https://doi.org/10.1093/nargab/lqac083

Gene co-expression analyses of health(span) across multiple species

Steffen Möller [1,*], Nadine Saul [2], Elias Projahn [1], Israel Barrantes [1], András Gézsi [3], Michael Walter [4], Péter Antal [3] and Georg Fuellen [1]

[1] Rostock University Medical Center, Institute for Biostatistics and Informatics in Medicine and Ageing Research, Rostock, Germany, [2] Humboldt-University of Berlin, Institute of Biology, Berlin, Germany, [3] Budapest University of Technology and Economics, Department of Measurement and Information Systems, Budapest, Hungary and [4] Rostock University Medical Center, Institute for Clinical Chemistry and Laboratory Medicine, Rostock, Germany

Received May 02, 2021; Revised August 20, 2022; Editorial Decision October 16, 2022; Accepted October 31, 2022

ABSTRACT

Health(span)-related gene clusters/modules were recently identified based on knowledge about the cross-species genetic basis of health, to interpret transcriptomic datasets describing health-related interventions. However, the cross-species comparison of health-related observations reveals a lot of heterogeneity, not least due to widely varying health(span) definitions and study designs, posing a challenge for the exploration of conserved healthspan modules and, specifically, their transfer across species. To improve the identification and exploration of conserved/transferable healthspan modules, here we apply an established workflow based on gene co-expression network analyses employing GEO/ArrayExpress data for human and animal models, and perform a comprehensive meta-study of the resulting modules related to health(span), yielding a small set of literature backed health(span) candidate genes. For each experiment, WGCNA (weighted gene correlation network analysis) was used to infer modules of genes which correlate in their expression with a 'health phenotype score' and to determine the most-connected (hub) genes (and their interactions) for each such module. After mapping these hub genes to their human orthologs, 12 health(span) genes were identified in at least two species (ACTN3, ANK1, MRPL18, MYL1, PAXIP1, PPP1CA, SCN3B, SDCBP, SKIV2L, TUBG1, TYROBP, WIPF1), for which enrichment analysis by g:profiler found an association with actin filament-based movement and associated organelles, as well as muscular structures. We conclude that a meta-study of hub genes from co-expression network analyses for the complex phenotype health(span), across multiple species, can yield molecular-mechanistic insights and can direct experimentalists to further investigate the contribution of individual genes and their interactions to health(span).

INTRODUCTION

Health and healthspan are gaining acceptance as central concepts in medicine, with a focus on (multi-)morbidity, aiming to delay the onset of disease and dysfunction for as long as possible. Health is difficult to describe and has different meanings to different people. Aging, and the deterioration of health that comes with it, affects nearly all species. But tissues that enable the systematic study of the underlying molecular processes are more easily available for animal models, especially for invertebrates, coming with further advantages such as controlled genetics and environments, and a much shorter lifespan. Thus, aging and health are frequently studied in animal models.

To support aging research, many databases are now available (1). Gene expression profiles across tissues of aging mice were already presented, e.g. by the AGEMAP (2) project in 2007 and recently by the Aging Atlas Consortium (3), but there is a lack of such data for health. Adding the dimension of health may amend the identification of molecular markers for aging and further support the identification of health-modulatory compounds (4).

An increasing number of transcriptomic data sets that can be used to compare young and old individuals are available on public repositories. The concept to derive aging-associated patterns from transcriptome repositories across species (5) already led to central elements of aging-related knowledge bases (1,6). Comprehensive analyses of transcriptome repositories were also expanded towards diseases in the context of aging (7). Yet, for expression profiles *per se*, there is a lack of gene expression co-regulation analyses across species with a focus on health(span). A ma-

[*] To whom correspondence should be addressed. Tel: +49 381 494 7361; Fax: +49 381 494 7203; Email: steffen.moeller@uni-rostock.de

© The Author(s) 2022. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

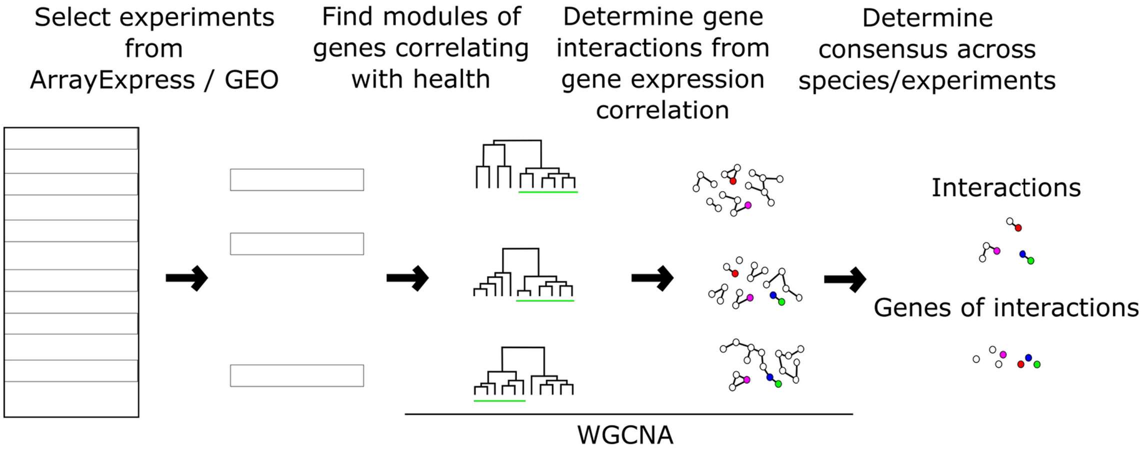

Select experiments from ArrayExpress / GEO
Find modules of genes correlating with health
Determine gene interactions from gene expression correlation
Determine consensus across species/experiments
Interactions
Genes of interactions
WGCNA

