and age-related diseases. *Translational Medicine of Aging* (2017). doi:10.1016/j.tma.2017.09.005

32. Angarola, B. L. & Anczuków, O. Splicing alterations in healthy aging and disease. *Wiley Interdisciplinary Reviews: RNA* (2021). doi:10.1002/wrna.1643

33. Somel, M. *et al.* MicroRNA, mRNA, and protein expression link development and aging in human and macaque brain. *Genome Res.* (2010). doi:10.1101/gr.106849.110

34. Dönertaş, H. M. *et al.* Gene expression reversal toward pre-adult levels in the aging human brain and age-related loss of cellular identity. *Sci. Rep.* **7**, 1–13 (2017).

35. Soreq, E. *et al.* Major Shifts in Glial Regional Identity Are a Transcriptional Hallmark of Human Brain Aging. *Cell Rep.* **18**, 557–570 (2017).

36. Enge, M. *et al.* Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. *Cell* (2017). doi:10.1016/j.cell.2017.09.004

37. Solé-Boldo, L. *et al.* Single-cell transcriptomes of the aging human skin reveal loss of fibroblast priming. *bioRxiv* (2019). doi:10.1101/633131

38. Salzer, M. C. *et al.* Identity Noise and Adipogenic Traits Characterize Dermal Fibroblast Aging. *Cell* 1–16 (2018). doi:10.1016/j.cell.2018.10.012

39. Anisimova, A. S. *et al.* Multifaceted deregulation of gene expression and protein synthesis with age. *Proc. Natl. Acad. Sci. U. S. A.* (2020). doi:10.1073/pnas.2001788117

40. Izgi, H. *et al.* Inter-tissue convergence of gene expression during ageing suggests age-related loss of tissue and cellular identity. *Elife* (2022). doi:10.7554/eLife.68048

41. Bocklandt, S. *et al.* Epigenetic predictor of age. *PLoS One* (2011). doi:10.1371/journal.pone.0014821

42. Hannum, G. *et al.* Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Mol. Cell* (2013). doi:10.1016/j.molcel.2012.10.016

43. Koch, C. M. & Wagner, W. Epigenetic-aging-signature to determine age in different tissues. *Aging (Albany. NY).* (2011). doi:10.18632/aging.100395

44. Weidner, C. I. *et al.* Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biol.* (2014). doi:10.1186/gb-2014-15-2-r24

45. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).

46. Marioni, R. E. *et al.* DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biol.* (2015). doi:10.1186/s13059-015-0584-6

47. Christiansen, L. *et al.* DNA methylation age is associated with mortality in a longitudinal Danish twin study. *Aging Cell* (2016). doi:10.1111/acel.12421

48. Perna, L. *et al.* Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. *Clin. Epigenetics* (2016). doi:10.1186/s13148-016-0228-z

49. Chen, B. H. *et al.* DNA methylation-based measures of biological age: Meta-analysis predicting time to death. *Aging (Albany. NY).* (2016). doi:10.18632/aging.101020

50. Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany. NY).* (2019). doi:10.18632/aging.101684

51. Levine, M. E. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany. NY).* (2018). doi:10.18632/aging.101414

52. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **19**, 371–384 (2018).

53. Peters, M. J. *et al.* The transcriptional landscape of age in human peripheral blood. *Nat. Commun.* **6**, (2015).

54. Mamoshina, P. *et al.* Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification. *Front. Genet.* (2018). doi:10.3389/fgene.2018.00242

55. Remondini, D. *et al.* Identification of a T cell gene expression clock obtained by exploiting a MZ twin design. *Sci. Rep.* (2017). doi:10.1038/s41598-017-05694-2

56. Fleischer, J. G. *et al.* Predicting age from the transcriptome of human dermal fibroblasts.

Genome Biol. **19**, 221 (2018).

57. Choukrallah, M. A., Hoeng, J., Peitsch, M. C. & Martin, F. Lung transcriptomic clock predicts premature aging in cigarette smoke-exposed mice. *BMC Genomics* (2020). doi:10.1186/s12864-020-6712-z

58. Meyer, D. H. & Schumacher, B. BiT age: A transcriptome-based aging clock near the theoretical limit of accuracy. *Aging Cell* **20**, (2021).

