Aus dem Institut für Bioinformatik

(Direktor Univ.- Prof. Dr. Lars Kaderali)

der Universitätsmedizin der Universität Greifswald

Thema:
Analyse intrinsischer und extrinsischer Aspekte von Hautalterung in hochdimensionalen biologischen Daten mittels bioinformatischer und maschineller Lernmethoden

Intrinsic and extrinsic aspects of skin aging in high-dimensional biological data analyzed using bioinformatic and machine-learning approaches

Kumulative Inauguraldissertation

zur

Erlangung des akademischen

Grades

Doctor of Philosophy
(PhD)

der

Universitätsmedizin

der

Universität Greifswald

2023

vorgelegt von:
Nicholas Jonas Holzscheck
geb. am: 26.06.1991
in: Hamburg

Dekan: Prof. Dr. Karlhans Endlich

1. Gutachter: Prof. Dr. Lars Kaderali

2. Gutachter: Prof. Dr. Frank Lyko

Datum der Verteidigung: 20.10.2023

Ort, Raum: Greifswald, Konferenzraum A/B C_FunGene

Die in dieser Dissertation vorgestellten Arbeiten wurden ausgeführt unter der Betreuung von
Prof. Dr. Lars Kaderali, Universitätsmedizin Greifswald,
Prof. Dr. Michael Jünger, Universitätsmedizin Greifswald,
Prof. Dr. Mario Stanke, Mathematisch-Naturwissenschaftliche Fakultät und
Prof. Dr. Johannes Hertel, Universitätsmedizin Greifswald,
der Universität Greifswald.

'No one is so old as to think that he cannot live one more year.'

— Cicero, 106–43 BC, Roman orator & statesman

Index

List of publications

Paper 1: Holzscheck N., Söhle J., Kristof B., Grönniger E., Gallinat S., Wenck H., Winnefeld M., Falckenhayn C. and Kaderali L., 2020. “**Multi-omics network analysis reveals distinct stages in the human aging progression in epidermal tissue**”, *Aging*.

Paper 2: Holzscheck N., Falckenhayn C., Söhle J., Kristof B., Siegner R., Werner A., Schössow J., Jürgens C., Völzke H., Wenck H., Winnefeld M., Grönniger E. and Kaderali L., 2021. “**Modeling transcriptomic age using knowledge-primed artificial neural networks**”, *npj Aging and Mechanisms of Disease*.

Paper 3: Holzscheck N., Söhle J., Schläger T., Falckenhayn C., Grönniger E., Kolbe L., Wenck H., Terstegen L., Kaderali L., Winnefeld M. and Gorges K., 2020. “**Concomitant DNA methylation and transcriptome signatures define epidermal responses to acute solar UV radiation**”, *Scientific Reports*.

List of abbreviations

bp	base pair
CPD	cyclobutane pyrimidine dimer
DNA	deoxyribonucleic acid
ER	endoplasmic reticulum
HGPS	Hutchinson-Gilford progeria syndrome
MED	minimal erythema dose
MP	molecular phototype
ROS	radical oxygen species
SASP	senescence-associated secretory phenotype
UV	ultra-violet
y	years

Short summary

Age is the single biggest risk factor for most major human diseases. As such, understanding the intricate molecular changes that drive biological aging holds great promise in attempting to slow the onset of systemic diseases and thereby increase the effective health-span in modern societies. This thesis explores several computational approaches to capture and analyze the molecular biological alterations triggered by intrinsic and extrinsic aging using skin as a model tissue to deliver genes and pathways as potential targets for intervention strategies.

Publication 1 demonstrates the utility of multi-omics data integration strategies for aging research, leading to the identification of four latent aging phases in skin tissue through an integrated cluster analysis of gene expression and DNA methylation data. The four phases improved the detection of molecular aging signals and were shown to be associated with sunbathing habits of the test subjects. Deeper analysis revealed extensive non-linear alterations in various biological pathways particularly at the transition into the fourth aging phase, coinciding with menopause, with potentially wide-reaching functional implications. Publication 2 describes the development of a novel type of age clock, that provides a new level of interpretability by embedding biological pathway information in the architecture of an artificial neural network. The clock not only generates meaningful biological age estimates from gene expression data, but further allows simultaneous monitoring of the aging states of various biological processes through the activations of intermediate neurons. Analyses of the inner workings of the clock revealed a wide-spread impact of aging on the global pathway landscape. Simulation experiments using the transcriptomic clock recapitulated known functional aging gene associations and allowed deciphering of the pathways by which accelerated aging conditions such as chronic sun exposure and Hutchinson-Gilford progeria syndrome exert their effects. Publication 3 further explores the molecular alterations caused by the pro-aging effector UV irradiation in the skin. The multi-omics data analysis of repetitively irradiated skin revealed signs of the immediate acquisition of aging- and cancer-related epigenetic signatures and concurrent wide-spread transcriptional changes across various biological processes. Investigations into the varying resilience to irradiation between subjects revealed prognostic biomarker signatures capable of predicting individual UV tolerances, with accuracies far surpassing the traditional Fitzpatrick classification scheme. Further analysis of the transcripts and pathways associated with UV tolerance identified a form of melanin-independent DNA damage protection in individuals with higher innate UV resilience.

Together, the approaches and findings described in this thesis explore several new angles to advance our understanding of aging processes and external drivers of aging such as UV irradiation in the human skin and deliver new insight on target genes and pathways involved.

Kurzzusammenfassung

Alterung ist der größte Risikofaktor für die meisten schweren Erkrankungen des Menschen. Das Verständnis der komplexen molekularen Veränderungen, die biologische Alterung vorantreiben, ist daher ein vielversprechender Ansatz, um das Auftreten systemischer Krankheiten zu verzögern und damit die effektive Gesundheitsspanne der Bevölkerung in modernen Gesellschaften zu erhöhen. In dieser Arbeit werden verschiedene computergestützte Ansätze zur Analyse altersassoziierter molekularbiologischer Veränderungen im Modellgewebe Haut exploriert, um so Gene und biologische Prozesse als neue Ziele für Interventionsstrategien zu identifizieren.

Publikation 1 zeigt den Nutzen von Multi-Omics-Daten und holistischer Analysestrategien für die Alterungsforschung auf, die zur Identifizierung von vier latenten Altersphasen in humanem Hautgewebe führten. Die Klassifizierung der Studienteilnehmer in diese Phasen führte zu einer verbesserten Erkennung molekularer Alterungssignale und zeigte zudem eine deutliche Assoziation mit der durchschnittlichen Sonnenexposition auf. Eine tiefergehende Analyse der molekularen Daten identifizierte umfangreiche nichtlineare Veränderungen in verschiedenen biologischen Signalwegen, insbesondere beim Übergang in die vierte, mit der Menopause koinzidierende Altersphase. Publikation 2 beschreibt die Entwicklung eines neuartigen Modells für die Altersvorhersage, das durch Einbettung von Informationen über zellbiologische Prozesse in seine Architektur eine neue Ebene an Interpretierbarkeit erreicht. Das Modell erlaubt dabei neben der Kalkulation des biologischen Alters der Haut gleichzeitig die Überwachung des Alterszustands biologischer Prozesse innerhalb des Gewebes. Analysen des Modells zeigten, dass Alterung mit weitreichenden Auswirkungen auf die globale biologische Prozesslandschaft einhergeht, und weitergehende Simulationsexperimente ermöglichten zudem die Identifizierung von Prozessen und Signalwegen, über die beschleunigte Alterungskonditionen wie chronische Sonnenexposition und das Hutchinson-Gilford-Progerie-Syndrom ihre Auswirkungen entfalten. Publikation 3 untersucht die molekularen Veränderungen, die durch den extrinsischen Alterungseffektor UV-Strahlung in der Haut verursacht werden. Die Multi-Omics-Analyse identifizierte auffällige frühe Ähnlichkeiten zu alters- und krebsassoziierten epigenetischen Signaturen in repetitiv bestrahlter Haut, die mit korrelierten Transkriptionsänderungen einhergingen. Untersuchungen von Unterschieden in der UV-Toleranz identifizierten Biomarker-Signaturen, die individuelle Toleranzen mit hohen Genauigkeiten vorhersagten sowie eine potentielle melaninunabhängige Schutzfunktion vor DNA-Schäden in Personen mit höherer angeborener UV-Resilienz.

Gemeinsam eröffnen die hier beschriebenen Ergebnisse neue Ansätze zur Untersuchung von Alterungsprozessen und externen Alterungstreibern wie UV-Bestrahlung in der menschlichen Haut, und liefern neue Erkenntnisse zu beteiligten Genen und biologischen Prozessen.

Introduction

The term aging generally describes the progressive impairment of normal physiological functioning over time, ultimately leading to an organism's increased vulnerability to death. Aging not only governs the extent of our life-spans however, it is also one of the greatest risk factors for most chronic human diseases, thereby drastically influencing human health-span. Understanding and addressing the complex molecular alterations driven by aging could thus be an effective way to slow the onset of the variety of systemic age-related diseases that have become ever so common particularly in western societies and that place an increasing burden on global healthcare systems[1–5].

Current theories on the phenomenon of aging largely attribute the process to a progressive accumulation of unavoidable deleterious influences that increasingly impair normal physiological processes and functioning[6–10]. Historically largely regarded as an intrinsic and invariable phenomenon inherent to complex lifeforms, more recent research is increasingly implicating the involvement of extrinsic factors modulating the progression, forming a picture of aging as a highly multifactorial process, more prone to modulation than previously believed[10–14].

In order to quantify the subtle alterations caused by extrinsic factors affecting the pace of aging, methods and approaches for capturing the actual biological aging state as opposed to mere chronological age have thus become an important cornerstone of aging research with human participants. One of the most prominent examples for such a method has been the development of 'age clocks', supervised machine learning models capable of accurately predicting the aging state of test subjects from quantitative biological data such as DNA methylation states, gene expression levels or metabolite abundances[15–24]. Age clocks have proven themselves as reliable biomarkers for aging state, even serving as predictors of remaining life-span and all-cause mortality, delivering quantifiable evidence of the uncoupling of chronological and biological age[18,19,25–29]. Analyses relying on the estimated ages from age clocks have also uncovered significant associations between accelerated aging rates and various systemic diseases, including cancer, cardiovascular and coronary heart disease as well as neurodegenerative disorders[28], demonstrating the importance of understanding aging as a process to slow the onset of disease and improve human health-span overall. In order to improve the sensitivity of these analyses further, a second generation of age clocks has been developed, trained on estimates of 'phenotypic' as opposed to chronological age. These phenotypic age estimates, calculated from descriptors of systemic health, such as levels of certain age-associated blood plasma proteins or mortality data from large longitudinal data sets, allowed the training of clocks that proved even more powerful for the identification of aberrant aging rates and for the prediction of remaining life- and health-span[27,29].

Despite their unquestionable utility however, current age clocks generate little insight into the actual biological processes driving aging, limiting their use to mere diagnostic or read-out tools. Transforming these ‘black box’ tools into transparent, interpretable models of how aging progresses on a molecular biological level, holds great promise then to further drive our understanding of aging as a whole. A new generation of interpretable age clocks could generate valuable insight on the genes and pathways driving their predictions, which could supply useful targets for the design of intervention strategies to help slow aging on a molecular level.

Another general perception about aging that has seen change over the years concerns the linearity of aging. Aging has traditionally been perceived as a linear process, and recent achievements such as the development of age clocks – in their earlier implementations usually variants of linear models – have generally not challenged this perception. Notably however, on a molecular biological level, several indications of non-linearity in the aging progression have recently been described[30–32]. Most indications of this are derived from model organisms such as fruit flies, recently however, data generated from human tissues have hinted that similar mechanisms might play a role in human aging as well. One of the most compelling of such reports describes the sudden loss of an age-associated transcriptional signature related to the known age-associated IGF-1/PI3K/mTOR-signaling pathway around the sixth decade of life in multiple human tissues[32]. The finding indicates a non-linear switch in an important regulatory pathway that has been well-described in the context of aging and longevity in various model organisms. Non-linear changes in gene regulation patterns might thus represent a previously overlooked feature of human aging that might warrant increased attention.

An attempt at categorizing the different biological processes that both drive and are themselves affected by age-related changes has been made in the form of the Hallmarks of Aging[33]. This conceptual categorization of biological phenomena that have been found associated with increasing age across different organisms and tissues proposes nine universal pillars of aging. These pillars are divided into three groups: primary Hallmarks with exclusively negative consequences (epigenetic alterations, telomere attrition, genomic instability and a loss of proteostasis), antagonistic Hallmarks with either beneficial or detrimental effects depending on their intensity (mitochondrial dysfunction, nutrient signaling – which prominently includes the previously mentioned IGF-1/PI3K/mTOR-signaling pathway – and cellular senescence) and finally the integrative Hallmarks (altered intercellular signaling and stem cell exhaustion), which are believed to emerge as consequences of the other Hallmarks and are hypothesized to be in great part driving the late detrimental effects observed in aging tissues[33].

Most Hallmarks represent highly interconnected processes, such as epigenetic alterations and telomere shortening strongly contributing to genomic instability, and genomic instability in turn marking one of the strongest factors finally driving cells into a state of senescence, an important phenomenon frequently encountered in aged tissues, in particular in the skin[34–38]. Cellular senescence, originally devised by nature as an important protective mechanism against cancer, describes the complete cessation of cell division upon persistent stresses such as DNA damage[34,39,40]. Early on an overwhelmingly beneficial response, the gradual accumulation of uncleared senescent cells in aging tissues leads to a severely deleterious microenvironment driven by the secretion of inflammatory cytokines, chemokines, and other soluble factors, which influence intercellular communication and interfere with normal tissue function[35,36,38,41].

The proposed sequence of emergence places the primary Hallmarks at the earlier stages of aging, leading to an accumulation of deleterious changes, that might serve as triggers for other age-related alterations. The antagonistic Hallmarks, which at the early stages pose no problem and at low intensity can even exert beneficial effects on the functioning of cells and tissues, can over time develop into progressively negative mechanisms, potentially exacerbated and accelerated under the influence of the primary Hallmarks. The integrative Hallmarks are hypothesized to finally manifest as a consequence of the accumulated damages of primary and antagonistic Hallmarks and are thus believed to follow them in temporal sequence[33]. The categorization gives a good overview over different cellular processes and their involvement with aging, however so far, their overarching classification and in particular the order of their temporal emergence have remained largely theoretical, with a lack of data-driven approaches available to quantify their emergence.

Of all organs in the human body, the skin represents a particularly well-suited tissue for studying aging in general. The skin is the largest organ of the human body, yet it is easily accessible, and biopsies can be taken using non- or minimally-invasive sampling methods, such as suction blisters. The skin is compartmentalized into three layers with distinct functions: the epidermis, representing the true barrier to the outside world and mainly consisting of specialized epithelial cells called keratinocytes; the dermis, a thicker layer of collagen-rich connective tissue largely populated by extracellular matrix producing fibroblasts providing mechanical cushioning as well as immune cells, specialized glands and hair follicles; and finally the hypodermis, a layer of subcutaneous fat, providing thermal insulation. The skin serves several crucial functions guaranteeing our survival, among it limiting the water loss from our bodies, regulating our body temperature, providing sensory functions and importantly as a barrier, protecting us from harmful pathogens and environmental insults such as hazardous solar irradiation. Owing to its exposed nature, the skin and in particular its outermost layer, the epidermis, is frequently subjected to extrinsic influences with potential impact on age-related processes, making it a highly suitable tissue to study the effects of accelerated aging through extrinsic factors.

The most prominent of such factors is solar irradiation. Exposure in particular to the ultra-violet fraction of natural sunlight is well-known to induce DNA damage and thereby contribute to genomic instability. This is in part caused directly by the dimerization of adjacent pyrimidine bases, as well as indirectly by increasing oxidative stress due to the formation of free radicals and reactive oxygen species within cells[33,42,43]. Apart from direct damage to DNA and proteins within cells, UV light also impacts the extracellular matrix within the skin by causing photodegradation of collagen fibers and modulating the expression of genes related to tissue structure and modeling by dermal fibroblasts[36,42,44]. Chronic exposure of the skin to solar irradiation then induces a skin phenotype characterized by wrinkling, dyspigmentation and a leathery appearance, in many traits resembling naturally aged tissue, leading to the description of this skin-specific phenomenon as 'photoaging'. If further left unprotected and untreated, photoaged skin can develop into actinic keratoses, pre-cancerous lesions frequently encountered in light-skinned individuals with a history of sun exposure, which present a significant risk to progress further into cutaneous squamous cell carcinoma[42,45–48]. This direct progression exemplifies the risks associated with accelerated aging and its impacts on human health and illustrates the importance of furthering our understanding of the intricate molecular processes driving biological aging.

Methods

In order to analyze the molecular changes triggered by intrinsic and extrinsic aging in the skin, a diverse set of statistical and computational methods was used, ranging from correlation and linear regression to different types of machine learning algorithms.

Most of the data analyzed for the purposes of this thesis can be classified as so called 'omics'-data, generated from high-throughput assays and technologies. The term 'omics' is generally used to describe several biological disciplines that aim at holistically describing biological processes or levels of biological organization, often involving the simultaneous profiling of tens to hundreds of thousand biological molecules or entities to characterize the complex dynamics that ultimately shape an organism's structure and function. In order to draw even more robust conclusions on biological processes, several assays profiling different layers of biology are frequently combined. These 'multi-omics' approaches have been shown to greatly improve the detection of biological signal amidst technical noise[49–54], and after being pioneered mainly in cancer research, are now quickly gaining traction in other research fields as well.

The high dimensionality of each data set poses significant challenges for the integrated analysis of several layers of biological regulation however, as do differing scales and distributions of parameters and assay-specific technical noise and bias[50,52,54,55]. In recent years, a number of algorithms have been developed to tackle these challenges, which allow the computational integration of multiple large data modalities for a simultaneous analysis of different biological layers. One possible way to achieve this – and the one specifically chosen for the integration of multi-omics data sets in this work – is using a network-based approach known as similarity network fusion[50]. Similarity network fusion relies on the calculation of pairwise similarities between individual samples, which are first determined for each of the data modalities separately. The resulting sample similarity matrices for each data type are subsequently transformed into networks, in which individual samples are represented by points while the edges between them code for their pairwise similarities. The advantage of this transformation of the data from diverse biological features to a similarity network space, is that it offers a way to avoid many of the aforementioned issues related to differing scales and distributions as well as other platform-specific biases. To integrate the individual sample networks, an iterative fusion algorithm performs stepwise updates on the edges, strengthening connections between samples that are pronounced in more than one data modality, thus making the networks more similar with each iteration, whilst unlocking synergistic information from the data sets in the process. The final fused similarity network then incorporates information from all separate input data sets and can be used for downstream analyses such as the identification of hidden subtypes among samples using unsupervised cluster analysis, as performed for the publications encompassed in this thesis. Clustering describes the task of grouping objects by similarity, yielding sets of objects that are more similar among themselves than they are to objects in other groups, which facilitates the detection of discrete groupings in the input data, and represents an example for an unsupervised machine learning technique. In this work, similarity network fusion and subsequent clustering were used for the identification of latent groupings along the aging progression in multi-omics data, leading to the identification of four discrete aging phases in epidermal tissue described in paper 1, as well as the identification of divergent biological responses among subjects to repeated UV irradiation in paper 2.

Machine learning algorithms can broadly be classified into supervised and unsupervised learning techniques, which differ in the type of data required and their area of application. Unsupervised learning algorithms are capable of learning patterns and representations from data without prior labeling of training cases, and frequent use cases include the aforementioned cluster analysis, outlier or anomaly detection, or the learning of latent representations. Supervised algorithms on the other hand learn representations from pairs of data and labels, essentially constructing functions that allow for the mapping of new data cases to their matching labels. Supervised machine learning tasks are further divided into either classification or regression tasks, depending on the type of label to predict, with classification models trained on predicting discrete classes and regression models used for predicting continuous numeric variables.

Age clocks, such as those employed in papers 1 and 2, represent prime examples for regression models. A variety of different age clocks has been proposed before, trained on different types of biomarkers, ranging from DNA methylation sites across the genome to metabolite abundances and more[15–22,24]. What all of these clocks have in common, no matter the data type used, is that they model the relationship between biological input features and age (either chronological or a proxy for phenotypic age[27,29]) of the sample donor as output, generating continuous biological age estimates, which have been shown numerous times to represent more accurate predictors of life-span and health status than chronological age itself, making these types of regression models very useful tools for aging research.

Various machine learning algorithms have been developed over the years to solve both types of supervised learning tasks, with most algorithms having different variants optimized for either classification or regression problems. The most notable algorithms used in this work for classification or regression tasks were implementations of random forests, support vector machines and artificial neural networks.

Random forests are an ensemble learning method based on a large number of individual decision trees[56]. Each decision tree represents a model that attempts to derive an optimal set of decision rules to predict new data points according to a number of features and training cases. Decision trees are constructed from nodes and branches, starting from a single root node. At each node, the available features are evaluated according to how well they split the cases in the training data, to derive the best prediction on the training set. This process is repeated recursively while constructing the tree, at each step using one of the remaining features that best splits the data, until reaching their final nodes (leaf nodes) at which label predictions are generated based on the decisions made along the branches of the tree. While very flexible and intuitive tools, a single decision tree can be prone to overfitting, especially as the number of splits increases, often leading to a biased model which performs exceptionally well during training, but poorly on new, unseen data. Random forests mitigate this problem by making use of a multitude of decision trees, each trained only on a randomly selected subset of the training data, with predictions aggregated from all individual decision trees. This process of randomly constraining the data that each individual model sees and aggregating the results is also known as bootstrap aggregating or bagging, and greatly helps reduce the bias associated with single decision trees, leading to models that generally perform very well on new data sets, even with minimal adjustments of model parameters. Random forests were used in this work for the training of classification models, allowing the grouping of new subjects to the aging phases identified in paper 1 and as part of the pathway predictivity analysis developed for ranking biological processes according to their predictivity for a given biological phenomenon, which in paper 1 was used to characterize aging phase identity.

Support vector machines on the other hand function by constructing a hyperplane or several hyperplanes between training cases which best separate them according to their labels[57]. The optimal hyperplane is determined by maximizing the margin, the distance between the hyperplane and the training cases closest to it. In cases where linear separability in a finite-dimensional space is not given, support vector machines are capable of mapping data points into a higher-dimensional space before hyperplane construction using kernel functions, which calculate the dot products representing the data points in higher-dimensional space, allowing non-linear classification and regression problems to be tackled. In this work, support vector machines were used as an alternative to random forests to rank the importance of various biological pathways in the pathway predictivity analysis used in paper 2.

Artificial neural networks represent a type of machine learning algorithm loosely inspired by the function of mammalian neural networks[58–61]. The building blocks for artificial neural networks are nodes or artificial neurons, and edges. The artificial neurons, typically organized in layers, perform non-linear mathematical operations on the sum of their inputs, while the edges form connections between neurons of different layers and thus allow numerical information to be passed from one neuron to the next. Every individual edge between two neurons across the network is assigned a weight, a randomly initialized numerical factor by which incoming information is multiplied before it is passed on to the next neuron, thereby increasing or decreasing the strength of individual connections. At the end of the receiving neuron, all incoming products are summed up and passed through a non-linear activation function, such as a sigmoid or rectified linear unit – the non-linearity of the activation at this step being crucial in allowing the network to learn complex non-linear relationships between input and output. During the learning phase, the weights between different neurons are adjusted at each epoch using backpropagation of errors to derive an optimal weight configuration across the network in order to predict a given label at the final node (or layer of nodes). This setup then, at least conceptually, bears resemblance to mammalian neural networks, in which signals are transmitted from one neuron to the next via synapses, where the incoming signals are registered, and the neurons reacts when a certain activation threshold is exceeded. Regular feedforward networks consist of several stacked layers of neurons between an input layer on one end, where information on features is fed into the network, and an output layer generating the predictions on the other. Usually in these networks, each neuron in a certain layer is connected to each neuron in the next, leading to their designation as ‘fully connected’ artificial neural networks. When these networks reach several layers in depth, the term ‘deep learning’ is often used to describe these types of machine learning models. In paper 3, an artificial neural network using a specifically designed architecture incorporating prior information on biological pathways into the model structure, was used to construct the first interpretable neural age clock.

Results

The publications detailed in the following describe several different aspects of aging in the skin with an emphasis on the close interplay between intrinsic and extrinsic factors driving aging. The latter notably includes solar irradiation, which also happens to constitute a major effector driving age-related carcinogenesis in the skin. A common theme across the publications is further the exploration of novel computational approaches and methods to assess biological aging state as well as genes and pathways involved with aging and age-related pathogenesis.

Investigation of non-linear aspects of aging in epidermal tissue

Following recent reports on non-linear changes in gene regulation within the nutrient sensing IGF-1/PI3K/mTOR-signaling pathway as a known Hallmark of Aging in human tissue[32], paper 1 was dedicated to investigating signs of non-linearity across multiple levels of molecular biological data at a broader scale using an unsupervised clustering approach. Borrowing from precision medicine, where the generation and analysis of multi-modal data sets has proven itself highly useful for enhancing the detection of hidden subtypes in heterogeneous biological data, multi-omics profiling was employed to investigate this question, in the hopes that a combination of different data modalities could uncover information on groupings that would be difficult to detect in any single data type. For this, genome-wide DNA methylation and gene expression data from 86 epidermal samples of female donors of ages between 21 and 76 years was generated and subsequently analyzed.

The data sets were computationally integrated into a single data modality using a similarity network fusion approach[50]. The resulting integrated network, representing sample similarities across the different data modalities, was then used as basis for an unsupervised cluster analysis, which identified four distinct subgroups among the samples. Analysis of the detected clusters showed a strong association of the groupings to chronological ages of the subjects, which led to the hypothesis that the clustering had identified stages in the multi-omics data capturing biological aging states of the study participants. This hypothesis was supported by subsequent analyses using transcriptomic and epigenetic age clocks. The association was particularly pronounced in the case of aging phase outliers, subjects which appeared to have prematurely migrated to the next aging phase relative to their chronological age, and presented significantly increased epigenetic and transcriptomic age estimates. Analysis of self-reports of the test subjects further revealed that these phase outliers were significantly more likely to have reported frequent sunbathing in the questionnaires, delivering evidence of how chronic solar irradiation drives aging of the skin on multiple biological levels and affirming the notion that the identified phases captured biological aging states of the skin.

In order to characterize which biological motifs characterize the different phases and explore the temporal sequence of emergence of the Hallmarks of Aging in a data-driven manner, a new approach for assessing pathway relevance was developed, based on machine learning classifiers quantifying the predictivity of a given set of genes for the phenotype in question. Analysis of the predictivity patterns of custom gene sets capturing the Hallmarks of Aging along the four aging phases interestingly revealed distinct succession patterns of the different Hallmarks over the phases. Even more, a clustering of the Hallmark emergence patterns showed strikingly close resemblance to their postulated classification into primary Hallmarks (genomic instability, telomere attrition, epigenetic alterations and originally loss of proteostasis), antagonistic or secondary Hallmarks (cellular senescence, deregulated nutrient sensing and mitochondrial dysfunction) and integrative Hallmarks (altered intercellular communication and stem cell exhaustion)[33]. The only divergence in classification with the postulated grouping was found for the proteostasis-Hallmark, which showed stronger similarity with the group of integrative Hallmarks in the data-driven clustering.

As stated earlier, reports on non-linearity in the regulation of IGF-1/PI3K/mTOR-signaling have been pinpointed to the mid-life transition[32], a period surprisingly very well captured in the identified aging phases in the transition of phases 2 to 3 (which in average chronological ages very well matches the transition into menopause) and phases 3 to 4 (coinciding with the transition into post-menopause). In order to assess the extent of non-linearity in biological processes around this point in life in more detail, a gene set predictivity analysis with increased resolution of pathways covered was performed. The analysis revealed a wide impact of aging across biological processes, with a number of pathways featuring non-linear predictivity patterns along the four phases. Prominently, a distinct loss in predictivity was detected for a lot of pathways in the transition from aging phases 3 to 4, matching strikingly well with the previously reported ‘mid-life switch’ in mTOR-related signaling[32] and the average transition out of menopause[62–65]. This loss in pathway state could also be detected using gene set enrichment methodology and very much appeared to be a global phenomenon across the pathway landscape. Among the pathways affected were several key regulatory pathways well-described in the context of aging, such as the previously reported mTOR-signaling pathway, confirming the previously cited finding[32] for skin tissue. The analysis identified a substantial number of further pathways showing similar patterns however, including ones related to cellular energy metabolism, stress response and also DNA repair processes.

The collective loss in pathway signal might be connected to an increase in transcriptional noise, a hypothesis that has been proposed before as a mechanism driving aging in multicellular organisms[66–68]. To investigate the involvement of stochasticity in gene expression in the mid-to-late-life transition, pairwise correlations between transcriptomic profiles of subjects in the respective aging phases were calculated to compare the degree of variation between transcriptomes which would be influenced by increases in transcriptional noise. The analysis of these inter-individual correlations showed a significant drop in transition to phase 4, in line with the transcriptional noise hypothesis and for the first time aligning increases in transcriptional noise with the menopausal transition.

One of the overall most predictive biological processes for aging phase identity in the analysis was notably found to be p53-signaling, a pathway that is not only a crucial part of the general DNA damage response and thus protection from carcinogenesis, but also strongly associated with the phenomenon of cellular senescence, as the irreversible growth arrest characteristic for this Hallmark of Aging is mainly mediated by the p53 as well as $p16^{Ink4a}$ pathways[69–74].

Together these findings then implicated previously overlooked, wide-spread non-linear changes at the very core of aging in the human skin, with potentially far-reaching implications, contributing to the functional deterioration of the tissue in old-age.

An interpretable neural age clock for transcriptomic data

The analyses in paper 1 had already demonstrated the utility of machine learning models as predictive tools in a research setting, as exemplified by the transcriptomic and epigenetic age clocks used to validate the identified aging phases. As previously stated however, first and second generation age clocks, including ones such as those used in paper 1, offer little insight on the biological processes actually driving their predictions. Paper 2 then describes the development of a new type of age clock, aimed to deliver both predictivity as well as interpretability of its inner workings. The concept of this new age clock is based on the incorporation of prior information on biological pathways into the architecture of an artificial neural network, controlling and guiding the flow of information across specific neurons and thereby forcing different parts of the network to model different biological processes. Artificial neural networks with sufficient depth are well suited to modeling arbitrarily complex non-linear relationships between input data and output, which given the non-linear alterations in the gene expression landscape observed in paper 1, might represent an important model property when designing age clocks, at least with transcriptomic data.

The network architecture was modeled according to the hallmark process gene set collection, which had previously served well for the characterization of aging-related pathway alterations in paper 1. The input layer was fed with transcript abundance data on the gene level, followed by several compartmentalized and pathway-centric layers propagating the information to a final neuron per pathway, which served as auxiliary output to monitor the respective pathway's aging state. The last layer, aggregating the individual pathway neuron outputs, generated the final age predictions. Training and testing of the network was performed on a large transcriptomic data set of epidermal skin samples (n = 887) from the SHIP-TREND study, a population-based longitudinal cohort study launched in 2008 to assess the prevalence of common clinical diseases, sub-clinical disorders and various risk factors amongst the general population in Northern Germany[75].

The final model achieved a median absolute error of 4.7 y on the independent test set, comparable in accuracy with other published transcriptomic age clocks[22,23,76,77], as well as a fully connected 'black box' model trained on the same data, indicating a relatively small tradeoff between performance and transparency.

One of the advantages of using skin as a tissue for studying aging is the ability to observe phenotypic manifestations of extrinsically accelerated aging with relative ease in the form of wrinkling and changes in texture and overall appearance of the skin. To investigate, if the assessments of biological age generated by the model correlated with any such phenotypic age markers, visual age estimates for a subset of the test cohort were generated using a blinded expert panel from standardized portrait photographs. Subsequent analysis revealed a significant association between the visual and transcriptomic age estimates after correcting for chronological age and sex, validating that the age clock indeed managed to capture biological aging states.

Observing the intermediate neuron activations while feeding the model with data from subjects of varying ages allowed a detailed assessment of the aging states of various biological processes for each individual sample. The data generated by analyzing neuron activation patterns for the test set revealed a diverse and wide-spread impact of aging on the pathway landscape, affecting most of the pathways covered by the model to at least some degree, in line with the wide-spread effects on pathways observed in paper 1. Correlation analysis of neuron activations with chronological ages of the test subjects revealed some distinct and reproducible differences in the extent of alterations these pathways underwent during aging however. The correlation-based ranking was led by p53- and TNFa-NFkB-related signaling pathways as the most strongly affected processes, followed by several other inflammatory and immune signaling pathways, as well as genes implicated in the response to UV irradiation and apoptosis. The top pathways p53- and TNFa-NFkB-signaling hinted towards a strong representation of cellular senescence signals captured by the model, again indicating a strong involvement of this Hallmark of Aging in the skin.

To further validate the model's utility, several virtual knockdowns of known and well-described aging target genes were performed *in silico* by downregulating these genes in the model's input data and assessing the impact of the artificially introduced alteration on age assessments generated by the clock. The simulated knockdowns of the respective genes chosen from the literature matched the reported functional impacts in all cases. Based on these results, systematic knockdown simulations of all genes covered by the model were performed. The simulations identified a number of well-described aging target genes among the strongest impact predictions, such as the SERPINE1, CDKN2A, IGFBP3 and TIMP1, several of which have previously been described as markers of senescence[37,69,70,74,78–82], in concordance with the pathway-level analysis of the inner neuron activations. The simulation experiments also revealed several genes related to metabolic processes that so far received less attention in terms of aging however, which were predicted to have substantial impact on the biological aging state, and that might represent interesting targets for future investigations.

As the individual gene knockdown simulations had accurately recapitulated known associations and revealed numerous other genes impacting the overall age prediction, we hypothesized that the clock could also be used to investigate the impact of more complex transcriptional signatures on biological aging state and the aging pathway landscape. In validation experiments using literature-derived gene expression signatures of accelerated aging conditions (Hutchinson-Gilford progeria syndrome or HGPS, photoaging), age-related processes and diseases (senescence, actinic keratosis and skin cancer), as well as established intervention strategies described from model organisms (caloric restriction), the impact of altered gene transcription levels on biological aging state was simulated using the model. While the predicted effects on biological age estimates varied widely between the signatures, the simulations generally recapitulated the overall conditions very well, with increased transcriptomic age estimates for accelerated aging conditions and pro-aging drivers such as HGPS and photoaging, and a modest rejuvenation observed under conditions of caloric restrictions, in line with findings from model animals[11,14,83,84].

To decode the biological processes associated with the observed changes in biological aging states, the alterations in the intermediate neuron activation patterns after gene perturbation with the respective signatures were analyzed. The analyses revealed several interesting aspects to the nature of the observed conditions, many in line with existing theories as well as some new findings, including pathways previously largely unrecognized in the pathophysiology of HGPS. Chronic sun exposure had already been shown to be a significant driver of aging in the epidermis in healthy human subjects in paper 1 and the simulations using a signature of chronically sun exposed skin using the clock had yielded similar results. Analysis of the pathways most strongly affected by photoaging revealed a particularly strong involvement of the oxidative stress response, with the neuron coding for the ROS pathway shifted particularly strongly towards a pro-aging state during the simulation of photoaging, which is very much in line current theories on the condition[42–44,85]. Further pathways altered in photoaged skin included several signal transduction and metabolic pathways, whose predicted implication in the emergence of the phenotype could provide starting points for future investigations.

Together these findings demonstrated the gain in utility awarded by designing predictive models with interpretability in mind, which expands their use from being biomarkers to even more powerful research tools, capable of generating important insight on the biological pathways underlying the modeled processes.

Acute effects of the pro-aging effector solar irradiation across multiple levels of biology

Several aspects of the publications 1 and 2 had already demonstrated the impact of chronic solar irradiation as an important extrinsic factor for aging. In paper 1 chronic sun exposure was identified as a significant factor driving the premature transition into higher aging phases, and the simulation experiments using signatures of photoaged skin in paper 2 had similarly demonstrated the impact of chronic sun exposure on biological aging state and the global pathway landscape and identified oxidative stress as a major driver of the condition. Solar irradiation, and in particular its ultraviolet components, not only drive photoaging however, they also present a significant risk for the development of skin cancer[42,45,86–91].

The long-term effects of solar irradiation are rather well-described then, however less is known on the short-term impact of repeated exposure to irradiation of the skin, in particular in terms of epigenetic alterations, which happen to represent a highly prognostic biomarker for both aging and carcinogenesis[76,92–95]. Considering the fact that non-cancerous chronically sun exposed skin already displays epigenetic features characteristic for squamous cell carcinomas[95,96] and that even few short-term UV exposures (i.e. severe sunburns during childhood and adolescence) are associated with late epidermal tumorigenesis[45,97–101], we were interested in examining how quickly such alterations could arise after sun exposure, as well as explore the heterogeneity in UV sensitivity observed across different phototypes on a biological level.