Figure 1. Workflow to determine cross-species consensus gene correlation networks, and subsequent analyses. WGCNA is applied independently for each selected experiment in ArrayExpress/GEO, defining modules and gene interactions. Gene interactions are filtered by experiment-specific thresholds. For each module, hub genes are retrieved and those with an ortholog found as a hub gene in another species are reported in Table 1. For each module, Table 2 lists the genes that correlate the most with its 'eigengene', i.e. that best represent the module's expression pattern across samples.

jor challenge for polygenic phenotypes in general is the heterogeneity of the underlying gene regulatory landscape (8), impeding the use of network-based methods for post-processing, i.e. smoothing, aggregating, and unifying, transcriptomic results (9,10). However, the power of the cross-species derivation of conserved co-regulation modules is becoming apparent, see, e.g. the CoCoCoNet database (11).

For prominent cellular characteristics of aging, such as cellular senescence, Avelar and coworkers (12) demonstrated how to integrate static data from public databases with insights from gene co-expression (https://coxpresdb.jp/) (13). Attempts have also been made to use known gene/protein interactions to describe age-induced expression profiles (14). The integration of co-expression data, also across species, could similarly be performed with GeneFriends (15) (for human and mouse) for RNA-seq or, for microarray data also with MIM (16). The latter also provides provenance information, i.e. the experimental context in which the correlation was found, to plan follow-up experiments.

We recently proposed an operational definition of health (17) and suggested that it may be applied across species. We then collected data on molecular contributions to health (18), with a focus on genetics. With the support of GeneMania (19) and the associated tool AutoAnnotate (20) we then constructed a map of network modules by clustering a functional interaction network of the genes implicated in health. Naturally, aging and health are complex phenotypes for which we still lack the means to single-out and investigate the contribution of individual genes. A detailed analysis is therefore expected to dissect a list of health-associated genes into gene sets that, in turn, can be understood as parts of the whole (that is, health), and these parts are distributed across diseases & dysfunctions, tissues, organs and species. The idea of identifying health-associated molecular patterns is at the root of molecular health research. Our efforts strived for a consensus across the species barrier between worms (*Caenorhabditis elegans*) and humans, and we investigated the transfer of findings from worms as a short-lived animal model of health to humans. A consensus in network modules of worm and human was thus determined (18), but it was small in relation to the much larger functional interaction networks that were the starting point for each species. However, functional interaction databases, upon which GeneMania is based, are woefully incomplete. Further, these databases do not usually consider the specific biological context of an interaction, but instead merge interaction data from very heterogeneous sets of experiments (8,21).

To harness the power of diverse transcriptomic experiments in the context of health(span), here we present a WGCNA-based meta-study for the exploration and characterization of health(span) related modules (Figure 1). WGCNA co-expression analyses have recently been used in aging research (22) to identify differences in old vs young and gene expression asymmetries in the brain that develop over time. In our study we integrated a highly diverse set of health(span) expression data across species from many different tissues. We manually derived a scoring for all the transcriptome samples we considered, based on a score combining quantitative and qualitative factors that the authors of the experiments provided, and refer to it as their 'health phenotype score'. WGCNA was found to be a competitive tool to find network modules reflecting such kinds of scores (23). This allows the filtering for health-associated modules generated by the WGCNA correlation analysis across tissues (or cell lines) and multiple species, and thus, the meta-study of health-associated most-connected genes (hubs) and their interactions, as presented here. We also collected the evidence for the implication of these genes in health(span) from the literature.

MATERIALS AND METHODS

External databases

All sets of transcriptomics experiments in the Gene Expression Omnibus (GEO) (24) and ArrayExpress (25) databases that mention 'healthspan' in the title or the description were included, if they featured more than six samples and a scale-free network could be derived from their correlation matrix (for the latter, see below). Experiments performed on *C. elegans* were added when these were alternatively annotated with the term 'health', to increase the number of datasets for the worm, since "healthspan and 'Caenorhabditis elegans' only finds the single entry E-GEOD-54853. We did not include non-worm experiments with 'health' in the title or the description, since the number of matches (specifically for human) turned out to be excessively large.

Log-transformation of expression levels was performed if not already performed for the data we retrieved. Suppl. Table 1 describes the experimental data and metadata which form the input to the following analyses. Our scripts relied directly on expression data provided by the individual uploader. Information on the way the data was normalized is made available in the description of the protocol. A transformation to *Z* scores made no difference to the analysis. We also tried out a quantile normalization (26) for each experiment. Genes still found significant with quantile normalization applied are indicated as such in Tables 1 and 2.

Data on gene essentiality was retrieved from the Online GEne Essentiality (OGEE) database (27). OGEE describes for several large initiatives what fraction of cell-culture experiments per tissue were inhibited in growth if a gene was disabled. If available, the minimal and maximal fractions are presented for the many human tissues of the 'avana' set of experiments, 'broad' or 'sanger' were alternatives. There was no gene for which OGEE did not have data provided by one of the sources.