59. Kenyon, C., Chang, J., Gensch, E., Rudner, A. & Tabtiang, R. A C. elegans mutant that lives twice as long as wild type. *Nature* (1993). doi:10.1038/366461a0

60. Kenyon, C. J. The genetics of ageing. *Nature* (2010). doi:10.1038/nature08980

61. Gems, D. & Partridge, L. Genetics of longevity in model organisms: Debates and paradigm shifts. *Annual Review of Physiology* (2013). doi:10.1146/annurev-physiol-030212-183712

62. Rando, T. A. & Chang, H. Y. Aging, rejuvenation, and epigenetic reprogramming: Resetting the aging clock. *Cell* **148**, 46–57 (2012).

63. Gurdon, J. B. Adult frogs derived from the nuclei of single somatic cells. *Dev. Biol.* (1962). doi:10.1016/0012-1606(62)90043-X

64. Campbell, K. H. S., McWhir, J., Ritchie, W. A. & Wilmut, I. Sheep cloned by nuclear transfer from a cultured cell line. *Nature* (1996). doi:10.1038/380064a0

65. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* (2006). doi:10.1016/j.cell.2006.07.024

66. Lapasset, L. *et al.* Rejuvenating senescent and centenarian human cells by reprogramming through the pluripotent state. *Genes Dev.* **25**, 2248–2253 (2011).

67. Petkovich, D. A. *et al.* Using DNA Methylation Profiling to Evaluate Biological Age and Longevity Interventions. *Cell Metab.* (2017). doi:10.1016/j.cmet.2017.03.016

68. Kurita, M. *et al.* In vivo reprogramming of wound-resident cells generates. *Nature* (2018).

69. Ocampo, A. *et al.* In Vivo Amelioration of Age-Associated Hallmarks by Partial Reprogramming. *Cell* **167**, 1719-1733.e12 (2016).

70. Rodríguez-Matellán, A., Alcazar, N., Hernández, F., Serrano, M. & Ávila, J. In Vivo Reprogramming Ameliorates Aging Features in Dentate Gyrus Cells and Improves Memory in Mice. *Stem Cell Reports* (2020). doi:10.1016/j.stemcr.2020.09.010

71. Chondronasiou, D. *et al.* Multi-omic rejuvenation of naturally aged tissues by a single cycle of transient reprogramming. *Aging Cell* **21**, e13578 (2022).

72. Gill, D. *et al.* Multi-omic rejuvenation of human cells by maturation phase transient reprogramming. *Elife* **11**, (2022).

73. Browder, K. C. *et al.* In vivo partial reprogramming alters age-associated molecular changes during physiological aging in mice. *Nat. Aging* (2022). doi:10.1038/s43587-022-00183-2

74. Lu, Y. *et al.* Reprogramming to recover youthful epigenetic information and restore vision. *Nature* (2020). doi:10.1038/s41586-020-2975-4

75. Gladyshev, V. N. The Ground Zero of Organismal Life and Aging. *Trends in Molecular Medicine* (2021). doi:10.1016/j.molmed.2020.08.012

76. Kerepesi, C., Zhang, B., Lee, S.-G., Trapp, A. & Gladyshev, V. N. Epigenetic clocks reveal a rejuvenation event during embryogenesis followed by aging. *Sci. Adv.* **7**, eabg6082 (2021).

77. Senís, E. *et al.* AAVvector-mediated in vivo reprogramming into pluripotency. *Nat. Commun.* (2018). doi:10.1038/s41467-018-05059-x

78. Ohnishi, K. *et al.* Premature termination of reprogramming in vivo leads to cancer development through altered epigenetic regulation. *Cell* (2014). doi:10.1016/j.cell.2014.01.005

79. Abad, M. *et al.* Reprogramming in vivo produces teratomas and iPS cells with totipotency

features. *Nature* (2013). doi:10.1038/nature12586

80. Olova, N., Simpson, D. J., Marioni, R. E. & Chandra, T. Partial reprogramming induces a steady decline in epigenetic age before loss of somatic identity. *Aging Cell* **18**, (2019).