The aim of the third study then was to generate a detailed image of the molecular biological events following acute repetitive irradiation of the skin, capturing alterations of both gene expression and DNA methylation patterns, their implications for aging- and cancer-related processes, and biomarkers and mechanisms defining differences in UV responses across subjects. For this, 32 female Caucasian volunteers with varying levels of innate UV tolerance (Fitzpatrick phototypes 1 to 4) were repeatedly subjected to individually calibrated doses of simulated solar irradiation reaching 0.9 MED (minimum erythema doses) on a sun-protected area on their lower backs. Epidermal samples were taken from irradiated and control areas and full-transcriptome gene expression as well as DNA methylation profiling performed from these samples. To assess the extent of molecular alterations introduced through the irradiation procedure, paired differential gene expression and methylation analyses were performed, which revealed extensive changes on both transcriptomic and epigenetic level.

In general, a tendency towards hypomethylation of CpGs was observed, which is in line with the detection of large hypomethylated blocks in photoaged skin. To investigate the extent by which the observed alterations match those found in chronically sun exposed skin, a more detailed analysis of genomic regions previously implicated with photoaging was performed. The analysis identified a strong linear correlation between hypomethylation patterns in a substantial fraction of the genomic regions investigated, hinting that even few repetitive short-term exposures to solar irradiation can already significantly impact long term epigenetic imprinting of the skin. Considering the risk factor that repetitive sunburns pose in terms of skin cancer development later in life, a comparative analysis of a large number of genomic regions described to be differentially methylated during tumorigenesis was performed, further revealing several regions across the genome with correlated methylation patterns post-irradiation and in cancerous tissue.

The identified alterations in DNA methylation patterns after repetitive irradiation were substantial, however not every methylation or demethylation implicates an immediate functional impact on the expression of a gene. To quantify the extent by which differential methylation in functional gene regions actually correlated with changes in expression, a transcriptome-wide association analysis was performed, modeling gene expression as a function of mean DNA methylation of CpGs annotated to various functional gene regions. The analysis identified a large number of significant associations, most of which were found to be located in enhancer regions. Among them were several well-known UV-responsive genes, commonly detected to be differentially expressed in *in vitro* irradiation experiments, but more interestingly also several previously unrecognized UV response genes, such as the immunomodulatory CARD14, a known positive regulator of NFkB-signaling, and the tumor suppressor gene IRF8, which notably is located within one of the genomic regions recently found differentially methylated in photoaged skin, thus potentially forming a direct link between extrinsically accelerated aging and carcinogenesis in the skin.

Inter-individual variation in UV sensitivity and tolerance of the human skin is substantial and of great importance for the risk of suffering from accelerated photoaging and developing skin cancer later in life[45,102–105]. As a means of estimating approximate UV sensitivity, the Fitzpatrick scale is frequently used, which classifies subjects into six distinct phototypes mainly based on skin complexion and hair color[106]. This classification, while useful due to its easy criteria, has been shown to only roughly correlate with actual UV sensitivity[107–109], which is more objectively quantifiable using the MED, i.e. the minimal dose of UV light required to evoke a visible erythema response in a test subject's skin[110,111]. As this more objective measurement procedure requires irradiating subjects with different dosages of UV with potentially harmful repercussions however, we investigated whether it might be possible to predict UV tolerance from molecular markers instead, that could be profiled from epidermal samples in a minimally-invasive manner. To assess the feasibility of this approach, the UV sensitivities of the test subjects in our collective as measured by their respective MED, were used as response variable in three different regression models trained on gene expression, DNA methylation and a data set combining both data types. The cross-validated models showed remarkably high accuracies for all data types, in particular those utilizing DNA methylation features, substantially surpassing Fitzpatrick classifications for UV sensitivity prediction.

The best predictor of individual UV tolerance was found in the combination of expression and methylation features, hinting that synergistic information might be present in the data, offering additional insight on the basis for the observed heterogeneity in UV sensitivity and potential variation in UV response mechanisms. To unlock any latent information from the combination of the data sets, the similarity network fusion approach as previously utilized in paper 1 was used to integrate the two data levels. To explicitly assess heterogeneity associated with the immediate biological UV response, only biological profiles from irradiated samples were used for the similarity network fusion process. Subsequent clustering on the fused network identified three latent subgroups of samples in the combined data. Notably, the three groups showed very strong correlation to the previously measured MED of the test subjects and allowed a better stratification of subjects according to UV sensitivity than Fitzpatrick phototypes, despite the unsupervised nature of the analysis.

In order to characterize the biology driving these molecular phototypes (MPs), pathway-based machine learning classifiers were used to assess the predictivity of various biological processes for UV irradiation status in the different groupings. Despite a large number of pathways showing similar predictivity patterns across the three groups, the analysis revealed a number of biological processes with divergent responses between the molecular phototypes. These included stronger signals detected for inflammatory and immune signaling in the less UV tolerant phototype MP 1 and to lesser extent MP 2, with a focus on cytokine signaling and a particularly defined inflammasome response in MP 1. MP 2 on the other hand showed a stronger regulation response in apoptotic and autophagy processes than both MP 1 and MP 3, potentially connected to p53-related signaling, which was also markedly more defined in this group. Both higher MED groupings MP 2 and MP 3 meanwhile presented stronger signals related to pigment metabolic processes. MP 3, the group containing the most UV tolerant subjects in the cohort, showed particularly pronounced signals related to DNA damage checkpoint and synthesis pathways, whereas inflammatory and immune response showed lower predictivities in this phototype in comparison. These findings were suggestive of a more sensitive DNA damage detection machinery in subjects with higher innate UV tolerance, which could in turn provide a quicker and more effective DNA damage repair. To test this hypothesis, the extent of photodamage in the form of cyclobutane pyrimidine dimers (CPDs), the most common type of UV-induced DNA damage[87,88,112], was profiled in a random subset of subjects using material left from the original sampling. Intriguingly, the analysis of CPD levels showed significantly lower extent of DNA damage in subjects from MP 3, in line with the molecular pathway data and the hypothesis of a pigmentation-independent UV protection mechanism in subjects with exceptionally high UV tolerance.

This presents an interesting subject for further research, as it may be related to the highly divergent carcinogenesis rates observed across different ethnicities, which are not fully explained using skin pigmentation alone[45,102–105].

Discussion

In this thesis, several aspects of intrinsic and extrinsic aging in the human skin were examined in high-dimensional biological data using computational approaches.

The first paper delivers evidence that aging in the skin should not be regarded as a strictly linear process, but instead involves a variety of non-linear changes on the molecular biological level, that allow the classification of subjects into several aging phases. The analysis shows the value of utilizing 'multi-omics' data and methodologies to explore biological aging, an approach that has not been widely adapted yet within the field, likely related to cost and effort of performing such studies as well as the increased computational efforts required to unlock its potential. In terms of skin aging in particular, no similar publications on the topic are available as of the writing of this thesis. As aging represents a highly complex phenomenon spanning many if not all levels of biological regulation, holistic 'multi-omics' approaches present a very promising tool to furthering our understanding of the mechanisms involved. New efforts such as the Aging Atlas[113] are now starting the process of collecting, curating and combining published data from various different biological levels, to provide easy access to the results from different 'omics' and 'multi-omics' studies exploring aging for biologists, to accelerate research going forward.

Two findings from the study exemplify the utility of multi-omics data in recovering subtle latent information from high-dimensional data and increasing the stability of inferred conclusions: The Hallmarks of Aging have widely been used as a theoretical framework aiming to not only classify and categorize the various age-related changes observed in aging cells, tissues and organisms, but also explain the order of their emergence along the aging progression. In paper 1, the detection of these conceptual cornerstones of aging and their emergence *in vivo* in the human skin were facilitated using the identified multi-omics aging phases as anchors for a pathway predictivity analysis, which revealed divergent predictivity patterns for the postulated primary, antagonistic and integrative Hallmarks, which were found in direct correlation with the identified phases, and strikingly close to the groupings proposed in the original publication[33]. The data-driven reconstruction confirmed for the first time the general order that was hypothesized by the authors, however it also pointed to a potential misclassification of the loss of proteostasis as a primary Hallmark. Based on the collected data a classification among the emergent integrative Hallmarks, appearing at a later stage of the aging progression would seem advisable. A growing body of evidence linking the proteostasis systems with inflammatory processes in the literature paints a picture of a complex and bidirectional relationship between these processes already, such as the modulation of proteasome and autophagy function by TNFa signaling[114–119] or the activation of NFkB and TNFa synthesis by the unfolded protein response under ER stress[120–122]. The multi-omics data and literature available would make a joint emergence of these Hallmarks during aging thus seem plausible indeed.

Secondly, the analysis of the identified aging phases also revealed a particularly interesting phase transition coinciding with the average chronological age commonly associated with menopause in women in industrialized Western countries[62–65]. Remarkably, a distinct non-linear change in the pathway states of a substantial number of processes could be pinpointed to this phase transition. This observation is of particular interest, as the entry into menopause has previously been associated with a significant epigenetic age acceleration[123]. On a phenotypic level, evidence also points to an increase in age-related changes at or around the menopause, such as an accelerated thinning, decreases in collagen content, loss in skin elasticity as well as increased wrinkling and dryness[124]. The identified non-linear molecular biological changes detected in the phase transition through the multi-omics analysis could provide a starting point to explain these signs of age acceleration.

Interestingly, the observed alterations were detectable across a large fraction of the analyzed pathways, very much resembling a global phenomenon captured in the multi-omics data. Among the pathways affected by the transition were both DNA damage sensing and repair pathways including p53-signaling, processes with crucial function in the skin in particular, as they serve as safeguards against solar irradiation-induced carcinogenesis, as was shown by analyses carried out in paper 3, where p53-signaling and DNA repair pathways were found associated with UV resilience after repeated irradiation. p53 serves multiple roles in the regulation of cell cycle, programmed cell death and genomic stability, and is widely regarded as one of the most important tumor suppressor genes in the human genome[125–129]. As such, it also directly governs DNA damage response and repair processes in the skin following exposure to damaging extrinsic influences such as solar UV irradiation and is essential for preventing genome mutations particularly in the relatively frequently and heavily exposed epidermal keratinocytes[127]. This links p53 directly to the accumulation of mutations in consequence of chronic sun exposure and the development of skin cancer, the risk of which notably also happens to increase substantially around the sixth decade of life, in temporal overlap with the observed phase shift from phase 3 to phase 4. The hypothesis that p53-signaling might be connected to the increased rate of cancer incidents in old age has been proposed before, as data from model animals had shown a similar correlation between diminished p53 activity and increased tumorigenesis in aged animals[130–132], and with animals in which the onset of diminishing activity was delayed interestingly exhibiting an increased life-span[132]. As such, these findings from *ex vivo* human skin could offer valuable insight on important regulatory processes occurring at the nexus between aging and carcinogenesis in human tissue that might well be worth following up on in future studies. Finally, the detected aging phases could also open up potential avenues towards personalized anti-aging intervention strategies. Such strategies could be tailored at ameliorating and slowing the age-associated changes characteristic for specific aging phases, akin to the use of molecular subtyping to inform therapy decisions and improve efficacy and treatment responses in personalized medicine.

The identified non-linear alterations in various age-related pathways were importantly also factoring into the second project, the design of an interpretable age clock. A number of design decisions were to be made for the construction of the neural age clock. One among them was the depth of layers, which has direct implications on whether or not the network is capable of modeling complex non-linear associations between its in- and outputs, with deeper networks consisting of two or more hidden layers and non-linear activation functions able to approximate arbitrarily complex functions. While earlier age clocks were often built using linear models[15–18,21,22], the findings from paper 1 suggested that more complex models might be useful to accurately capture the non-linear alterations in genes and pathways in transcriptomic data. This guided the decision to opt for a deeper architecture that could take advantage of the non-linear and discontinuous regulation patterns observed along the age gradient.

The newly designed neural age clock with its pathway-centric design offers a defining advantage compared to earlier generation clocks, and that is the added interpretability of the model, greatly expanding the utility of the tool. The analysis of intermediate neuron activation patterns revealed wide-spread alterations of the pathway-landscape with increasing age, with no single pathway greatly dominating the ranking. This is very much in line with observations from paper 1, where the pathway predictivity analysis of the aging phases revealed a similarly broad impact of increasing biological age on various cellular processes. Both of these findings concur with most current hypotheses on aging, that in general consensus attribute the progression to a plethora of individual detrimental alterations driving the overall process, and that predict no single master switch for aging.

Nonetheless however, the pathway analyses showed subtle differences in the ranking of different processes. Interestingly, while very different analytical approaches were chosen to assess the impact of aging on the skin in the two aforementioned publications and different data sets were used, some significant overlap in the top most strongly age-associated biological pathways was observed across the studies:

Most strikingly, these concerned the ranking of p53- and TNFa-NFkB-signaling as the overall most predictive processes for both aging phase identity and the most highly ranked pathways in the neural age clock. Both processes have individually been described in the context of aging before. p53 has been hypothesized to be an important regulatory factor in aging, based on its crucial role in DNA damage repair processes as well as its involvement in the oxidative stress response[133], which simulations in paper 2 in turn identified as the major driver of photoaging, in line with general literary consensus[134]. Similarly, TNFa-NFkB-signaling has been implicated in aging as well, prominently as one of the pathways driving inflammaging, the chronic low-grade pro-inflammatory state observed in aging tissues[135–137].

Interestingly, a cooperation of NFkB and p53 is also well-described as a component in one particular Hallmark of Aging – cellular senescence. p53 is activated in response to numerous external stressors and is capable of stopping cellular proliferation, activating DNA repair machineries and driving cells into either senescence or apoptosis – depending on type, intensity and persistence of the endured stress[71–73,125,127]. Activation of p53 has been shown to orchestrate the induction of senescence together with the $p16^{INK4a}$ pathway, inducing permanent growth arrest after sustained stress, such as DNA damage following irradiation or exposure to oncogenic stimuli[138–140]. Notably, components of $p16^{INK4a}$ signaling, including p16 (CDKN2A) itself, were also found among the most important features for the prediction of transcriptomic age of the skin, as identified by the neural age clock in paper 2.

p53 was also found among the processes shifted towards a rejuvenated state in the simulation experiments performed in paper 2 using gene signatures of caloric restriction, the most well-documented and reliable approach to increase life-span across numerous model organisms[11,14,83,84,141]. This observed shift might be connected to the process of cellular senescence here as well, as caloric restriction has been shown to reliably protect from an accumulation of senescent cells[142–144]. This effect is hypothesized to be facilitated through decreases in oxidative stress levels[11,13], which would again be very much in line with the other pathways found modulated by caloric restriction in the simulation experiments, which also strongly pointed to an involvement of oxidative stress relief as the main mechanism driving the effect of caloric restriction, which is supported by findings from model animals[11,13,14,84,141,144–147].

Like p53, TNFa and the downstream NFkB-signaling pathway have been shown to be deeply involved with cellular senescence. Activation of NFkB-signaling by the pro-inflammatory cytokine TFNa promotes senescence across different cell types by numerous accounts[139,148–150], whereas inhibition of NFkB-signaling could be shown to have inverse effects and delay the onset of senescence as well as other age-related pathologies, at least in model animals[151].

As such, both TNFa-NFkB and p53 not only serve as markers for senescence but play important roles in the establishment of the phenomenon. Crosstalk between the two pathways has been observed in the establishment of cellular senescence before[138–140], as well as in other processes, such as the regulation of pro-inflammatory gene programs[152] or p53-mediated programmed cell death[138]. While p53, together p16^{INK4a} and further cyclin-dependent kinase inhibitors, are mainly implicated in the early stages of senescence, such as the first initiation of cell cycle arrest and transition to a persistent senescence-like state, TNFa- and NFkB-signaling have been identified as a major regulator for an important later stage of senescence, namely the development of the senescence-associated secretory phenotype (SASP)[139]. This phenotype, which describes fully senescent, growth-arrested cells secreting a complex mixture of proinflammatory cytokines, growth factors and proteases, is the main culprit of senescence in aging tissue, as it negatively impacts the surrounding microenvironment and drives tissue dysfunction by altering intercellular communication. Effects range from the disruption of local stem cell niches to alterations in tissue organization through secreted proteases and promotion of a chronic pro-inflammatory milieu, all of which interferes with normal tissue function[35,38]. NFkB has been shown to be one of the master regulators of SASP induction[139], cooperating with p53 in promotion of these detrimental effects of senescence observed in aging tissues. The joint occurrence of the two pathways across the studies and analyses performed thus appears far from coincidental, but instead points to a strong involvement of cellular senescence in the biological aging of the skin.

Another common theme across the studies and analyses conducted and findings made, concerns the strong impact solar irradiation as a driver of accelerated aging of the skin. In paper 1, analyses of meta data showed a significant impact of sun bathing on biological aging phase status, and simulation experiments performed using the neural age clock in paper 2 similarly showed a strong impact on transcriptomic age by a gene signature of chronic sun exposure. These findings are unsurprising considering the wealth of studies showing how sun exposure can accelerate the aging of the skin phenotypically, resulting in increased wrinkling, loss of elasticity, dyspigmentations and an overall aged, leathery appearance[44].

Photoaging by chronic sun exposure is widely hypothesized to mediate its effects through oxidative stress from UV-induced formation of reactive oxygen species[134], which was supported by the simulation experiments using the neural age clock in paper 2. UV irradiation is also known to directly and indirectly cause damage to the DNA in this manner, which as discussed marks one of the strongest triggers driving cells into a state of senescence.

Given the exposed nature of the skin and these well-established links between chronic sun exposure, oxidative stress, DNA damage and senescence, a strong accumulation of senescent cells in the aging skin presents a plausible consequence of chronic sun exposure. Indeed, the tissue dysfunction seen in photoaged skin is believed to be at least partly mediated by UV-induced senescent cells[38], and this hypothesis is also supported by experimental findings showing accumulations of senescent cells in sun exposed skin[153,154] and the upregulation of typical senescence markers such as p16^{INK4a} following irradiation[155,156]. Interestingly in this context, the analyses exploring the biological responses to repetitive UV irradiation in subjects with varying innate UV resilience in paper 3 identified divergent responses related to p53-signaling and DNA damage recognition pathways between the groupings, which might be directly connected to UV-resilience and the lower rates of skin cancer observed in more UV-tolerant individuals, which are not explained by differences in melanin alone[45,102–105]. A deeper mechanistic understanding of the identified differences might potentially be exploited to develop novel approaches reducing the risk of carcinogenesis by reproducing the apparent protective biological conditions in normally cancer-prone skin by targeted intervention strategies.

As these particular pathways are intricately linked to aging and the emergence of senescence as well, the findings would also invite explorations of the identified differences concerning UV-tolerance in relation to biological aging rates as well, particularly in chronically sun exposed skin, which shares several features with aged skin even beyond the already mentioned phenotypical similarities, including delayed wound healing[157,158] and importantly an increased susceptibility to cancer[45,159]. Recent findings from model animal experiments, which showed how increased DNA damage in the form of double-strand breaks alone is capable of driving an aging phenotype through gradual erosion of the epigenetic landscape and subsequent loss of epigenetic information, further solidify this link[160].

Overall then, the many overlaps among genes, pathways and biological processes identified across the studies and discussed in this work illustrate how crucially aging, extrinsic effectors such as solar irradiation and severe diseases such as skin cancer are connected, underlining the aforementioned importance of understanding biological aging on a molecular level to derive intervention strategies capable of successfully slowing the onset of age-related disease and extending human health span in the future.

References

1. World Health Organization (WHO). *Preventing Chronic Diseases: A Vital Investment*. (2005).
2. Christensen, K., Doblhammer, G., Rau, R. & Vaupel, J. W. Ageing populations: the challenges ahead. *The Lancet* **374**, 1196–1208 (2009).
3. Fontana, L., Kennedy, B. K., Longo, V. D., Seals, D. & Melov, S. Medical research: Treat ageing. *Nature* **511**, 405–407 (2014).
4. Chang, A. Y., Skirbekk, V. F., Tyrovolas, S., Kassebaum, N. J. & Dieleman, J. L. Measuring population ageing: an analysis of the Global Burden of Disease Study 2017. *The Lancet Public Health* **4**, e159–e167 (2019).
5. Atella, V. *et al.* Trends in age-related disease burden and healthcare utilization. *Aging Cell* **18**, e12861 (2019).
6. Harman, D. Aging: A Theory Based on Free Radical and Radiation Chemistry. *Journal of Gerontology* **11**, 298–300 (1956).
7. Kirkwood, T. B. L. Evolution of ageing. *Nature* **270**, 301–304 (1977).
8. Kirkwood, T. B. L. & Austad, S. N. Why do we age? *Nature* **408**, 233–238 (2000).
9. Karnaukhov, A. V. & Karnaukhova, E. V. Informational hypothesis of aging: How does the germ line “avoid” aging? *BIOPHYSICS* **54**, 531 (2009).
10. Gladyshev, V. N. Aging: progressive decline in fitness due to the rising deleteriome adjusted by genetic, environmental, and stochastic processes. *Aging Cell* **15**, 594–602 (2016).
11. Sohal, R. S. & Weindruch, R. Oxidative Stress, Caloric Restriction, and Aging. *Science* **273**, 59–63 (1996).
12. Krutmann, J., Bouloc, A., Sore, G., Bernard, B. A. & Passeron, T. The skin aging exposome. *Journal of Dermatological Science* **85**, 152–161 (2017).
13. Redman, L. M. *et al.* Metabolic Slowing and Reduced Oxidative Damage with Sustained Caloric Restriction Support the Rate of Living and Oxidative Damage Theories of Aging. *Cell Metabolism* **27**, 805-815.e4 (2018).
14. Colman, R. J. *et al.* Caloric restriction delays disease onset and mortality in rhesus monkeys. *Science* **325**, 201–204 (2009).
15. Bocklandt, S. *et al.* Epigenetic predictor of age. *PLoS ONE* **6**, e14821 (2011).
16. Koch, C. M. & Wagner, W. Epigenetic-aging-signature to determine age in different tissues. *Aging (Albany NY)* **3**, 1018–1027 (2011).
17. Hannum, G. *et al.* Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Molecular Cell* **49**, 359–367 (2013).

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

19. Marioni, R. E. *et al.* DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biology* **16**, 25 (2015).

20. Hertel, J. *et al.* Measuring Biological Age via Metabonomics: The Metabolic Age Score. *J. Proteome Res.* **15**, 400–410 (2016).

21. Rist, M. J. *et al.* Metabolite patterns predicting sex and age in participants of the Karlsruhe Metabolomics and Nutrition (KarMeN) study. *PLoS ONE* **12**, e0183228 (2017).

22. Peters, M. J. *et al.* The transcriptional landscape of age in human peripheral blood. *Nat Commun* **6**, 1–14 (2015).

23. Mamoshina, P. *et al.* Machine Learning on Human Muscle Transcriptomic Data for Biomarker Discovery and Tissue-Specific Drug Target Identification. *Front Genet* **9**, 242 (2018).

24. Fleischer, J. G. *et al.* Predicting age from the transcriptome of human dermal fibroblasts. *Genome Biology* **19**, 221 (2018).

25. Chen, B. H. *et al.* DNA methylation-based measures of biological age: meta-analysis predicting time to death. *Aging* **8**, 1844–1865 (2016).

26. Quach, A. *et al.* Epigenetic clock analysis of diet, exercise, education, and lifestyle factors. *Aging* **9**, 419–446 (2017).

27. Levine, M. E. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging* **10**, 573–591 (2018).

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

29. Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany NY)* **11**, 303–327 (2019).

30. Tricoire, H. & Rera, M. A New, Discontinuous 2 Phases of Aging Model: Lessons from Drosophila melanogaster. *PLOS ONE* **10**, e0141920 (2015).

31. Rana, A. *et al.* Promoting Drp1-mediated mitochondrial fission in midlife prolongs healthy lifespan of Drosophila melanogaster. *Nature Communications* **8**, 448 (2017).

32. Timmons, J. A. *et al.* Longevity-related molecular pathways are subject to midlife “switch” in humans. *Aging Cell* **18**, (2019).

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

34. Coppé, J.-P., Desprez, P.-Y., Krtolica, A. & Campisi, J. The Senescence-Associated Secretory Phenotype: The Dark Side of Tumor Suppression. *Annu Rev Pathol* **5**, 99–118 (2010).

35. van Deursen, J. M. The role of senescent cells in ageing. *Nature* **509**, 439–446 (2014).

36. Cavinato, M. & Jansen-Dürr, P. Molecular mechanisms of UVB-induced senescence of dermal fibroblasts and its relevance for photoaging of the human skin. *Experimental Gerontology* **94**, 78–82 (2017).

37. Wang, A. S. & Dreesen, O. Biomarkers of Cellular Senescence and Skin Aging. *Frontiers in Genetics* **9**, 247 (2018).

38. Fitsiou, E., Pulido, T., Campisi, J., Alimirah, F. & Demaria, M. Cellular Senescence and the Senescence-Associated Secretory Phenotype as Drivers of Skin Photoaging. *Journal of Investigative Dermatology* **141**, 1119–1126 (2021).

39. Bartek, J., Hodny, Z. & Lukas, J. Cytokine loops driving senescence. *Nat Cell Biol* **10**, 887–889 (2008).

40. Malavolta, M. *et al.* Inducers of Senescence, Toxic Compounds, and Senolytics: The Multiple Faces of Nrf2-Activating Phytochemicals in Cancer Adjuvant Therapy. *Mediators of Inflammation* vol. 2018 e4159013 https://www.hindawi.com/journals/mi/2018/4159013/ (2018).

41. Ho, C. Y. & Dreesen, O. Faces of cellular senescence in skin aging. *Mechanisms of Ageing and Development* **198**, 111525 (2021).

42. Scharffetter-Kochanek, K. *et al.* UV-induced reactive oxygen species in photocarcinogenesis and photoaging. *Biol. Chem.* **378**, 1247–1257 (1997).

43. Rinnerthaler, M., Bischof, J., Streubel, M. K., Trost, A. & Richter, K. Oxidative Stress in Aging Human Skin. *Biomolecules* **5**, 545–589 (2015).

44. Scharffetter-Kochanek, K. *et al.* Photoaging of the skin from phenotype to mechanisms. *Exp. Gerontol.* **35**, 307–316 (2000).

45. Armstrong, B. K. & Kricker, A. The epidemiology of UV induced skin cancer. *Journal of photochemistry and photobiology. B, Biology* **63**, 8–18 (2001).

46. Röwert-Huber, J. *et al.* Actinic keratosis is an early in situ squamous cell carcinoma: a proposal for reclassification. *Br. J. Dermatol.* **156 Suppl 3**, 8–12 (2007).

47. Singh, A. *et al.* Ultraviolet radiation-induced tumor necrosis factor alpha, which is linked to the development of cutaneous SCC, modulates differential epidermal microRNAs expression. *Oncotarget* **7**, 17945–17956 (2016).

48. Aunan, J. R., Cho, W. C. & Søreide, K. The Biology of Aging and Cancer: A Brief Overview of Shared and Divergent Molecular Hallmarks. *Aging Dis* **8**, 628–642 (2017).

49. Meng, C., Kuster, B., Culhane, A. C. & Gholami, A. M. A multivariate approach to the integration of multi-omics datasets. *BMC bioinformatics* **15**, 162 (2014).

50. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods* **11**, 333–337 (2014).

51. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biology* **18**, 83 (2017).

52. Huang, S., Chaudhary, K. & Garmire, L. X. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Frontiers in Genetics* **8**, 84 (2017).

53. Francescatto, M. *et al.* Multi-omics integration for neuroblastoma clinical endpoint prediction. *Biology Direct* **13**, 5 (2018).

54. Argelaguet, R. *et al.* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology* **14**, e8124 (2018).

55. Bersanelli, M. *et al.* Methods for the integration of multi-omics data: mathematical aspects. *BMC bioinformatics* **17 Suppl 2**, 15 (2016).

56. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).

57. Cortes, C. & Vapnik, V. Support-vector networks. *Mach Learn* **20**, 273–297 (1995).

58. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* **5**, 115–133 (1943).

59. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* **65**, 386–408 (1958).

60. Hebb, D. O. *The Organization of Behavior: A Neuropsychological Theory*. (Psychology Press, 2005).

61. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks* **61**, 85–117 (2015).

62. Stanford, J. L., Hartge, P., Brinton, L. A., Hoover, R. N. & Brookmeyer, R. Factors influencing the age at natural menopause. *Journal of Chronic Diseases* **40**, 995–1002 (1987).

63. McKinlay, S. M., Brambilla, D. J. & Posner, J. G. The normal menopause transition. *Maturitas* **14**, 103–15 (1992).

64. Ossewaarde, M. E. *et al.* Age at Menopause, Cause-Specific Mortality and Total Life Expectancy. *Epidemiology* **16**, 556–562 (2005).

65. Gold, E. B. The Timing of the Age at Which Natural Menopause Occurs. *Obstetrics and Gynecology Clinics of North America* **38**, 425 (2011).

66. Martinez-Jimenez, C. P. *et al.* Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science* **355**, 1433–1436 (2017).

67. Enge, M. *et al.* Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. *Cell* **171**, 321-330.e14 (2017).

68. Angelidis, I. *et al.* An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nat Commun* **10**, 963 (2019).

69. Sherr, C. J. The INK4a/ARF network in tumour suppression. *Nat Rev Mol Cell Biol* **2**, 731–737 (2001).

70. Baker, D. J., Jin, F. & van Deursen, J. M. The yin and yang of the Cdkn2a locus in senescence and aging. *Cell Cycle* **7**, 2795–2802 (2008).

71. Rufini, A., Tucci, P., Celardo, I. & Melino, G. Senescence and aging: the critical roles of p53. *Oncogene* **32**, 5129–5143 (2013).

72. Qian, Y. & Chen, X. Senescence Regulation by the p53 Protein Family. *Methods Mol Biol* **965**, 37–61 (2013).

73. Mijit, M., Caracciolo, V., Melillo, A., Amicarelli, F. & Giordano, A. Role of p53 in the Regulation of Cellular Senescence. *Biomolecules* **10**, 420 (2020).

74. Azazmeh, N. *et al.* Chronic expression of p16 INK4a in the epidermis induces Wnt-mediated hyperplasia and promotes tumor initiation. *Nature Communications* **11**, 2711 (2020).

75. Völzke, H. *et al.* Cohort profile: the study of health in Pomerania. *Int J Epidemiol* **40**, 294–307 (2011).

76. Bormann, F. *et al.* Reduced DNA methylation patterning and transcriptional connectivity define human skin aging. *Aging cell* **15**, 563–71 (2016).

77. Holzscheck, N. *et al.* Multi-omics network analysis reveals distinct stages in the human aging progression in epidermal tissue. *Aging (Albany NY)* **12**, (2020).

78. Khan, S. S. *et al.* A null mutation in SERPINE1 protects against biological aging in humans. *Science Advances* **3**, eaao1617 (2017).

79. Jiang, C. *et al.* Serpine 1 induces alveolar type II cell senescence through activating p53-p21-Rb pathway in fibrotic lung disease. *Aging Cell* **16**, 1114–1124 (2017).

80. Baege, A. C., Disbrow, G. L. & Schlegel, R. IGFBP-3, a Marker of Cellular Senescence, Is Overexpressed in Human Papillomavirus-Immortalized Cervical Cells and Enhances IGF-1-Induced Mitogenesis. *J Virol* **78**, 5720–5727 (2004).

81. Hong, S. & Kim, M.-M. IGFBP-3 plays an important role in senescence as an aging marker. *Environ Toxicol Pharmacol* **59**, 138–145 (2018).

82. Guccini, I. *et al.* Senescence Reprogramming by TIMP1 Deficiency Promotes Prostate Cancer Metastasis. *Cancer Cell* **39**, 68-82.e9 (2021).

83. Weindruch, R. & Walford, R. L. Dietary restriction in mice beginning at 1 year of age: effect on life-span and spontaneous cancer incidence. *Science* **215**, 1415–1418 (1982).

84. Mattison, J. A. *et al.* Impact of caloric restriction on health and survival in rhesus monkeys: the NIA study. *Nature* **489**, (2012).

85. Wondrak, G. T., Jacobson, M. K. & Jacobson, E. L. Endogenous UVA-photosensitizers: mediators of skin photodamage and novel targets for skin photoprotection. *Photochem. Photobiol. Sci.* **5**, 215–237 (2006).

86. Glogau, R. G. The risk of progression to invasive disease. *J. Am. Acad. Dermatol.* **42**, 23–24 (2000).

87. Protić-Sabljić, M. *et al.* UV light-induced cyclobutane pyrimidine dimers are mutagenic in mammalian cells. *Molecular and cellular biology* **6**, 3349–56 (1986).

88. Drouin, R. & Therrien, J.-P. UVB-induced Cyclobutane Pyrimidine Dimer Frequency Correlates with Skin Cancer Mutational Hotspots in p53. *Photochemistry and Photobiology* **66**, 719–726 (1997).

89. Hill, H. Z. & Hill, G. J. UVA, pheomelanin and the carcinogenesis of melanoma. *Pigment Cell Res.* **13 Suppl 8**, 140–144 (2000).

90. D'Orazio, J., Jarrett, S., Amaro-Ortiz, A. & Scott, T. UV radiation and the skin. *International journal of molecular sciences* **14**, 12222–48 (2013).

91. Brash, D. E. UV signature mutations. *Photochemistry and photobiology* **91**, 15–26 (2015).

92. Grönniger, E. *et al.* Aging and Chronic Sun Exposure Cause Distinct Epigenetic Changes in Human Skin. **6**, e1000971 (2010).

93. Easwaran, H., Tsai, H.-C. & Baylin, S. B. B. Cancer epigenetics: tumor heterogeneity, plasticity of stem-like states, and drug resistance. **54**, (2014).

94. Greenberg, E. S., Chong, K. K., Huynh, K. T., Tanaka, R. & Hoon, D. S. B. Epigenetic biomarkers in skin cancer. *Cancer letters* **342**, 170–7 (2014).

95. Vandiver, A. R. *et al.* Age and sun exposure-related widespread genomic blocks of hypomethylation in nonmalignant skin. *Genome biology* **16**, 80 (2015).

96. Rodríguez-Paredes, M. *et al.* Methylation profiling identifies two subclasses of squamous cell carcinoma related to distinct cells of origin. *Nature Communications* **9**, 577 (2018).

97. Kennedy, C. *et al.* The influence of painful sunburns and lifetime sun exposure on the risk of actinic keratoses, seborrheic warts, melanocytic nevi, atypical nevi, and skin cancer. *J Invest Dermatol* **120**, 1087–1093 (2003).

98. Gandini, S. *et al.* Meta-analysis of risk factors for cutaneous melanoma: II. Sun exposure. *Eur J Cancer* **41**, 45–60 (2005).

99. Dennis, L. K. *et al.* Sunburns and risk of cutaneous melanoma, does age matter: a comprehensive meta-analysis. *Ann Epidemiol* **18**, 614–627 (2008).

100. Wu, S., Han, J., Laden, F. & Qureshi, A. A. Long-term Ultraviolet Flux, Other Potential Risk Factors, and Skin Cancer Risk: A Cohort Study. *Cancer Epidemiology Biomarkers & Prevention* **23**, 1080–1089 (2014).

101. Wu, S. *et al.* History of Severe Sunburn and Risk of Skin Cancer Among Women and Men in 2 Prospective Cohort Studies. *American Journal of Epidemiology* **183**, 824–833 (2016).

102. Kaidbey, K. H., Agin, P. P., Sayre, R. M. & Kligman, A. M. Photoprotection by melanin—a comparison of black and Caucasian skin. *Journal of the American Academy of Dermatology* **1**, 249–260 (1979).

103. Halder, R. M. & Bang, K. M. Skin cancer in blacks in the United States. *Dermatologic clinics* **6**, 397–405 (1988).

104. Cress, R. D. & Holly, E. A. Incidence of cutaneous melanoma among non-Hispanic whites, Hispanics, Asians, and blacks: an analysis of california cancer registry data, 1988-93. *Cancer causes & control : CCC* **8**, 246–52 (1997).

105. National Cancer Institute (NCI). *SEER Cancer Statistics Review (CSR) 1975-2014*. https://seer.cancer.gov/archive/csr/1975_2014/ (2018).

106. Fitzpatrick, T. B. Soleil et peau. *Journal de Médecine Esthétique* 33–34 (1975).

107. Rampen, F. H. J., Fleuren, B. A. M., de Boo, T. M. & Lemmens, W. A. J. G. Unreliability of Self-reported Burning Tendency and Tanning Ability. *Archives of Dermatology* **124**, 885 (1988).