Each experiment's metadata was inspected to manually derive a health(span) phenotype score designed to reflect the health status of the individual(s) from which the sample(s) were taken. This 'health phenotype score' was manually tailored for each experiment by a custom formula that takes the experiment's factor annotation as an input and thus consistently annotates each sample: For our analysis, if an experiment provides phenotype scores that reflect the health of the individual samples, then these scores are taken. If such a score is not available for an experiment, it is derived by a minimal number of operations from the phenotypes or from the experimental factors describing the experiment. Such a minimal operation could be the inversion of a factor describing accumulated damage. If multiple factors are given (like a combination of treatments), then we inspected the paper in detail to derive a combined score. Gene knock-downs, knock-outs, gene transfers or natural genetic variations are all interpreted as if they were a treatment. The young untreated wildtype individual was assigned a health phenotype score of 1. If a treatment was reported in the paper to improve or reduce health, by default we added or subtracted 0.2, for example yielding a score of 1.2 for the healthier individual. Importantly, only the scores of the individuals providing the samples within one experiment must be comparable to each other and we never relate the scores in one experiment to the scores in another experiment. This allows us to use a rather arbitrary scorecard that has to be consistent only within one experiment. Within such an experiment, scores may be adjusted further considering the relative strengths of the effects reported in the paper. Samples from old individuals are assigned a healthspan phenotype score of 0, which may be increased or decreased by 0.2 as before. The details can be inspected in the 'Data_parameters' folder (see Availability).

RNA-seq data re-analysis

Gene expression levels were typically not available for the RNA-seq data. Therefore, the RNA-seq datasets were all reanalyzed based on the raw data by the following protocol. All target RNA-seq datasets were retrieved from the European Nucleotide Archive (28), and the corresponding FASTQ files were filtered for Illumina adapters, phage PhiX sequences and quality (Phred score over 25) using BBTools version 38.49 (29). Gene expression was then quantified for each RNA-seq run. To this end, the filtered outputs were mapped against the corresponding target genomes from the Ensembl database release 98 (30), using the STAR program version 2.7.3a (31). This program also enabled us to assign uniquely mapped reads to individual genes from the short read alignments. Finally, the mapped read counts were normalized as transcripts per million (32).

Network analysis

In WGCNA, the network construction is independent from the separation of genes into modules: The modules are a means to perform a gene selection and prioritization with respect to an observed association with the phenotype. The interaction network is defined via a threshold on a score on pairs of genes, which is not the Pearson correlation itself. Instead, it is a score that indicates to what degree a pair of genes correlates in the same fashion (in terms of direction and magnitude) with other genes (33,34). The scored gene pairs are filtered for a minimum interaction score of 0.2.

The WGCNA analysis was performed for undirected interactions. Parameters were set as instructed by the WGCNA standard protocol, as follows. For every experiment the *cutHeight* was manually set to remove outliers and the exponent/power was manually determined to ensure that the network is a scale-free network (see below). An experiment is skipped if that is not possible and then marked with 'no modules found' in Supplementary Table S1. For RNA-seq, prior to the removal of outliers, low-count genes were removed by a manual setting of the parameter *cutHeight* so that the separation of the samples reflects their phenotypes and could no longer be improved, based on the clustering of the genes by expression data with the R function *hclust* as performed as part of the WGCNA protocol.

The WGCNA protocol proposes to apply an experiment-specific exponent to the correlation coefficients (WGCNA calls it 'power') to strengthen the differences in the correlation data. This power is set, for each experiment, just large enough so that in the derived correlation network, the fraction of genes that have *k*-many interactions with other genes is proportional to $k^{-\gamma}$ with $\boldsymbol{\gamma}$ being a small positive parameter. Networks with that property are called scale-free; the parameter $\boldsymbol{\gamma}$ describes how quickly this fraction gets smaller when the number of connections increases. Genes with a high number of connections are rare in scale-free networks, but they exist, and these 'hub' genes are considered highly influential on the expression levels of genes in that module.

Table 1. Hub genes in health-associated WGCNA network modules, found in at least two species. Orthologs were mapped to the human gene name using Ensembl. The human gene names also correspond to the names in mouse and rat, whereas the names of the orthologs in worms based on the Ensembl database are given in brackets. The last column summarizes the report by the OGEE database labeling the gene as essential. If available, it presents what fraction of cell-culture experiments showed the gene to be essential across the many human tissues of the avana set of experiments. PPP1CA and SDCBP (marked by an asterisk) have also been found as hub genes when starting with quantile-normalized data, see Supplement Table S3, which also lists other members of the MRPL and SCN families