81. Manukyan, M. & Singh, P. B. Epigenetic rejuvenation. *Genes to Cells* **17**, 337–343 (2012).

82. Roux, A. E. *et al.* Diverse partial reprogramming strategies restore youthful gene expression and transiently suppress cell identity. *Cell Syst.* (2022). doi:10.1016/j.cels.2022.05.002

83. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

84. Tacutu, R. *et al.* Human Ageing Genomic Resources: new and updated databases. *Nucleic Acids Res.* **46**, D1083–D1090 (2018).

85. Rackham, O. J. L. *et al.* A predictive computational framework for direct reprogramming between human cell types. *Nat. Genet.* (2016). doi:10.1038/ng.3487

86. Jung, S., Appleton, E., Ali, M., Church, G. M. & del Sol, A. A computer-guided design tool to increase the efficiency of cellular conversions. *Nat. Commun.* (2021). doi:10.1038/s41467-021-21801-4

87. Fabris, F., Palmer, D., Salama, K. M., De Magalhães, J. P. & Freitas, A. A. Using deep learning to associate human genes with age-related diseases. *Bioinformatics* (2020). doi:10.1093/bioinformatics/btz887

88. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nature Methods* (2014). doi:10.1038/nmeth.3047

89. Chavez, A. *et al.* Highly efficient Cas9-mediated transcriptional programming. *Nat. Methods* (2015). doi:10.1038/nmeth.3312

90. Yeo, N. C. *et al.* An enhanced CRISPR repressor for targeted mammalian gene regulation. *Nat. Methods* (2018). doi:10.1038/s41592-018-0048-5

91. Yang, X. *et al.* A public genome-scale lentiviral expression library of human ORFs. *Nat. Methods* (2011). doi:10.1038/nmeth.1638

92. Janssens, G. E. *et al.* Transcriptomics-Based Screening Identifies Pharmacological Inhibition of Hsp90 as a Means to Defer Aging. *Cell Rep.* (2019). doi:10.1016/j.celrep.2019.03.044

93. Kramme, C. *et al.* An integrated pipeline for mammalian genetic screening. *Cell Reports Methods* 100082 (2021). doi:10.1016/j.crmeth.2021.100082

94. Kramme, C. *et al.* MegaGate: A toxin-less gateway molecular cloning tool. *STAR Protoc.* **2**, 100907 (2021).

95. Wei, L. *et al.* Genome-wide CRISPR/Cas9 library screening identified PHGDH as a critical driver for Sorafenib resistance in HCC. *Nat. Commun.* (2019). doi:10.1038/s41467-019-12606-7

96. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science (80-.).* (2020). doi:10.1126/science.aba7721

97. Domcke, S. *et al.* A human cell atlas of fetal chromatin accessibility. *Science (80-.).* (2020). doi:10.1126/science.aba7612

98. Ng, A. H. M. *et al.* A comprehensive library of human transcription factors for cell fate engineering. *Nat. Biotechnol.* (2021). doi:10.1038/s41587-020-0742-6

99. Schlitt, T. & Brazma, A. Current approaches to gene regulatory network modelling. *BMC Bioinformatics* (2007). doi:10.1186/1471-2105-8-S6-S9

100. Mardis, E. R. A decade's perspective on DNA sequencing technology. *Nature* (2011). doi:10.1038/nature09796

101. Rual, J.-F. *et al.* Human ORFeome Version 1.1: A Platform for Reverse Proteomics. *Genome Res.* **14**, 2128–2135 (2004).

102. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion

for RNA-seq data with DESeq2. *Genome Biol.* (2014). doi:10.1186/s13059-014-0550-8

103. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* (2009). doi:10.1093/bioinformatics/btp616

104. Kamimoto, K., Hoffmann, C. M. & Morris, S. A. CellOracle: Dissecting cell identity via network inference and in silico gene perturbation. *bioRxiv* (2020). doi:https://doi.org/10.1101/2020.02.17.947416

105. Jiang, P. *et al.* Network analysis of gene essentiality in functional genomics experiments. *Genome Biol.* (2015). doi:10.1186/s13059-015-0808-9

106. Szklarczyk, D. *et al.* STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* (2015). doi:10.1093/nar/gku1003

107. Tacutu, R. *et al.* Human Ageing Genomic Resources: New and updated databases. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gkx1042