108. Ravnbak, M. H. Objective determination of Fitzpatrick skin type. *Danish medical bulletin* **57**, B4153 (2010).

109. Wulf, H. C., Philipsen, P. A. & Ravnbak, M. H. Minimal erythema dose and minimal melanogenesis dose relate better to objectively measured skin type than to Fitzpatricks skin type. *Photodermatology, Photoimmunology & Photomedicine* **26**, 280–284 (2010).

110. Harrison, G. I., Young, A. R. & McMahon, S. B. Ultraviolet radiation-induced inflammation as a model for cutaneous hyperalgesia. *The Journal of investigative dermatology* **122**, 183–9 (2004).

111. International Organization for Standardization. *DIN EN ISO 24444 – In vivo determination of the sun protection factor (SPF)*. (2010).

112. Pfeifer, G. P. & Besaratinia, A. UV wavelength-dependent DNA damage and human non-melanoma and melanoma skin cancer. *Photochemical & photobiological sciences : Official journal of the European Photochemistry Association and the European Society for Photobiology* **11**, 90–7 (2012).
113. Aging Atlas: a multi-omics database for aging biology. *Nucleic Acids Res* **49**, D825–D830 (2020).
114. Xue, X. *et al.* Tumor Necrosis Factor α (TNFα) Induces the Unfolded Protein Response (UPR) in a Reactive Oxygen Species (ROS)-dependent Fashion, and the UPR Counteracts ROS Accumulation by TNFα *. *Journal of Biological Chemistry* **280**, 33917–33925 (2005).
115. Hu, P., Han, Z., Couvillon, A. D., Kaufman, R. J. & Exton, J. H. Autocrine tumor necrosis factor alpha links endoplasmic reticulum stress to the membrane death receptor pathway through IRE1alpha-mediated NF-kappaB activation and down-regulation of TRAF2 expression. *Mol Cell Biol* **26**, 3071–3084 (2006).
116. Salminen, A., Kaarniranta, K. & Kauppinen, A. Inflammaging: disturbed interplay between autophagy and inflammasomes. *Aging* **4**, 166–175 (2012).
117. Shim, S. M. *et al.* Role of S5b/PSMD5 in Proteasome Inhibition Caused by TNF-α/NFκB in Higher Eukaryotes. *Cell Reports* **2**, 603–615 (2012).
118. Zheng, L. *et al.* Role of autophagy in tumor necrosis factor-α-induced apoptosis of osteoblast cells. *J Investig Med* **65**, 1014–1020 (2017).
119. Tyciakova, S., Valova, V., Svitkova, B. & Matuskova, M. Overexpression of TNFα induces senescence, autophagy and mitochondrial dysfunctions in melanoma cells. *BMC Cancer* **21**, 507 (2021).
120. Tam, A. B., Mercado, E. L., Hoffmann, A. & Niwa, M. ER Stress Activates NF-κB by Integrating Functions of Basal IKK Activity, IRE1 and PERK. *PLoS One* **7**, e45078 (2012).
121. Schmitz, M. L., Shaban, M. S., Albert, B. V., Gökçen, A. & Kracht, M. The Crosstalk of Endoplasmic Reticulum (ER) Stress Pathways with NF-κB: Complex Mechanisms Relevant for Cancer, Inflammation and Infection. *Biomedicines* **6**, 58 (2018).
122. Chipurupalli, S., Samavedam, U. & Robinson, N. Crosstalk Between ER Stress, Autophagy and Inflammation. *Frontiers in Medicine* **8**, (2021).
123. Levine, M. E. *et al.* Menopause accelerates biological aging. *Proceedings of the National Academy of Sciences* **113**, 9327–9332 (2016).
124. Thornton, M. J. Estrogens and aging skin. *Dermato-Endocrinology* **5**, 264–270 (2013).
125. Kuerbitz, S. J., Plunkett, B. S., Walsh, W. V. & Kastan, M. B. Wild-type p53 is a cell cycle checkpoint determinant following irradiation. *Proc Natl Acad Sci U S A* **89**, 7491–7495 (1992).

126. Lane, D. P. p53, guardian of the genome. *Nature* **358**, 15–16 (1992).

127. Jiang, W., Ananthaswamy, H. N., Muller, H. K. & Kripke, M. L. p53 protects against skin cancer induction by UV-B radiation. *Oncogene* **18**, 4247–4253 (1999).

128. Soussi, T. The history of p53. *EMBO Rep* **11**, 822–826 (2010).

129. Toufektchan, E. & Toledo, F. The Guardian of the Genome Revisited: p53 Downregulates Genes Required for Telomere Maintenance, DNA Repair, and Centromere Structure. *Cancers (Basel)* **10**, 135 (2018).

130. Tyner, S. D. *et al.* p53 mutant mice that display early ageing-associated phenotypes. *Nature* **415**, 45–53 (2002).

131. Maier, B. *et al.* Modulation of mammalian life span by the short isoform of p53. *Genes Dev* **18**, 306–319 (2004).

132. Feng, Z. *et al.* Declining p53 function in the aging process: A possible mechanism for the increased tumor incidence in older populations. *PNAS* **104**, 16633–16638 (2007).

133. Liu, D. & Xu, Y. p53, Oxidative Stress, and Aging. *Antioxid Redox Signal* **15**, 1669–1678 (2011).

134. Miyachi, Y. Photoaging from an oxidative standpoint. *Journal of Dermatological Science* **9**, 79–86 (1995).

135. Franceschi, C. *et al.* Inflamm-aging. An evolutionary perspective on immunosenescence. *Annals of the New York Academy of Sciences* **908**, 244–54 (2000).

136. De Martinis, M., Franceschi, C., Monti, D. & Ginaldi, L. Inflamm-ageing and lifelong antigenic load as major determinants of ageing rate and longevity. *FEBS Lett* **579**, 2035–2039 (2005).

137. Zhuang, Y. & Lyga, J. Inflammaging in skin and other tissues - the roles of complement system and macrophage. *Inflammation & allergy drug targets* **13**, 153–161 (2014).

138. Ryan, K. M., Ernst, M. K., Rice, N. R. & Vousden, K. H. Role of NF-kappaB in p53-mediated programmed cell death. *Nature* **404**, 892–897 (2000).

139. Chien, Y. *et al.* Control of the senescence-associated secretory phenotype by NF-κB promotes senescence and enhances chemosensitivity. *Genes Dev* **25**, 2125–2136 (2011).

140. Iannetti, A. *et al.* Regulation of p53 and Rb Links the Alternative NF-κB Pathway to EZH2 Expression and Cell Senescence. *PLOS Genetics* **10**, e1004642 (2014).

141. Colman, R. J. *et al.* Caloric restriction reduces age-related and all-cause mortality in rhesus monkeys. *Nat Commun* **5**, 3557 (2014).

142. Wang, C. *et al.* Adult-onset, short-term dietary restriction reduces cell senescence in mice. *Aging (Albany NY)* **2**, 555–566 (2010).

143. Fontana, L. *et al.* The effects of graded caloric restriction: XII. Comparison of mouse to human impact on cellular senescence in the colon. *Aging Cell* **17**, e12746 (2018).

144. Fontana, L., Nehme, J. & Demaria, M. Caloric restriction and cellular senescence. *Mechanisms of Ageing and Development* **176**, 19–23 (2018).

145. Youngman, L. D., Park, J. Y. & Ames, B. N. Protein oxidation associated with aging is reduced by dietary restriction of protein or calories. *Proceedings of the National Academy of Sciences* **89**, 9112–9116 (1992).

146. Sohal, R. S., Ku, H. H., Agarwal, S., Forster, M. J. & Lal, H. Oxidative damage, mitochondrial oxidant generation and antioxidant defenses during aging and in response to food restriction in the mouse. *Mech Ageing Dev* **74**, 121–133 (1994).

147. Walsh, M. E., Shi, Y. & Van Remmen, H. The effects of dietary restriction on oxidative stress in rodents. *Free Radic Biol Med* **66**, 88–99 (2014).

148. Rovillain, E. *et al.* Activation of Nuclear Factor-kappa B signalling promotes cellular senescence. *Oncogene* **30**, 2356–2366 (2011).

149. Vaughan, S. & Jat, P. S. Deciphering the role of Nuclear Factor-κB in cellular senescence. *Aging (Albany NY)* **3**, 913–919 (2011).

150. Kandhaya-Pillai, R. *et al.* TNFα-senescence initiates a STAT-dependent positive feedback loop, leading to a sustained interferon signature, DNA damage, and cytokine secretion. *Aging (Albany NY)* **9**, 2411–2435 (2017).

151. Tilstra, J. S. *et al.* NF-κB inhibition delays DNA damage-induced senescence and aging in mice. *J Clin Invest* **122**, 2601–2612 (2012).

152. Lowe, J. M. *et al.* p53 and NF-κB Co-regulate Pro-inflammatory Gene Responses in Human Macrophages. *Cancer Res* **74**, 2182–2192 (2014).

153. Waaijer, M. E. C. *et al.* P16INK4a Positive Cells in Human Skin Are Indicative of Local Elastic Fiber Morphology, Facial Wrinkling, and Perceived Age. *J Gerontol A Biol Sci Med Sci* **71**, 1022–1028 (2016).

154. Wang, A. S., Ong, P. F., Chojnowski, A., Clavel, C. & Dreesen, O. Loss of lamin B1 is a biomarker to quantify cellular senescence in photoaged skin. *Sci Rep* **7**, 15678 (2017).

155. Pavey, S., Conroy, S., Russell, T. & Gabrielli, B. Ultraviolet radiation induces p16CDKN2A expression in human skin. *Cancer Res* **59**, 4185–4189 (1999).

156. Ahmed, N. U., Ueda, M. & Ichihashi, M. Induced expression of p16 and p21 proteins in UVB-irradiated human epidermis and cultured keratinocytes. *J Dermatol Sci* **19**, 175–181 (1999).

157. Davidson, S. F., Brantley, S. K. & Das, S. K. The effects of ultraviolet radiation on wound healing. *Br J Plast Surg* **44**, 210–214 (1991).
158. Gerstein, A. D., Phillips, T. J., Rogers, G. S. & Gilchrest, B. A. Wound healing and aging. *Dermatol Clin* **11**, 749–757 (1993).
159. White, M. C. *et al.* Age and Cancer Risk. *Am J Prev Med* **46**, S7-15 (2014).
160. Yang, J.-H. *et al.* Loss of epigenetic information as a cause of mammalian aging. *Cell* **186**, 305-326.e27 (2023).

Publication abstracts and contributions

Abstract paper 1:

In recent years, reports of non-linear regulations in age- and longevity-associated biological processes have been accumulating. Inspired by methodological advances in precision medicine involving the integrative analysis of multi-omics data, we sought to investigate the potential of multi-omics integration to identify distinct stages in the aging progression from ex vivo human skin tissue. For this we generated transcriptome and methylome profiling data from suction blister lesions of female subjects between 21 and 76 years, which were integrated using a network fusion approach. Unsupervised cluster analysis on the combined network identified four distinct subgroupings exhibiting a significant age-association. As indicated by DNAm age analysis and Hallmark of Aging enrichment signals, the stages captured the biological aging state more clearly than a mere grouping by chronological age and could further be recovered in a longitudinal validation cohort with high stability. Characterization of the biological processes driving the phases using machine learning enabled a data-driven reconstruction of the order of Hallmark of Aging manifestation. Finally, we investigated non-linearities in the mid-life aging progression captured by the aging phases and identified a far-reaching non-linear increase in transcriptional noise in the pathway landscape in the transition from mid- to late-life.

Contributions to the publication: Raw data processing, contribution to the design of the analysis approach, computational work including data analysis and interpretation, figure design and writing of the original manuscript.

Abstract paper 2:

The development of ‘age clocks’, machine learning models predicting age from biological data, has been a major milestone in the search for reliable markers of biological age and has since become an invaluable tool in aging research. However, beyond their unquestionable utility, current

generation clocks offer little insight into the molecular biological processes driving aging, and their inner workings often remain non-transparent. Here we propose a new type of age clock, one that couples predictivity with interpretability of the underlying biology, achieved through the incorporation of prior knowledge into the model design. The clock, an artificial neural network constructed according to well-described biological pathways, allows the prediction of age from gene expression data of skin tissue with high accuracy, while at the same time capturing and revealing aging states of the pathways driving the prediction. The model recapitulates known associations of aging gene knockdowns in simulation experiments and demonstrates its utility in deciphering the main pathways by which accelerated aging conditions such as Hutchinson Gilford progeria syndrome, as well as pro longevity interventions like caloric restriction, exert their effects.

Contributions to the publication: Raw data processing, design and construction of the neural network, computational work including data analysis and interpretation, figure design and writing of the original manuscript.

Abstract paper 3:

The simultaneous analysis of different regulatory levels of biological phenomena by means of multi-omics data integration has proven an invaluable tool in modern precision medicine, yet many processes ultimately paving the way towards disease manifestation remain elusive and have not been studied in this regard. Here we investigated the early molecular events following repetitive UV irradiation of in vivo healthy human skin in depth on transcriptomic and epigenetic level. Our results provide first hints towards an immediate acquisition of epigenetic memories related to aging and cancer and demonstrate significantly correlated epigenetic and transcriptomic responses to irradiation stress. The data allowed the precise prediction of inter-individual UV sensitivity, and molecular subtyping on the integrated post-irradiation multi-omics data established the existence of three latent molecular phototypes. Importantly, further analysis suggested a form of melanin-independent DNA damage protection in subjects with higher innate UV resilience. This work establishes a high-resolution molecular landscape of the acute epidermal UV response and demonstrates the potential of integrative analyses to untangle complex and heterogeneous biological responses.

Contributions to the publication: Raw data processing, contribution to the design of the analysis approach, computational work including data analysis and interpretation, figure design and writing of the original manuscript.

Publications

Research Paper

Multi-omics network analysis reveals distinct stages in the human aging progression in epidermal tissue

Nicholas Holzscheck[1,2], Jörn Söhle[1], Boris Kristof[1], Elke Grönniger[1], Stefan Gallinat[1], Horst Wenck[1], Marc Winnefeld[1], Cassandra Falckenhayn[1,*], Lars Kaderali[2,*]

[1]Front End Innovation, Beiersdorf AG, Hamburg, Germany
[2]Institute for Bioinformatics, University Medicine Greifswald, Greifswald, Germany
*Co-last authors

Correspondence to: Nicholas Holzscheck; **email:** nicholas.holzscheck@beiersdorf.com

Keywords: multi-omics, biological age, aging phases, hallmarks of aging, transcriptional noise
Abbreviations: AUC: area under the curve (receiver operating characteristic curve); HoA: Hallmark(s) of Aging; SNF: similarity network fusion
Received: February 20, 2020 **Accepted:** June 1, 2020 **Published:** June 18, 2020

ABSTRACT

In recent years, reports of non-linear regulations in age- and longevity-associated biological processes have been accumulating. Inspired by methodological advances in precision medicine involving the integrative analysis of multi-omics data, we sought to investigate the potential of multi-omics integration to identify distinct stages in the aging progression from *ex vivo* human skin tissue. For this we generated transcriptome and methylome profiling data from suction blister lesions of female subjects between 21 and 76 years, which were integrated using a network fusion approach. Unsupervised cluster analysis on the combined network identified four distinct subgroupings exhibiting a significant age-association. As indicated by DNAm age analysis and Hallmark of Aging enrichment signals, the stages captured the biological aging state more clearly than a mere grouping by chronological age and could further be recovered in a longitudinal validation cohort with high stability. Characterization of the biological processes driving the phases using machine learning enabled a data-driven reconstruction of the order of Hallmark of Aging manifestation. Finally, we investigated non-linearities in the mid-life aging progression captured by the aging phases and identified a far-reaching non-linear increase in transcriptional noise in the pathway landscape in the transition from mid- to late-life.

INTRODUCTION

Biological age represents the main risk factor for most chronic human pathologies, which is why therapies slowing the aging progression and postponing the onset of age-driven disease manifestation have frequently been suggested as major interventions to improve human health span. Chronological age has long been utilized as a proxy for biological aging state, in recent years however, the heterogeneity of biological aging rates for individuals of the same chronological age has become increasingly apparent. The most prominent example for this decoupling has probably been delivered in the wake of the discovery of the "epigenetic clock" in both mouse and human tissues [1–7], which revealed accelerated aging rates associated with various disease states and all-cause mortality [8–11], and is measured by DNA methylation state.

The notion of aging being a continuous process meanwhile remained. Lately though, this view has been questioned by reports on non-linearity and discontinuities in biological processes associated with aging and longevity. Early indications included the identification of two distinguishable phases in the aging progression of *Drosophila melanogaster* [12]. The transition to the second aging phase, marked by decreased motor activity and heightened inflammation, was accompanied by an exponentially increased mortality risk. Remarkably, this 2-phased model was able to reproduce a variety of experimental longevity curves [12]. More recently, evidence of the existence of non-linear switches, capable of extending model animal lifespan *in vivo*, has been presented concerning mitochondrial function, further implicating discontinuous biological processes in aging [13]. Not long ago now, the report of a mid-life switch involving a longevity-associated signaling pathway in aging human muscle and brain tissue was published [14]. Using gene and long non-coding RNA expression profiling, the authors observed that an age-related IGF-1/PI3K/mTOR-related RNA response signature was essentially lost with the start of the sixth decade of life. The report provides compelling evidence that discontinuous processes might be a previously overlooked feature of human aging as well and indicate that the progression of biological aging on a molecular level might be even more intricately regulated and complex than previously assumed.

The different biological processes driving aging meanwhile are manifold. The Hallmarks of Aging [15] provide a description of nine common denominators of aging in different tissues and organisms, attempting a categorization of various biological pathways into conceptual cornerstones of aging. Based on extensive literature review, the authors not only grouped, but also postulated the order of emergence for the different hallmarks. While the theoretical depiction of these hallmarks is detailed and comprehensive, a data-driven characterization of their importance to the aging phenotype and the actual disentanglement of the timely order of their occurrence in *in vivo* human tissue have remained elusive.

Recent years have seen a continuous decline in costs for genome-wide analyses, leading to an increasing feasibility of multi-omics profiling studies. Simultaneously monitoring multiple different omics levels in a living system holds great promises in generating a holistic understanding of phenotypical manifestations and might prove beneficial for aging research, as it has for the medical sciences. However, the integration of multi-omics data also brings tremendous novel statistical and computational challenges. These are related to the properties of many omics datasets, which include high dimensionality with often low sample counts, differing scales and distributions of measurements, as well as platform specific bias and technical noise [16]. In order to tackle these challenges and to uncover complementary information from multi-omics data, an increasing number of algorithmic approaches have been developed in the past years [17]. Network based methods such as similarity network fusion (SNF) offer an elegant solution to the problem, by transferring the feature-patient data for each dataset into featureless patient-patient space before their integration [18]. From every dataset a similarity network is created with patients represented as nodes and similarities between patients as edges. The separate networks are then integrated through an iterative fusion algorithm, which strengthens edges present in several data views, and finally converges into a fused network. This final network incorporates similarities from all omics data views and can be used for downstream analyses such as subtype identification through clustering.

In an effort to further explore the discontinuities in the aging progression using multi-omics methodology, we generated gene expression and DNA methylation data from *ex vivo* samples of aging skin. Skin represents an extraordinarily well-suited tissue for studying aging, owing to its well-documented aging phenotype and the ease of sampling using well-established non- or minimally invasive procedures. Using similarity network fusion, we integrated and clustered the multi-omics data to identify discrete stages along the aging progression. We validated the latent stages using DNA methylation age, the detection of Hallmark of Aging signals, and using a longitudinal validation cohort. Finally, we deployed machine learning to elucidate the order of Hallmark of Aging manifestation throughout the aging phases, and characterized the phases regarding pathway importance, which subsequently revealed a distinctly non-linear decrease in pathway enrichment at the mid- to late-life transition from aging phase 3 to phase 4.

RESULTS

Identification of latent age-associated molecular stages

To identify distinct stages in aging skin tissue, we examined 86 female subjects between 21 and 76 years. Subjects were chosen so that all ages were represented evenly and were required to be in good health. From each subject we sampled epidermal tissue from the subject's volar forearms via the suction blister method. 31 of the original subjects were further re-invited for a longitudinal second measurement, which took place three years later (Figure 1). From the epidermal samples we generated gene expression and DNA methylation data, which were computationally integrated using the similarity network fusion approach. The resulting network, incorporating information from both transcriptomes and methylomes, was then used to identify hidden subtypes via unsupervised spectral clustering. The clustering revealed four distinct subgroups in our data with roughly equal sizes of 22 (cluster 1), 20 (cluster 2), 18 (cluster 3) and 26 (cluster 4) subjects, that captured the multi-omics similarity structure between the samples more clearly than either chronological or DNA methylation (DNAm) age (Figure 2A-2C). Association analysis to subject metadata showed that the clusters were significantly associated with chronological age (p = 5.8e-12, Figure 3A), whilst not being confounded by BMI (p = 0.71, Supplementary Figure 1A).

Latent stages associate more strongly with DNAm age rather than chronological age

As the unsupervised and purely data-driven clustering had identified groupings with strong association to chronological age, we explored the possibility that the clusters might capture hidden stages in the aging progression. We hypothesized that if this were so, the groupings ought to be more strongly associated with the actual biological aging state of our subjects, rather than their chronological age. To test this, DNAm ages of all subjects were calculated as previously described [19], to serve as a proxy measure for biological age (Supplementary Figure 1B). The comparison revealed a stronger association of the identified stages to DNAm age (p = 3.9e-13) as opposed to chronological age (p = 5.2e-12), strengthening the hypothesis that the clusters captured multi-omics aging stages.

Aging phase outliers are also biological age outliers in the sense of DNAm age

As subjects within the phases still presented considerable variation in chronological age, and the most proven approximation to biological age available to date is through the use of DNAm age, we further explored if the chronological outliers in the different aging phases were also outliers in the sense of DNAm age. We defined individuals as outliers for every aging phase (Figure 3B) if their chronological age either exceeded the 3rd quartile by at least one third of the interquartile range ("young-like") or subceeded the 1st quartile by said amount ("old-like"). Analysis of the deviation of chronological to DNAm ages revealed that phase outliers were indeed biologically significantly younger (mean/median = -3.8/-4.5 y) or older (mean/median = +4.2/+4.3 y) than average and also than each other respectively (Figure 3C). Association testing using age-adjusted logistic regression models further revealed that subjects assigned to the "old-like" group were also significantly more likely to have reported frequent sun bathing in the questionnaires (p = 0.0336), delivering evidence of photoaging factoring into aging phase assignment. We further repeated the analysis using age estimates from a transcriptomic clock [19], which on average showed lower accuracy than its DNAm counterpart (Supplementary Figure 1C), in concordance with previous reports [19]. Association testing of phase outlier status to transcriptomic age revealed the same trends observed with DNAm age, albeit in this case without reaching statistical significance (Supplementary Figure 1D). Notably, the correlation of the two biological age markers was lower than their respective correlations to chronological age (Supplementary Figure 1E), indicating that the clocks capture at least partly independent features of aging, again underlining the importance of multi-omics approaches for aging research.

Aging phases show improved detection of Hallmark of Aging signals

The Hallmarks of Aging (HoA) describe nine main biological motives and processes that are believed to be driving the aging progression. We hypothesized that if

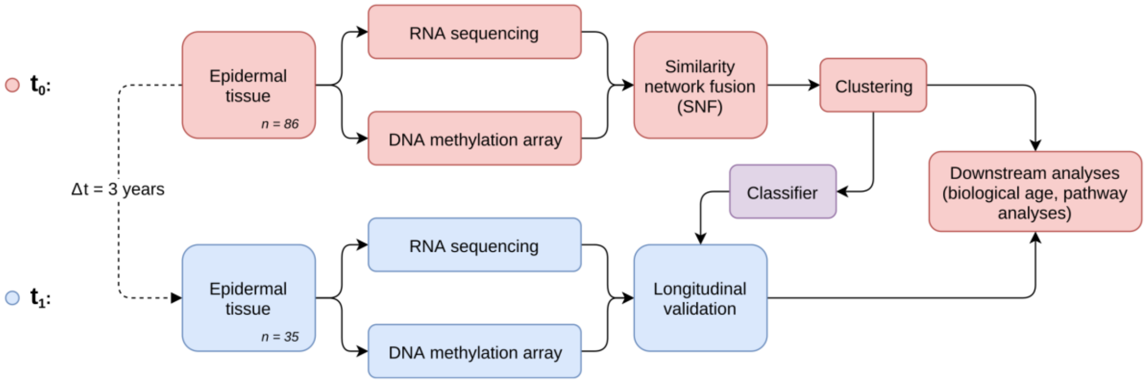

Figure 1. Study and analysis setup. Workflow diagram depicting the two-stage longitudinal study setup and the main steps of multi-omics data generation, integration and analysis.

the identified aging phases captured stages of biological aging, this ought to be reflected in gene expression patterns related to the known aging cornerstone processes, as summarized by the HoA. We therefore generated lists of genes involved in each of the nine HoA, by selecting GO and Reactome pathways which captured the essence of the respective hallmarks and combining them to novel gene sets (Supplementary Figure 2A). We then used the sets to test if the aging phases allowed better detection of HoA-enrichment signals than chronological age groups. ANOVA analyses using gene set enrichment scores indeed showed stronger discrimination based on aging stage for all HoA gene sets (Figure 3D). This further extends the evidence of the phases capturing biological multi-omics age to the level of gene expression.

Longitudinal validation of aging phases over three-year period

To assess if the phases could be longitudinally reproduced, 31 subjects from the original cohort were re-invited three years later for a second measurement. To assess the aging phase of the new samples, a random forest classification model was built on both expression and methylation features from data of the original cohort. The classifier demonstrated high accuracy (AUC = 0.95) in discriminating between the four phases in repeated cross-validation on the original data (Supplementary Figure 2B) and was subsequently used to predict the aging stages of all subjects at the second time point. For most subjects the aging phase did not change within the 3-year-period, indicating high stability of the identified groupings (Figure 3E). This finding is not unexpected, considering the time span of only three years past since original sampling, relative to the much larger average phase windows with a standard deviation of between 8.2-11.5 chronological years. Nonetheless five subjects could be observed migrating from one aging phase to another, all of them transitioning naturally along the age gradient to the next phase (Figure 3E). Notably, four of these five subjects were previously classified as chronological outliers at the upper end of their age phase (Supplementary Figure 2C).

Data-driven ranking of the Hallmarks of Aging along the phases reveals distinct succession patterns

In order to identify the most important biological motives driving the aging phases, we resorted to the use of machine learning models. For this we implemented a method based on classifiers that learn to distinguish between aging phases, whilst taking advantage of biological pathway information in the training process. The workflow consisted of a stepwise training of classifiers only on subsets of genes annotated to pathways to predict aging phase from gene expression. By restricting the training to these genes, the cross-validated accuracies of these classifiers allow the assessment of how well a given gene set enables the differentiation between aging phases, thus resulting in a score for each pathway's relevance. This score can intuitively be interpreted as a measure of how important

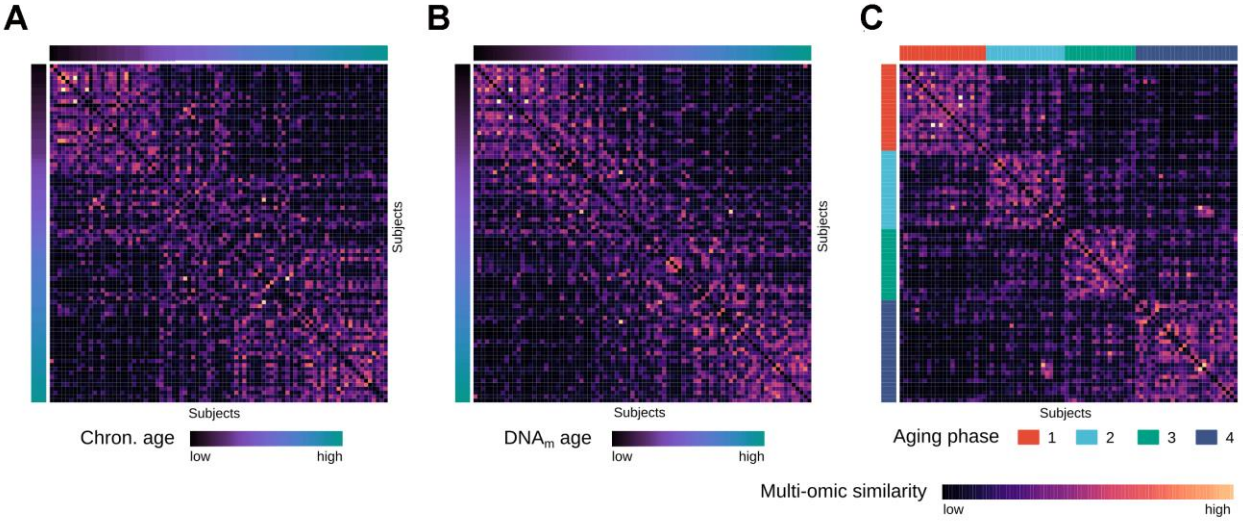

Figure 2. Multi-omics similarity between subjects in the integrated network. (**A**) Heatmap visualization showing similarities between subjects in the fused multi-omics similarity network generated from gene expression and methylation data, with subjects ordered by increasing chronological age. (**B**) Same heatmap visualization of multi-omics similarity as in (**A**), with subjects ordered by increasing DNAm age. (**C**) Same heatmap visualization of multi-omics similarity as in (**A**) and (**B**), with the subjects ordered by the identified aging phases.

or predictive a gene set is to the grouping of interest. In order to derive a data-driven ranking of the Hallmarks of Aging along the aging phases, we performed this pathway predictivity analysis using the aforementioned HoA gene sets, calculating 100 permutations for each pathway model. The predictivity scores revealed a clear patterning of the HoA along the four phases that allowed a grouping of the hallmarks using hierarchical clustering (Figure 4A). Strikingly, the hallmarks clustered almost in the exact constellations postulated in their original description [15], namely into primary hallmarks (genomic instability, telomere attrition, epigenetic alterations and originally loss of proteostasis), antagonistic or secondary hallmarks (cellular senescence, deregulated nutrient sensing and mitochondrial dysfunction) and integrative hallmarks (altered intercellular communication and stem cell exhaustion). Our analysis did however reveal a divergence in the classification of the proteostasis-hallmark, which clustered more strongly with the group of integrative hallmarks. Examination of the HoA predictivity patterns based on the newly generated classification revealed that the predictivity peaks for the respective hallmark classes extracted through our analysis (Figure 4B) also precisely match the temporal manifestation sequence postulated in the original description of the hallmarks as well [15]: Namely, primary hallmarks peaked in aging phases 2

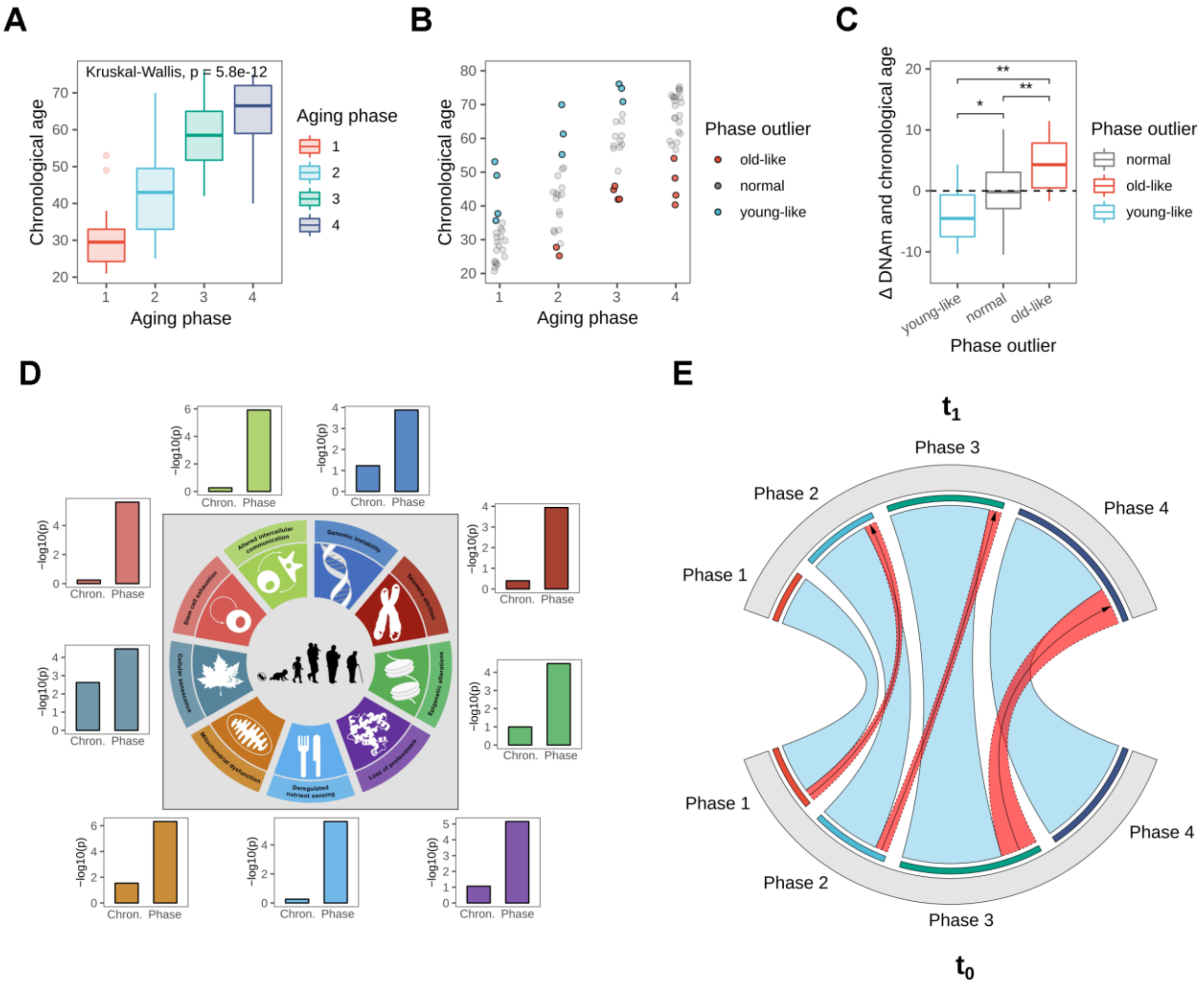

Figure 3. Biological age validation of the identified phases. (**A**) Boxplot showing chronological age distributions among the four identified aging phases. (**B**) Chronological age outliers among the aging phases, denoted as "old-like" for subjects that appeared to prematurely cluster into a higher aging phase, and "young-like" for subjects that were classified into a lower aging phase relative to their chronological age. (**C**) Boxplot showing the deviation of DNAm from chronological age based on aging phase outlier status, revealing a divergence in DNAm aging rate for aging phase outliers. Statistical significance determined using pairwise T-tests. (**D**) Hallmark of Aging signal strengths in gene expression data, comparing chronological age groups to the biological aging phases. Shown are the adjusted p-values from Anova comparisons, testing the segregation of the groupings among gene set enrichment scores. Figure adapted from the original Hallmark of Aging publication [15]. (**E**) Longitudinal validation after three-year period. The chord diagram shows aging phase classification of re-invited subjects at both time points, with phase transitions highlighted in red.

and 3, followed by a sharp drop in predictivity thereafter. Meanwhile the importance of the secondary hallmarks increased notably in aging phase 3. The integrative hallmarks, postulated to be emerging as a consequence of primary and secondary manifestation, increased slowly along the phases, while peaking in the late aging phases 3 and 4, again in concordance with the original postulation [15]. To our knowledge this is the first data-driven validation of the overarching sequence in which these cornerstones of aging manifest in human tissue.

Pathway predictivity analysis reveals a distinctly non-linear loss in pathway predictivity in old-age

As our analyses on the succession of the HoA already indicated a distinct shift in predictivities towards phase 4 of the identified aging stages, and in light of the recent publication of a sharp loss of signature identity for longevity-associated mTOR-signaling in human tissue around 60 years [14], we decided to expand our pathway analysis and explore the mid-to-late-life transition in more detail. For this we utilized the hallmark gene sets defined by the Broad Institute into the analysis, a set of conserved and highly refined gene sets created to improve pathway inference by reducing variance and gene overlap [20]. The predictivity analysis using these gene sets revealed a number of pathways that showed distinctly non-linear predictivity patterns along the four aging phases (Figure 5A). Notably, the most prominent global pattern was a sharp loss in predictivity observed in the transition from aging phase 3 to aging phase 4 in many pathways. In line with the loss of predictivity in nutrient sensing signaling hallmark observed in the HoA analysis and recent reports [14], mTOR-related signaling was among the pathways undergoing this distinct transition in the transition from phase 3 to phase 4 (Figure 5B), which, biological age outliers aside, matches the chronological age threshold of 60 identified in recent reports [14].