Gene human (worm)	Human	Mouse	Rat	Worm	Description in context of healthspan	Essential (OGEE)
ACTN3 (atn-1)			x	x	Expressed in muscle, known marker for healthspan and athletes' muscle phenotypes (36). Localized to Z-discs, anchoring to actin filaments.	no
ANK1 (unc-44)		x		x	Ankyrin 1 (ANK1) is associated genetically with Diabetes type 2 (37), spherocytosis (38) and epigenetically with neurological diseases, likely triggered by ApoE with effect on TNFalpha and Akt (39).	conditional (mouse)
MRPL18 (mrpl-18)		x		x	The mitochondrial ribosomal protein L18 (MRPL18) is involved in the cytosolic stress response and promotes the translation of Hsp70 (40).	10–27%
MYL1 (mlc-6 & mlc-5)	x		x		MYL1 encodes the myosin light chain 1 expressed in fast-twitch skeletal muscle fibers (41). Human ageing is associated with lower MYL1 content and higher MYL3 content (42).	4–38% (broad)
PAXIP1 (pis-1)		x		x	The PAX interacting protein 1 (PAXIP1) contributes to DNA repair and correlates with breast cancer staging (43).	2–18%
PPP1CA* (C06A1.3 & 26 others)	x			x	PPP1CA is one of three catalytic subunits of the serine/threonine specific protein phosphatase 1 (PP1), which is known to be involved in the regulation of glycogen metabolism, cell division, muscle contractility and protein synthesis (44). PPP1CA itself is linked to diverse tumor entities and is involved in ERK/MAPK signaling (45, 46), TGFβ signaling (47), Ras signaling and Ras-induced senescence (48), spermatogenesis (49) as well as in tau hyperphosphorylation leading to Alzheimer's disease (50).	1–21%
SCN3B (-)		x	x		The sodium voltage-gated channel beta subunit 3 (SCN3B) controls electrolytes and contributes to the pacemaking in the heart and has an effect on intracellular trafficking (51). It also suppresses senescence and apoptosis via its interaction with p53 and thus, is considered to be an oncogenic factor (52).	no
SDCBP* (lin-10)		x		x	Syntenin-1 (formerly Syndecan(SDC)-binding protein) regulates autophagy (53) and together with Syndecan contributes to exosome formation (54) also in cancer cells (55).	3–21%
SKIV2L (skih-2)	x		x		The Ski2-like RNA helicase (SKIV2L) is part of the Super killer (SKI) complex and involved in mRNA degradation, DNA-RNA hybrid control, and telomere stability (56). SKIV2L is also known to contribute to inflammatory bowel disease (57) and macular degeneration (58). Furthermore, SKIV2L features antiviral capacities and plays a role in innate immunity (59) associated with RNA exosomes (60).	no
TUBG1 (-)	x		x		TUBG1 encodes the tubulin gamma 1 protein, which, when mutated, can lead to brain malformations (61) with clinical features such as motor and intellectual disabilities and epilepsy. Moreover, TUBG1 is involved in tumor diseases, as shown for breast cancer (62), lung cancer (63) and medulloblastomas (64).	common 60–100%
TYROBP (-)		x	x		The transmembrane immune signaling adaptor TYROBP is considered to be involved in Alzheimer's disease (65,66) and as a target of TERC in inflammatory processes (67). In addition, TYROBP is suggested as a prognostic marker for gastric cancer and renal cell carcinoma (68, 69).	9–20%
WIPF1 (wip-1)		x		x	The WAS/WASL interacting protein family member 1 (WIPF1) regulates actin, phagocytosis, and neurotransmission and is among the top-3 genes upregulated by caloric restriction in the hypothalamus of wild-type mice (70). Furthermore, overexpression of WIPF1, triggered by BRAF-mutation activated MAP kinase pathway, promotes aggressiveness of thyroid cancer and thus acts like an oncoprotein (71). Its oncoprotein character was also described for pancreatic adenocarcinoma (72) as well as breast cancer, glioma and colorectal cancer (73).	no

Table 2. Genes correlating the strongest with the module's eigengene (quantifying module membership) in at least two species. Genes in this table are among the top-30 of the module membership and found in experiments of at least two species. The gene name is marked in bold if that gene was listed as a hub gene in Table 1. The column 'Correlation' flags 'positive' (or 'negative') to refer to an observed positive (or negative) correlation with the 'health phenotype score' when the gene is upregulated. 'mixed' indicates that the experiments did not yield a consensus direction of correlation. Supplement Table 2 extends this list to all genes that appear in the top 30 of modules of two or more experiments. The '#Experiments' column indicates the number of experiments with a module for which the gene was identified as a member. The last column summarizes the report by the OGEE database labeling the gene as essential as in Table 1. *PPP1CA and SQSTM1 (marked by an asterisk) have also been found as hub genes when starting with quantile-normalized data, see Supplement Table S3, which also lists other members of the MRPL and SCN families. $^+$CEBPB, PAXIP1 and PPP1CA (marked by a plus sign) are confirmed by the same analysis on quantile-normalized data presented in Supplement Table S4