108. Moerman, T. *et al.* GRNBoost2 and Arboreto: Efficient and scalable inference of gene regulatory networks. *Bioinformatics* (2019). doi:10.1093/bioinformatics/bty916

109. Page, L., Brin, S., Motwani, R. & Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. *World Wide Web Internet Web Inf. Syst.* (1998). doi:10.1.1.31.1768

110. Burke, E. E. *et al.* Dissecting transcriptomic signatures of neuronal differentiation and maturation using iPSCs. *Nat. Commun.* (2020). doi:10.1038/s41467-019-14266-z

111. Choi, I. Y. *et al.* Transcriptional landscape of myogenesis from human pluripotent stem cells reveals a key role of TWIST1 in maintenance of skeletal muscle progenitors. *Elife* (2020). doi:10.7554/eLife.46981

112. Liu, L. P. *et al.* Therapeutic Potential of Patient iPSC-Derived iMelanocytes in Autologous Transplantation. *Cell Rep.* (2019). doi:10.1016/j.celrep.2019.03.046

113. Chavez, A. *et al.* Comparison of Cas9 activators in multiple species. *Nat. Methods* (2016). doi:10.1038/nmeth.3871

114. Arnoldo, A. *et al.* A genome scale overexpression screen to reveal drug activity in human cells. *Genome Med.* (2014). doi:10.1186/gm549

115. Tak, Y. E. *et al.* Inducible and multiplex gene regulation using CRISPR-Cpf1-based transcription factors. *Nat. Methods* (2017). doi:10.1038/nmeth.4483

116. Zhang, X. *et al.* Multiplex gene regulation by CRISPR-ddCpf1. *Cell Discov.* (2017). doi:10.1038/celldisc.2017.18

117. Shimojima, M. *et al.* Efficient functional screening of a cellular cDNA library to identify severe fever with thrombocytopenia syndrome virus entry factors. *Sci. Rep.* (2020). doi:10.1038/s41598-020-62876-1

118. Parekh, U. *et al.* Mapping Cellular Reprogramming via Pooled Overexpression Screens with Paired Fitness and Single-Cell RNA-Sequencing Readout. *Cell Syst.* (2018). doi:10.1016/j.cels.2018.10.008

119. Konermann, S. *et al.* Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* (2015). doi:10.1038/nature14136

120. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–83 (2013).

121. Horlbeck, M. A. *et al.* Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *Elife* (2016). doi:10.7554/eLife.19760

122. Bikard, D. *et al.* Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Res.* (2013). doi:10.1093/nar/gkt520

123. Gilbert Luke A *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* (2013).

124. Zalatan, J. G. *et al.* Engineering complex synthetic transcriptional programs with CRISPR RNA scaffolds. *Cell* (2015). doi:10.1016/j.cell.2014.11.052

125. Tanenbaum, M. E., Gilbert, L. A., Qi, L. S., Weissman, J. S. & Vale, R. D. A protein-tagging system for signal amplification in gene expression and fluorescence imaging. *Cell* (2014). doi:10.1016/j.cell.2014.09.039

126. Randolph, L. N., Bao, X., Zhou, C. & Lian, X. An all-in-one, Tet-On 3G inducible PiggyBac system for human pluripotent stem cells and derivatives. *Sci. Rep.* (2017). doi:10.1038/s41598-017-01684-6

127. Kim, S. Il *et al.* Inducible transgene expression in human iPS cells using versatile all-in-one piggybac transposons. in *Methods in Molecular Biology* (2016). doi:10.1007/7651_2015_251

128. Lian, J., Hamedirad, M., Hu, S. & Zhao, H. Combinatorial metabolic engineering using an orthogonal tri-functional CRISPR system. *Nat. Commun.* (2017). doi:10.1038/s41467-017-01695-x

129. Hazelbaker, D. Z. *et al.* A multiplexed gRNA piggyBac transposon system facilitates efficient induction of CRISPRi and CRISPRa in human pluripotent stem cells. *Sci. Rep.* (2020). doi:10.1038/s41598-020-57500-1

130. Vad-Nielsen, J., Lin, L., Bolund, L., Nielsen, A. L. & Luo, Y. Golden Gate Assembly of CRISPR gRNA expression array for simultaneously targeting multiple genes. *Cell. Mol. Life Sci.* (2016). doi:10.1007/s00018-016-2271-5

131. Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PLoS One* (2008). doi:10.1371/journal.pone.0003647

132. Kirchmaier, S., Lust, K. & Wittbrodt, J. Golden GATEway cloning--a combinatorial approach to generate fusion and recombination constructs. *PLoS One* (2013). doi:10.1371/journal.pone.0076117

133. Magnani, E., Bartling, L. & Hake, S. From Gateway to MultiSite Gateway in one recombination event. *BMC Mol. Biol.* (2006). doi:10.1186/1471-2199-7-46

134. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* (2009). doi:10.1038/nmeth.1318

135. Zufferey, R., Nagy, D., Mandel, R. J., Naldini, L. & Trono, D. Multiply attenuated lentiviral vector achieves efficient gene delivery in vivo. *Nat. Biotechnol.* (1997). doi:10.1038/nbt0997-871

136. Yusa, K., Zhou, L., Li, M. A., Bradley, A. & Craig, N. L. A hyperactive piggyBac transposase for mammalian applications. *Proc. Natl. Acad. Sci. U. S. A.* (2011). doi:10.1073/pnas.1008322108

137. Vargas, J. E. *et al.* Retroviral vectors and transposons for stable gene therapy: Advances, current challenges and perspectives. *Journal of Translational Medicine* (2016). doi:10.1186/s12967-016-1047-x

138. Martin, D. P., Miya, J., Reeser, J. W. & Roychowdhury, S. Targeted RNA sequencing assay to characterize gene expression and genomic alterations. *J. Vis. Exp.* (2016). doi:10.3791/54090

139. Smith, A. M. *et al.* Quantitative phenotyping via deep barcode sequencing. *Genome Res.* (2009). doi:10.1101/gr.093955.109

140. Horvath, S. *et al.* Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. *Aging (Albany. NY).* (2018). doi:10.18632/aging.101508

141. Tarkhov, A. E. *et al.* A universal transcriptomic signature of age reveals the temporal scaling of Caenorhabditis elegans aging trajectories. *Sci. Rep.* (2019). doi:10.1038/s41598-019-43075-z

142. Hansen, J., Meretzky, D., Woldesenbet, S., Stolovitzky, G. & Iyengar, R. A flexible ontology for inference of emergent whole cell function from relationships between subcellular processes. *Sci. Rep.* **7**, 17689 (2017).

143. Busskamp, V. *et al.* Rapid neurogenesis through transcriptional activation in human stem cells. *Mol. Syst. Biol.* (2014). doi:10.15252/msb.20145508

144. Matsuyama, M., WuWong, D. J., Horvath, S. & Matsuyama, S. Epigenetic clock analysis of human fibroblasts in vitro: Effects of hypoxia, donor age, and expression of hTERT and SV40 largeT. *Aging (Albany. NY).* (2019). doi:10.18632/aging.101955

145. Sturm, G. *et al.* A Multi-Omics and Bioenergetics Longitudinal Aging Dataset in Primary Human Fibroblasts with Mitochondrial Perturbations. *BioRxiv* 1–35 (2021).

146. Wang, W. *et al.* A genome-wide CRISPR-based screen identifies KAT7 as a driver of cellular senescence. *Sci. Transl. Med.* **13**, (2021).

147. Rogina, B. & Helfand, S. L. Sir2 mediates longevity in the fly through a pathway related to calorie restriction. *Proc. Natl. Acad. Sci. U. S. A.* (2004). doi:10.1073/pnas.0404184101

148. Tissenbaum, H. A. & Guarente, L. Increased dosage of a sir-2 gene extends lifespan in Caenorhabditis elegans. *Nature* (2001). doi:10.1038/35065638

149. Burnett, C. *et al.* Absence of effects of Sir2 overexpression on lifespan in C. elegans and Drosophila. *Nature* (2011). doi:10.1038/nature10296

150. Huang, J. *et al.* SIRT1 overexpression antagonizes cellular senescence with activated ERK/S6k1 signaling in human diploid fibroblasts. *PLoS One* **3**, e1710 (2008).

151. Rapoport, S. I., Primiani, C. T., Chen, C. T., Ahn, K. & Ryan, V. H. Coordinated Expression of Phosphoinositide Metabolic Genes during Development and Aging of Human Dorsolateral Prefrontal Cortex. *PLoS One* **10**, e0132675 (2015).