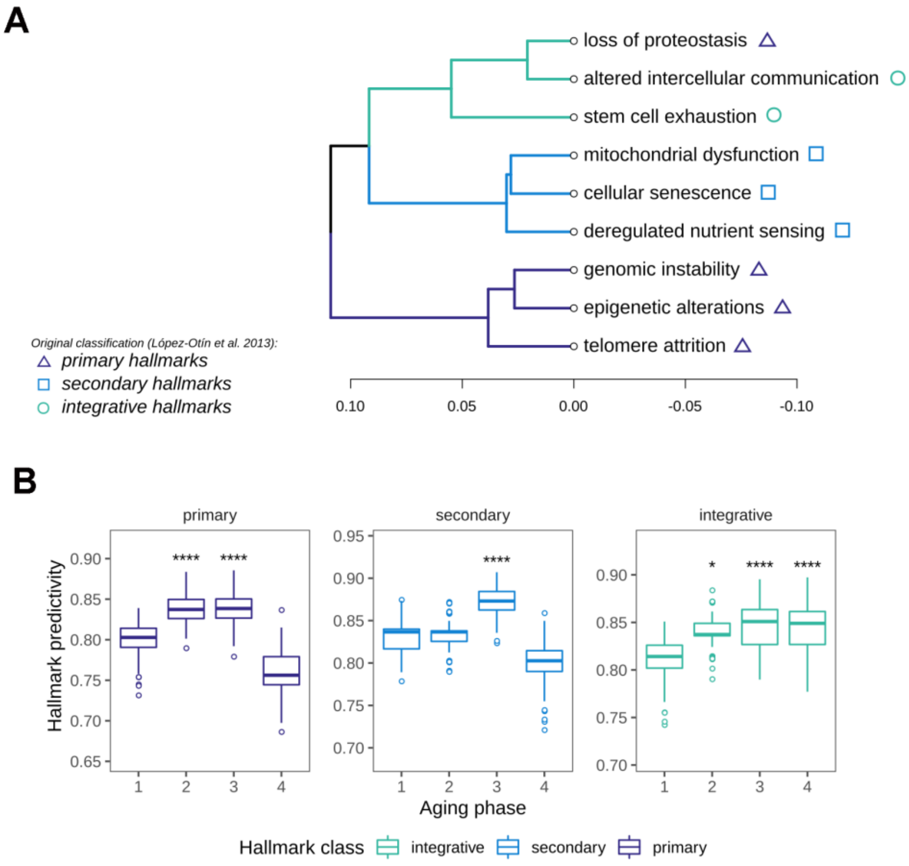

Figure 4. Characterization of Hallmark of Aging predictivity within the aging phases. (**A**) Hierarchical clustering of the nine Hallmarks of Aging based on their gene set predictivity analysis along the four aging phases. Predictivity was determined using cross-validated random forest classifiers, trained to distinguish each of the aging phases from the others. (**B**) Predictivity of the Hallmark of Aging gene sets along the four aging phases, grouped into primary, secondary and integrative hallmarks. Statistical testing was performed using one-sided Wilcoxon tests. All predictivity scores were derived from 100 permutations.

Other pathways exhibiting this pattern included oxidative phosphorylation and fatty acid metabolism, and notably also DNA repair pathways (Figure 5B). Exceptions from this trend included interferon and interleukin signaling, which increased steadily in predictivity along the phases, in line with the inflammaging theory of aging [21, 22], and the previously observed patterns in the HoA analysis. Apart from these exceptions, statistical analysis of all pathway predictivity signals between aging phases 3 and 4 still revealed a significant decrease in pathway predictivity, that is replicated using gene set enrichment analysis, also showing a distinct loss in pathway enrichment in transition to phase 4 (Figure 5C). As this finding potentially points to an increase in transcriptional noise, we investigated whether there was a change in the transcriptional similarity between subjects with transition into aging phase 4. For this we calculated pairwise correlations between the full transcriptomes of all subjects. In line with the results from the pathway analysis, we observed a significant drop in transcriptional similarity in the transition from phase 3 to phase 4. Notably, a similar effect can be observed in the methylation data, where a concomitant decrease in correlation between methylation profiles is observed (Figure 5D).

Together these findings indicate a distinctly non-linear increase in biological noise in the transition from mid-to-late-life, likely to contribute to the deterioration of human tissue function in old age.

DISCUSSION

In this study we applied network based multi-omics integration to investigate non-linearity in the *in vivo* human aging progression. Similarity network fusion has so far mostly seen use in cancer research, in other fields there have only been rare applications of this methodology so far. To the best of our knowledge this publication represents the first documented use of a network based multi-omics integration and cluster analysis in the context of aging. The four aging stages that we identified in the integrated similarity network were more strongly associated with measures for biological age as opposed to chronological age, demonstrating the use of unsupervised integration and clustering in approximating biological aging and elucidating discrete stages from multi-omics data in the process.

To characterize the aging stages, we turned to the conceptual cornerstones that are believed to drive organismal aging, the so-called Hallmarks of Aging. For this we devised a novel approach to rank the gene sets according to their importance for the aging phases using machine learning methodology. The approach allowed us to validate the originally proposed classification of the hallmarks in a data-driven way and further to elucidate the order of their occurrence from the molecular data. The overall concordance of the data-driven reconstruction of the order of hallmark manifestation to the postulated sequence of succession is striking. All hallmarks clustered according to the proposed classification into primary, antagonistic and integrative hallmarks, with the only notable exception being the loss of proteostasis hallmark, which somewhat deviated from its postulated order. Based on this, a reclassification of this hallmark might be advisable. To resolve this, further investigations greatly expanding the width of studied tissues will be required though. The order of succession reconstructed from our data matches the proposed order of primary, antagonistic or secondary and integrative hallmarks almost perfectly. The earliest recorded peak is observed for the primary hallmarks in aging phase 2 and is followed by a significant increase in predictivity in phase 3 for the secondary hallmarks. This also includes an increased importance of mitochondrial processes around mid-life, a finding that is especially interesting in light of recent reports that a mid-life intervention alleviating mitochondrial dysfunction is sufficient to significantly increase health span in model animals [13]. The integrative hallmarks and in particular altered intercellular communication, a hallmark strongly based on immune and inflammatory signaling pathways, slowly increases in predictivity and peaks in the late aging phases, supporting the inflammaging theory of aging [22, 23]. Inflammaging describes the process by which immunosenescence and thus reduced ability to deal with stressors lead to a chronic low grade proinflammatory state in aged tissue, in turn deregulating the immune response and increasing vulnerability to pathologies with inflammatory genesis or progression.

The transitions to phase 3 and phase 4 were marked by various changes in the predictivity ranking of the hallmarks for each phase, indicating substantial rearrangements in the biological processes around mid-life. We investigated these transitions further by expanding our analysis to a wider range of conserved biological pathways. The analysis results showed a marked and global loss in pathway predictivity and pathway enrichment in the transition out of phase 3, indicative of an increase in transcriptional noise in aging phase 4. Pairwise correlation of all subjects further confirmed a significant deterioration of both transcriptomic and epigenetic patterning in the passage from aging phase 3 to aging phase 4. Chronological outliers aside, the onset of phase 3 exactly matches the median age at onset of menopause, which is around 51

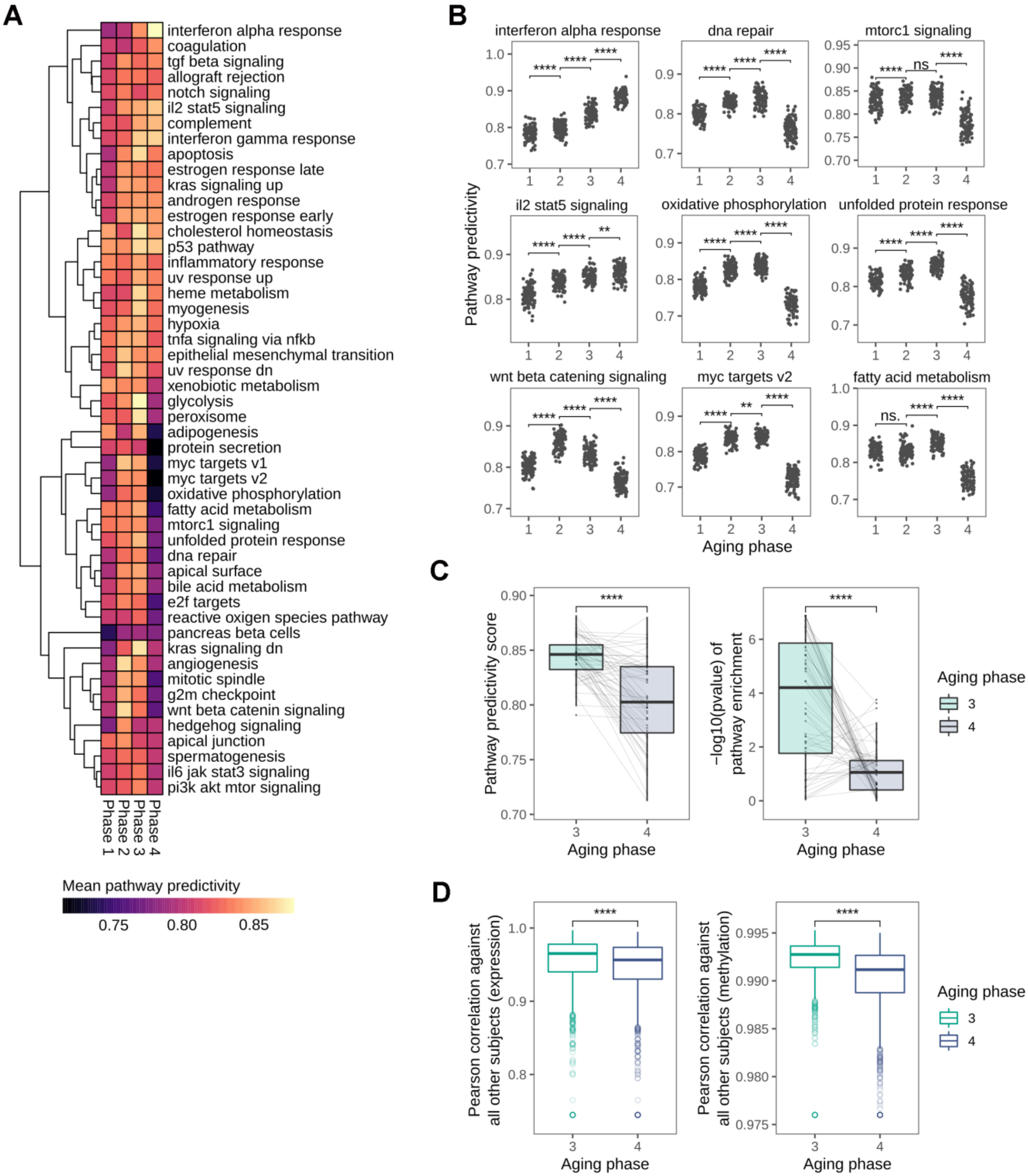

Figure 5. Global loss in pathway predictivity in the transition from mid- to late-life. (**A**) Heatmap showing the changes in pathway predictivity along the identified aging phases. The predictivities shown are the average predictivities calculated from 100 permutations for every pathway. (**B**) Scatterplots visualizing the changes in predictivity along the aging phases for selected pathways, several of which show distinctly non-linear patterns. (**C**) Overall loss in pathway predictivity observed in the transition from aging phase 3 to phase 4 is also detectable using gene set enrichment analysis. (**D**) Pairwise Pearson correlation between all subjects based on transcriptional and DNA methylation patterns.

years for Caucasian women in industrialized countries [24–26], a period that is indeed known to have substantive effects on the biology of the female body through wide-reaching hormonal adjustments. This is particularly interesting as the menopausal transition has been shown to accelerate biological aging based on large-scale analyses of blood derived DNAm age [27]. This finding has thus far lacked mechanistic explanation, the significant mid-life shift and loss in predictivity we observed in the pathway landscape in the transition from phase 3 to post-menopausal phase 4 might be a connected phenomenon and serve as starting point for further investigations into this matter. Notably, one of the hallmarks suffering a sharp loss in predictivity in this phase is the epigenetic alterations hallmark, linking the loss in transcriptional pathway state to an epigenetic age acceleration. Meanwhile, in a skin-specific context, the reports of accelerated skin aging following menopause are also manifold [28] and might equally be connected to our findings. A direct coupling between the identified aging phases 3 and 4 and the menopausal transition might explain yet another interesting epidemiological finding: the fact that higher age at onset of natural menopause has frequently been associated with greater remaining life expectancy and reduced all-cause mortality [26, 29–31]. Considering menopause as a distinct stage in the natural aging progression would allow the interpretation that women entering it later (at a higher chronological age) are biologically younger or “young-like” in the sense of the outlier classification proposed earlier (Figure 3B and 3C). The observed greater remaining life expectancy would then present itself as a plausible consequence of their lower biological age entering menopause.

One of the pathways that notably lost predictivity at the beginning of aging phase 4 was PI3K-mTOR-signaling, a known longevity-associated pathway, whose regulation has recently been reported to be largely lost around the chronological age of 60 [14]. Among the other pathways affected by a similar decrease in pathway enrichment were also DNA repair pathways. This might present a finding with significant impact to health in aging phase 4 onwards, as these pathways are crucial for cancer protection, and mutations and dysregulation in these pathways have been identified numerous times as drivers of tumorigenesis. The observed loss in pathway enrichment in the transition to phase 4 could be a worrying sign of decreased safeguarding ability towards carcinogenesis in this later aging phase, which is especially relevant in the skin, a tissue that is frequently exposed to mutagenic solar irradiation. The transition to phase 4 happens to coincide with epidemiological observations that pinpoint a strongly increasing risk of developing cancer from the chronological age range of 60 upwards [32]. Naturally further studies will be required to evaluate if any causal relationship between aging phase and cancer risk exists indeed, but the overlap in the chronological age ranges is intriguing and might warrant further investigations (Supplementary Figure 2D).

In summary, using multi-omics analysis we identified four aging phases in *ex vivo* human skin tissue of female participants over a wide age range. The phases appeared to be driven by actual biological age rather than chronological age, capturing distinct stages along the aging progression and allowed the data-driven reconstruction of the manifestation sequence postulated for the Hallmarks of Aging. Characterization of the mid- to late-life transition identified an extensive loss in pathway enrichment, with potential implications for life- and health-span in old age.

MATERIALS AND METHODS

Recruiting

The study was performed in agreement with the recommendations of the Declaration of Helsinki and all test subjects provided written, informed consent. Subjects were recruited in the age range of 20 to 80 years, with equal numbers of participants within each decade. Subjects were required to be female, in good health and belonging to phototypes II or III according to the Fitzpatrick scale [33], to limit non-age related variability in the data. Exclusion criteria included tattoos or scars in the test area, pigmentation disorders, pregnancy and medication such as anti-histamines or anti-inflammatory drugs within two weeks prior to study start. A detailed listing of exclusion criteria can be found in the Supplementary. Participants were further required to complete a self-assessment questionnaire on age, weight, height, smoker status, sun bathing habits, as well as food and drinking habits upon study start.

Tissue sample preparation

The suction blister method applied in this study has been approved by the Ethics Commission of the University of Freiburg (general approval Dec 8, 2008; Beiersdorf AG No. 28857). Three suction blisters of 7 mm diameter were taken from the volar forearms of all subjects as previously described [34].

Nucleic acid extraction

Tissue samples were suspended in the respective lysis buffers for DNA or RNA extraction and homogenized using an MM 301 bead mill (Retsch). DNA was then extracted using the QIAamp DNA Investigator Kit (Qiagen) according to manufacturer’s instructions. RNA was extracted using the RNeasy Fibrous Tissue Mini Kit (Qiagen) according to manufacturer's instructions.

Transcriptome sequencing

Transcriptome libraries were prepared using TruSeq Library Prep Kit (Illumina) and sequencing performed at 1x50 bp on Illumina's HiSeq system to a final sequencing depth of 100 million reads per sample. Sequencing data was processed using a custom pipeline including Fastqc v0.11.7 [35] for quality control, Trimmomatic v0.36 [36] for trimming and Salmon v0.8.1 [37] for mapping and read quantification.

Array based methylation profiling

Methylation profiling was performed using Illumina 450k (first time point) and EPIC (second time point) arrays. In order to ensure comparability of measurements, EPIC arrays were computationally reduced to include only probes present on the original 450k array using the minfi package [38] in R [39]. Methylation data was processed in minfi using the funnorm normalization method.

Similarity network fusion and clustering

Prior to integration, the gene expression (log2 transformed transcripts per million) and CpG methylation data (M values) were batch corrected using the Combat algorithm [40] implemented in the sva package [41], following a feature selection step via filtering by median absolute deviation, retaining 10 % of the most informative features. The data was then integrated as previously described [18] using parameter settings of k = 10 (number of neighbors), t = 20 (number of iterations) and *alpha* = 0.5 (hyperparameter). Clustering on the fused network was performed via spectral clustering as previously described [18]. Measures used for the selection of cluster numbers were the eigen-gap statistic and rotation cost as proposed in the original method description [18], as well as visual inspection using heatmaps.

Age clock analyses

Analyses of DNAm and transcriptomic age were performed as previously described [19]. DNAm age was calculated from M values, whereas transcriptomic age was predicted based on log2 transformed transcripts per million.

Hallmark of aging gene sets

The HoA gene sets were generated from GO [42] and Reactome [43] gene sets by manually selecting matching pathways assigned to the nine Hallmarks of Aging [15]. A detailed list of genes annotated to each hallmark is provided in the Supplementary Material in .gmt format.

Enrichment analyses

Enrichment analyses were performed using the PLAGE algorithm based on singular value decomposition as described in [44] and implemented in the GSVA [45] R package.

Classification model to predict aging phase in longitudinal validation

To predict aging phase of re-invited subjects at the second time point, a random forest classifier was trained on the samples from the original cohort. Features were selected as the top 50 hits derived from differential gene expression analysis using DESeq2 [46] and differential methylation analysis using limma [47] from pairwise aging phase comparisons. The model was trained within the machine learning framework mlr [48], using the algorithm implemented in the original randomForest package [49]. Adjusted model hyperparameters included *ntree* = 1000 and $mtry = \sqrt{features}$. Accuracy of prediction was calculated as the area under the receiver operating characteristic curve (AUC) for multi-class comparisons, as implemented in the pROC package [50], and was derived from 5 x 5-fold repeated cross-validation.

Pathway predictivity analysis

Pathway predictivity was assessed using random forest pathway classifiers, constructed using the gene sets generated in this study and using the Hallmark Process [20] gene sets downloaded from the Molecular Signatures Database v6.2 [51]. The models were trained by restricting the molecular data to that of genes annotated within a given hallmark and trained to predict the aging phase of every sample. Predictivity was determined as the accuracy of correct classification derived from 5 x 5-fold repeated cross-validation for each pathway model, giving insight on how well genes within the gene set allow a discrimination between the phases and was thus used as a measure of importance of the respective pathway. Samples were stratified with respect to the target variable in the cross-validation process in order to avoid unbalanced proportions in any fold that might lead to bloated accuracy measures. Hyperparameters of all models were adjusted to *ntree* = 1000 and $mtry = \sqrt{number\ of\ genes\ in\ pathway}$. To determine the predictivity of the HoA stratified for each of four aging phases, the classifiers were separately trained in a one-against-all type of setup, learning to distinguish a phase from all the others. Modeling parameters and cross-validation were chosen as described above, and results for the four phases were aggregated afterwards.

General data analysis and visualization

Data analysis in R further included the usage of the package data.table [52], dplyr [53] and Hmisc [54] for data handling and general purpose functions, as well as the packages ggplot2 [55], ggpubr [56], ggsci [57], circlize [58] and pheatmap [59] for data visualization. Workflow diagrams were built using draw.io [60].

AUTHOR CONTRIBUTIONS

MW, CF and LK conceived the original idea for the study. SG and HW provided funding for the study. CF and BK planned the study. SJ carried out the wet lab experiments. NH, CF and LK conceived the analysis. NH performed the computations. NH, EG, CF, MW, SG, HW and LK discussed and contributed to the interpretation of the results. NH wrote the manuscript. All authors discussed and commented on the manuscript.

CONFLICTS OF INTEREST

NH, CF, JS, BK, EG, SG, HW and MW are employees of Beiersdorf AG. LK received consultation fees from Beiersdorf AG.

FUNDING

This research received no external funding.

REFERENCES

1. Bocklandt S, Lin W, Sehl ME, Sánchez FJ, Sinsheimer JS, Horvath S, Vilain E. Epigenetic predictor of age. PLoS One. 2011; 6:e14821. https://doi.org/10.1371/journal.pone.0014821 PMID:21731603

2. Koch CM, Wagner W. Epigenetic-aging-signature to determine age in different tissues. Aging (Albany NY). 2011; 3:1018–27. https://doi.org/10.18632/aging.100395 PMID:22067257

3. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan JB, Gao Y, Deconde R, Chen M, Rajapakse I, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013; 49:359–67. https://doi.org/10.1016/j.molcel.2012.10.016 PMID:23177740

4. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013; 14:R115. https://doi.org/10.1186/gb-2013-14-10-r115 PMID:24138928

5. Weidner CI, Lin Q, Koch CM, Eisele L, Beier F, Ziegler P, Bauerschlag DO, Jöckel KH, Erbel R, Mühleisen TW, Zenke M, Brümmendorf TH, Wagner W. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. Genome Biol. 2014; 15:R24. https://doi.org/10.1186/gb-2014-15-2-r24 PMID:24490752

6. Petkovich DA, Podolskiy DI, Lobanov AV, Lee SG, Miller RA, Gladyshev VN. Using DNA methylation profiling to evaluate biological age and longevity interventions. Cell Metab. 2017; 25:954–60.e6. https://doi.org/10.1016/j.cmet.2017.03.016 PMID:28380383

7. Meer MV, Podolskiy DI, Tyshkovskiy A, Gladyshev VN. A whole lifespan mouse multi-tissue DNA methylation clock. Elife. 2018; 7:e40675. https://doi.org/10.7554/eLife.40675 PMID:30427307

8. Chen BH, Marioni RE, Colicino E, Peters MJ, Ward-Caviness CK, Tsai PC, Roetker NS, Just AC, Demerath EW, Guan W, Bressler J, Fornage M, Studenski S, et al. DNA methylation-based measures of biological age: meta-analysis predicting time to death. Aging (Albany NY). 2016; 8:1844–65. https://doi.org/10.18632/aging.101020 PMID:27690265

9. Quach A, Levine ME, Tanaka T, Lu AT, Chen BH, Ferrucci L, Ritz B, Bandinelli S, Neuhouser ML, Beasley JM, Snetselaar L, Wallace RB, Tsao PS, et al. Epigenetic clock analysis of diet, exercise, education, and lifestyle factors. Aging (Albany NY). 2017; 9:419–46. https://doi.org/10.18632/aging.101168 PMID:28198702

10. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, Hou L, Baccarelli AA, Stewart JD, Li Y, Whitsel EA, Wilson JG, Reiner AP, et al. An epigenetic biomarker of aging for lifespan and healthspan. Aging (Albany NY). 2018; 10:573–91. https://doi.org/10.18632/aging.101414 PMID:29676998

11. Marioni RE, Shah S, McRae AF, Chen BH, Colicino E, Harris SE, Gibson J, Henders AK, Redmond P, Cox SR, Pattie A, Corley J, Murphy L, et al. DNA methylation age of blood predicts all-cause mortality in later life. Genome Biol. 2015; 16:25. https://doi.org/10.1186/s13059-015-0584-6 PMID:25633388

12. Tricoire H, Rera M. A new, discontinuous 2 phases of aging model: lessons from drosophila melanogaster. PLoS One. 2015; 10:e0141920.

https://doi.org/10.1371/journal.pone.0141920 PMID:26528826

13. Rana A, Oliveira MP, Khamoui AV, Aparicio R, Rera M, Rossiter HB, Walker DW. Promoting Drp1-mediated mitochondrial fission in midlife prolongs healthy lifespan of drosophila melanogaster. Nat Commun. 2017; 8:448. https://doi.org/10.1038/s41467-017-00525-4 PMID:28878259

14. Timmons JA, Volmar CH, Crossland H, Phillips BE, Sood S, Janczura KJ, Törmäkangas T, Kujala UM, Kraus WE, Atherton PJ, Wahlestedt C. Longevity-related molecular pathways are subject to midlife "switch" in humans. Aging Cell. 2019; 18:e12970. https://doi.org/10.1111/acel.12970 PMID:31168962

15. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. Cell. 2013; 153:1194–217. https://doi.org/10.1016/j.cell.2013.05.039 PMID:23746838

16. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, Milanesi L. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics. 2016 (Suppl 2); 17:15. https://doi.org/10.1186/s12859-015-0857-9 PMID:26821531

17. Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. Front Genet. 2017; 8:84. https://doi.org/10.3389/fgene.2017.00084 PMID:28670325

18. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014; 11:333–37. https://doi.org/10.1038/nmeth.2810 PMID:24464287

19. Bormann F, Rodríguez-Paredes M, Hagemann S, Manchanda H, Kristof B, Gutekunst J, Raddatz G, Haas R, Terstegen L, Wenck H, Kaderali L, Winnefeld M, Lyko F. Reduced DNA methylation patterning and transcriptional connectivity define human skin aging. Aging Cell. 2016; 15:563–71. https://doi.org/10.1111/acel.12470 PMID:27004597

20. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database (MSigDB) hallmark gene set collection. Cell Syst. 2015; 1:417–25. https://doi.org/10.1016/j.cels.2015.12.004 PMID:26771021

21. Salminen A, Kaarniranta K, Kauppinen A. Inflammaging: disturbed interplay between autophagy and inflammasomes. Aging (Albany NY). 2012; 4:166–75. https://doi.org/10.18632/aging.100444 PMID:22411934

22. Franceschi C, Garagnani P, Parini P, Giuliani C, Santoro A. Inflammaging: a new immune-metabolic viewpoint for age-related diseases. Nat Rev Endocrinol. 2018; 14:576–90. https://doi.org/10.1038/s41574-018-0059-4 PMID:30046148

23. Franceschi C, Bonafè M, Valensin S, Olivieri F, De Luca M, Ottaviani E, De Benedictis G. Inflamm-aging. An evolutionary perspective on immunosenescence. Ann N Y Acad Sci. 2000; 908:244–54. https://doi.org/10.1111/j.1749-6632.2000.tb06651.x PMID:10911963

24. Stanford JL, Hartge P, Brinton LA, Hoover RN, Brookmeyer R. Factors influencing the age at natural menopause. J Chronic Dis. 1987; 40:995–1002. https://doi.org/10.1016/0021-9681(87)90113-5 PMID:3654908

25. McKinlay SM, Brambilla DJ, Posner JG. The normal menopause transition. Maturitas. 1992; 14:103–15. https://doi.org/10.1016/0378-5122(92)90003-m PMID:1565019

26. Ossewaarde ME, Bots ML, Verbeek AL, Peeters PH, van der Graaf Y, Grobbee DE, van der Schouw YT. Age at menopause, cause-specific mortality and total life expectancy. Epidemiology. 2005; 16:556–62. https://doi.org/10.1097/01.ede.0000165392.35273.d4 PMID:15951675

27. Levine ME, Lu AT, Chen BH, Hernandez DG, Singleton AB, Ferrucci L, Bandinelli S, Salfati E, Manson JE, Quach A, Kusters CD, Kuh D, Wong A, et al. Menopause accelerates biological aging. Proc Natl Acad Sci USA. 2016; 113:9327–32. https://doi.org/10.1073/pnas.1604558113 PMID:27457926

28. Thornton MJ. Estrogens and aging skin. Dermatoendocrinol. 2013; 5:264–70. https://doi.org/10.4161/derm.23872 PMID:24194966

29. Snowdon DA, Kane RL, Beeson WL, Burke GL, Sprafka JM, Potter J, Iso H, Jacobs DR Jr, Phillips RL. Is early natural menopause a biologic marker of health and aging? Am J Public Health. 1989; 79:709–14. https://doi.org/10.2105/ajph.79.6.709 PMID:2729468

30. Jacobsen BK, Heuch I, Kvåle G. Age at natural menopause and all-cause mortality: a 37-year follow-up of 19,731 norwegian women. Am J Epidemiol. 2003; 157:923–29.

https://doi.org/10.1093/aje/kwg066 PMID:12746245

31. Shadyab AH, Macera CA, Shaffer RA, Jain S, Gallo LC, Gass ML, Waring ME, Stefanick ML, LaCroix AZ. Ages at menarche and menopause and reproductive lifespan as predictors of exceptional longevity in women: the women's health initiative. Menopause. 2017; 24:35–44. https://doi.org/10.1097/GME.0000000000000710 PMID:27465713

32. National Cancer Institute (NCI). SEER Cancer Statistics Review (CSR) 1975–2014. 2018.

33. Fitzpatrick TB. Soleil et peau. Journal de Médecine Esthétique. 1975; 2:4.

34. Südel KM, Venzke K, Knussmann-Hartig E, Moll I, Stäb F, Wenck H, Wittern KP, Gercken G, Gallinat S. Tight control of matrix metalloproteinase-1 activity in human skin. Photochem Photobiol. 2003; 78:355–60. https://doi.org/10.1562/0031-8655(2003)078<0355:tcomma>2.0.co;2 PMID:14626663

35. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

36. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 2014; 30:2114–20. https://doi.org/10.1093/bioinformatics/btu170 PMID:24695404

37. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017; 14:417–19. https://doi.org/10.1038/nmeth.4197 PMID:28263959

38. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA. Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. Bioinformatics. 2014; 30:1363–69. https://doi.org/10.1093/bioinformatics/btu049 PMID:24478339

39. R Development Core Team. R: The R Project for Statistical Computing. 2008. https://www.r-project.org/

40. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics. 2007; 8:118–27. https://doi.org/10.1093/biostatistics/kxj037 PMID:16632515

41. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012; 28:882–83. https://doi.org/10.1093/bioinformatics/bts034 PMID:22257669

42. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet. 2000; 25:25–29. https://doi.org/10.1038/75556 PMID:10802651

43. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, Milacic M, Roca CD, Rothfels K, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2018; 46:D649–55. https://doi.org/10.1093/nar/gkx1132 PMID:29145629

44. Tomfohr J, Lu J, Kepler TB. Pathway level analysis of gene expression using singular value decomposition. BMC Bioinformatics. 2005; 6:225. https://doi.org/10.1186/1471-2105-6-225 PMID:16156896

45. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013; 14:7. https://doi.org/10.1186/1471-2105-14-7 PMID:23323831

46. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15:550. https://doi.org/10.1186/s13059-014-0550-8 PMID:25516281

47. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43:e47. https://doi.org/10.1093/nar/gkv007 PMID:25605792

48. Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, Casalicchio G, Jones ZM, Casalicchio G, Gallo M, Schratz P. mlr: Machine Learning in R. Journal of Machine Learning Research. 2016; 17:1–5.

49. Breiman L. Random Forests. Machine Learning. Kluwer Academic Publishers; 2001; 45:5–32. https://doi.org/10.1023/A:1010933404324

50. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011; 12:77. https://doi.org/10.1186/1471-2105-12-77 PMID:21414208

51. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011; 27:1739–40. https://doi.org/10.1093/bioinformatics/btr260 PMID:21546393

52. Dowle M, Srinivasan A. data.table: Extension of ‘data.frame’. 2018.

53. Wickham H, Francois R, Henry L, Müller K. Package ‘dplyr’. A Grammar of Data Manipulation. R package version 0801. 2019; 1–88.

54. Harrell FE. CRAN - Package Hmisc. Hmisc: Harrell Miscellaneous. 2019.

55. Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, Woo K. ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. 2018.

56. Kassambara A. ggpubr: “ggplot2” Based Publication Ready Plots. 2018.

57. Xiao N. ggsci: Scientific Journal and Sci-Fi Themed Color Palettes for “ggplot2.” 2018.

58. Gu Z. circlize: Circular Visualization. 2018.

59. Kolde R. Package ‘pheatmap’. Bioconductor. 2012; 1–6.

60. Alder G, Benson D, and JGraph Ltd. draw.io. 2005. https://www.draw.io/

npj | Aging and Mechanisms of Disease

www.nature.com/npjamd

ARTICLE OPEN

Modeling transcriptomic age using knowledge-primed artificial neural networks

Nicholas Holzscheck[1,2]✉, Cassandra Falckenhayn[1], Jörn Söhle[1], Boris Kristof[1], Ralf Siegner[1], André Werner[3], Janka Schössow[3], Clemens Jürgens[3], Henry Völzke[3], Horst Wenck[1], Marc Winnefeld[1], Elke Grönniger[1] and Lars Kaderali[2]✉

The development of 'age clocks', machine learning models predicting age from biological data, has been a major milestone in the search for reliable markers of biological age and has since become an invaluable tool in aging research. However, beyond their unquestionable utility, current clocks offer little insight into the molecular biological processes driving aging, and their inner workings often remain non-transparent. Here we propose a new type of age clock, one that couples predictivity with interpretability of the underlying biology, achieved through the incorporation of prior knowledge into the model design. The clock, an artificial neural network constructed according to well-described biological pathways, allows the prediction of age from gene expression data of skin tissue with high accuracy, while at the same time capturing and revealing aging states of the pathways driving the prediction. The model recapitulates known associations of aging gene knockdowns in simulation experiments and demonstrates its utility in deciphering the main pathways by which accelerated aging conditions such as Hutchinson–Gilford progeria syndrome, as well as pro-longevity interventions like caloric restriction, exert their effects.

npj Aging and Mechanisms of Disease (2021)7:15 ; https://doi.org/10.1038/s41514-021-00068-5

INTRODUCTION

In recent years the increasing availability of large-scale molecular biological data from high-throughput experiments, in parallel with technological advancements in machine learning and bioinformatics, have greatly accelerated the discovery of biomarkers and fueled the use of computational modeling to unravel complex biological phenomena. In aging research particularly, the discovery of the 'epigenetic clock'—a machine learning model predicting individual age using genome-wide DNA methylation data—as a highly accurate and reliable biomarker of biological age, has understandably sparked immense interest in the research community. Since then, numerous age clocks have been developed and the concept expanded to further levels of biological data, using transcriptomic, proteomic, and metabolic features[1–9]. While no other data type thus far allowed prediction accuracies quite on par with those achievable using DNA methylation data, features based on metabolite production or gene expression are arguably causally a step closer to the aging phenotype, thereby—at least conceptually—increasing the interpretability of the biomarker. Previously published age clocks based on these data types have not been capitalizing on this conceptual advantage however. On the contrary, interpretability has frequently been neglected as a property in these models so far, no matter the type of data used.

We argue that increasing the interpretability of age clocks may unlock unprecedented utility of these machine learning models in aging research and help expand their use in applied research, e.g. in a human cell-culture-based screening setting, where finding suitable holistic cellular read-outs for the biological aging state is not an easy task and added interpretability could offer additional insight on potential mechanisms of action for given treatment approaches. The concept we propose to achieve this is based on a knowledge-primed artificial neural network, in which information on biological pathways in the form of gene-pathway annotations is incorporated into the architecture of the model. A similar approach has recently been shown to be effective in the modeling of yeast growth from transcriptomic data[10]. Normally, artificial neural networks feature densely connected layers of neural units, in which every neuron in a given layer is connected to every neuron of the next layer. As the information flow through the network is not linked to any particular processes and connections between neurons are essentially interchangeable, it is inherently hard to interpret, which is why deep learning models are frequently quoted as examples of 'black box' models. A defining feature of artificial neural networks however, is the flexibility they offer to implement architectures with unique properties. Omitting the fully connected design and restricting the connections between neurons as implemented for the proposed new age clock can be used to guide the flow of information within the network, thereby augmenting and controlling the way the model learns. Importantly, this allows for the embedding of prior information on biological processes, such as the pathway annotation of genes, directly into the model architecture and therefore ties the model's learning process to known biological processes. Such a design thus enables the model to learn pathway-based representations of the molecular data, which—through the inspection of neuron activations in the pathway layers—allows the monitoring of pathway aging-states and delivers interpretability to the clock's inner workings.