			Correlation with healthspan phenotype					
Gene		#Experiments	All species	Human	Mouse	Rat	Worm	Essential (OGEE)
AC068831.7	vps-33.2	2	*negative*		*negative*		*negative*	conditional (worm)
ADAM10	sup-17	2	*mixed*		*negative*		*positive*	no
APBB1IP	mig-10	2	*mixed*		*negative*		*positive*	no
CEBPB$^+$	cebp-1	2	*mixed*		*negative*	*positive*		no
CREBBP	cbp-1	3	*negative*		*negative*		*negative*	2–11%
EIF3F	eif-3.F	2	*positive*		*positive*		*positive*	yes (sanger)
INTS12	F53H1.4	2	*mixed*		*negative*		*positive*	1–21%
KPNA3	ima-3	2	*mixed*		*negative*		*positive*	11–25%
MEX3C	mex-3	2	*negative*	*negative*			*negative*	no
MRPL19	mrpl-19	2	*mixed*		*positive*		*negative*	2–26% (broad)
MYL1	mlc-6	2	*positive*	*positive*		*positive*		4–38% (broad)
PAXIP1$^+$	pis-1	2	*positive*		*positive*		*positive*	1–21%
PCNX2	B0511.12	2	*negative*		*negative*		*negative*	conditional (mouse)
PPP1CA^{*+}	C06A1.3	4	*mixed*	*positive*			*mixed*	8–20%
PPP1CB	gsp-1	2	*positive*			*positive*	*positive*	52–96% (sanger)
PPP2R3C	-	2	*negative*	*negative*	*negative*			10–27%
RAB2A	unc-108	2	*negative*		*negative*		*negative*	no
RAB31	-	2	*negative*		*negative*	*negative*		no
RPL29	rpl-29	2	*positive*		*positive*	*positive*		2–30% (broad)
RTN2	-	2	*mixed*	*positive*	*negative*			no
RYR1	unc-68	2	*positive*	*positive*	*positive*			no
SCN3B	-	2	*positive*		*positive*	*positive*		no
SIX4	ceh-32	2	*negative*		*negative*		*negative*	no
SNRPD1	snr-3	2	*negative*		*negative*		*negative*	common 100%
TMEM70	F32D8.5	2	*mixed*		*positive*		*negative*	no
TUBG1	-	2	*mixed*	*positive*	*negative*			common 60–100%
WIPF1	wip-1	3	*negative*		*negative*		*negative*	no
ZC3H15	F27D4.4	2	*negative*		*negative*		*negative*	no
SQSTM1*	sqst-1	3	*negative*		*negative*	*negative*	*negative*	2–25% (broad)

Further, we filtered for modules that are associated with the health(span) phenotype (see next paragraph), and the hub genes are likely to also have a strong effect on this phenotype.

Only network modules whose WGCNA eigengene correlated with the 'health phenotype score' (P value < 0.05) were retained further. Then, the 30 genes (see the WGCNA tutorial https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/) most connected in a module according to the WGCNA *softConnectivity* function were considered for subsequent consensus analyses, and called 'hub' genes hereafter. Besides the modules, output of the WGCNA workflow is the topological overlap matrix with a quantitative description (termed *adjacency*) of all interactions between any pair of genes of an experiment. For each experiment, we determined a threshold at the 95% quantile of all the adjacency values. Only gene interactions with an adjacency above that experiment-wide threshold contribute to our analysis of interactions of genes in the health-associated modules. For the 30 hub genes, all pairwise interactions above that experiment-wide 95% quantile were thus exported, i.e. subjected to a pan-module search for consensus genes and consensus interactions, also across species by considering orthologs, presented in Tables 1 and 2. Orthologs were determined based on Ensembl version 101 (30).

Further, for each health-associated module, the 30 genes correlating the strongest with the module's eigengene (reflecting average module behavior, also called 'module membership' in WGCNA) were retrieved. Those found in at least two species are presented in Table S3. The correlation is taken in relation to the eigengene, and not in relation to the health phenotype score. Either would be fine for a ranking of the hub genes within a module, and the ranking is expected to be identical for the genes most central to a module. However, our particular interest was to abstract from the phenotypes of the experiment and thus utilize the WGCNA-performed modularization to influence the ranking. This is assumed to be particularly useful for experiments with multiple health-associated modules, each of which we expect to focus on a different aspect of health and for which the constituent genes should thus be ranked differently, in order to analyze that particular module most appropriately, and without any particularities pertaining to the health phenotype score. Multiple probesets describing the same gene, or its splice variants, were not distinguished and mapped to the same human gene, resulting in genes interacting with themselves. Such self-interactions were removed.

Network figures

The network figures were created with the R igraph package. A spring-embedding layout was chosen for the plots and manually refined. Figure 2 shows an overview on all hub genes from Table 1, their direct interactions and genes found in any species that connect to at least two hub genes. Additionally, Figure 2a–d in the Supplement were prepared separately for each species, i.e. they show only interactions from modules that WGCNA identified for an experiment based on samples from that species. Input to these supplement figures are the hub genes from Table 1 and all genes that are reachable from the hub genes which are no more than two transitions away. All interactions between the selected (reachable) genes were also added. The resulting graphs were simplified with the igraph minimum spanning tree implementation that maintains the connectivity of the graph but removes all redundant paths between genes. The spanning tree retains the stronger of two alternative paths between genes. A gene connected to a hub gene with a low adjacency value will thus lose that direct link if it is correlating strongly with another gene that has a strong correlation with that hub gene. Hub genes were determined from within WGCNA considering all interactions, not only the ones above the 95th percentile. Hub genes that are strongly connected for experiments in one species may not be equally dominating in another species. This possibility and the competitive effect on directly connected hub genes (one cross-species, the other only observed for one species) imposed by the spanning tree give the impression that the cross-species consensus hub genes are marginalized in Supplemental Figure S2a–d, albeit these graphs are seeded from the consensus hub genes and their interactions.