152. Li, C. Y. *et al.* Mutation analysis of seven SLC family transporters for early-onset Parkinson's

disease in Chinese population. *Neurobiol. Aging* (2021). doi:10.1016/j.neurobiolaging.2021.02.022

153. Holly, A. C. *et al.* Changes in splicing factor expression are associated with advancing age in man. *Mech. Ageing Dev.* (2013). doi:10.1016/j.mad.2013.05.006

154. Lee, B. P. *et al.* Changes in the expression of splicing factor transcripts and variations in alternative splicing are associated with lifespan in mice and humans. *Aging Cell* (2016). doi:10.1111/acel.12499

155. Wang, Y. T. *et al.* Human serum RNase-L level is inversely associated with metabolic syndrome and age. *Cardiovasc. Diabetol.* (2017). doi:10.1186/s12933-017-0522-x

156. Lai, C. Q. *et al.* WDTC1, the ortholog of drosophila adipose gene, associates with human obesity, modulated by MUFA intake. *Obesity* (2009). doi:10.1038/oby.2008.561

157. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* (2018). doi:10.21105/joss.00861

158. Ahadi, S. *et al.* Personal aging markers and ageotypes revealed by deep longitudinal profiling. *Nature Medicine* (2020). doi:10.1038/s41591-019-0719-5

159. Waddington, C. H. The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser. *Strateg. genes A Discuss. some …* (1957).

160. Doeser, M. C., Schöler, H. R. & Wu, G. Reduction of Fibrosis and Scar Formation by Partial Reprogramming In Vivo. *Stem Cells* **36**, 1216–1225 (2018).

161. Zhao, W. *et al.* POSTAR3: An updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res.* (2022). doi:10.1093/nar/gkab702

162. Yu, T. *et al.* SRSF1 governs progenitor-specific alternative splicing to maintain adult epithelial tissue homeostasis and renewal. *Dev. Cell* (2022). doi:10.1016/j.devcel.2022.01.011

163. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci.* **112**, 7285–7290 (2015).

164. Cahan, P. *et al.* CellNet: Network biology applied to stem cell engineering. *Cell* (2014). doi:10.1016/j.cell.2014.07.020

165. Schaum, N. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* (2018). doi:10.1038/s41586-018-0590-4

166. Tabula Sapiens Consortium* *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).

167. GTEx Project. GTEx portal. *GTEx Analysis Release V6p (dbGaP Accession phs000424.v6.p1)* (2017).

168. Szklarczyk, D. *et al.* STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gky1131

169. Grath, A. & Dai, G. Direct cell reprogramming for tissue engineering and regenerative medicine. *Journal of Biological Engineering* (2019). doi:10.1186/s13036-019-0144-9

170. Kallunki, T., Barisic, M., Jäättelä, M. & Liu, B. How to choose the right inducible gene expression system for Mammalian studies? *Cells* (2019). doi:10.3390/cells8080796

171. Mauger, D. M. *et al.* mRNA structure regulates protein expression through changes in functional half-life. *Proc. Natl. Acad. Sci. U. S. A.* (2019). doi:10.1073/pnas.1908052116

172. Wang, K. *et al.* Comprehensive map of age-associated splicing changes across human tissues and their contributions to age-associated diseases. *Sci. Rep.* (2018). doi:10.1038/s41598-018-29086-2

173. Sneha Dutta, Caroline Heintz, Maria C. Perez-Matos, Ayse Sena Mutlu, Mary E. Piper, Meeta Mistry, Arpit Sharma, Hannah J. Smith, Porsha Howell, Rohan Sehgal, Anne

Lanjuin, View ORCID ProfileMeng C. Wang, V. O. P. B. M. Efficacy of longevity interventions in C. elegans is determined by early life activity of RNA splicing factors. *bioRxiv* (2021).

174. Marco Mariotti, Csaba Kerepesi, Winona Oliveros, Marta Mele, V. N. G. Deterioration of the human transcriptome with age due to increasing intron retention and spurious splicing. *bioRxiv* (2022).