In order to evaluate the utility of this approach for aging research, we constructed a pathway-based artificial neural network and trained it for age prediction based on a large transcriptomic dataset from epidermal skin samples ($n = 887$). Skin represents an extraordinarily well-suited tissue for studying aging, owing to its well-documented aging phenotype and the ease of sampling using non-invasive procedures. As it represents the body's outermost layer, shielding other tissues from hazardous external influence, it also offers the unique possibility to study

[1]Front End Innovation, Beiersdorf AG, Hamburg, Germany. [2]Institute for Bioinformatics, University Medicine Greifswald, Greifswald, Germany. [3]Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany. ✉email: nicholas.holzscheck@beiersdorf.com; lars.kaderali@uni-greifswald.de

extrinsically accelerated aging, phenotypically well-documented in the form of photoaging[11]. The data used to construct the model was derived from the latest iteration of the ongoing Study of Health in Pomerania (SHIP), SHIP-TREND, a longitudinal cohort study generating a broad population-based picture of health and disease in northeastern Germany[12]. Owing to its unbiased observational design, the study is particularly well-suited to investigate the natural aging progression.

RESULTS AND DISCUSSION

Architecture of the neural age clock

The architecture of the artificial neural network was modeled based on the 'Hallmark' pathway collection, a selection of 50 conserved and highly refined gene sets, capturing essential biological processes, created to improve pathway inference by reducing variance and gene overlap, as it is often found in larger pathway collections such as GO terms[13]. The pathway-guided design generates a compartmentalized neural network, in which different parts of the network model distinct pathways, enabling the activations of intermediate neurons to be interpreted to generate insight on the aging states of diverse biological processes. As such the network consists of a single input layer for the gene expression data, followed by four hidden pathway layers and two separate output layers (Fig. 1a), the main output generating the final age estimate, the auxiliary output providing summarized information on the aging states of the respective biological pathways.

To improve both reproducibility and accuracy of the age clock, an ensemble learning approach was implemented. For the final model, a stacked ensemble was constructed from 10 individually trained networks, which shared input and output layers (Fig. 1b). Ensemble stacking is a popular approach to improve the generalization ability of machine learning models by combining the strengths of different model instances, such as those awarded by different weight configurations learned in individual training reboots of neural networks[14]. We found that stacking several models improved prediction accuracy by around 0.3 years, and importantly further cemented the reproducibility of the learning process.

Model training and testing

As a basis for model training, gene expression data were generated via RNA sequencing from epidermal skin samples collected from 887 subjects aged between 30 and 89 years in the SHIP-TREND cohort study (Supplementary Fig. 1a and b). The data were randomly split into independent training and test sets (70/30), with the test set of 267 samples reserved for accuracy assessment and further in silico experiments, leaving 640 samples for model training. The 10 neural networks making up the final model were trained separately for 200 epochs each (Fig. 2a) until no further substantial improvements were detectable without risking overfitting, and then combined into an ensemble by fusing their input and output layers. Assessment of the final age clock's accuracy on the independent test set revealed a median absolute error of 4.7 years (Fig. 2b). This is similar in performance to published 'black box' clocks on transcriptomic data[5,7,8,15,16], which generally tend to perform worse in terms of pure accuracy compared to their DNA methylation-based counterparts[17]. We additionally trained a fully connected "black box" neural network with a comparable number of parameters in the same ensemble approach on the same data, which slightly outperformed its pathway-based counterpart with a median absolute error of 4.4 years (Supplementary Fig. 2a). Based on our data, this suggests that there is a small trade-off between transparency and precision, albeit at a rate that might well be tolerable in practice.

Transcriptomic age is associated with visual age estimates

As the skin presents a well-suited tissue to observe the phenotypic manifestations of aging, we investigated if the transcriptomic age estimates generated by our pathway-based age clock were associated with any phenotypic markers of age. For this, we used standardized portrait images of a random subset of 154 subjects from the test set and generated visual age estimates using a blinded expert panel, tasked to assess the age of the test subjects from the portrait photographs. Linear modeling identified a significant association between the average visual age estimates of this panel and the transcriptomic age predictions ($p = 0.016$) after adjusting for chronological age and gender (Supplementary Table 1), delivering not only a validation of the clock's capabilities to detect biological aging state but also evidence of a direct link between phenotypic manifestations of aging and the molecular alterations in aging skin, captured by the model.

Model reveals the wide-spread impact of aging on the global pathway landscape

Visualizing the intermediate pathway neuron activation for samples of different ages in the pathway-based age clock shows increasing activations for older subjects, allowing not only a general glimpse into the inner workings of the clock but also the detailed assessment of aging states of single biological pathways (Fig. 2c). Ranking the pathways based on a correlation analysis of the intermediate neuron outputs with the actual ages of the subjects revealed p53- and TNFa/NFkB-signaling as the pathways that most clearly captured the aging state out of all modeled processes (Fig. 2d and Supplementary Table 2). However, the margin to the rest of the pathways was rather small and most of the processes showed a significantly higher age association than an artificially introduced control pathway consisting of randomly sampled genes, indicating that the impact of age on gene expression is indeed a global phenomenon, rather than being restricted to a few pathways. The most notable exception to this finding was the low correlation of the pancreas beta-cell pathway at the other end of the spectrum. This might be explained by the low overlap in gene function between pancreas and skin however, given that this gene set mainly describes the differentiation process of beta cells.

The wide-spread impact of increasing age on biological processes meanwhile is in line with the general aging hypothesis of the deleteriome[18]. The deleteriome hypothesis attempts to unify a variety of previous theories of aging under a common motif, the eponymous accumulation of deleterious effects over the lifetime, which are amplified by the inherent imperfection of biochemical processes and reactions. The theoretical framework encompasses previously proposed theories such as the free radical theory of aging[19] but further expands the scope to include observations and theories from evolutionary biology such as the existence of antagonistically pleiotropic genes[20]. The key feature of the theory, despite managing to unify the various explanatory approaches to how the process of aging arises, is that it importantly predicts no single 'master switch' gene or biological process that drives the natural aging progression, but rather a plethora of small individually detrimental alterations to cellular and organismal function accumulating over time. The model's estimates on biological pathway relevance would seem to support this.

In silico gene knockdowns recapitulate associations from the literature

Seeing that the performance of our clock compared reasonably well to 'black box' models and achieved transparency on the biological processes affected, we next set out to test how well the clock actually captured known aging mechanisms and

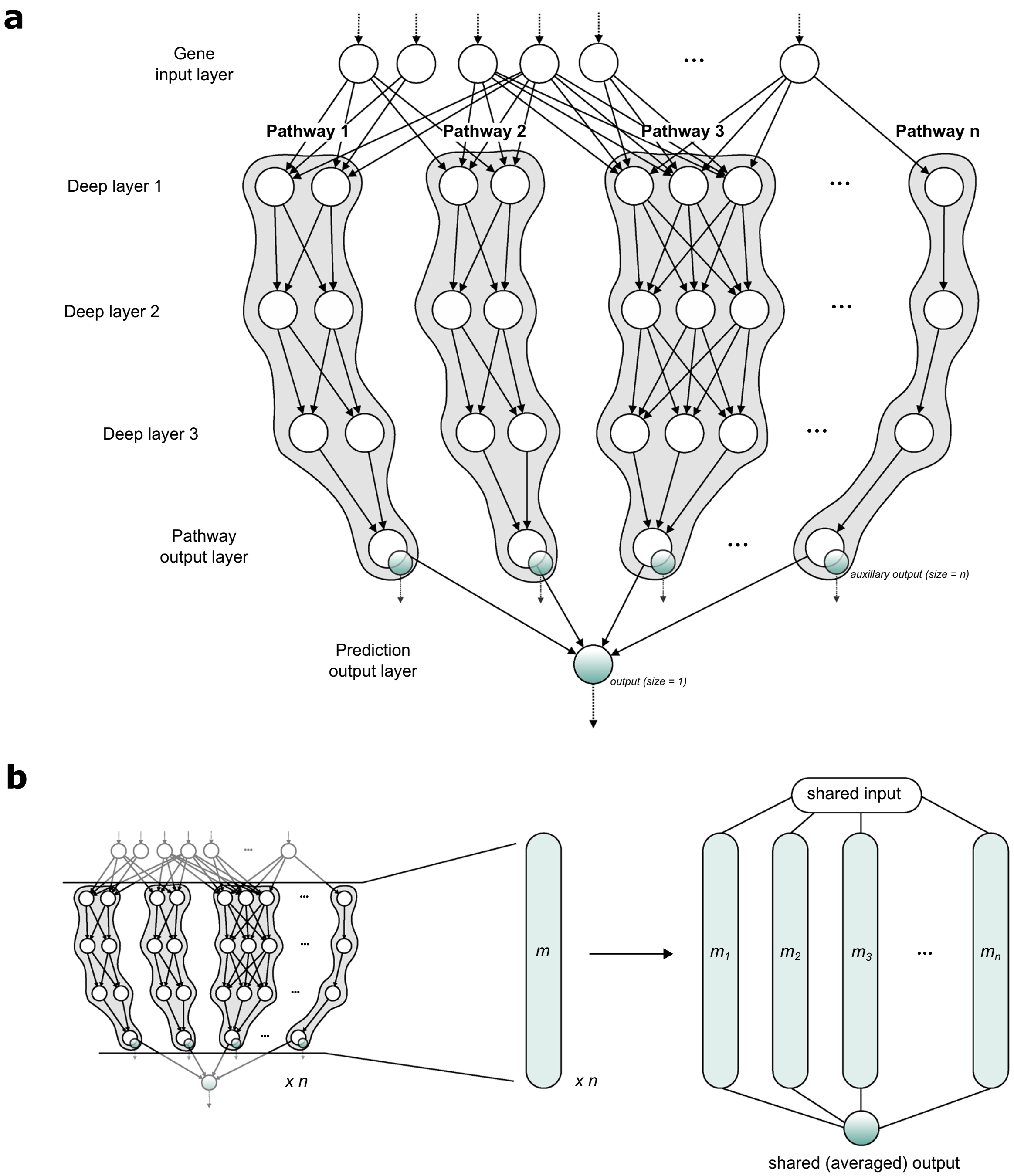

Fig. 1 Model architecture and setup. a Schematic of the artificial neural network architecture. Gene expression data is fed to the input layer, which is connected to the following hidden layer through gene-specific edges that are constructed based on pathway affiliation. In the following hidden layers, information is processed by the network in a pathway-centric manner culminating into a final linear pathway layer with one neuron per pathway, which also serves as an auxiliary output to monitor pathway aging states. Finally, the information from all pathway neurons is aggregated in the main output neuron, which generates the age prediction. **b** Ensemble setup. To improve the stability and accuracy of the final model, an ensemble model was constructed from individually trained networks by joining the separate models to the common input and output layers.

associations through a series of in silico experiments. As discussed above, past research has not identified a single 'master switch' gene or pathway driving aging, nonetheless, several genes have been identified over the years, whose deregulation is associated with changes in lifespan in model organisms or the manifestation of aging phenotypes. To test if the model could recapitulate such associations, we performed virtual gene knockdowns of known aging target genes with a history of experimental data available from model organisms and human genome-wide association studies, to evaluate if the predictions accurately replicated the effects of the perturbation (Fig. 3a). The knockdown of SIRT1 for example, a widely studied NAD-dependent deacetylase with various conserved pro-longevity functions, has been shown to have detrimental effects on the lifespan of several models organisms[21–24]. Indeed, simulation of a decreased SIRT1 regulation by a log2 fold-change of −2 using our model predicted a significant age increase for all subjects in our test set, in concordance with expectations and data from the literature. In

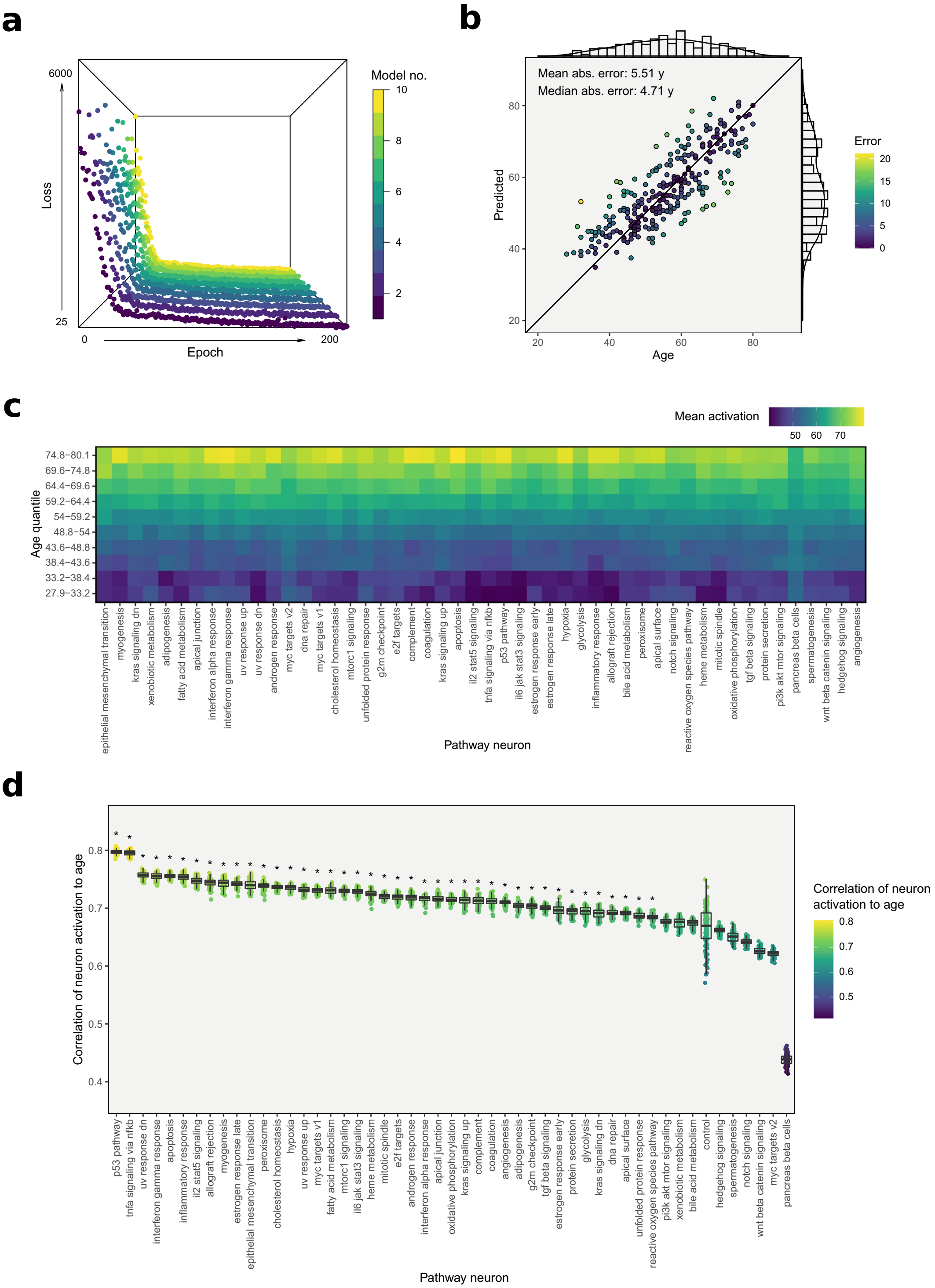

Fig. 2 **Training and performance testing of the neural age clock.** **a** Training history of the 10 individual neural networks. Depicted is the loss on the held-out testing set, over the full range of 200 training epochs. **b** Predicted against actual chronological age for the held-out test set, with observations colored by absolute prediction error. **c** Heatmap showing distinct activations of pathway neurons for the test set samples stratified by age quantiles. **d** Pathway ranking based on Pearson correlation coefficients of pathway neuron activations and chronological age over the test set. The results shown are based on 100 permutations calculated for a model including an artificial control pathway consisting of 150 randomly sampled, unrelated genes as a baseline. Significance was determined using one-sided Wilcoxon rank-sum tests comparing the correlation estimates of the various pathways to the introduced control pathway, adjusted for multiple testing.

contrast, the knockdown of thioredoxin-interacting protein TXNIP, a major player in maintaining cellular redox-status and recently implied in the induction of senescence by its role in antagonizing AKT-signaling[25], reduced predicted ages significantly, in line with experimental data that shows that knockdowns of TXNIP increase life-span by reducing reactive oxygen species (ROS)-mediated stress in model organisms[26]. Moving away from model organisms, a null-mutation of SERPINE1 is one of the few causal associations

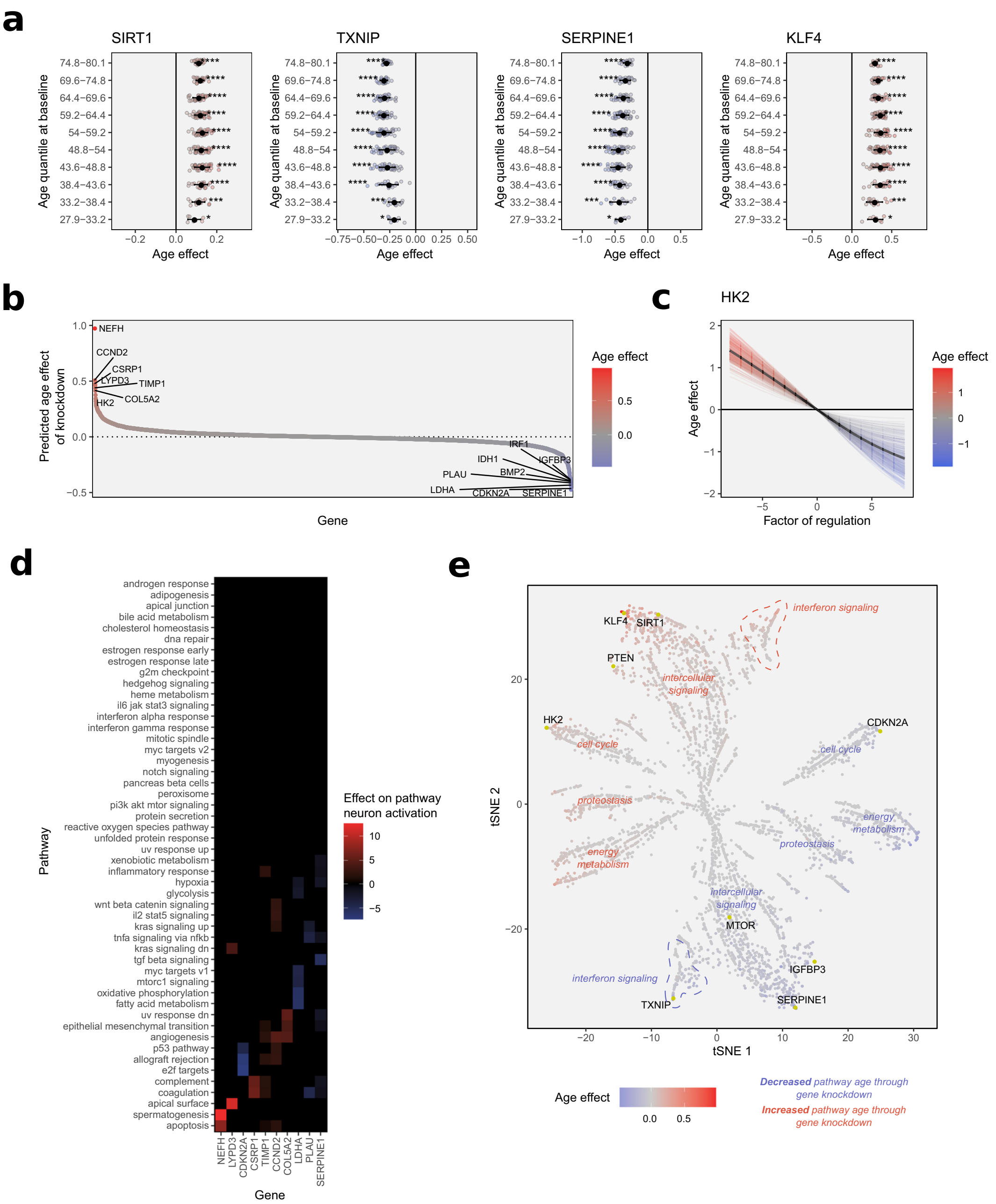

Fig. 3 Identifying relevant aging target genes through simulated gene perturbation. a Predicted effects of the in silico knockdowns of SIRT1, TXNIP, SERPINE1, and KLF4. Effect on age is stratified by chronological age quantiles of the test subjects. Significance was determined using one-sample Wilcoxon rank-sum tests, testing for the difference in medians from an effect size of 0, with *p*-values adjusted for multiple testing. Error bars show standard deviations. **b** Distribution of the predicted age effects of the in silico knockdown of all genes covered by the model. Genes at the upper and lower extremes can be regarded as the most important features of the model. **c** Predicted age effect of the simulated continuous knockdown and overexpression of HK2, one of the most important genes in the model. **d** Effects of simulated gene knockdowns by the most important genes according to impact on age estimation upon the activation of pathway neurons. **e** Two-dimensional embedding of the aging pathway landscape. The map was generated by assessing the effects of all gene knockdowns on pathway neuron activation and calculating a lower-dimensional embedding of the data using the tSNE algorithm. Genes are colored by the overall impact of their knockdown on the final age estimate. Thereby clustering of genes according to strength and direction of correlation to age, as well as functional pathway annotation, can be observed.

discovered so far, that links a single gene loss-of-function mutation directly with increased longevity in humans[27]. In line with the literature, simulated knockdown of the senescence-associated gene lead to a significant decrease in transcriptomic age predicted by the model. These simulations, while intended mainly as validation of the associations learned by the model, also highlight the utility of computational models for translational research, in this case, the ability to test the relevance of target genes identified in a systemic context or in other tissues to the biology of aging skin, which the model was trained on. An example of a gene association more specific to the skin however, is the knockdown of Krueppel-like Factor 4. KLF4 is, among others, a stemness factor and direct regulator of telomerase expression[28], as well as importantly a regulator of keratinocyte senescence[29]. As

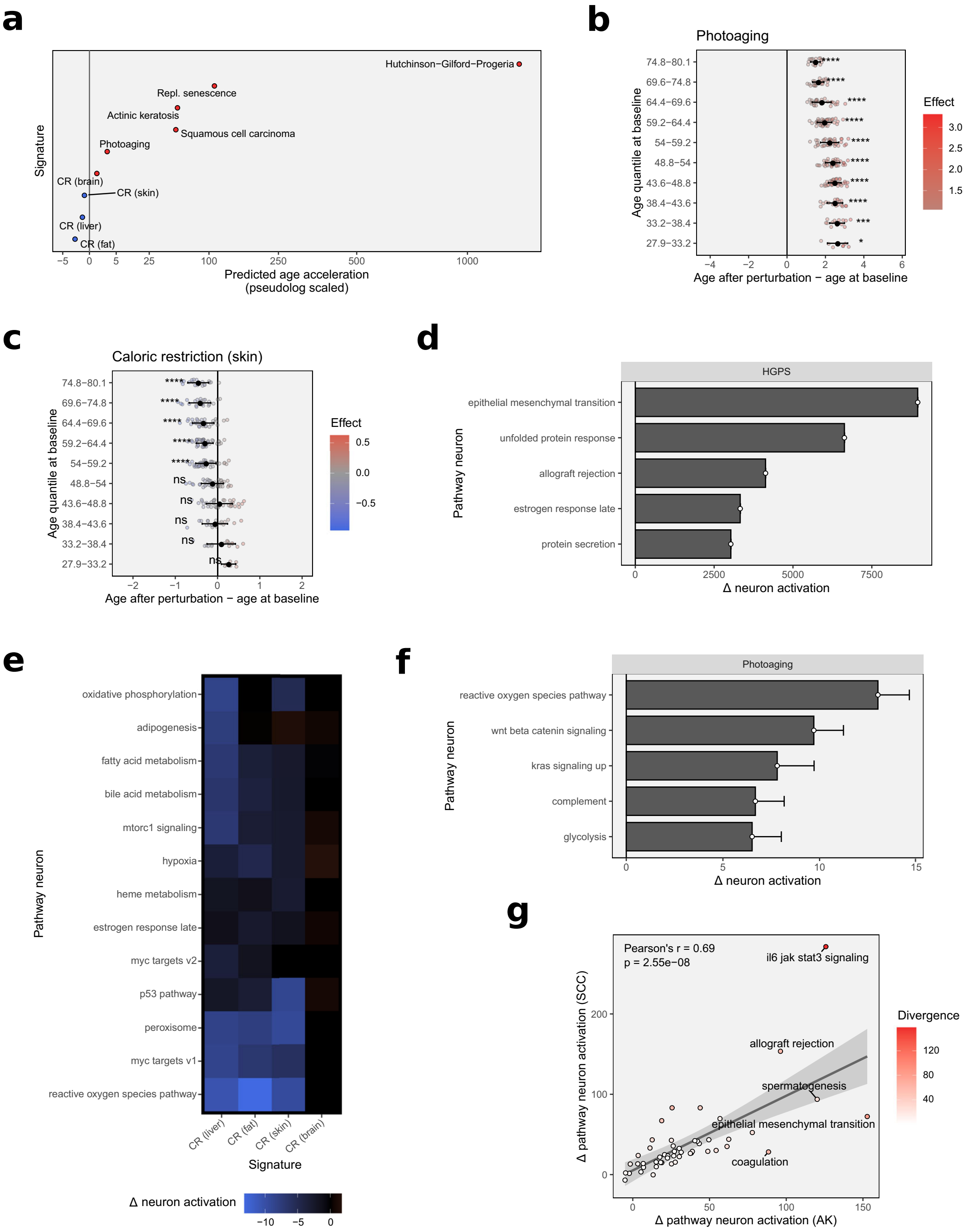

Fig. 4 Assessing the impact of age- and disease-related gene expression signatures on transcriptomic age and pathway aging states. **a** Overview of the average predicted effect of the transcriptomic perturbation using multiple age- and disease-related signatures. **b** Predicted effect of the transcriptomic perturbation using a signature of chronically sun-exposed skin, stratified by chronological age quantiles. Significance was determined using one-sample Wilcoxon rank-sum tests, testing for the difference in medians from an effect size of 0, with *p*-values adjusted for multiple testing. Error bars show standard deviations. **c** Predicted effect of the transcriptomic perturbation using a caloric restriction signature, stratified by chronological age quantiles. Significance was determined using one-sample Wilcoxon rank-sum tests, testing for the difference in medians from an effect size of 0, with *p*-values adjusted for multiple testing. Error bars show standard deviations. **d** Effect of the transcriptional signature of Hutchinson–Gilford progeria syndrome on pathway neuron activation. Shown are the five most strongly affected pathways. Error bars show standard deviations. **e** Heatmap showing the effects of tissue-specific caloric restriction signatures on pathway neuron activation. **f** Effect of the transcriptional signature of photoaging on pathway neuron activation. Shown are the five most strongly affected pathways. Error bars show standard deviations. **g** Impact of actinic keratosis (AK) and cutaneous squamous cell carcinoma (SCC) signatures on pathway neuron activations.

such, KLF4 silencing alone has been shown to be sufficient to induce a senescent phenotype in human keratinocytes[29]. In line with these findings, the simulated knockdown resulted in an increased age prediction across subjects of all ages.

Systematic knockdown simulations identify known and novel aging target genes

As the knockdowns of selected literature-based aging target genes had recapitulated experimental findings, we then extended the knockdown to the rest of the transcriptome, at least insofar as it was covered by the Hallmarks pathway annotation database and therefore represented in the model. Simulating the knockdown of all genes by a log2 fold-change of −2 revealed an approximately equal distribution of age increasing and decreasing knockdowns, ranging from around +1 to −0.5 years in effect sizes (Fig. 3b). Among the highest-scoring knockdowns of all genes were several well-described aging marker genes, such as SERPINE1, IGFBP3, CDKN2A, and TIMP1, as well as some less intensely studied genes such as HK2, a hexokinase whose expression has previously been reported to diminish with increasing age in the skin, with potentially detrimental effects on energy metabolism and epidermal cell proliferation[30]. The simulated overexpression of HK2 on the other hand was concordantly predicted by the model as a rejuvenating intervention (Fig. 3c), highlighting the utility of interpretable machine learning models to discover novel angles and targets for potential intervention strategies.

Observing the effects of the most influential gene knockdowns on pathway neuron activation revealed that interestingly all of them mediated their effect via at least two distinct pathways (Fig. 3d), indicating that genes at the crossroads of several pathways might exert a larger influence on the final age estimate, which was confirmed by association testing (Supplementary Fig. 3a) for both positive ($p = 2.6e{-}116$) and negative impact genes ($p = 2.4e{-}153$). This indicates that the network architecture organically increases the impact of master regulators and genes which act as effectors in several different biological processes. This emergent property is very much desirable, as it reflects the underlying biology more closely than other machine learning models that tend to weight features purely based on predictivity or correlation to the modeled phenotype, rather than by the breadth of their biological impact. We subsequently expanded the pathway impact analysis to all genes covered by the model and found that using the single-gene knockdown data allowed reconstruction of the aging pathway landscape, with genes arranged by similarity in effect as well as capturing the structure of the diverse biological motifs and processes. The resulting map (Fig. 3e) demonstrates the gain in interpretability awarded by this new type of clock, allowing the visual inspection of gene–pathway relationships in the context of aging, unlike any previous age clock.

Predicting the impact of complex transcriptional signatures on biological aging state

We then set out to evaluate the impact of more complex aging-related transcriptional signatures on model prediction. This analysis served two purposes: (i) investigate if the model recapitulates the overall effect of the signature and (ii) demonstrate the use of an interpretable machine learning model in deciphering the biological processes driving accelerated aging or rejuvenating conditions. For this, we searched the literature for gene expression data or published signatures of diverse aging-related conditions and simulated their impact on the predicted age of the test set (Fig. 4a).

The most prominent example of an accelerated aging disorder is the Hutchinson–Gilford progeria syndrome (HGPS). HGPS is a rare autosomal dominant genetic disorder that manifests very early in life, with symptoms that strikingly resemble those of natural aging particularly in regards to the skin, including wrinkle formation, the emergence of dyspigmentations (age spots), and a general thinning of the skin including a loss of subcutaneous fat, as well as alopecia[31]. The condition is caused by mutations leading to incorrectly processed forms of lamin A that weaken the nucleus structure with diverse detrimental consequences. The overall effects of this are severe, and the average life expectancy for patients is only between 13 and 15 years[31,32]. Simulating the effect of the transcriptomic signature of HGPS[33] likewise has a heavy impact on age estimation, with the clock putting out predictions beyond 1200 years after signature application (Fig. 4a). Though these numbers might at first seem absurdly high, they are easily explained considering the clock was trained on data of a natural aging progression. The fact that predictions are exceeding this scale is caused by the underlying learned mathematical model and signals that, while the model clearly assesses HGPS or aspects of HGPS as an accelerated aging condition, the transcriptomic state seen in HGPS is shifted far beyond that of the natural physiological aging progression. The effect size can therefore be interpreted as a manifestation of the pathophysiology of the underlying condition, in sync with the low life expectancy of individuals suffering from HGPS.

A milder form of accelerated aging, one that specifically affects the skin, can be observed in the form of photoaging. Caused by the chronic exposure of the skin to solar irradiation, photoaging is an extrinsically accelerated aging phenotype, characterized by wrinkling, dyspigmentation, and a leathery appearance of the skin[11]. Simulating the impact of the signature of chronically sun-exposed skin[34] increases the predicted age by around 2.1 years on average (Fig. 4b). This result is again in line with expectations but importantly demonstrates that the clock is sensitive enough to be used to detect smaller transcriptional alterations caused by exogenous stressors that affect aging, such as chronic sun exposure.

Further unprotected from the damages of solar irradiation, photoaged skin can over time develop into scaly pre-cancerous lesions known as actinic keratoses (AKs). AKs, caused by the intraepidermal proliferation of atypical keratinocytes, are a frequently diagnosed skin condition in light-skinned individuals with a history of sun exposure[35]. Although themselves often asymptomatic, around 10% of all AK lesions progress into cutaneous squamous cell carcinoma (SCCs) if left untreated[35–37], one of the most common types of cancer in developed countries with predominantly fair-skinned populations[38]. Due to the direct link between photoaging and the emergence of AKs, and the direct progression path from AKs to SCCs, we decided to include signatures from these pre-cancerous and cancerous tissues into the analysis. Interestingly, both signatures[39] induced substantial increases in predicted age across all samples, on average by 54 and 52 years, respectively (Fig. 4a). Considering the hyperproliferative traits of both disorders this might appear somewhat counter-intuitive, then again, the relationship between aging and cancer is complex, and several shared mechanisms between the two have been identified over the years[40], let alone the fact, that age remains one of the greatest single risk factors for the development of cancer overall[41].

A key feature of aging that is lately receiving increasing attention, and also happens to play an important role in tumorigenesis, is the accumulation of senescent cells in aging tissues. Likely evolved as a cancer protection mechanism, senescence describes the cessation of cell division, induced by extrinsic stress or replicative exhaustion. Senescent cells influence their surrounding tissue by secreting a complex proinflammatory mixture of cytokines, growth factors, and proteases[42]. This senescence-associated secretory phenotype (SASP) plays an important role in the recruiting of immune cells to the tissue, and as such has beneficial functions in wound healing and tissue regeneration[43]. In aging tissues however, the increasing accumulation of senescent cells impairs normal tissue function, and SASP has been proposed as one of the mechanisms that drive inflammation, the chronic low-grade inflammatory state of aging tissue[44,45]. As senescence is an important aspect of aging and also a common in vitro model of aging, we tested the signature of replicative exhaustion-induced senescence using the model. The simulations showed an increase in age of over 100 years on average (Fig. 4a), which is the strongest impact of any signature we recorded apart from HGPS. It should be noted here, that previous experiments calculating the DNA methylation age of fibroblasts in culture have estimated cells aging around 62× faster in vitro[46], which could factor into these predictions as well. Irrespective of this, the data shows that the clock not only accurately captures aging in vivo but also models processes that define aging in vitro, adding to its utility. The sensitivity of the model towards senescence also delivers one potential mechanism explaining the pronounced age acceleration predicted by the model for the AK and SCC signatures, as an accumulation of senescent cells is not only a feature of aging tissues but also frequently observed in precancerous and cancerous lesions, including AKs and SCCs[47,48].

Next, we were interested in seeing if the model was also capable of recapitulating the positive effects of lifespan-extending intervention strategies. Most data on pro-longevity interventions stem from experiments with model organisms, but one of the advantages of the presented in silico approach is the opportunity to transfer such settings into a human model and simulate the effects of such treatment in human tissue. The most reliable and well-documented form of pro-longevity intervention is caloric restriction[49]. The reduction of caloric intake has been shown to increase health- and life span in a large number of organisms of varying size and complexity, including roundworms, flies, mice, rats, and even non-human primates[50]. It is therefore believed to be a conserved mechanism among animals, although its effectiveness in terms of lifespan extension has yet to be proven in humans. Data from model animals are generally amply available, we did however only identify a single recently published dataset that included the transcriptional patterns triggered by caloric restriction in skin tissue, which was based on *Rattus norvegicus* samples[51]. Mapping the gene signatures from this dataset to their human homologs allowed testing the signature with the age clock and simulate its effects on human aging. The signature indeed shifted the aging transcriptome landscape to a younger state by around 0.2 years on average, although the effect was only statistically significant for subjects above 50 years (Fig. 4c). Despite its low effect size, this indicates that caloric restriction might indeed have beneficial effects in humans, and ones that might favorably affect skin biology. The data also points to the existence of an age-dependency of these effects, a theory that has interestingly been proposed before and is backed by experimental data from mice showing that the beneficial impact of the intervention, while significant in adult animals, is lacking in younger specimens[52]. Conceptually this has been explained with caloric restriction largely mediating alterations to biological processes that accumulate throughout age, therefore lacking an impact on young organisms, when these processes still operate smoothly, and scarcity is more likely to impair normal functioning[53]. The age-dependency predicted by our model would further seem to support these hypotheses. As most molecular analyses of the effects of caloric restriction have been performed in other tissues though, we expanded our simulations to the signatures generated from liver, fat, and brain tissue[51]. The predicted rejuvenation of both liver, as well as fat signatures, was greater, reducing age estimates by 0.4 and 1.5 years, respectively (Fig. 4a). As these tissues are more immediately involved with and affected by caloric restriction schemes, this appears plausible. Surprisingly however, the brain signature lead to divergent results and caused the model to predict a small but significant age acceleration by 0.4 years on average. While this may simply be an artifact of tissue-specific gene regulation, one might speculate on the involvement of a biological component as well. Being the most demanding organ in terms of energy needs in most animals, it is conceivable that the brain would be the organ most immediately affected by negative repercussions of decreased caloric intake, which could help explain the finding. This theory is supported by data from non-human primates under caloric restriction, that—despite showing significant life-span extension—suffered from an accelerated loss of gray brain matter, albeit without affecting cognitive performance[54].