Permutation tests

The question whether the selected hub genes (see Table 1) are special with respect to their number of interactions in protein-protein interaction databases was addressed empirically. Hereto, all genes in the STRING (35) database were ordered by their degree (sum of incoming and outgoing connections). For single genes a rank in that order can be determined, and for any set of genes the sum of these ranks can be derived. 10 000 random gene selections of equal size each yield their respective rank sum. The fraction of these random gene selections with a higher rank sum score than the gene sets in Tables 1 and 2 estimates the probability to make that respective finding by chance.

Similarly, the significance of the number of consensus genes of Tables 1 and 2 was assessed by repeating the consensus determination with an equal number of randomly picked genes from the respective same species.

RESULTS

We analyzed all experiments listed in Supplemental Table S1 with WGCNA. This analysis provided a modularization by an expression-based clustering of genes and allowed to describe the association of each module with the 'health phenotype score'. WGCNA also quantified the strength of gene correlations and determined hub genes for each module. We identified 12 genes (Table 1) that are among the 30 hub genes in health(span)-associated modules from at least two species. In total (Supplement Table 2), 658 different genes were found among these top-30 hub genes of all modules as determined by WGCNA. An interaction network of the genes from Table 1, based on correlation of gene expression, is presented in Figure 2.

To prioritize the cross-species hub genes of Table 1, we also looked at the module membership of all genes for each module. The genes most correlating with the module's eigengene are reported and, analogous to Table 1, the genes that are found in multiple species were determined and listed in Table 2. This table further indicates whether a gene's change in expression is positively or negatively correlated with the eigengene of the WGCNA module to which it belongs, which in turn may be positively or negatively correlated with the health(span) phenotype. Supplement Table 2 shows the WGCNA module details from which Table 2 was derived. To allow for a direct comparison of the genes' correlation with health(span), not quantitatively but in terms of direction (that is, up- or downregulation in relation to the health phenotype score), Table 2 presents a gene's inverted direction if the gene's module is already negatively correlated with the health(span) phenotype. Its column ' Correlation' presents the direction that all experiments are in agreement with or 'mixed' if the experiments differ in terms of their correlation with the health phenotype score. This information can be calculated for all genes, which we consider to help interpreting a module. The provenance of each module is described in the supplement (Supplement Table S1).

The intersection of Tables 1 (hub genes) and 2 (genes correlating with the health phenotype score) points to a subset of genes that are considered both influential and directly associated with health, i.e. MYL1, PAXIP1, PPP1CA, SCN3B, TUBG1 and WIPF1. The enrichment by g:profiler for the genes of Table 1 are shown in Figure 3. Supplement Figure S1a shows an enrichment analysis for the intersection of Tables 1 and 2 which is matching closely the enrichments in Figure 3, except that it does not feature the terms associated with muscle. Supplement Figure S1b shows the enrichment for all genes in Table 2. The latter is the least robust since the enriched terms do not cover a large fraction of the genes.

Tables 1 and 2 were both created by taking the 30 top-ranked genes in each module. The gene set enrichment analyses for Table 2 (Supplement Figure S1b) were not conclusive but the 29 genes (Table 2) in multiple species are unlikely to be achieved by chance, as we demonstrate next. The quantiles after 10000 permutations with an equal number of random genes selected for a module are

```
0%  1%  5%  10%  25%  50%  75%  90%  95%  97.5%  99%  99.5%  99.9%  100%
0   3   4   5    7    9    11   13   14   15     16   17     19     21
```

which indicates a P value < 0.0001 to find 29 genes confirmed as healthspan-associated across species with modules of the same size and a maximal selection of 30 genes