175. Blanco, F. J. & Bernabéu, C. The splicing factor SRSF1 as a marker for endothelial senescence. *Frontiers in Physiology* (2012). doi:10.3389/fphys.2012.00054

176. Fregoso, O. I., Das, S., Akerman, M. & Krainer, A. R. Splicing-Factor Oncoprotein SRSF1 Stabilizes p53 via RPL5 and Induces Cellular Senescence. *Mol. Cell* (2013). doi:10.1016/j.molcel.2013.02.001

177. Das, S., Anczuków, O., Akerman, M. & Krainer, A. R. Oncogenic Splicing Factor SRSF1 Is a Critical Transcriptional Target of MYC. *Cell Rep.* (2012). doi:10.1016/j.celrep.2011.12.001

178. Gonçalves, V. & Jordan, P. Posttranscriptional regulation of splicing factor SRSF1 and its role in cancer cell biology. *BioMed Research International* (2015). doi:10.1155/2015/287048

179. Replogle, J. M. *et al.* Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell* (2022). doi:10.1016/j.cell.2022.05.013

180. Hanczar, B., Zehraoui, F., Issa, T. & Arles, M. Biological interpretation of deep neural network for phenotype prediction based on gene expression. *BMC Bioinformatics* (2020). doi:10.1186/s12859-020-03836-4

181. Nuñez, J. K. *et al.* Genome-wide programmable transcriptional memory by CRISPR-based epigenome editing. *Cell* (2021). doi:10.1016/j.cell.2021.03.025

182. Sarkar, T. J. *et al.* Transient non-integrative expression of nuclear reprogramming factors

promotes multifaceted amelioration of aging in human cells. *Nat. Commun.* (2020). doi:10.1038/s41467-020-15174-3

183. Hysolli, E. *et al.* Regulation of the DNA Methylation Landscape in Human Somatic Cell Reprogramming by the miR-29 Family. *Stem cell reports* **7**, 43–54 (2016).

184. Church, G. M. & Kieffer-Higgins, S. Multiplex DNA sequencing. *Science (80-.).* (1988). doi:10.1126/science.3353714

185. Wetterstrand, K. A. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). (2021). Available at: www.genome.gov/sequencingcostsdata. (Accessed: 13th June 2022)

186. Levine, M., Higgins-Chen, A. T., Thrush, K., Minteer, C. J. & Niimi, P. Clock Work: Deconstructing the Epigenetic Clock Signals in Aging, Disease, and Reprogramming. *bioRxiv* (2022).

187. Almanzar, N. *et al.* A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* (2020). doi:10.1038/s41586-020-2496-1

188. Zhang, M. J., Pisco, A. O., Darmanis, S. & Zou, J. Mouse aging cell atlas analysis reveals global and cell type-specific aging signatures. *Elife* (2021). doi:10.7554/ELIFE.62293

189. Matthew T. Buckley, Eric Sun, Benson M. George, Ling Liu, Nicholas Schaum, Lucy Xu, Jaime M. Reyes, Margaret A. Goodell, Irving L. Weissman, View ORCID ProfileTony Wyss-Coray, Thomas A. Rando, A. B. Cell type-specific aging clocks to quantify aging and rejuvenation in regenerative regions of the braintle. *bioRxiv* (2022).

190. Angermueller, C. *et al.* Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* (2016). doi:10.1038/nmeth.3728

191. Argelaguet, R. *et al.* Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* (2019). doi:10.1038/s41586-019-1825-8

192. Hernando-Herraez, I. *et al.* Ageing affects DNA methylation drift and transcriptional cell-to-cell variability in mouse muscle stem cells. *Nat. Commun.* (2019). doi:10.1038/s41467-019-12293-4

193. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* (2014). doi:10.1038/nmeth.3035

194. Trapp, A., Kerepesi, C. & Gladyshev, V. N. Profiling epigenetic age in single cells. *Nat. Aging* (2021). doi:10.1038/s43587-021-00134-3

195. Slavov, N. Unpicking the proteome in single cells. *Science* (2020). doi:10.1126/science.aaz6695

196. Shrestha, B. Single-Cell Metabolomics by Mass Spectrometry. *Methods Mol. Biol.* **2064**, 1–8 (2020).

197. Alon, S. *et al.* Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. *Science (80-.).* (2021). doi:10.1126/science.aax2656