Decoding the pathways implicated in accelerated aging and pro-longevity phenotypes

Seeing that the model was capable of recapitulating both accelerated aging and pro-longevity interventions in the form of caloric restriction, we were interested in establishing the network's utility in deciphering the biological processes by which these conditions exerted their effects. For this, we analyzed the activations of the pathway neurons in the intermediate pathway output layer before and after perturbation with the respective signatures and monitored the changes induced in neuron activation.

The most substantial alterations to the pathway landscape were caused by the transcriptional signature of HGPS (Fig. 4d). The effects were dominated by a massively increased positive activation in the epithelial–mesenchymal transition pathway neuron, indicating a substantial shift in pathway states towards an older transcriptome, but far surpassing the originally modeled range. Epithelial–mesenchymal transition describes the process of epithelial cells losing their polarity and gaining functions allowing them to migrate and gain mesodermal character. This process, while originally observed during embryogenesis, has since been shown to be a crucial mechanism in the metastasis of cancers, during wound healing, and—importantly—in the manifestation of fibrosis[55,56]. The cause of death in patients suffering from HGPS is usually found in cardiovascular complications from substantial levels of atherosclerosis, but interestingly in the absence of typical risk factors such as increased L-LDL or C-reactive protein[57], and with more prominent signs of vascular fibrosis than typically observed in patients suffering from cardiovascular disease[58]. Interestingly then, the most strongly affected pathway identified by the model is one with a direct connection to the most severe clinical feature of HGPS, which might warrant further investigation, especially since this pathway has not received a lot of attention in studying the disease progression of HGPS thus far. Other noteworthy pathways that were strongly affected by the signature were related to proteostasis and protein secretion, immune signaling, and the estrogen response (Fig. 4d), several of which are not only well described Hallmarks of Aging[59] but have also previously been associated with HGPS[60].

In contrast to the HGPS signature, analyzing the pathways impacted by caloric restriction revealed a number of processes shifted towards a younger state (Fig. 4e). The effects were generally similar between tissues, with the exception of the brain-derived signature, which showed no substantially rejuvenated pathways at all. The processes that were most prominently shifted towards a favorable state were related to ROS, peroxisome pathways, and to a lower extent mTOR-signaling and general metabolism across all tissues. Reduced production of ROS through a slowing of the metabolic rate, thereby reducing the load of oxidative stress, is one of the very key mechanisms proposed by which caloric restriction is believed to exert its life-span extending effects, the observed changes in pathway states are therefore very much in line with existing theories and reports[61,62]. Another well-described effect of restricting caloric intake is the reduction of mTOR activity, marking one of the most reliable single mechanisms to prolong lifespan in various model organisms from fruit flies to non-human primates[63–65]. The rejuvenating impact on mTOR-signaling predicted by the model is therefore again very much in line with existing data, as are naturally the observed effects on metabolic pathways, including oxidative phosphorylation and fatty acid oxidation in mitochondria and peroxisomes. Interestingly though, the skin-derived signature appeared to have a lower impact on metabolic pathways but instead showed a more strongly rejuvenated profile associated with p53-signaling, which is an interesting finding considering its crucial role in cancer protection in the skin[66]. Notably, caloric restriction has been shown to delay carcinogenesis and tumor-related mortality in rodents[67,68] and rhesus monkeys[69,70], this finding could therefore be suggestive of another potential benefit of caloric restriction for skin biology. It should be noted that as these results represent a translation from rodents to human biology, so a margin of error is to be expected. The analysis does however highlight the potential of interpretable machine learning to use available data from animal experiments and to explore the translation of findings to a model of human biology in a virtual setting.

The effects of the photoaging signature were similarly diverse, with the strongest impact also recorded on the ROS pathway (Fig. 4f), here substantially shifting the pathway towards an older state. Further processes altered in this direction were related to Wnt and Kras signaling, and metabolic pathways such as glycolysis. Interestingly a couple of pathway states appeared shifted towards a younger profile, most notably involving the G2 damage checkpoint and the estrogen response pathways. The effects of a chronic exposure to solar irradiation that over time lead to the manifestation of photoaging, are believed to be primarily driven by oxidative damage resulting from the UV-induced formation of ROS[11,71–73]. The predominant pathway identified by the model very much supports this hypothesis. Metabolic changes in photoaged skin have likewise been reported[34]. Data on Wnt modulation in association with photoaging is sparser, but recent reports implicate the pathway in the response following UVB irradiation in keratinocytes in vitro[74]. Given its function as an important mediator of cell proliferation and differentiation and importantly its essential role in regulating adult epidermal stem cell reservoirs, regulatory alterations in Wnt signaling could potentially be an important mechanism driving the gradual thinning of the epidermis frequently observed in (photo-)aged skin[11,75].

Finally, we investigated the similarity in pathway neuron activation following perturbation using the AK and SCC signatures. Although the progression from AKs to SCCs, in general, is well-described, only around 10% of all AK lesions develop into actual carcinoma[35–37]. The exact mechanisms determining which AKs progress meanwhile remain elusive. Analyzing the predicted pathway perturbations revealed a substantial correlation between pathway patterns induced by AK and SCC signatures (Fig. 4g). Given the reported progression path, this finding seems conclusive. The analysis also revealed a number of pathways that were notably more strongly deregulated than others, mainly related to IL6-JAK-STAT-signaling, immune pathways and coagulation, a gene set that contains many genes related to the complement system as well as senescence-associated genes such as SERPINE1. The latter is particularly interesting, as the prolonged expression of the senescence marker gene CDKN2A has very recently been shown to induce hyperplasia in the epidermis of mice very similar to the early stages of AKs by increasing proliferation of surrounding keratinocytes, implicating senescent cells as one of the early mechanisms in epidermal tumorigenesis[76]. The comparably lower activation in the SCC signature suggests that the impact of senescence-associated genes is higher in the early stages leading to AK lesions though, which fits the experimental data available[76]. Among the processes that showed notable divergences between AKs and SCCs as well were immune and JAK-STAT-signaling, both found more strongly altered by the SCC signature. The involvement of immune-related genes contained in the allograft rejection gene set is of little surprise given that alterations to immune signaling in cancer are well-documented, the increased activation induced by the SCC signature does however highlight a very important characteristic of SCCs, which is their ability to evade immune surveillance, setting it apart from pre-cancerous AK lesions[77]. Aberrant activation of JAK-STAT-signaling is a frequently reported feature in human cancers as well[78], and SCCs are no exception[79]. Constitutive activation of STAT3 has in fact been shown to be a key event in the SCC tumorigenesis[80], validating the model's predictions. Surprisingly little is known about the state of the IL6-JAK-STAT axis in AKs however and seeing the diverging pathway patterns uncovered by our model and the documented importance of the pathway in tumorigenesis would therefore encourage further investigations into this pathway in AK lesions to help explain the observed heterogeneity in AK to SCC progression.

Despite their popularity and unquestionable utility as biomarkers, age clocks have thus far generated little insight into the processes that actually drive the aging progression or provoke phenotypical manifestations of biological aging. Here we present a new type of age clock, that delivers unprecedented interpretability to its inner workings. Through the incorporation of prior information on pathways into the structure of the model, the learning process is tied to known biological processes, allowing their states to be interpreted in the activation of intermediate neurons in the neural network. While not surpassing other age clocks in terms of sheer accuracy, the model's performance is comparable with other published as well as a 'black box' transcriptomic age clock trained on the same data and offers greatly expanded utility beyond the use as a readout tool. We would argue that this property is more desirable in a research setting than mere predictivity and would like to see more efforts to increase the interpretability of machine learning models applied in aging research and biological research in general. Neural networks, in particular, present themselves as a very promising technology to fully unlock the potential of such approaches in an area of research that, due to the inherent breadth and complexity of the biological processes involved and ever-increasing amounts of high-throughput data available, is predestined to benefit from further technological advancements in machine learning.

METHODS

Study of Health in Pomerania (SHIP)

SHIP was designed as a population-based study to assess the prevalence and incidence of common clinical diseases, subclinical disorders, and risk factors among the population of the Federal State of Mecklenburg/West Pomerania in Northeastern Germany[12]. Examinations of the original cohort of 4308 randomly sampled subjects between 20 and 79 years started in 1997, with two follow-up examinations being performed after intervals of 5 and 11 years. The second cohort (SHIP-TREND), comprising another random sample of 4420 adults aged 20–79 years, started in 2008, again designed with regular follow-ups. The data used in this study consisting of 887 epidermal samples were collected during the first follow-up of the SHIP-TREND cohort, with subjects aged between 30 and 89 years (Supplementary Fig. 1a and b). The study was approved by the ethics committee of the University Medicine Greifswald (ethics approval number BB 39/08). All participants signed an informed consent form and all investigations were undertaken in accordance with the ethical principles outlined in the Declaration of Helsinki.

Tissue sample preparation

The suction blister method applied in this study has been approved by the Ethics Commission of the University of Freiburg (general approval December 8, 2008; Beiersdorf AG No. 28857). Suction blisters of 7 mm diameter were taken from the volar forearms of all subjects as previously described[81].

Nucleic acid extraction

As previously described[16], tissue samples were suspended in the respective lysis buffers for DNA or RNA extraction and homogenized using an MM 301 bead mill (Retsch). DNA was then extracted using the QIAamp DNA Investigator Kit (Qiagen) according to the manufacturer's instructions. RNA was extracted using the RNeasy Fibrous Tissue Mini Kit (Qiagen) according to the manufacturer's instructions.

Transcriptome sequencing

Transcriptome libraries were prepared using the TruSeq Library Prep Kit (Illumina) and sequencing performed at 1×50 bp on Illumina's HiSeq system to a final sequencing depth of 100 million reads per sample. Sequencing data were processed using a custom pipeline including Fastqc 0.11.7[82] for quality control, Trimmomatic 0.36[83] for trimming, and Salmon 0.8.1[84] for read mapping against the GRCh38 build of the human transcriptome and read quantification in the form of transcripts per million (TPM).

Pathway-based neural network architecture

The network was implemented using keras[85] with a tensorflow[86] backend and fully coded in R 3.6.1[87]. In the following and for the purpose of this work, we will use the term "pathway" to denote any gene sets or knowledge-guided collections of genes involved in distinct biological processes. In order to embed this pathway information into the network, first a binary 'gene × pathway' filter matrix was constructed based on gene annotations to the Hallmark pathway collection[13]. This filter matrix was used to set the crucial gene-specific connections between input neurons and the neurons in the first pathway layer. The following hidden layers operated in a pathway-centric manner. Neurons assigned to the same pathway were densely connected to each other to allow the network maximum flexibility to process and learn pathway representations from the data, while no connections to neurons of other pathways were allowed, as this would break the chain of interpretability. Information of each pathway was then aggregated in a final neuron, serving a dual purpose as both a step to condense the pathway information in one neuron per pathway and as an auxiliary output of pathway neuron activations to update the network loss during training and for further analysis purpose during inference. Finally, this pathway output layer was connected to a common output neuron in the last layer, tasked with aggregating the information passed by the pathway neurons to a final age estimate. The number of neurons within the hidden layers was adjusted to the number of genes in each pathway and thus determined for every pathway individually as shown in Eq. (1):

$$\text{number of neurons} = 5 + \left(\frac{\text{number of genes}}{f}\right) \tag{1}$$

This established a minimal size of 5 neurons per layer for each pathway, with additional neurons awarded with increasing pathway size to accommodate an increase in regulatory complexity. The neuron scaling factor *f* that determined the number of neurons added per additional gene was set to 2 in the final model (Supplementary Fig. 5a). The number of hidden layers was set to 4, as testing with more layers showed no additional gains in accuracy justifying a further increase in complexity (Supplementary Fig. 5b). Taken together, this setup resulted in a final network with 1,740,858 trainable parameters. In order to improve generalization ability and control overfitting of the model, dropout layers were inserted between the hidden layers, randomly dropping connections between the hidden layers in the training phase. Furthermore, global weight decay (regularization factor = 0.01) was implemented as another form of regularization, improving generalization ability of the model.

The model used 'elu' (exponential linear units) activation functions[88] in all hidden layers, and was accordingly initialized using the He-initialization, a weight initialization scheme optimized for 'relu'-like activation functions[89].

The loss function for model training combined two individual losses, calculated from the mean squared error (MSE) of the main and auxiliary outputs of the network, joined together by a balancing hyperparameter *alpha* as shown in Eq. (2):

$$\text{loss} = (1 - alpha) * \text{MSE}_{\text{main}} + alpha * \text{MSE}_{\text{auxiliary}} \tag{2}$$

The advantages of this are two-fold: (i) It forces all parts of the network to be trained, ensuring that the all encoded information is utilized, and all pathway neurons are active. This is critical, as early testing showed that without the added auxiliary loss, the network would heavily rely on only one or few pathways, the selection of which varied greatly depending on initial weight configuration (Supplementary Fig. 4a). This resulted in very poor reproducibility between network reboots and only a fraction of the available information being utilized. (ii) All pathway neurons now generate a positive continuous output, which is essentially an age estimate based on the information encoded in the pathway or 'pathway age'. This has clear benefits for the interpretability of the neuron activations, whose scale and direction could otherwise vary greatly between network reboots and which stabilized significantly through the addition of the auxiliary loss (Supplementary Fig. 4b). *Alpha* was set to 0.4 in the final model after testing different configurations (Supplementary Fig. 5c).

The training of the model was performed using stochastic gradient descent with Adam[90] and a learning rate of 0.001, with a mini-batch size of 16 samples for a total of 200 epochs. Table 1 summarizes the parameters of the pathway-based neural network.

Table 1. Pathway-based neural network parameters.

Parameter	Value
Number of input genes	4359
Number of input pathways	50
Number of parameters	1,740,858
Activation function	elu
Weight initialization	He
L2-regularization (weight decay)	0.01
Dropout rate (drop probability)	0.1
Loss calculation	Mean squared error (main and aux. output)
Hyperparameter alpha	0.4
Optimizer	Adam
Learning rate	0.001
Mini-batch size	16
Training epochs	200
Training samples	620
Test samples	267

Ensemble setup

In order to further improve both reproducibility and accuracy of the model, the final setup was designed as an ensemble of several individual networks. For this, 10 single networks were trained separately, and then joined to a common input layer and a shared main and auxiliary output. In the shared output layers, individual outputs by the 10 networks are averaged to generate the final model estimates. The ensemble setup proved successful in further stabilizing the intermediate neuron activations and thereby improving reproducibility (Supplementary Fig. 4c).

Fully connected neural network

To assess any potential trade-off between transparency and model precision, we trained an ensemble of 10 fully connected neural networks with the same number of layers per network, a comparable number of parameters, trained for the same number of epochs on the same data with the same training/test split as used for our pathway-based model. Table 2 summarizes the parameters of the fully connected neural network.

Assessment of visual age and association analysis

In order to generate estimates of phenotypic aging state to compare with the transcriptomic age estimates by the model, we used portrait images of 154 randomly sampled subjects from the test set. The images were captured in a standardized setup, taking evenly lit (through the use of a flash diffuser), non-polarized and color-controlled frontal portrait images of the test subjects with their eyes closed, any hair (except facial hair) covered to reduce the impact of features unrelated to the skin, and any make-up removed beforehand. The images were then presented to a blinded panel of 31 experts that were asked to estimate the ages of the subjects based on these photographs. The individual age estimates were then averaged over the panel, which resulted in the final visual age estimates, which showed generally very good concordance with chronological ages with a median absolute error of 4.38 years. Linear models were then employed in R[87] to test for an association between transcriptomic and visual age estimates, whilst adjusting for chronological age and gender (Supplementary Table 1).

Table 2. Fully connected neural network parameters.

Parameter	Value
Number of input genes	4359
Number of neurons per hidden layer	[350,350,350,50]
Number of parameters	1,789,301
Activation function	elu
Weight initialization	He
L2-regularization (weight decay)	0.01
Dropout rate (drop probability)	0.1
Loss calculation	Mean squared error
Optimizer	Adam
Learning rate	0.001
Mini-batch size	16
Training epochs	200
Training samples	620
Test samples	267

Table 3. Signatures used for perturbation experiments.

Signature	Species	Tissue	Technology	Ref.
Photoaging	*Homo sapiens*	Skin	Microarray	34
Hutchinson–Gilford Progeria	*Homo sapiens*	Skin	Microarray	33
Replicative senescence	*Homo sapiens*	Skin	RNA seq	99
Actinic keratosis	*Homo sapiens*	Skin	RNA seq	39
Cutaneous squamous cell carcinoma	*Homo sapiens*	Skin	RNA seq	39
Caloric restriction	*Rattus norvegicus*	Skin Liver Fat Brain	RNA seq	51

In silico gene knockdown and overexpression experiments

The perturbation of single genes was performed by up- or downregulating gene expression by a common log2 fold-change (which was −2 for all knockdown experiments, unless otherwise specified) in all samples of the test set ($n = 267$) and comparing the model's predictions with the unperturbed baseline predictions per sample. For the assessment of age impact, the changes in the main output neuron generating the overall age estimate were analyzed. For assessing the impact on the aging state of the biological pathways, the activity of the auxiliary output neurons was monitored instead, and the generated outputs of these neurons were similarly analyzed by comparing the 'pathway age' estimates with the unperturbed baseline estimates per sample.

The map of the aging pathway landscape shown in Fig. 3e was generated by embedding the perturbation effects from all gene knockdowns on each of the auxiliary pathway neurons using t-distributed stochastic neighbor embedding (tSNE) into two new dimensions[91], using the implementation of the algorithm in the routine R package[92].

Mapping *Rattus norvegicus* genes to human homologs

Rattus norvegicus genes from the caloric restriction signatures (genome build Rnor_6.0) were mapped to their human homologs (genome build GRCh38) using the biomaRt R package[93].

Perturbation experiments using complex gene expression signatures

Assessing the impact of more complex transcriptional signatures was performed by up- or downregulating each significantly differentially regulated gene (cutoff was an FDR < 0.05) in the signature by the exact effect size (determined by its log2 fold-change) recorded by the differential gene expression analysis. The analysis was again performed using all samples of the test set ($n = 267$) and comparing the predictions of the perturbed data with the unperturbed baseline predictions per sample, as with the single gene knockdowns. Significance of impact was determined using one-sample Wilcoxon rank-sum tests, testing for the difference in medians from an effect size of 0. When more than one comparison was performed, *p*-values were adjusted for multiple testing using the Holm–Bonferroni method[94]. Table 3 shows a summary of the signatures used for the perturbation experiments.

General data analysis and visualization

Data analysis in R[87] further included the usage of the packages data.table[95] and dplyr[96] for data handling, as well as the packages ggplot2[97] and ggpubr[98] for data visualization.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

Due to German and EU privacy legislation and the sensitive nature of the SHIP data, the RNA sequencing data generated in this study are available from the SHIP consortium[12] upon official request only (https://www.fvcm.med.uni-greifswald.de/dd_service/data_use_intro.php). For any questions or assistance with the process of accessing the data please contact the consortium via transfer@uni-greifswald.de.

CODE AVAILABILITY

All code used to generate the results shown in this study is available for research purposes from the corresponding authors upon request.

Received: 12 August 2020; Accepted: 26 April 2021;
Published online: 01 June 2021

REFERENCES

1. Bocklandt, S. et al. Epigenetic predictor of age. *PLoS ONE* **6**, e14821 (2011).
2. Hannum, G. et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **49**, 359–367 (2013).
3. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).
4. Holly, A. C. et al. Towards a gene expression biomarker set for human biological age. *Aging Cell* **12**, 324–326 (2013).
5. Peters, M. J. et al. The transcriptional landscape of age in human peripheral blood. *Nat. Commun.* **6**, 1–14 (2015).
6. Hertel, J. et al. Measuring biological age via metabonomics: the metabolic age score. *J. Proteome Res.* **15**, 400–410 (2016).
7. Mamoshina, P. et al. Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification. *Front. Genet.* **9**, 242 (2018).
8. Fleischer, J. G. et al. Predicting age from the transcriptome of human dermal fibroblasts. *Genome Biol.* **19**, 221 (2018).
9. Tanaka, T. et al. Plasma proteomic signature of age in healthy humans. *Aging Cell* **17**, 5 (2018).
10. Ma, J. et al. Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods* **15**, 290–298 (2018).
11. Scharffetter-Kochanek, K. et al. Photoaging of the skin from phenotype to mechanisms. *Exp. Gerontol.* **35**, 307–316 (2000).
12. Völzke, H. et al. Cohort profile: the study of health in Pomerania. *Int. J. Epidemiol.* **40**, 294–307 (2011).
13. Liberzon, A. et al. The molecular signatures database hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
14. Hansen, L. K. & Salamon, P. Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.* **12**, 993–1001 (1990).
15. Bormann, F. et al. Reduced DNA methylation patterning and transcriptional connectivity define human skin aging. *Aging Cell* **15**, 563–71 (2016).
16. Holzscheck, N. et al. Multi-omics network analysis reveals distinct stages in the human aging progression in epidermal tissue. *Aging* **12**, 12393–12409 (2020).
17. Galkin, F. et al. Biohorology and biomarkers of aging: current state-of-the-art, challenges and opportunities. *Ageing Res. Rev.* **60**, 101050 (2020).

18. Gladyshev, V. N. Aging: progressive decline in fitness due to the rising deleteriome adjusted by genetic, environmental, and stochastic processes. *Aging Cell* **15**, 594–602 (2016).
19. Harman, D. Aging: a theory based on free radical and radiation chemistry. *J. Gerontol.* **11**, 298–300 (1956).
20. Williams, G. C. Pleiotropy, natural selection, and the evolution of senescence. *Evolution* **11**, 398–411 (1957).
21. Boily, G. et al. SirT1 regulates energy metabolism and response to caloric restriction in mice. *PLoS ONE* **3**, e1759 (2008).
22. Herranz, D. et al. Sirt1 improves healthy ageing and protects from metabolic syndrome-associated cancer. *Nat. Commun.* **1**, 3 (2010).
23. Satoh, A. et al. Sirt1 extends life span and delays aging in mice through the regulation of Nk2 homeobox 1 in the DMH and LH. *Cell Metab.* **18**, 416–430 (2013).
24. Kim, D. H., Jung, I. H., Kim, D. H. & Park, S. W. Knockout of longevity gene Sirt1 in zebrafish leads to oxidative injury, chronic inflammation, and reduced life span. *PLOS ONE* **14**, e0220581 (2019).
25. Huy, H. et al. TXNIP regulates AKT-mediated cellular senescence by direct interaction under glucose-mediated metabolic stress. *Aging Cell* **17**, 6 (2018).
26. Oberacker, T. et al. Enhanced expression of thioredoxin-interacting-protein regulates oxidative DNA damage and aging. *FEBS Lett.* **592**, 2297–2307 (2018).
27. Khan, S. S. et al. A null mutation in SERPINE1 protects against biological aging in humans. *Sci. Adv.* **3**, eaao1617 (2017).
28. Wong, C.-W. et al. Krüppel-like transcription factor 4 contributes to maintenance of telomerase activity in stem cells. *Stem Cells* **28**, 1510–1517 (2010).
29. Panatta, E. et al. Kruppel-like factor 4 regulates keratinocyte senescence. *Biochem. Biophys. Res. Commun.* **499**, 389–395 (2018).
30. Kuehne, A. et al. An integrative metabolomics and transcriptomics study to identify metabolic alterations in aged skin of humans in vivo. *BMC Genomics* **18**, 169 (2017).
31. Hennekam, R. C. M. Hutchinson–Gilford progeria syndrome: review of the phenotype. *Am. J. Med. Genet. A* **140**, 2603–2624 (2006).
32. Gordon, L. B. et al. Impact of farnesylation inhibitors on survival in Hutchinson–Gilford progeria syndrome. *Circulation* **130**, 27–34 (2014).
33. Csoka, A. B. et al. Genome-scale expression profiling of Hutchinson–Gilford progeria syndrome reveals widespread transcriptional misregulation leading to mesodermal/mesenchymal defects and accelerated atherosclerosis. *Aging Cell* **3**, 235–243 (2004).
34. Yan, W. et al. Transcriptome analysis of skin photoaging in Chinese females reveals the involvement of skin homeostasis and metabolic changes. *PLOS ONE* **8**, e61946 (2013).
35. Röwert-Huber, J. et al. Actinic keratosis is an early in situ squamous cell carcinoma: a proposal for reclassification. *Br. J. Dermatol.* **156**, 8–12 (2007).
36. Glogau, R. G. The risk of progression to invasive disease. *J. Am. Acad. Dermatol.* **42**, 23–24 (2000).
37. Lambert, S. R. et al. Key differences identified between actinic keratosis and cutaneous squamous cell carcinoma by transcriptome profiling. *Br. J. Cancer* **110**, 520–529 (2014).
38. Armstrong, B. K. & Kricker, A. The epidemiology of UV induced skin cancer. *J. Photochem. Photobiol. B* **63**, 8–18 (2001).
39. Hoang, V. L. T. et al. RNA-seq reveals more consistent reference genes for gene expression studies in human non-melanoma skin cancers. *PeerJ* **5**, e3631 (2017).
40. Aunan, J. R., Cho, W. C. & Søreide, K. The biology of aging and cancer: a brief overview of shared and divergent molecular hallmarks. *Aging Dis.* **8**, 628–642 (2017).
41. National Cancer Institute (NCI). *SEER Cancer Statistics Review (CSR) 1975–2014*. https://seer.cancer.gov/archive/csr/1975_2014/ (2018).
42. Krtolica, A., Parrinello, S., Lockett, S., Desprez, P. Y. & Campisi, J. Senescent fibroblasts promote epithelial cell growth and tumorigenesis: a link between cancer and aging. *Proc. Natl Acad. Sci. USA* **98**, 12072–12077 (2001).
43. Demaria, M. et al. An essential role for senescent cells in optimal wound healing through secretion of PDGF-AA. *Dev. Cell* **31**, 722–733 (2014).
44. Salminen, A., Kaarniranta, K. & Kauppinen, A. Inflammaging: disturbed interplay between autophagy and inflammasomes. *Aging* **4**, 166–175 (2012).
45. Franceschi, C., Garagnani, P., Parini, P., Giuliani, C. & Santoro, A. Inflammaging: a new immune–metabolic viewpoint for age-related diseases. *Nat. Rev. Endocrinol.* **14**, 576–590 (2018).
46. Sturm, G. et al. Human aging DNA methylation signatures are conserved but accelerated in cultured fibroblasts. *Epigenetics* **14**, 961–976 (2019).
47. Hodges, A. & Smoller, B. R. Immunohistochemical comparison of p16 expression in actinic keratoses and squamous cell carcinomas of the skin. *Mod. Pathol.* **15**, 1121–1125 (2002).
48. Toutfaire, M. et al. Unraveling the interplay between senescent dermal fibroblasts and cutaneous squamous cell carcinoma cell lines at different stages of tumorigenesis. *Int. J. Biochem. Cell Biol.* **98**, 113–126 (2018).
49. Weindruch, R. & Walford, R. L. Dietary restriction in mice beginning at 1 year of age: effect on life-span and spontaneous cancer incidence. *Science* **215**, 1415–1418 (1982).
50. Lee, C. & Longo, V. Dietary restriction with and without caloric restriction for healthy aging. *F1000Res* **5**, 117 (2016).
51. Ma, S. et al. Caloric restriction reprograms the single-cell transcriptional landsc. *Rattus Norvegicus-. Aging Cell* **180**, 984–1001.e22 (2020).
52. Chen, C.-N. J., Lin, S.-Y., Liao, Y.-H., Li, Z.-J. & Wong, A. M.-K. Late-onset caloric restriction alters skeletal muscle metabolism by modulating pyruvate metabolism. *Am. J. Physiol. Endocrinol. Metab.* **308**, E942–949 (2015).
53. Chen, C.-N., Liao, Y.-H., Tsai, S.-C. & Thompson, L. V. Age-dependent effects of caloric restriction on mTOR and ubiquitin-proteasome pathways in skeletal muscles. *GeroScience* **41**, 871–880 (2019).
54. Pifferi, F. et al. Caloric restriction increases lifespan but affects brain integrity in grey mouse lemur primates. *Commun. Biol.* **1**, 1–8 (2018).
55. Kalluri, R. & Neilson, E. G. Epithelial–mesenchymal transition and its implications for fibrosis. *J. Clin. Investig.* **112**, 1776–1784 (2003).
56. Hill, C., Jones, M. G., Davies, D. E. & Wang, Y. Epithelial–mesenchymal transition contributes to pulmonary fibrosis via aberrant epithelial/fibroblastic cross-talk. *J. Lung Health Dis.* **3**, 31–35 (2019).
57. Gordon, L. B., Harten, I. A., Patti, M. E. & Lichtenstein, A. H. Reduced adiponectin and HDL cholesterol without elevated C-reactive protein: clues to the biology of premature atherosclerosis in Hutchinson–Gilford Progeria Syndrome. *J. Pediatr.* **146**, 336–341 (2005).
58. Olive, M. et al. Cardiovascular pathology in Hutchinson–Gilford progeria: correlation with the vascular pathology of aging. *Arterioscler. Thromb. Vasc. Biol.* **30**, 2301–2309 (2010).
59. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194–1217 (2013).
60. Vidak, S. & Foisner, R. Molecular insights into the premature aging disease progeria. *Histochem. Cell Biol.* **145**, 401–417 (2016).
61. Sohal, R. S. & Weindruch, R. Oxidative stress, caloric restriction, and aging. *Science* **273**, 59–63 (1996).
62. Redman, L. M. et al. Metabolic slowing and reduced oxidative damage with sustained caloric restriction support the rate of living and oxidative damage theories of aging. *Cell Metab.* **27**, 805–815.e4 (2018).
63. Harrison, D. E. et al. Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. *Nature* **460**, 392–395 (2009).
64. Taormina, G. & Mirisola, M. G. Calorie restriction in mammals and simple model organisms. *Biomed. Res. Int.* **2014**, 308690 (2014).
65. Unnikrishnan, A., Kurup, K., Salmon, A. B. & Richardson, A. Is rapamycin a dietary restriction mimetic? *J. Gerontol. A* **75**, 4–13 (2020).
66. Jiang, W., Ananthaswamy, H. N., Muller, H. K. & Kripke, M. L. p53 protects against skin cancer induction by UV-B radiation. *Oncogene* **18**, 4247–4253 (1999).
67. Hursting, S. D., Perkins, S. N. & Phang, J. M. Calorie restriction delays spontaneous tumorigenesis in p53-knockout transgenic mice. *Proc. Natl Acad. Sci. USA* **91**, 7036–7040 (1994).
68. Lv, M., Zhu, X., Wang, H., Wang, F. & Guan, W. Roles of caloric restriction, ketogenic diet and intermittent fasting during initiation, progression and metastasis of cancer in animal models: a systematic review and meta-analysis. *PLOS ONE* **9**, e115147 (2014).
69. Colman, R. J. et al. Caloric restriction delays disease onset and mortality in rhesus monkeys. *Science* **325**, 201–204 (2009).
70. Mattison, J. A. et al. Impact of caloric restriction on health and survival in rhesus monkeys: the NIA study. *Nature* **489**, 318–321 (2012).
71. Scharffetter-Kochanek, K. et al. UV-induced reactive oxygen species in photocarcinogenesis and photoaging. *Biol. Chem.* **378**, 1247–1257 (1997).
72. Wondrak, G. T., Jacobson, M. K. & Jacobson, E. L. Endogenous UVA-photosensitizers: mediators of skin photodamage and novel targets for skin photoprotection. *Photochem. Photobiol. Sci.* **5**, 215–237 (2006).
73. Rinnerthaler, M., Bischof, J., Streubel, M. K., Trost, A. & Richter, K. Oxidative stress in aging human skin. *Biomolecules* **5**, 545–589 (2015).
74. Michalczyk, T. et al. UVB exposure of a humanized skin model reveals unexpected dynamic of keratinocyte proliferation and Wnt inhibitor balancing. *J. Tissue Eng. Regener. Med.* **12**, 505–515 (2018).
75. Rittié, L. & Fisher, G. J. Natural and sun-induced aging of human skin. *Cold Spring Harb. Perspect. Med.* **5**, a015370 (2015).
76. Azazmeh, N. et al. Chronic expression of p16 INK4a in the epidermis induces Wnt-mediated hyperplasia and promotes tumor initiation. *Nat. Commun.* **11**, 2711 (2020).
77. Clark, R. A. et al. Human squamous cell carcinomas evade the immune response by down-regulation of vascular E-selectin and recruitment of regulatory T cells. *J. Exp. Med.* **205**, 2221–2234 (2008).

78. Thomas, S. J., Snowden, J. A., Zeidler, M. P. & Danson, S. J. The role of JAK/STAT signalling in the pathogenesis, prognosis and treatment of solid tumours. *Br. J. Cancer* **113**, 365–371 (2015).
79. Sriuranpong, V. et al. Epidermal growth factor receptor-independent constitutive activation of STAT3 in head and neck squamous cell carcinoma is mediated by the autocrine/paracrine stimulation of the interleukin 6/gp130 cytokine system. *Cancer Res.* **63**, 2948–2956 (2003).
80. Grandis, J. R. et al. Constitutive activation of Stat3 signaling abrogates apoptosis in squamous cell carcinogenesis in vivo. *PNAS* **97**, 4227–4232 (2000).
81. Südel, K. M. et al. Tight control of matrix metalloproteinase-1 activity in human skin. *Photochem. Photobiol.* **78**, 355–60 (2003).
82. Andrews, S. s-andrews/FastQC. *GitHub* https://github.com/s-andrews/FastQC.
83. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
84. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
85. Chollet, F. keras-team/keras. *GitHub* https://github.com/keras-team/keras.
86. Abadi, M. et al. TensorFlow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* 265–283 (The Advanced Computing Systems Association, 2016).
87. R Core Team. R: A Language and Environment for Statistical Computing. *The R Foundation* (2018).
88. Clevert, D. -A., Unterthiner, T. & Hochreiter, S. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). Preprint at arXiv:1511.07289 [cs] (2015).
89. He, K., Zhang, X., Ren, S. & Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. Preprint at arXiv:1502.01852 [cs] (2015).
90. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at arXiv:1412.6980 [cs] (2017).
91. Maaten, L. vander & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
92. Krijthe, J. jkrijthe/Rtsne. *GitHub* https://github.com/jkrijthe/Rtsne.
93. Durinck, S. & Huber, W. biomaRt: Interface to BioMart databases (i.e. Ensembl). *Bioconductor* (2019).
94. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).
95. Dowle, M. & Srinivasan, A. data.table: Extension of 'data.frame'. (CRAN, 2018).
96. Wickham, H., François, R., Henry, L. & Müller, K. dplyr: A Grammar of Data Manipulation. (CRAN, 2019).
97. Wickham, H. et al. ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. (CRAN, 2018).
98. Kassambara, A. ggpubr: 'ggplot2' Based Publication Ready Plots. (CRAN, 2018).
99. Casella, G. et al. Transcriptome signature of cellular senescence. *Nucleic Acids Res.* **47**, 7294–7305 (2019).

ACKNOWLEDGEMENTS

The authors would like to wholeheartedly thank the SHIP study team for all the assistance rendered in realizing this project, and overall for the truly cordial cooperation. The authors would also like to thank the Biophysics Department at Beiersdorf AG for providing both the required equipment and assistance in setting up the imaging setup for the portrait photography of study participants, in particular Thorsten Bretschneider and Stefan Hoppe.

AUTHOR CONTRIBUTIONS

L.K. conceived the original idea for the presented work. M.W. and H.W. provided funding for the experiments. B.K. planned and organized the skin sample collection with assistance of R.S., in close coordination with the SHIP study team. A.W., J.S., C.J., and H.V. coordinated and conducted the examinations and handled the data management. J.S. performed all wet lab work. N.H. designed and implemented the model and performed all computational work. N.H. wrote the manuscript with input of C.F., E.G., and L.K. All authors read and discussed the manuscript.

COMPETING INTERESTS

Skin examinations within SHIP were supported by Beiersdorf AG. N.H., C.F., J.S., B.K., R. S., H.W., M.W., and E.G. are employees of Beiersdorf AG. L.K. received consultation fees from Beiersdorf AG. The remaining authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41514-021-00068-5.