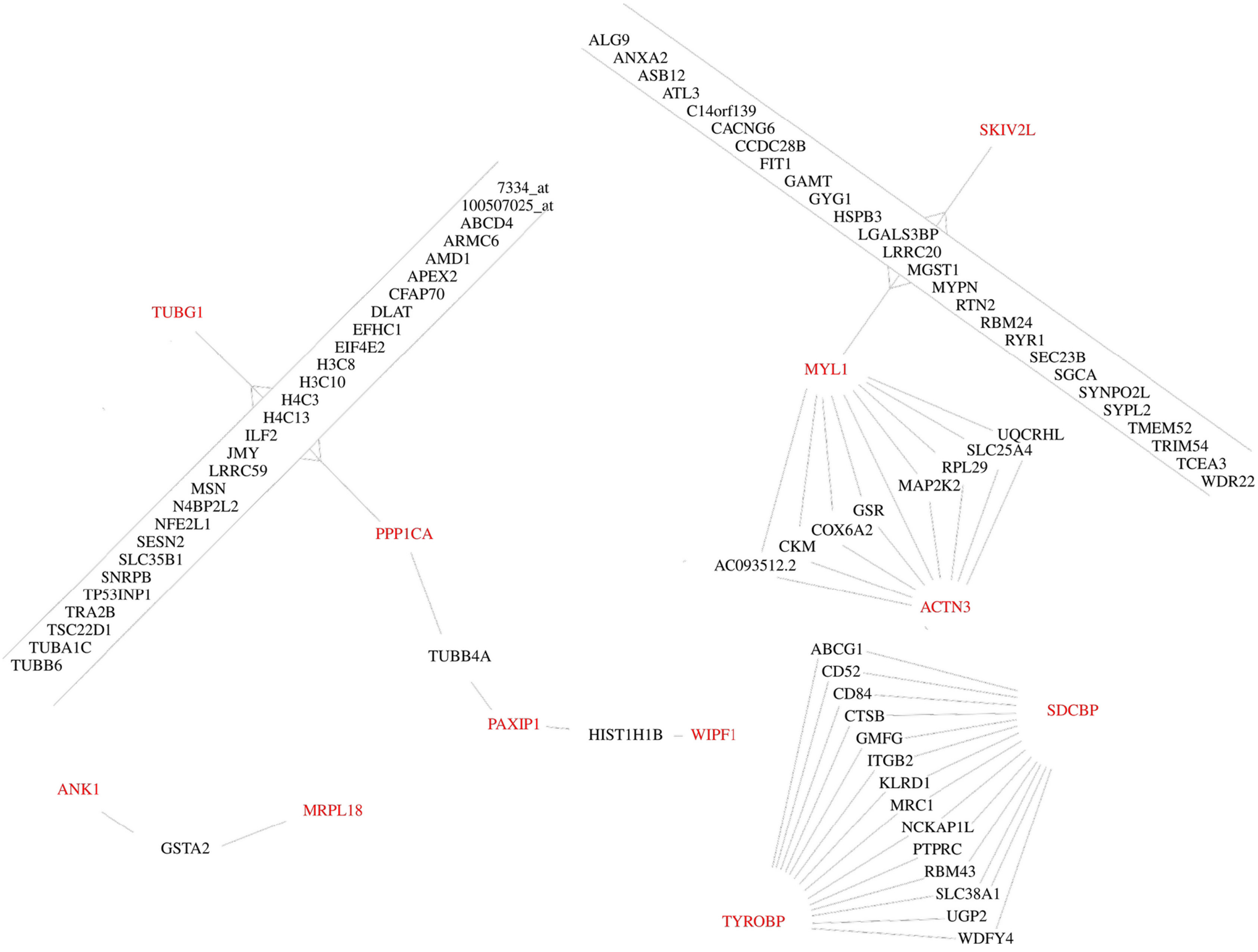


Figure 2. Cross-species conserved hub genes observed in health(span)-associated WGCNA modules, and genes that connect these hub genes. Connections are interactions taken from the WGCNA adjacency matrix if the adjacency is above the 95th percentile of all interactions of that experiment and if for that experiment the interaction is in a health(span)-associated module. The only direct interaction between hub genes is between MYL1 and ACTN3.

per module. Analogously, for the selection of hub genes, an equal selection of random genes after 10 000 iterations shows this distribution:

```
0%  1%  5%  10%  25%  50%  75%  90%  95%  97.5%  99%  99.5%  99.9%  100%
0   1   2   3    4    5    7    8    9    10     11   12     13     16
```

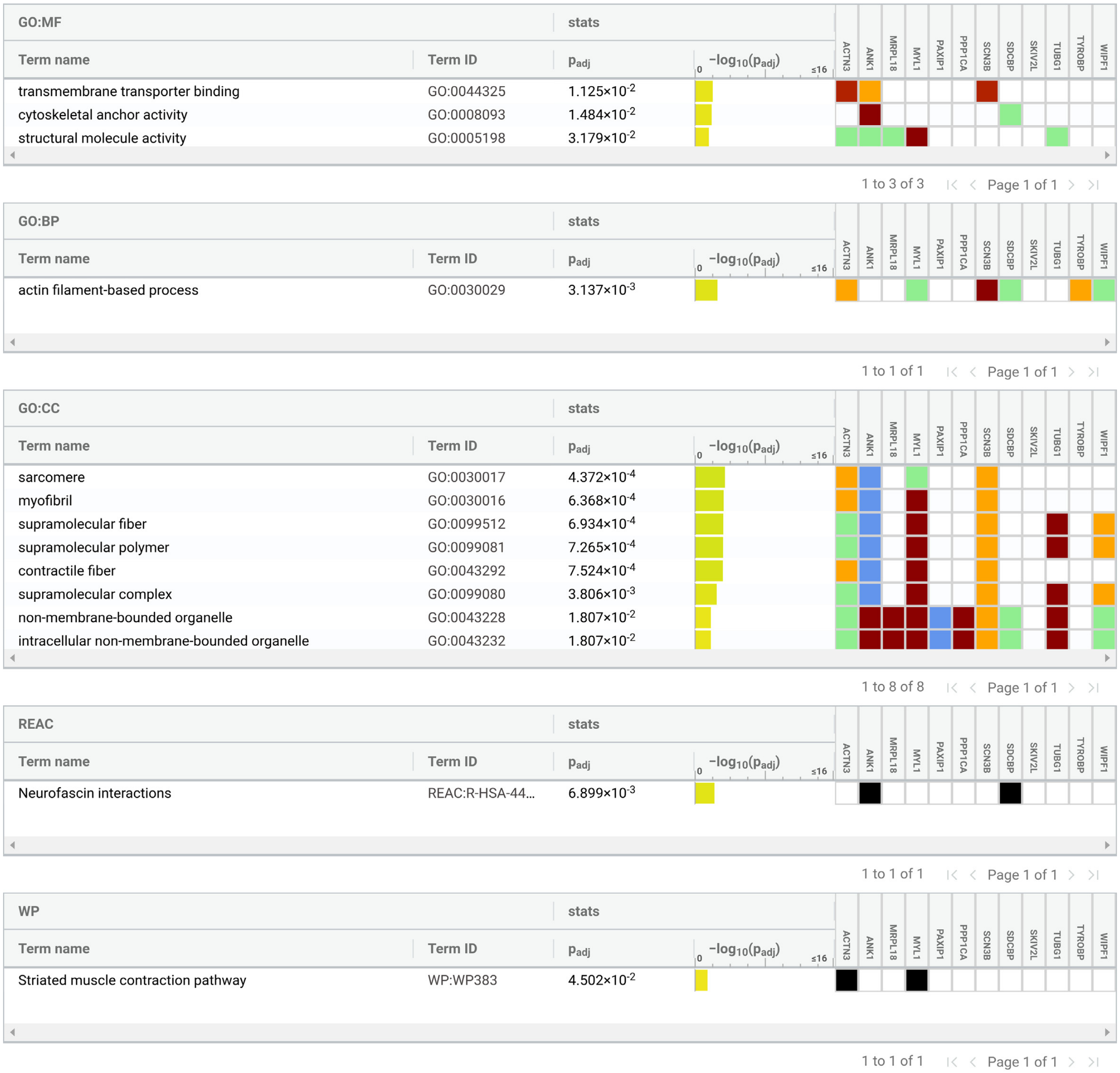
This indicates a P value < 0.01 to confirm 12 hub genes across species by chance from modules of equal size.