198. Borch Jensen, M. & Marblestone, A. In vivo Pooled Screening: A Scalable Tool to Study the Complexity of Aging and Age-Related Disease. *Front. Aging* (2021). doi:10.3389/fragi.2021.714926

199. Simpson, D. J., Olova, N. N. & Chandra, T. Cellular reprogramming and epigenetic rejuvenation. *Clinical Epigenetics* (2021). doi:10.1186/s13148-021-01158-7

200. Alberto Parras, Alba Vílchez-Acosta, Gabriela Desdín-Micó, Calida Mrabti, Cheyenne Rechsteiner, Fabrice Battiston, Clémence Branchina, Kevin Pérez, Christine Sempoux, A. O. In vivo reprogramming leads to premature death due to hepatic and intestinal failure. *bioRxiv* (2022).

201. Hall, H. *et al.* Glucotypes reveal new patterns of glucose dysregulation. *PLoS Biol.* (2018). doi:10.1371/journal.pbio.2005143

202. Dunn, J. *et al.* Wearable sensors enable personalized predictions of clinical laboratory measurements. *Nat. Med.* (2021). doi:10.1038/s41591-021-01339-0

203. Qi, Z. *et al.* SRSF1 serves as a critical posttranscriptional regulator at the late stage of thymocyte development. *Sci. Adv.* (2021). doi:10.1126/sciadv.abf0753

204. Liu, Y. *et al.* Splicing Factor SRSF1 Is Essential for Satellite Cell Proliferation and Postnatal Maturation of Neuromuscular Junctions in Mice. *Stem Cell Reports* (2020). doi:10.1016/j.stemcr.2020.08.004

205. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

206. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* (2014). doi:10.1093/bioinformatics/btt656

207. Team, N. D. NetworkX. *NetworkX* (2014).

208. McKinney, J. R. W., @rstudio, @ursa-labs / & Dong;, ; jbrockmendel; Joris Van den Bossche; Tom Augspurger; Phillip Cloud; Simon Hawkins; gfyoung; Sinhrks; Matthew Roeschke; Adam Klein; Terji Petersen; Jeff Tratner; Chang She; William Ayd; Shahar Naveh; patrick; Marc Garcia; Jeremy Schendel; Andy Hayden; Da. Pandas-Dev/pandas: Pandas 1.2.4. *Zenodo* (2021).

209. Dreos, R., Ambrosini, G., Groux, R., Perier, R. C. & Bucher, P. The eukaryotic promoter database in its 30th year: Focus on non-vertebrate organisms. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkw1069

210. Andrews, S. FastQC - A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. *Babraham Bioinforma.* (2010).

211. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* (2021). doi:10.1093/gigascience/giab008

212. Li, H. seqtk Toolkit for processing sequences in FASTA/Q formats. *GitHub* (2012).

213. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* (2016). doi:10.1038/nbt.3519

214. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* (2015). doi:10.12688/f1000research.7563.1

215. Wu, S. & Mamber, U. Agrep - a fast approximate pattern matching tool. *Proceedings of the Winter 1992 USENIX Conference San Francisco USA. Berkeley* (1992).

216. Trincado, J. L. *et al.* SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.* **19**, 40 (2018).

217. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).

218. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).

219. LeDell, E. *et al.* R Interface for the 'H2O' Scalable Machine Learning Platform. *April* (2020).

220. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19**, 716–723 (1974).

221. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).

222. Tarazona, S. *et al.* Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res.* **43**, e140 (2015).

223. Jacob, L., Gagnon-Bartsch, J. A. & Speed, T. P. Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics* **17**, 16–28 (2016).

224. Schindelin, J. *et al.* Fiji: An open-source platform for biological-image analysis. *Nature Methods*

(2012). doi:10.1038/nmeth.2019

225. Preibisch, S., Saalfeld, S. & Tomancak, P. Globally optimal stitching of tiled 3D microscopic image acquisitions. *Bioinformatics* (2009). doi:10.1093/bioinformatics/btp184

226. Engel, M. *et al.* Bowhead: Bayesian modelling of cell velocity during concerted cell migration. *PLoS Comput. Biol.* (2018). doi:10.1371/journal.pcbi.1005900