Correspondence and requests for materials should be addressed to N.H. or L.K.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

SCIENTIFIC REPORTS
natureresearch

Check for updates

OPEN

Concomitant DNA methylation and transcriptome signatures define epidermal responses to acute solar UV radiation

Nicholas Holzscheck[1,2], Jörn Söhle[1], Torsten Schläger[1], Cassandra Falckenhayn[1], Elke Grönniger[1], Ludger Kolbe[1], Horst Wenck[1], Lara Terstegen[1], Lars Kaderali[2], Marc Winnefeld[1] & Katharina Gorges[1]

The simultaneous analysis of different regulatory levels of biological phenomena by means of multi-omics data integration has proven an invaluable tool in modern precision medicine, yet many processes ultimately paving the way towards disease manifestation remain elusive and have not been studied in this regard. Here we investigated the early molecular events following repetitive UV irradiation of in vivo healthy human skin in depth on transcriptomic and epigenetic level. Our results provide first hints towards an immediate acquisition of epigenetic memories related to aging and cancer and demonstrate significantly correlated epigenetic and transcriptomic responses to irradiation stress. The data allowed the precise prediction of inter-individual UV sensitivity, and molecular subtyping on the integrated post-irradiation multi-omics data established the existence of three latent molecular phototypes. Importantly, further analysis suggested a form of melanin-independent DNA damage protection in subjects with higher innate UV resilience. This work establishes a high-resolution molecular landscape of the acute epidermal UV response and demonstrates the potential of integrative analyses to untangle complex and heterogeneous biological responses.

Solar UV irradiation has complex and ambivalent effects on the human organism. Beneficial effects of sun exposure are thought to be mainly mediated by vitamin D, which is synthesized in the skin through a photosynthetic reaction triggered by exposure to UVB. Vitamin D was primarily acknowledged for its importance in bone formation, increasing evidence however points to its influence on the proper functioning of nearly every tissue in our bodies[1]. In contrast to this however, solar UV irradiation is also the most abundant risk factor for skin cancer and other extrinsically influenced skin disorders[2,3]. It is well established that UV irradiation both directly and indirectly induces DNA damage. Direct damage is mainly a result of UVB and to lesser extent UVA irradiation, causing dimerization of adjacent pyrimidine bases, a frequent cause of mutations during replication[4]. Indirect DNA damage results mainly from oxidative stress, caused by free radicals and cellular reactive oxygen species, which increase after UV irradiation[5]. Damaged DNA, if not properly repaired, interferes with many cellular mechanisms such as transcription, the cell cycle and replication and can give rise to mutations and epigenetic alterations, driving genomic instability and ultimately carcinogenesis.

Human skin has developed several defense systems to guard against the damaging effects of UV: Prominently these include structural changes to the tissue such as epidermal thickening and the synthesis of melanin, but they also comprise quick molecular adaptations like the suspension of cell cycle and gene transcription, as well as the activation of DNA repair pathways. The extent of protection afforded by these mechanisms however is characterized by high inter-individual variation[6]. The stratification of individual UV response is thus highly important for risk assessment in cancer prevention (UV-protection), therapeutic dose determination (PUVA therapy) and in the understanding of the biological processes leading to malignancies (e.g. squamous skin cancers). Fitzpatrick skin type categories[7] have been widely used as an indicator and predictor of sun sensitivity in epidemiology and experimental photobiology. However, this categorization is hampered by subjectivity and is prone to recall

[1]Beiersdorf AG, Research and Development, Troplowitzstraße 17, 22529 Hamburg, Germany. [2]Institute for Bioinformatics, University Medicine Greifswald, Greifswald, Germany. email: nicholas.holzscheck@beiersdorf.com; katharina.roeck@beiersdorf.com

error[8]. In a study assessing the reliability of Fitzpatrick skin type classifications for instance, only ~ 60% of all study participants self-identified as the same skin type after repeated questioning a few months later[9]. Several authors have investigated the relationship between Fitzpatrick skin type and minimal erythema dose (MED)[10–13], a more objective measure of UV sensitivity frequently used in clinical or research settings, showing increasing MED with higher Fitzpatrick classification in general, but with considerable intergroup variation.

DNA methylation is a covalent epigenetic modification of cytosine to 5-methylcytosine, occurring within CpG dinucleotides[14,15]. Although methylations of adenine have been reported as well, these have so far received considerably less attention. Methylation of CpG sites in the human genome is an important regulatory mechanism that can lead to the activation or repression of gene transcription. Modifications are established and maintained by a set of specific enzymes called DNA methyltransferases. DNA methylation is generally considered to represent a regulatory interface between environmental cues and the genome and might cause or allow long-lasting changes in gene transcriptional activity[16]. Our current knowledge about epigenetic changes associated with acute UV irradiation, its contribution to transcriptomic alterations and implication in skin photobiology, remains very limited. Previous studies have shown however, that chronic solar UV gives rise to large hypomethylated blocks of DNA in the healthy epidermis and that these blocks are conserved in cutaneous squamous skin carcinomas[17], underlining the importance of studying DNA methylation in the context of solar irradiation. In addition, a multi-omics analysis of UV irradiated keratinocytes recently identified several new UV target genes including CYP24A1, GJA5, SLAMF7 and ETV1[18], demonstrating the value of multi-layered omics analyses in unraveling biological phenomena and enabling more reliable biomarker detection, as it has similarly been shown in cancer research, allowing molecular diagnosis and prognosis, often utilizing DNA methylation markers[19,20].

Here we hypothesized that integrative analysis of UV induced epigenetic and transcriptomic alterations in vivo might help to decipher inter-individual responses to environmental challenges and give hints towards early pathogenesis. For this reason, we generated high-resolution multi-omics molecular profiles of the in vivo irradiated epidermis. Our results provide evidence that a UV induced epigenetic memory might be established already after short term repetitive UV irradiation. Integrative analyses of methylation and expression data reveal previously unnoted pathways involved in the acute epidermal UV response and allow the precise inter-individual prediction of MED without the need for prior UV irradiation. Finally, analysis of these molecular phototypes indicates the existence of a melanin-independent form of damage protection in individuals with higher innate resilience to UV irradiation.

Results

UV irradiated epidermis shows genome-wide aberrant methylation patterns and substantial transcriptomic reprogramming. Elucidating the complex molecular mechanisms underlying UV-gene interaction might offer new insights into how UV modulates skin homeostasis and disease pathogenesis to help improve the prevention of UV-induced skin aging and related pathologies. In order to obtain a comprehensive picture of the molecular events regulating acute epidermal photobiology, 32 female Caucasian volunteers (Fitzpatrick phototypes 1–4) where irradiated with individually calibrated doses of 0.9 MED using a full spectrum solar simulator on three subsequent days on a sun-protected area on their lower backs. 24 h after the last irradiation, suction blister roofs were extracted from irradiated and control sites of each subject and gene expression profiling (Illumina RNA seq) and concomitant DNA methylation profiling (Illumina EPIC Arrays) were performed. Paired differential expression and methylation analyses between irradiated and control areas revealed that in total 20.5% (FDR < 0.05) of all interrogated CpGs and 32.4% (FDR < 0.05) of all detected gene transcripts were significantly altered in response to irradiation. These considerable changes were spread over the whole genome, with notable exceptions only occurring in the regions around the centromeres and in some constitutively heterochromatic regions e.g. on chromosome 13 (Fig. 1a). In general, a tendency towards hypomethylation was detected with 65.1% of all significant CpGs decreasing in methylation. Notably, the tendency towards hypomethylation increased from open sea regions to CpG-islands (Fig. S1 a).

Large blocks of the genome have previously been shown to be hypomethylated in chronically sun-exposed epidermal samples in comparison to protected skin[17,21] and have also been associated with clinical measures of photoaging[17]. How quickly this epigenetic imprinting in response to UV exposure occurs however, is so far unknown. We thus investigated whether early indications of photoaging were already detectable after acute repetitive UV irradiation and analyzed the methylation status within the previously reported regions[17] in our data. We found that in over a fifth of the originally described genomic blocks (49/224) the observed methylation changes after acute irradiation correlated very well with the reported patterns (Fig. 1b), differing mainly in magnitude in comparison to chronically exposed skin. This delivers evidence that epigenetic alterations in response to extrinsic stimuli can manifest quickly after external stimulation and suggests that even few repetitive sunburns can be sufficient to impact epigenetic imprinting in genomic regions associated with extensive photoaging.

Considering the extent of alterations in methylation patterns in response to acute irradiation and the universal role of DNA methylation in cancer biology, we then also performed a comparison of pan-cancer methylation signatures[22] to our data, to establish if any overlap in signatures could be observed. The analysis revealed a small number of genomic regions with methylation changes post-irradiation very much reminiscent of those found aberrantly methylated in cancerous tissue. Most of these showed extensive hypomethylation (Supplementary Fig. S2a,b). Whether these alterations are in fact linked to carcinogenesis or purely a product of stochasticity will remain to be determined, but the overlap and extent of correlation raises concern and might warrant further investigation.

Genome-wide correlative analyses of gene expression and methylation reveal coordinated changes in known and novel players of the UV response. Methylation can lead to long-lasting acti-

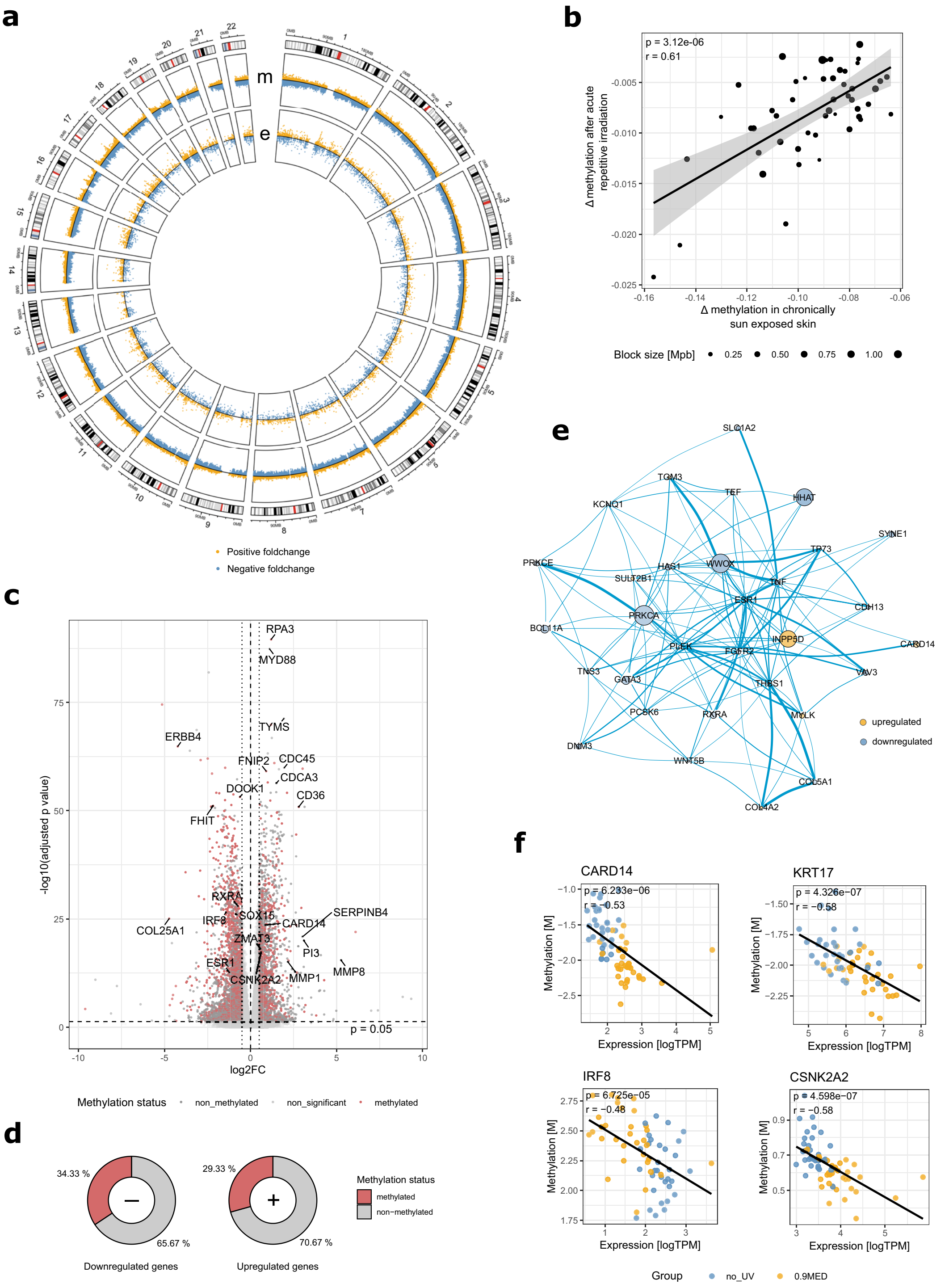

Figure 1. Epigenetic and transcriptomic changes of irradiated samples compared to non-irradiated controls: (**a**) Circos plot showing differential methylation (m, outer circle) and expression (e, inner circle) in response to irradiation to 0.9 MED in a genomic context (FDR < 0.05). Amplitude of points corresponds to log2 fold-change with the solid black line representing no change. Hypomethylated CpGs and downregulated genes are colored in blue, hypermethylated CpGs and upregulated genes in yellow. Colored bands in the karyogram mark centromeres (red) and heterochromatin status (grey to black). (**b**) Differential methylation of 49 genomic regions previously associated with chronic sun-exposure[17] compared to differential methylation after acute repetitive irradiation. (**c**) Volcano plot of differential gene expression in response to irradiation. Differentially expressed genes with ≥ 3 differentially methylated CpGs are marked in red. (**d**) Genome-wide ratio of differentially up- and downregulated genes with concomitant change in methylation (≥ 3 CpGs). (**e**) Protein–protein-interaction network between the most interconnected differentially expressed and methylated genes. Points are scaled by the negative logarithmized FDR of differential expression and colored by log2 fold-changes. Edges are scaled by confidence of interaction. (**f**) Significantly correlated differential expression and enhancer methylation of CARD14, expression and TSS200 methylation of IRF8, expression and TSS200 methylation of CSNK2A2, and expression and TSS200 methylation of KRT17. Plots were generated using R v3.6.1[76] software.

vation or repression of gene transcription and analysis of simultaneously regulated genes on methylation and transcription level have been shown to yield higher prognostic values in several pathophysiological states[23,24]. We thus mapped the most stably differentially expressed genes and CpGs by genomic position and observed a high proportion of genes with equally pronounced methylation changes. Analysis revealed that 29.3% of all upregulated genes (FDR < 0.05 and $\log 2FC_{abs} > 0.5$) harbored at least three differentially methylated CpGs (FDR < 0.05 and $\log 2FC_{abs} > 0.2$), whereas for downregulated genes this number increased significantly further to 34.3% (Fig. 1c,d), with a high number of genes exhibiting inverse correlations to their methylation state. Stratifying the significant CpGs within these genes by regulatory regions revealed they were most frequently located in enhancers and more seldom in exon regions (Supplementary Fig. S1d). Analyzing known protein–protein-interactions between these differentially regulated and methylated genes (DEMGs) using STRING[25] revealed a network of highly interconnected proteins surrounding ESR1, the estrogen receptor α, which was found downregulated following repetitive irradiation (Fig. 1e). Estrogen receptor α expression has been shown to be reduced following UV irradiation in vitro before[26], and its activity has been linked to photoimmune suppression in animal studies. In mice, estrogen receptor antagonists were found to exacerbate immune suppressive action in a dose-dependent manner with estradiol treatment exerting protective effects respectively[27], and the estrogenic compound equol protecting against irradiation-induced carcinogenesis[28]. Notably, the core network also involved the similarly downregulated retinoic × receptor α, previously linked to a functional vitamin A deficiency in the skin following UV irradiation[29] and thereby contributing to photoaging.

We next performed genome wide correlation analyses of all annotated genes and interrogated CpGs to identify significant linear correlations between gene expression and methylation in annotated functional gene regions. Modeling gene expression as a function of mean methylation for all CpGs in potentially regulatory gene regions (enhancers, 1,500 bp and 200 bp upstream of the TSS as well as exon regions) revealed 2,267 significant associations after multiple testing correction. Again, most of these associations were found with alterations in methylation patterns in enhancer regions. Among these highly correlated differentially expressed and methylated genes we identified several known and previously described actors in the UV response, such as CYP24A1, BRCA2, NOTCH2, FOXO3 and GATA3. Examples also included the observed hypomethylation and upregulation of CSNK2A2, a catalytic subunit of Casein kinase II, a ubiquitous serine/threonine protein kinase involved with a manifold of cellular processes, such as cell cycle control and apoptosis and the immune-modulatory keratin KRT17 (Fig. 1f), both of which have previously been associated with UV response and tumorigenesis. Remarkably some of the identified genes, e.g. CARD14 or IRF8 (Fig. 1f), have thus far not been associated with UV irradiation, possibly reflecting the variance between in in vivo and in vitro generated data. Interestingly however, CARD14 mutations have been observed previously in psoriasis patients. Gain-of-function CARD14 mutations in mice lead to spontaneous psoriasis-like skin inflammation by inducing activation of the IL-23-IL-17 axis in keratinocytes and thereby immune cell infiltration[30]. In contrast $CARD14^{-/-}$ mice displayed attenuated skin inflammation in murine psoriasis models[31]. Demethylation-driven CARD14 activation in irradiated cells of the human epidermis might thus present a hitherto undiscovered mechanism of epidermal UV response. The transcription factor IRF8 was found concomitantly significantly hypermethylated and downregulated, which is significant given its function as a tumor suppressor and its frequent downregulation in various cancer types through epigenetic silencing[32–34]. Recently IRF8 has further been implicated in cutaneous wound healing[35], the methylation-driven downregulation of IRF8 might therefore constitute a novel mechanism contributing to the observed impairment of wound healing following irradiation. Notably, IRF8 is located within one of the genomic regions differentially methylated in photoaged skin[17], it would therefore be interesting to investigate its functional role in photoaging, even more so considering the age-associated impairment of wound healing in the skin and the increased risk of developing skin cancer that is associated with both chronic sun-exposure and higher age.

Pathway analysis shows distinct functional enrichments for methylation-associated transcriptional alterations. Since the dissection of DEMGs revealed several genes which had previously not been connected to epidermal UV responses, we performed pathway analyses by means of gene set enrichment. Multiple pathways were strongly enriched with DEMGs, including DNA repair, immune signaling and stress response, strengthening the notion that DEMGs are at the heart of known and key response mechanisms to UV irradiation (Fig. 2a,b). In addition, a high number of enriched pathways were involved in metabolic processes, including some that had previously not been assigned to the canonical UV response pathways. Prominently these were linked to lipid biosynthetic and cofactor metabolic processes (Fig. 2c). Lipid synthesis in the epidermis is vital to skin permeability and barrier function, one of the skin's most crucial functions. Outer epidermal keratinocytes secrete lamellar bodies, which are unique to the epidermis[36] and contain phospholipids, glycosyl-ceramides, sphingomyelin, as well as cholesterol and numerous enzymes, including lipid hydrolases, such as β-glucocerebrosidase, acidic sphingomyelinase, secretory phospholipase A2 (sPLA2), and acidic/neutral lipases[37,38]. When the permeability barrier is perturbed, both the secretion and synthesis of lamellar bodies is stimulated, which allows for the rapid repair and normalization of permeability barrier function[39]. So far only few studies have evaluated the effect of UV on the stratum corneum. They provide evidence for an increased epidermal lipid synthesis in response to UV radiation and alterations of lipid profiles[40–42], however these studies gave no functional correlation to genes or mechanisms involved. In addition, atopic dermatitis and psoriasis patients display modified lipid profiles and both groups are known to benefit from UV therapy. Induced DEMGs related to lipid biosynthetic processes might therefore provide evidence of an understudied UV response mechanism and potentially aid in identifying novel targets to help the regeneration of diseased skin. Differentially upregulated genes involved with other notably positively enriched pathways, such as TYMS and DHFR (Fig. 2c), are mainly involved in nucleotide synthesis and alterations to their increased expression may be part of important cellular responses that ensure proper DNA repair through the replenishment of DNA precursor molecules.

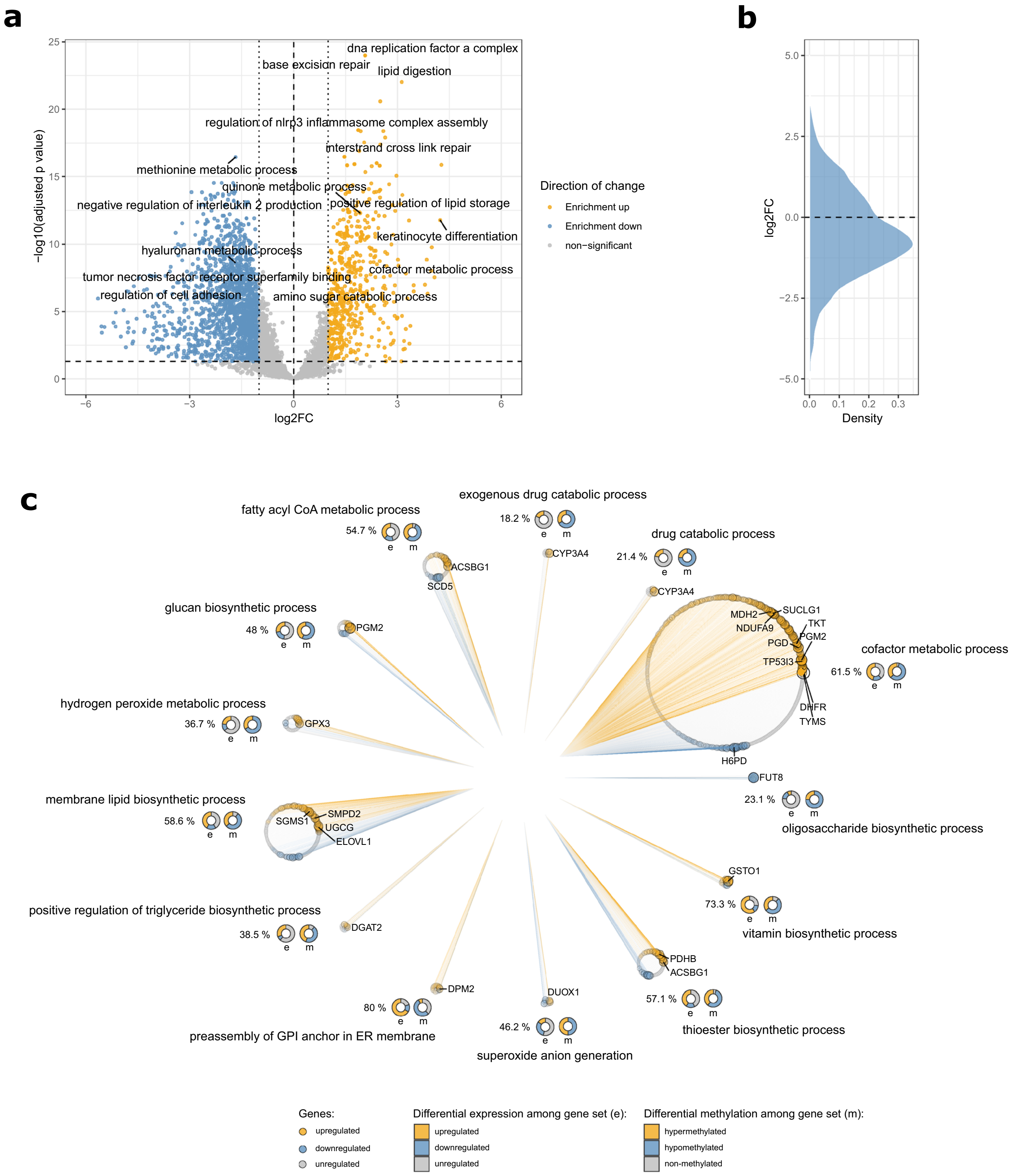

Figure 2. Biological pathways affected by simultaneous changes in both methylation and expression patterns in response to UV irradiation: (**a**) Volcano plot of enriched GO terms based on the analysis of differentially expressed genes with concomitant changes in methylation patterns (≥ 3 CpGs) with positively enriched pathways colored in yellow and negatively enriched pathways in blue (**b**) Distribution of log2 fold-changes in pathway enrichment after irradiation. (**c**) Enriched pathways involved with lipid biosynthesis and cofactor metabolic processes. Pathways are shown as circles with points corresponding to genes annotated to each respective pathway. Genes are colored by up- (yellow) or downregulation (blue) with size scaled to the negative log10 of the FDR derived from differential expression analysis and ordered by log2 fold-changes. Circles underneath pathway names represent proportions of differentially regulated (e) and methylated (m) genes within each gene set. Numbers to the left of the circles summarize the overall percentage of differentially expressed genes per pathway (FDR < 0.05). Plots were generated using R v3.6.1[76] software.

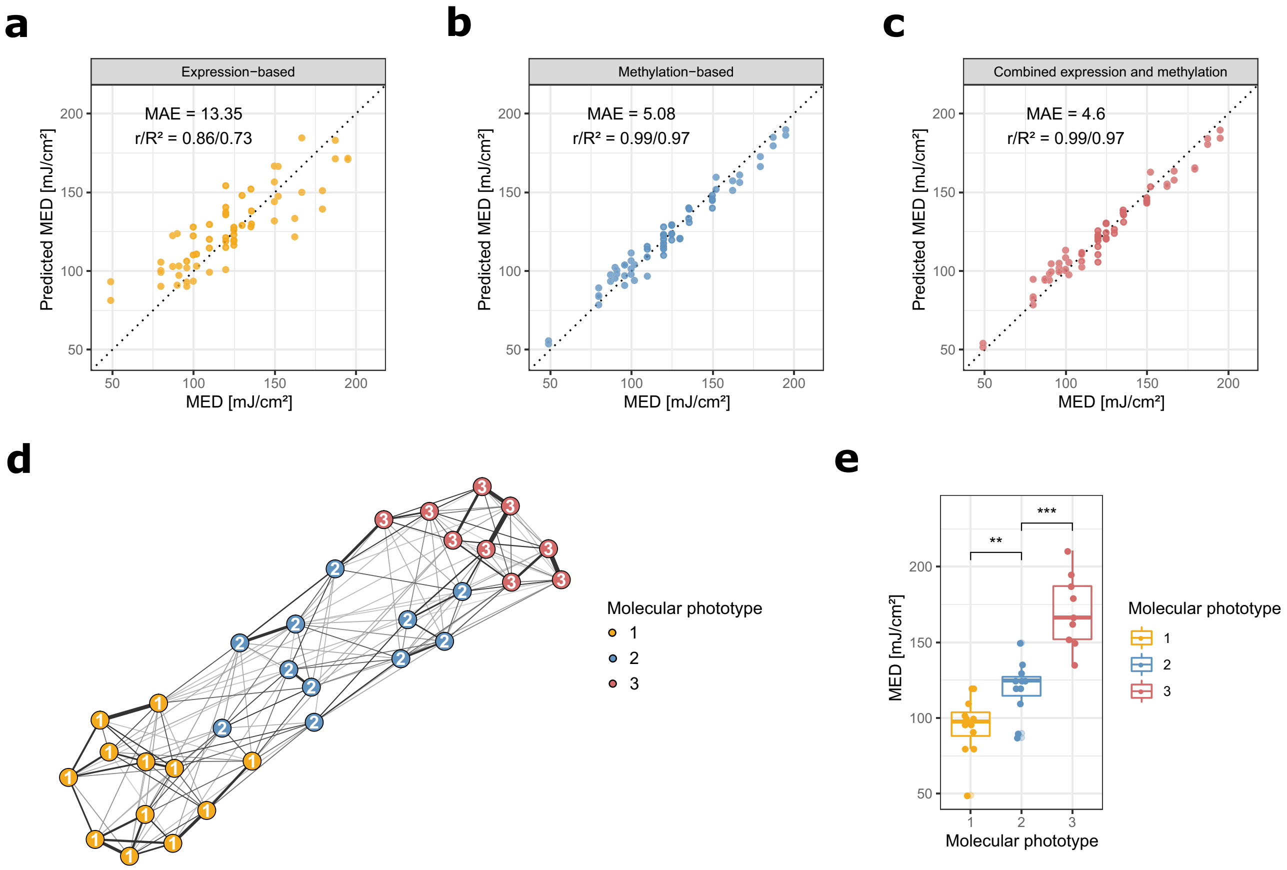

Figure 3. Prediction of inter-individual UV sensitivity from molecular data and identification of MED-correlated molecular phototypes: (**a**) Cross-validated predictions of MED from gene expression data using lasso regression models. (**b**) Cross-validated predictions of MED from DNA methylation data. (**c**) Cross-validated predictions of MED from combination of gene expression and DNA methylation data. (**d**) Fused similarity network generated from gene expression and DNA methylation data of irradiated samples, with nodes colored by molecular phototypes identified through spectral clustering. (**e**) Distribution of MED stratified by the molecular phototypes identified through spectral clustering on the fused similarity network. Statistical comparison was performed using unpaired two-sided t-tests. Plots were generated using R v3.6.1[76] software.

The extent and magnitude of differential regulation in these pathways indicates high cellular priorities of these processes. These findings might warrant further investigation, as these pathways may be vital to maintaining genomic stability after UV irradiation.

Molecular data allow precise inter-individual prediction of UV tolerance without experimental irradiation.

Prediction of UV response is an important tool for risk assessment and prognostication of sun tolerance, photoaging, skin cancer and phototherapy. As a proxy for UV sensitivity, the Fitzpatrick scale is often used[7]. The Fitzpatrick scale or Fitzpatrick phototypes are a subjective, semi-quantitative scale made up of six phototypes that describe skin color by basal complexion, melanin level, and subjective assessment of inflammatory response to UV[43]. A more accurate way to measure UV tolerance is the experimental determination of the MED, which includes the acute irradiation of a test area with different UV dosages and a subsequent assessment of the minimal dose leading to erythema manifestation[13,44]. This method produces accurate and objective results, but is potentially harmful as it exposes the test subject to UV irradiation during the assessment. In the present study, subjects ranging from Fitzpatrick phototype 1 to 4 were analyzed and their MED assessed. As expected from previously published data[8,9,11], stratification of donors using the Fitzpatrick classification was a relatively poor predictor of MED. For instance, the measured MED values for subjects of Fitzpatrick phototype 4 varied from 99.7 up to 210.4 mJ/cm^2. We thus set out to explore if the assessment of individual UV sensitivity could be improved using molecular markers, forgoing the necessity to expose test subjects to harmful UV irradiation in the first place. We employed lasso regression models to attempt the prediction of individual UV sensitivity, as measured by MED, based on gene expression and DNA methylation data and a dataset combining expression and methylation features. The data included both irradiated and control samples, in order for the models to select features that would allow reliable estimation of UV tolerance irrespective of prior sun exposure of the tissue. The tenfold cross-validated predictions showed a high accuracy achieved by both expression- and methylation-based models (Fig. 3a,b), far outperforming the Fitzpatrick classifications with median absolute errors of 13.35 mJ/cm^2 (expression-based) and 5.08 mJ/cm^2 (methylation-based). Models built using DNA methylation features in particular were able to predict individual UV sensitivity to a remarkable degree, indicating a strong epigenetic component associated with UV tolerance. The combination of both expression and methylation data yielded the most accurate prediction model with a median absolute error of 4.6 mJ/cm^2, suggesting further complementarity in the two data levels (Fig. 3c). Model performance was similar on irradiated and control samples, demonstrating the utility of the method irrespective of exposure status (Supplementary Fig. S3a–c). To our knowledge this is the first attempt to derive an accurate estimation of UV sensitivity purely from molecular data, which provides a reliable tool to assess individual UV tolerance, which importantly does not necessitate putting patients at risk of prior irradiation of the skin, as is the case with regular MED assessment.

Multi-omics integration allows the identification of latent subgroups among irradiated samples. Considering the predictivity of the multi-omics data with regard to UV sensitivity, we decided to use an integrative approach to search for heterogeneity in the biological UV response. For this we integrated gene expression and methylation data from irradiated samples using similarity network fusion[45]. Similarity network fusion is a flexible network-based method for integrating different levels biological data, otherwise mostly employed in cancer research: First, a separate similarity network is created from each data level, with samples represented as nodes and similarities in profiles as edges. In a second step, the separate networks are then integrated using an iterative algorithm that strengthens edges between individual samples present in several levels of data, and finally converges into a fused similarity network that incorporates information from all the different data levels. In our case, this lead to a fused network incorporating information from both gene expression and DNA methylation data of the irradiated samples (Fig. 3d). Spectral clustering on the fused network then identified three latent subgroups in the multi-omics data, indicating differences in the biological responses to UV by different test subjects, and allowing their classification into distinct subtypes. The identified subtypes showed very high association to the MED (Fig. 3e) and allowed a better stratification of subjects based on UV sensitivity than Fitzpatrick phototypes, especially in the higher MED range (Supplementary Fig. S4a,b). Molecular subtyping of the skin with regards to UV response using these molecular phototypes (MPs) could prove helpful in developing preventive interventions, stratifying patients for risk factors (e.g. skin cancer and disease) and yielding deeper insights into molecular response mechanisms to irradiation.

Molecular phototypes reveal divergent biological responses to UV irradiation connected to cytokine response, programmed cell death and DNA damage sensing and repair. To characterize the biological processes underpinning the variability of UV responses exhibited by the identified MPs, we assessed the importance of all pathways in the GO term collection with regard to UV response. For this we employed pathway-based machine-learning classifiers that were based on support vector machines using radial basis function kernels, capable of learning non-linear patterns from high-dimensional data. These classifiers were trained to predict irradiation status of a sample using gene expression data from a given pathway, and each "pathway model" was subsequently scored for how well it enabled discrimination between the groups in a repeated cross-validation scheme. This yielded a predictivity score for every gene set, ranking all pathways on a common scale whilst also capturing non-linear gene regulation patterns. Predictivity was assessed for all pathways stratified by the three identified MPs, allowing the identification of biological processes predictive for the UV response for a given subtype and thus also revealing pathways whose regulation diverges between the three subgroups. This resulted in a mapping of the whole pathway landscape with regard to UV response relevance within the three subtypes (Supplementary Fig. S5).

The on average most predictive pathways were involved with DNA damage response mechanisms such as cell cycle transition, DNA replication and chromosome condensation in concordance with the top pathways obtained using gene set enrichment earlier. Further analysis of the involved pathways however revealed divergent patterns between the three molecular subtypes (Fig. 4a). MP 1 and 2 for instance exhibited stronger signals in pathways associated with inflammatory and immune signaling in comparison to MP 3. In case of MP 1, the subgroup with the lowest average MED, these related strongly to inflammasome activation and cytokine response, both generally well-described responses in regards to UV irradiation in human skin[46–49]. In comparison, MP 2 exhibited decreased inflammasome predictivity scores but on the other hand a stronger type I interferon response than either MP 1 or MP 3. MP 2 was further singled out by stronger signals detected in apoptotic and autophagy pathways compared to the other subgroups. This might be connected to a stronger regulation in p53 related signaling pathways, as signaling by p53 class mediators showed increased predictivity in this subtype accordingly. Taken together this could indicate a higher efficiency in clearing cells with unrepairable DNA damage from the tissue. Both MP 2 and MP 3 further showed higher activities in pigment metabolic processes, which is in concordance with the stronger tanning responses observed in more UV tolerant skin[50]. MP 3 on the other hand, incorporating subjects with the highest recorded UV resilience in our cohort, was defined by the strongest pathway signals detected in cell cycle checkpoint and DNA synthesis pathways, as well as genes involved with chromosome condensation. These findings are indicative of a higher sensitivity of the DNA damage sensing machinery in MP 3 subjects in response to irradiation, which would provide a more tightly regulated cessation of DNA replication and thus more time for the repair of UV-induced DNA damage. This hypothesis led us to investigate the extent of DNA photodamage in the samples of study subjects from the different molecular phototypes. We profiled the most common and important form of UV-induced damage to the DNA, the formation of cyclobutane pyrimidine dimers (CPDs), a frequent cause of mutation in the skin after UV irradiation, that directly links UV damage to carcinogenesis[51]. Analysis of the extent of CPDs detectable in the samples revealed lower abundances of CPD-alterations in the DNA of MP 3 subjects compared to the other molecular phototypes (Fig. 4b). This supports not only the model predictions but also the hypothesis of a pigmentation-independent UV protective mechanism in highly UV tolerant skin after repetitive irradiation. The identification of the direct

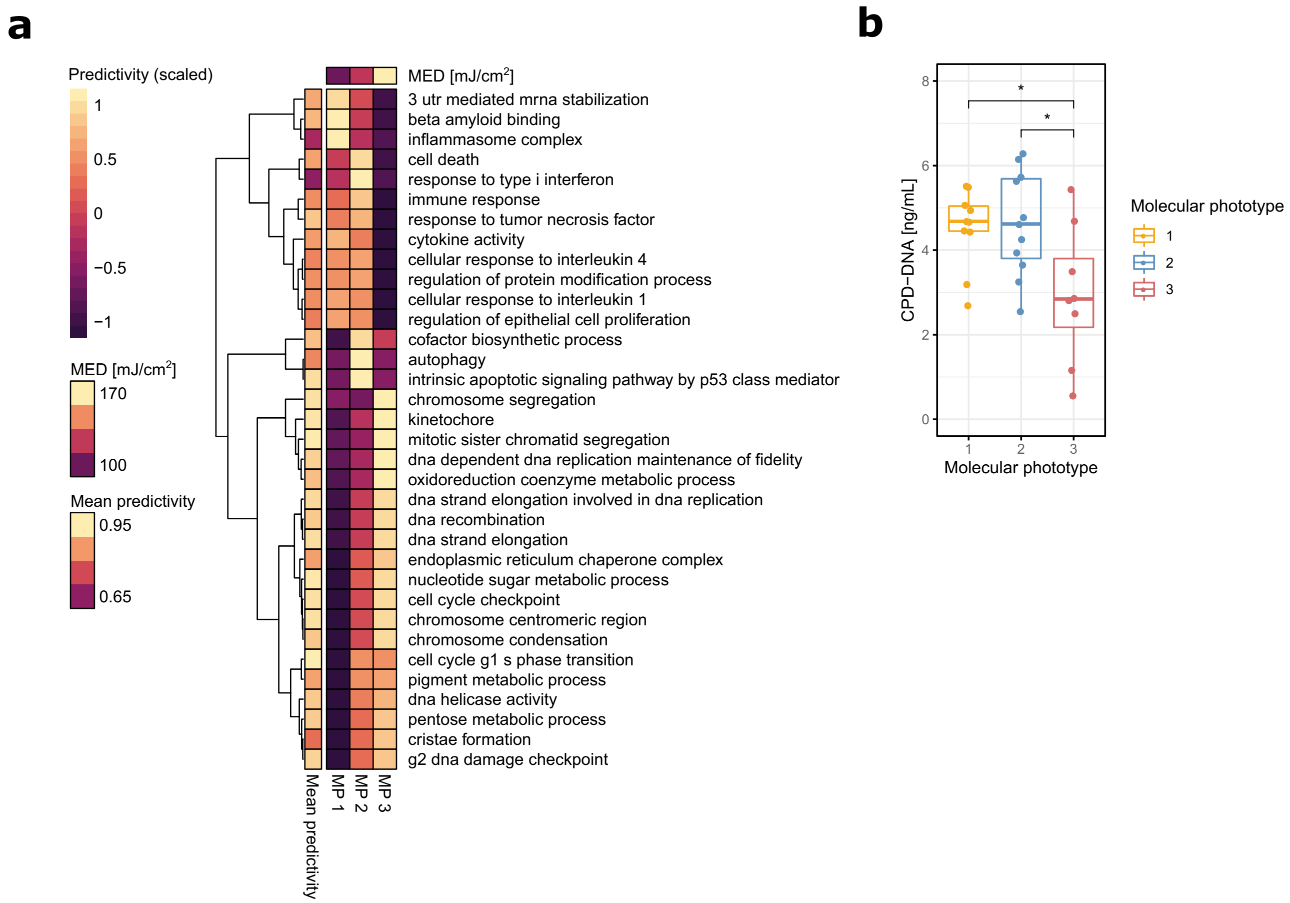

Figure 4. Molecular subtyping identifies heterogeneous biological responses to irradiation that correlate with innate UV sensitivity: (**a**) Heatmap showing the predictivity of the most defining pathways for each of the molecular phototypes to UV irradiation. The heatmap is scaled by pathway to enhance readability, average predictivity of a given pathway over all three molecular phototypes is shown to the left of the heatmap in original scale. (**b**) Extent of DNA damage in the form of cyclobutane pyrimidine dimers (CPDs) measured in the molecular phototypes 24 h after the last irradiation. Statistical comparison was performed using unpaired two-sided t-tests. Plots were generated using R v3.6.1[76] software.

mechanics and the elucidation of key players involved with this response will be important directions for future studies, as they may have potential implications for skin cancer prevention.

Discussion

Epigenetic changes are considered to play a fundamental role in establishing gene expression patterns and providing a genomic response mechanism towards extrinsic influences. However, the experimental evidence describing the extent of this response still remains somewhat limited in many biological processes. We have generated comprehensive methylation and expression profiling data to enable a more comprehensive examination of the intricacies of epidermal UV responses. Our results provide first hints towards an immediate acquisition of aging and cancer related epigenetic patterns in response to UV irradiation. In accordance with these findings, epidemiological studies have previously established a causal role for short term UV exposure (e.g. blistering sunburns) during childhood and adolescence in the late epidermal cancer pathogenesis[52,53]. The spectrum of driver mutations related to skin cancer provides unequivocal genomic evidence for a direct mutagenic role of UV light in carcinogenesis[54–56]. Meanwhile, genomic sites of mutation in skin cancer frequently coincide with CpG-islands[57,58], regions of high DNA methylation density, which has been attributed to the higher vulnerability of 5-methylcytosine bases to CPD-formation[59,60]. Apart from potential mutagenic effects, recent publications also revealed that actinic keratosis samples already bear the classical methylation features of cutaneous squamous cell carcinomas[61]. These reports are consistent with the notion that epigenetic imprinting might present another common mechanism of both photoaging and carcinogenesis.

In general, substantial inter-individual variation in UV tolerability and cancer risk can be observed among Caucasian subjects. Genetic factors like polymorphisms of the melanocortin 1 receptor (MC1R) gene correlate with fairness of skin, UV sensitivity and enhanced cancer risk, however do not fully explain the diversity of UV responses, suggesting the possibility of epigenetic involvement in UV sensitivity and pathogenesis[44]. In support of this, the molecular data and methylation-based features in particular allowed the highly precise prediction of individual UV sensitivity without any prior irradiation, delivering new and strong evidence of an epigenetic component to individual UV tolerance. Further molecular evidence of the heterogeneity in response to irradiation was delivered by the molecular subtyping analysis, where clustering on the integrated multi-omics data revealed three different molecular phototypes (MPs) among irradiated samples with distinct biological signatures. The MPs differed most prominently in their association with pathways regarding cellular stress response, apoptosis/autophagy and DNA damage sensing and repair (MP 1–3, respectively). In an attempt to validate the predicted improved damage sensing and DNA repair of the most UV resilient MP 3, we analyzed the extent of cyclobutane pyrimidine dimers (CPDs) at 24 h after the last irradiation, as readout of UV-induced DNA damage and mutagenic potential. Significantly, this data indeed revealed a decreased amount of CPDs after UV irradiation in MP 3 compared to lower MPs. This is in line with some previous studies which showed lower CPD counts in irradiated samples derived from higher phototypes[62,63], although no molecular mechanisms could so far be elucidated. Interestingly, studies of this type often suffered from substantial biological variation within each phototype as well, once more highlighting the need for better stratification and potentially explaining why the mechanisms leading to these observations could so far not be elucidated in ex vivo tissue. Notably, our study setup differed slightly from most of the previous by making use of a repetitive irradiation scheme, potentially widening the window for detecting inter-individual differences in response mechanisms.

The most intensely explored UV-protection mechanism of the human skin is melanin pigmentation. Melanin serves as a physical barrier that scatters UVR and as an absorbent filter that reduces the penetration of UV through the epidermis[64]. The efficacy of melanin as a sunscreen in darker skin is two- to four-fold higher compared to Caucasians[65]. However individuals with highly pigmented skin have been found 16–500 times less likely to present with skin cancer compared to individuals with fair skin[53,66–68]. The type of melanin produced also plays an important role in skin cancer risk determination. The photoprotective effects of melanin are mainly attributed to eumelanin. Pheomelanin on the other hand has only weak photoprotective properties and has even been found to contribute to carcinogenesis by a mechanism of oxidative damage[69]. Still, even less deeply pigmented ethnicities such as Asians present far lower skin cancer rates compared to Caucasians[53], hinting towards the existence of additional cancer protective mechanisms apart from melanin. One possible explanation involves MC1R variants, which have been shown to confer an increased risk of melanoma and non-melanoma skin cancers, independently of skin pigment (including red hair phenotype)[70]. The increased expression of transcripts which are associated with nucleotide metabolism and DNA repair in our dataset might present another previously uncharacterized mechanism leading to higher cancer protection afforded by skin with high UV tolerance. The detailed characterization of these biological pathways and the analysis of their clinical significance will be important aspects for future studies.

Taken together, our analyses demonstrate the benefit of using multi-omics integration for elucidating complex and diverse responses by disentangling inter-individual variation caused by insufficiently precise subject groupings, such as Fitzpatrick phototypes. The presented data illuminates the diverse and interconnected impacts of repetitive UV irradiation on both transcriptomic and epigenetic patterning in the human skin and provides new insights on protective mechanisms of subjects with high innate UV resilience, that might have further-reaching implications for UV-induced carcinogenesis.

Material and methods

Recruiting. 32 healthy female Caucasian subjects belonging to Fitzpatrick phototypes 1–4 were recruited, with twelve subjects belonging to phototype 1 + 2, ten to phototype 3 and ten to phototype 4. Subjects were aged between 30 and 65 years, with homogenous age distributions in each phototype group. Similar to previous studies[71], exclusion criteria included tattoos or scars in the test area, pigmentation disorders, pregnancy and medication such as anti-histamines or anti-inflammatory drugs within two weeks prior to study start. A detailed listing of exclusion criteria can be found in the Supplementary information.

Minimal erythema dose determination. Minimal erythema dose (MED) estimation is a quantitative method to report the amount of UV (particularly UVB) needed to induce sunburn in the skin 24–48 h after exposure, by determining erythema (redness) and edema (swelling) as endpoints. Individual MED was determined for every subject on the first day of the study following the protocols described in DIN EN ISO 24444[13].

Repetitive irradiation of test sites and sampling. The study sites were located in a sun-protected area on the subjects' lower backs and were randomly split into control and test areas. On the second day of the study, the first irradiation of the test sites was performed using a SOL 500 full spectrum solar simulator (Hönle UV Technology). Intensities were chosen individually to reach 0.9 MED for all subjects or in other words 90% of the required minimal dose causing erythema in a given test subject. Irradiation to 0.9 MED was repeated in the same manner on the third day and once again on the fourth day of the study, leading to a cumulative irradiation of all test sites three times. On the fifth day of the study and 24 h after the last irradiation session of each subject, epidermal samples were taken using the suction blister method, as previously described[72]. For each subject, two suction blister roofs of 7 mm diameter were extracted from both control and test sites, one of each to be used to extract RNA for sequencing, the other to extract DNA for the DNA methylation analysis. This amounted to four suction blister roofs extracted per subject and a total of 128 samples.

Nucleic acid extraction. Nucleic acid extraction was performed as previously described[71]. Tissue samples were suspended in the respective lysis buffers for DNA or RNA extraction and homogenized using an MM 301 bead mill (Retsch). DNA was then extracted using the QIAamp DNA Investigator Kit (Qiagen) according to manufacturer's instructions. RNA was extracted using the RNeasy Fibrous Tissue Mini Kit (Qiagen) according to manufacturer's instructions.

Transcriptome sequencing. Transcriptome libraries were prepared using TruSeq Library Prep Kit (Illumina) and sequencing was performed at 1 × 50 bp on Illumina's HiSeq system to a final sequencing depth of approximately 100 million reads per sample. Sequencing data was processed using a pipeline including Fastqc v0.11.7[73] for quality control, Trimmomatic v0.36[74] for quality based read trimming and Salmon v0.8.1[75] for read mapping and quantification of transcript expression in the form of read counts and transcripts per million (TPM).

Differential expression analysis. Differential gene expression analysis was performed based on the quantified read counts in R v3.6.1[76] using DESeq2[77]. Linear models were fitted using a paired design matrix to account for inter-individual variation unrelated to the irradiation treatment. Genes were considered significantly differentially regulated with $FDR < 0.05$ after multiple testing adjustment by the Benjamini–Hochberg procedure.

Array based methylation profiling. Methylation profiling was performed using Infinium MethylationEPIC arrays (Illumina)[76]. Methylation data was processed using the minfi package[78] in R. Normalization was carried out using the functional normalization method[79], which makes use of internal control probes present on the array to infer and correct for technical variation between arrays. Subsequent analyses used M values to describe CpG methylation levels, as their approximate homoscedasticity renders them superior for statistical testing compared to Beta values[80].

Differential methylation analysis. Differential CpG methylation analysis was performed in R using limma[81]. Linear models were fitted using a paired design matrix to account for inter-individual variation unrelated to the irradiation treatment. CpGs were considered significantly differentially methylated with $FDR < 0.05$ after multiple testing adjustment by the Benjamini–Hochberg procedure. To compare DNA methylation patterns with those previously described in chronically sun-exposed skin and cancer, we used lists of the respective genomic regions and their methylation status in photoaged skin[17] and different types of cancer[22], which were available from the Supplementary information. The originally reported methylation changes within these regions were then compared to the average difference in methylation of all significantly differentially methylated CpGs ($FDR < 0.05$) annotated to the respective genomic regions in our data, allowing for a region-wise comparison of differential methylation.

Gene expression and methylation overlap and correlation analysis. For the calculation of overlap between genes and CpGs, only differentially expressed genes with absolute log2 fold-changes above 0.5 with at least three differentially methylated CpGs with absolute log2 fold-changes above 0.2 were considered, in order to uncover the most reliably differentially expressed and methylated genes. Pearson's correlation coefficients of gene-CpG pairs were calculated as the sum of all gene transcripts for a given gene correlated with the mean of all CpGs belonging to a functionally annotated group (i.e. all enhancer CpGs) annotated to a given gene. Annotations such enhancer status, location in transcription start sites or within exons were extracted from the official manifest files for the Infinium MethylationEPIC array provided by Illumina via their website. Significance was assessed using linear models in R, with p-values being adjusted for multiple testing using the Benjamini–Hochberg procedure.

Protein–protein-interaction analysis. Information on protein–protein-interactions (PPI) retrieved from the STRING[25] database, accessed through the STRINGdb package[25] in R. PPI information was retrieved for all differentially expressed genes ($FDR < 0.05$) with absolute log2 fold-changes above 0.7 with at least three differentially methylated CpGs ($FDR < 0.05$) with absolute log2 fold-changes above 0.5. The interaction query was performed using the standard combined interaction score threshold of 400. The resulting network was refined using in R using igraph[82] by retaining only the top 20% of the most reliable edges based on the combined interaction score, with nodes disconnected from the core network being trimmed in the process. The resulting PPI-network was then visualized using igraph[82].

GO term enrichment analysis. Enrichment analyses were performed using the z-score method[83] as implemented in the GSVA R package[84], applied to the log2 transformed TPMs. GO term gene sets[85] were downloaded originated from the Molecular Signatures Database v6.2[86] and included all three sub-ontologies: biological processes (BP), molecular functions (MF) and cellular compartments (CC).

MED regression models. Lasso regression models[87] for the prediction of MED were built in R using the implementation provided in the glmnet package[88] and interfaced using the machine learning framework mlr[89]. As lasso regression models perform automatic feature weighting by regularizing the absolute magnitude of coefficients, the models were trained on the full datasets, forgoing the necessity of prior feature selection. Furthermore, data from both control and irradiated samples was included in the training process, in order to allow accurate predictions irrespective of previous UV or sun exposure of a given sample. Model predictions and accuracy scores were extracted from tenfold cross-validation to avoid overfitting and derive unbiased predictions and estimates for the quality of model fit. Metrics used for judging model performance were median absolute error (MAE), as well as the Pearson correlation coefficient (r) and the coefficient of determination (R^2).

Similarity network fusion and clustering.

After a filtering step, removing features which showed little correlation to MED and reducing feature matrices to 10% of their original size, gene expression (log2 transformed TPMs) and CpG methylation data (M values) from irradiated samples were integrated via similarity network fusion as previously described[45] using parameter settings of $k = 10$(number of neighbors), $t = 20$ (number of iterations) and $alpha = 0.5$ (hyperparameter). Clustering on the fused network was then performed via spectral clustering as previously described[45]. Measures used for the selection of cluster numbers were the eigen-gap statistic and rotation cost as proposed in the original method description[45].

Pathway predictivity analysis.

Pathway predictivity analysis was performed using GO term gene sets[85] downloaded from the Molecular Signatures Database v6.2[86]. The pathway models were based on the support vector machine (SVM) implementation from the e1071 R package[90], interfaced via the mlr[89] machine learning framework. The models were trained by restricting the expression data (log2 transformed TPMs) to that of genes annotated within a given pathway and trained to predict sample irradiation status (control or irradiated to 0.9 MED) stratified by molecular phototype. The SVMs used the radial basis function kernel with hyperparameters set to $gamma = \frac{1}{sizeofgeneset}$ and $C = 1$. Accuracy of prediction was derived from 5 × 5-fold repeated cross-validation for each pathway model, giving insight on how well genes within the gene set allow a discrimination between UV irradiated and control samples while controlling for overfitting, and used as a measure of predictivity of the respective pathway to irradiation status.

Profiling of cyclobutane pyrimidine dimers (CPDs).

CPD concentrations were determined using the OxiSelect UV-Induced DNA Damage ELISA Kit (Cell Biolabs) according to the manufacturer's instructions.

General data analysis and visualization.

Data analysis in R further included the usage of the package data.table[91] and dplyr[92] for diverse data handling tasks, as well as the packages ggplot2[93], ggpubr[94], circlize[95], and pheatmap[96] for visualization. Mapping and annotation of gene identifiers was performed using the biomaRt[97] and org.Hs.eg.db[98] packages, utilizing the GRCh37 (hg19) human genome build.

Ethics.

The study was performed in agreement with the recommendations of the Declaration of Helsinki and all test subjects provided written, informed consent. Approval of the study protocol was granted by the Ethics Committee of the University of Freiburg (study code 016/1672).

Data availability

Data generated within this study has been deposited online at ArrayExpress, under the accessions E-MTAB-9251 and E-MTAB-9249.

Received: 8 April 2020; Accepted: 16 July 2020
Published online: 31 July 2020

References

1. Mostafa, W. Z. & Hegazy, R. A. Vitamin D and the skin: Focus on a complex relationship: A review. *J. Adv. Res.* **6**, 793–804 (2015).
2. Leccia, M.-T., Lebbe, C., Claudel, J.-P., Narda, M. & Basset-Seguin, N. New vision in photoprotection and photorepair. *Dermatol. Ther.* **9**, 103–115 (2019).
3. Guerra, K. C. & Crane, J. S. Sunburn. (2019).
4. Ikehata, H. & Ono, T. The mechanisms of UV mutagenesis. *J. Radiat. Res. (Tokyo)* **52**, 115–125 (2011).
5. Mohania, D. *et al.* Ultraviolet radiations: Skin defense-damage mechanism. *Adv. Exp. Med. Biol.* **996**, 71–87 (2017).
6. Costin, G.-E. & Hearing, V. J. Human skin pigmentation: Melanocytes modulate skin color in response to stress. *FASEB J.* **21**, 976–994 (2007).
7. Fitzpatrick, T. B. Soleil et peau. *J. Méd. Esthét.* **2**, 33–34 (1975).
8. Rampen, F. H. J., Fleuren, B. A. M., de Boo, T. M. & Lemmens, W. A. J. G. Unreliability of self-reported burning tendency and tanning ability. *Arch. Dermatol.* **124**, 885 (1988).
9. Ravnbak, M. H. Objective determination of Fitzpatrick skin type. *Dan. Med. Bull.* **57**, B4153 (2010).
10. Hemminki, K. & Snellman, E. How fast are UV-dimers repaired in human skin DNA in situ?. *J. Investig. Dermatol.* **119**, 699 (2002).
11. Wulf, H. C., Philipsen, P. A. & Ravnbak, M. H. Minimal erythema dose and minimal melanogenesis dose relate better to objectively measured skin type than to Fitzpatricks skin type. *Photodermatol. Photoimmunol. Photomed.* **26**, 280–284 (2010).
12. Harrison, G. I., Young, A. R. & McMahon, S. B. Ultraviolet radiation-induced inflammation as a model for cutaneous hyperalgesia. *J. Investig. Dermatol.* **122**, 183–189 (2004).
13. International Organization for Standardization. *DIN EN ISO 24444—In vivo determination of the sun protection factor (SPF)*. (2010).
14. Smith, Z. D. & Meissner, A. DNA methylation: Roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).
15. Baubec, T. & Schübeler, D. Genomic patterns and context specific interpretation of DNA methylation. *Curr. Opin. Genet. Dev.* **25**, 85–92 (2014).
16. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nat. Genet.* **33**(Suppl), 245–254 (2003).
17. Vandiver, A. R. *et al.* Age and sun exposure-related widespread genomic blocks of hypomethylation in nonmalignant skin. *Genome Biol.* **16**, 80 (2015).

18. Shen, Y., Stanislauskas, M., Li, G., Zheng, D. & Liu, L. Epigenetic and genetic dissections of UV-induced global gene dysregulation in skin cells through multi-omics analyses. *Sci. Rep.* **7**, 42646 (2017).
19. Stirzaker, C., Taberlay, P. C., Statham, A. L. & Clark, S. J. Mining cancer methylomes: Prospects and challenges. *Trends Genet.* **30**, 75–84 (2014).
20. Costa-Pinheiro, P., Montezuma, D., Henrique, R. & Jerónimo, C. Diagnostic and prognostic epigenetic biomarkers in cancer. *Epigenomics* **7**, 1003–1015 (2015).
21. Grönniger, E. *et al.* Aging and chronic sun exposure cause distinct epigenetic changes in human skin. *PLoS Genet.* **6**, e1000971 (2010).
22. Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.* **43**, 768–775 (2011).
23. Liu, D. *et al.* Discovery and validation of methylated-differentially expressed genes in *Helicobacter pylori*-induced gastric cancer. *Cancer Gene Ther.* https://doi.org/10.1038/s41417-019-0125-7 (2019).
24. Tong, Y., Song, Y. & Deng, S. Combined analysis and validation for DNA methylation and gene expression profiles associated with prostate cancer. *Cancer Cell Int.* **19**, 50 (2019).
25. Szklarczyk, D. *et al.* STRING v10: Protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2015).
26. Toillon, R. A. *et al.* Interaction between estrogen receptor alpha, ionizing radiation and (anti-) estrogens in breast cancer cells. *Breast Cancer Res. Treat.* **93**, 207–215 (2005).
27. Widyarini, S., Domanski, D., Painter, N. & Reeve, V. E. Estrogen receptor signaling protects against immune suppression by UV radiation exposure. *Proc. Natl. Acad. Sci.* **103**, 12837–12842 (2006).
28. Widyarini, S., Husband, A. J. & Reeve, V. E. Protective effect of the isoflavonoid equol against hairless mouse skin carcinogenesis induced by UV radiation alone or with a chemical cocarcinogen. *Photochem. Photobiol.* **81**, 32–37 (2005).
29. Wang, Z., Boudjelal, M., Kang, S., Voorhees, J. J. & Fisher, G. J. Ultraviolet irradiation of human skin causes functional vitamin A deficiency, preventable by all-trans retinoic acid pre-treatment. *Nat. Med.* **5**, 418–422 (1999).
30. Mellett, M. *et al.* CARD14 gain-of-function mutation alone is sufficient to drive IL-23/IL-17-mediated psoriasiform skin inflammation in vivo. *J. Investig. Dermatol.* **138**, 2010–2023 (2018).
31. Wang, M. *et al.* Gain-of-function mutation of Card14 leads to spontaneous psoriasis-like skin inflammation through enhanced keratinocyte response to IL-17A. *Immunity* **49**, 66-79.e5 (2018).
32. Mattei, F. *et al.* IRF-8 controls melanoma progression by regulating the cross talk between cancer and immune cells within the tumor microenvironment. *Neoplasia N. Y. N* **14**, 1223–1235 (2012).
33. Tshuikina, M., Jernberg-Wiklund, H., Nilsson, K. & Oberg, F. Epigenetic silencing of the interferon regulatory factor ICSBP/IRF8 in human multiple myeloma. *Exp. Hematol.* **36**, 1673–1681 (2008).
34. Luo, X. *et al.* The tumor suppressor interferon regulatory factor 8 inhibits β-catenin signaling in breast cancers, but is frequently silenced by promoter methylation. *Oncotarget* **8**, 48875–48888 (2017).
35. Guo, Y. *et al.* Inhibition of IRF8 Negatively regulates macrophage function and impairs cutaneous wound healing. *Inflammation* **40**, 68–78 (2017).
36. van Smeden, J. & Bouwstra, J. A. Stratum corneum lipids: Their role for the skin barrier function in healthy subjects and atopic dermatitis patients. *Curr. Probl. Dermatol.* **49**, 8–26 (2016).
37. Feingold, K. R. Thematic review series: Skin lipids. The role of epidermal lipids in cutaneous permeability barrier homeostasis: Fig. 1. *J. Lipid Res.* **48**, 2531–2546 (2007).
38. Feingold, K. R. The outer frontier: The importance of lipid metabolism in the skin. *J. Lipid Res.* **50**(Suppl), S417–S422 (2009).
39. Lehmann, P., Hölzle, E., Melnik, B. & Plewig, G. Effects of ultraviolet A and B on the skin barrier: A functional, electron microscopic and lipid biochemical study. *Photodermatol. Photoimmunol. Photomed.* **8**, 129–134 (1991).
40. Biniek, K., Levi, K. & Dauskardt, R. H. Solar UV radiation reduces the barrier function of human skin. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17111–17116 (2012).
41. Wefers, H. *et al.* Influence of UV irradiation on the composition of human stratum corneum lipids. *J. Investig. Dermatol.* **96**, 959–962 (1990).
42. Jungersted, J. M., Høgh, J. K., Hellgren, L. I., Jemec, G. B. E. & Agner, T. The impact of ultraviolet therapy on stratum corneum ceramides and barrier function. *Photodermatol. Photoimmunol. Photomed.* **27**, 331–333 (2011).
43. Scherer, D. & Kumar, R. Genetics of pigmentation in skin cancer: A review. *Mutat. Res.* **705**, 141–153 (2010).
44. D'Orazio, J., Jarrett, S., Amaro-Ortiz, A. & Scott, T. UV radiation and the skin. *Int. J. Mol. Sci.* **14**, 12222–12248 (2013).
45. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333–337 (2014).
46. Sand, J. *et al.* Expression of inflammasome proteins and inflammasome activation occurs in human, but not in murine keratinocytes. *Cell Death Dis.* **9**, 24 (2018).
47. Hasegawa, T., Nakashima, M. & Suzuki, Y. Nuclear DNA damage-triggered NLRP3 inflammasome activation promotes UVB-induced inflammatory responses in human keratinocytes. *Biochem. Biophys. Res. Commun.* **477**, 329–335 (2016).
48. Faustin, B. & Reed, J. C. Sunburned skin activates inflammasomes. *Trends Cell Biol.* **18**, 4–8 (2008).
49. Sontheimer, C., Liggitt, D. & Elkon, K. B. Ultraviolet B irradiation causes stimulator of interferon genes-dependent production of protective type I interferon in mouse skin by recruited inflammatory monocytes. *Arthritis Rheumatol. Hoboken NJ* **69**, 826–836 (2017).
50. Miller, S. A. *et al.* Evidence for a new paradigm for ultraviolet exposure: A universal schedule that is skin phototype independent. *Photodermatol. Photoimmunol. Photomed.* **28**, 187–195 (2012).
51. Pfeifer, G. P. & Besaratinia, A. UV wavelength-dependent DNA damage and human non-melanoma and melanoma skin cancer. *Photochem. Photobiol. Sci. Off. J. Eur. Photochem. Assoc. Eur. Soc. Photobiol.* **11**, 90–97 (2012).
52. Henrikson, N. B. *et al.* Behavioral counseling for skin cancer prevention. *JAMA* **319**, 1143 (2018).
53. Armstrong, B. K. & Kricker, A. The epidemiology of UV induced skin cancer. *J. Photochem. Photobiol. B* **63**, 8–18 (2001).
54. Garibyan, L. & Fisher, D. E. How sunlight causes melanoma. *Curr. Oncol. Rep.* **12**, 319–326 (2010).
55. Wu, S., Han, J., Laden, F. & Qureshi, A. A. Long-term ultraviolet flux, other potential risk factors, and skin cancer risk: A cohort study. *Cancer Epidemiol. Biomark. Prev.* **23**, 1080–1089 (2014).
56. Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150**, 251–263 (2012).
57. Drouin, R. & Therrien, J.-P. UVB-induced cyclobutane pyrimidine dimer frequency correlates with skin cancer mutational hotspots in p53. *Photochem. Photobiol.* **66**, 719–726 (1997).
58. You, Y.-H. & Pfeifer, G. P. Similarities in sunlight-induced mutational spectra of CpG-methylated transgenes and the p53 gene in skin cancer point to an important role of 5-methylcytosine residues in solar UV mutagenesis11 Edited by J. Miller. *J. Mol. Biol.* **305**, 389–399 (2001).
59. Tommasi, S., Denissenko, M. F. & Pfeifer, G. P. Sunlight induces pyrimidine dimers preferentially at 5-methylcytosine bases. *Cancer Res.* **57**, 4727–4730 (1997).
60. Martinez-Fernandez, L., Banyasz, A., Esposito, L., Markovitsi, D. & Improta, R. UV-induced damage to DNA: Effect of cytosine methylation on pyrimidine dimerization. *Signal Transduct. Target. Ther.* **2**, 17021 (2017).
61. Rodríguez-Paredes, M. *et al.* Methylation profiling identifies two subclasses of squamous cell carcinoma related to distinct cells of origin. *Nat. Commun.* **9**, 577 (2018).

62. Cragg, N., Chadwick, C. A., Potten, C. S., Sheehan, J. M. & Young, A. R. Repeated ultraviolet exposure affords the same protection against DNA photodamage and erythema in human skin types II and IV but is associated with faster DNA repair in skin type IV. *J. Investig. Dermatol.* **118**, 825–829 (2002).
63. Tadokoro, T. *et al.* UV-induced DNA damage and melanin content in human skin differing in racial/ethnic origin. *FASEB J.* **17**, 1177–1179 (2003).
64. Kollias, N., Sayre, R. M., Zeise, L. & Chedekel, M. R. Photoprotection by melanin. *J. Photochem. Photobiol. B* **9**, 135–160 (1991).
65. Kaidbey, K. H., Agin, P. P., Sayre, R. M. & Kligman, A. M. Photoprotection by melanin: A comparison of black and Caucasian skin. *J. Am. Acad. Dermatol.* **1**, 249–260 (1979).
66. Halder, R. M. & Bang, K. M. Skin cancer in blacks in the United States. *Dermatol. Clin.* **6**, 397–405 (1988).
67. Cress, R. D. & Holly, E. A. Incidence of cutaneous melanoma among non-Hispanic whites, Hispanics, Asians, and blacks: An analysis of california cancer registry data, 1988–93. *Cancer Causes Control CCC* **8**, 246–252 (1997).
68. National Cancer Institute (NCI). *SEER Cancer Statistics Review (CSR) 1975–2014*. (2018).
69. Mitra, D. *et al.* An ultraviolet-radiation-independent pathway to melanoma carcinogenesis in the red hair/fair skin background. *Nature* **491**, 449–453 (2012).
70. Demenais, F. *et al.* Association of MC1R variants and host phenotypes with melanoma risk in CDKN2A mutation carriers: A GenoMEL study. *JNCI J. Natl. Cancer Inst.* **102**, 1568–1583 (2010).
71. Holzscheck, N. *et al.* Multi-omics network analysis reveals distinct stages in the human aging progression in epidermal tissue. *Aging* **12**, 12393–12409 (2020).
72. Südel, K. M. *et al.* Tight control of matrix metalloproteinase-1 activity in human skin. *Photochem. Photobiol.* **78**, 355–360 (2003).
73. Andrews, S. FastQC: A quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).
74. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
75. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
76. R Development Core Team. R: The R Project for Statistical Computing. https://www.r-project.org/ (2008).
77. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
78. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
79. Fortin, J.-P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.* **15**, 503 (2014).
80. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinform.* **11**, 587 (2010).
81. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).
82. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJ. Complex Syst.* **1695**, 1–9 (2006).
83. Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T. & Lee, D. Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.* **4**, e1000217 (2008).
84. Hänzelmann, S., Castelo, R., Guinney, J. & Castelo, R. GSVA: Gene set variation analysis for microarray and RNA-seq data. *BMC Bioinform.* **14**, 7 (2013).
85. Liberzon, A. *et al.* The molecular signatures database hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
86. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
87. Tibshirani, R. & Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288 (1994).
88. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **39**, 1–13 (2011).
89. Bischl, B. *et al.* mlr: Machine learning in R. *J. Mach. Learn. Res.* **17**, 1–5 (2016).
90. Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D. & Weingessel, A. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. *R package version* vol. 1 (2011).
91. Dowle, M. & Srinivasan, A. data.table: Extension of 'data.frame'. (2018).
92. Wickham, H., François, R., Henry, L. & Müller, K. dplyr: A grammar of data manipulation. (2019).
93. Wickham, H. *ggplot2: elegant graphics for data analysis* (Springer-Verlag, New York, 2016).
94. Kassambara, A. ggpubr: 'ggplot2' Based Publication Ready Plots. (2018).
95. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. Circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
96. Kolde, R. Package 'pheatmap'. *Bioconductor* 1–6 (2012).
97. Smedley, D. *et al.* BioMart—Biological queries made easy. *BMC Genom.* **10**, 22 (2009).
98. Carlson, M. org.Hs.eg.db: Genome wide annotation for Human. (2018).

Author contributions

T.S., L.T. and M.W. conceived the original idea for the study and planned the experiments. J.S. carried out the experiments. N.H. and K.G. conceived the analysis and performed the computations. N.H., K.G., L.K., M.W., T.S., C.F., E.G., L.K. and H.W. contributed to the interpretation of the results. N.H. and K.G. wrote the manuscript with contributions from all other authors. All authors discussed the results and commented on the manuscript.

Competing interests

All authors apart from LK are employees of Beiersdorf AG. LK received consultation fees from Beiersdorf AG.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-69683-8.

Correspondence and requests for materials should be addressed to N.H. or K.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements

I would here like to take this paragraph and the reader's time to express my sincere gratitude to those who supported me throughout the creation of this thesis.

Firstly, my deepest thanks to Prof. Dr. Lars Kaderali for the scientific supervision of this entire endeavour. Thank you for the numerous exchanges, the invaluable guidance and encouragement you gave me, for challenging me whenever necessary and for always having good advice whenever that was not enough. My gratitude also goes out to Prof. Dr. Michael Jünger, Prof. Dr. Mario Stanke and Prof. Dr. Johannes Hertel, who generously agreed to serve as members of my thesis committee, and Prof. Dr. Frank Lyko for serving as secondary assessor to this thesis.

Next I would like to thank those who provided me with the possibility to indulge in the creation of this thesis in the first place by providing the necessary funding for this work at Beiersdorf AG. Thank you to Dr. May Shana'a, Dr. Gitta Neufang, Dr. Horst Wenck, and lastly and especially – Dr. Stefan Gallinat – for allowing me to take on this challenge. My particular gratitude in this regard goes to Dr. Marc Winnefeld for taking me aboard and into his team, thank you for all the support and the wealth of good advice shared along this journey. I would also like to wholeheartedly thank Dr. Elke Grönniger for all the lively discussions, encouragement and motivation, as well as all the mentoring I was lucky enough to receive along the way.

Importantly, my sincerest thanks to Dr. Cassandra Falckenhayn for serving as the unofficial second supervisor of all my work, thank you for your ever honest feedback and for the countless hours of productive discussions we had, as well as for always having an open ear.

The entire Skin Aging Team at Beiersdorf I would like to thank for providing a genuinely amazing work atmosphere and team spirit throughout the years that made every day a joy to come to work, so thank you so much to Dr. Marc Winnefeld, Dr. Elke Grönniger, Dr. Cassandra Falckenhayn, Dr. Katharina Gorges, Dr. Annette Siracusa, Juliane Ahlers, Leonie Gather, Ronny Kaufmann, Boris Kristof, Kathrin Schmidt, Ralf Siegner and Jörn Söhle.

Finally, my deepest gratitude goes out to my family. My parents, Danielle and Knut, how have shown me nothing but unwavering support in so many ways throughout my life and who enabled and encouraged me repeatedly to chase my passion in science. Thank you also to Ula, my life partner and companion, for putting up with me and my idiosyncrasies throughout the creation of this thesis as well as naturally all the years before and after.

Finally, I would like to dedicate this thesis to my late grandparents, who unfortunately did not live to see it finished, but would have surely been overflowing with pride and happiness at its completion – as was their nature.