We also tested if the genes in Tables 1 and 2 have more interactions than one could expect by random selections of an equal number of genes in STRING. For interactions with a quality score of 900, which excludes genes that have only support by gene-correlation analyses, a P value of 0.0022 was empirically determined with 10000 repeated attempts to outperform the hub gene selection of Table 1, while the P value was found to be 0.019 (not significant) for the genes of Table 2. When lowering the quality score threshold to 200, as suggested in the tutorial of the R interface of the STRING database, then both gene selections are found to be significant with $P < 0.0001$.

Effect of data normalization

WGCNA is considered robust towards linear transformations that affect all genes in an equal fashion since only the correlation between genes affects its data processing. External factors that increase or lower the expression of all genes in a dataset will only increase the threshold above which a correlation becomes significant. Thus, a shift of the mean, also combined with a division to derive Z scores does not affect the analysis.

We also performed quantile normalization which, for each gene, examines the data across samples and which therefore does not necessarily maintain the order of expression levels for each sample. This affected the correlation scores, the modules and the module membership ranking for each module, and the results. Data for the same analysis performed on quantile normalized expression levels are shown in the supplement. For our analysis we used the data as normalized by the authors of the respective dataset. This should enable the straightforward comparison of the analyses with the publication accompanying the data and it leaves the decision regarding the normalization method and correction for batch effects to the authors of the dataset, who are much more familiar with it.



GO:MF		stats													
Term name	Term ID	p_{adj}	0 $-\log_{10}(p_{adj})$ ≤16	ACTN3	ANK1	MRPL18	MYL1	PAXIP1	PPP1CA	SCN3B	SDCBP	SKIV2L	TUBG1	TYROBP	WIPF1
transmembrane transporter binding	GO:0044325	1.125×10^{-2}	🟨	🟥	🟧					🟥					
cytoskeletal anchor activity	GO:0008093	1.484×10^{-2}	🟨		🟥						🟩				
structural molecule activity	GO:0005198	3.179×10^{-2}	🟨	🟩	🟩	🟩	🟥						🟩		

1 to 3 of 3 |< < Page 1 of 1 > >|

GO:BP		stats													
Term name	Term ID	p_{adj}	0 $-\log_{10}(p_{adj})$ ≤16	ACTN3	ANK1	MRPL18	MYL1	PAXIP1	PPP1CA	SCN3B	SDCBP	SKIV2L	TUBG1	TYROBP	WIPF1
actin filament-based process	GO:0030029	3.137×10^{-3}	🟨	🟧			🟩			🟥	🟩			🟧	🟩

1 to 1 of 1 |< < Page 1 of 1 > >|

GO:CC		stats													
Term name	Term ID	p_{adj}	0 $-\log_{10}(p_{adj})$ ≤16	ACTN3	ANK1	MRPL18	MYL1	PAXIP1	PPP1CA	SCN3B	SDCBP	SKIV2L	TUBG1	TYROBP	WIPF1
sarcomere	GO:0030017	4.372×10^{-4}	🟨	🟧	🟦		🟩			🟧					
myofibril	GO:0030016	6.368×10^{-4}	🟨	🟧	🟦		🟥			🟧					
supramolecular fiber	GO:0099512	6.934×10^{-4}	🟨	🟩	🟦		🟥			🟧			🟥		🟧
supramolecular polymer	GO:0099081	7.265×10^{-4}	🟨	🟩	🟦		🟥			🟧			🟥		🟧
contractile fiber	GO:0043292	7.524×10^{-4}	🟨	🟧	🟦		🟥			🟧					
supramolecular complex	GO:0099080	3.806×10^{-3}	🟨	🟩	🟦		🟥			🟧			🟥		🟧
non-membrane-bounded organelle	GO:0043228	1.807×10^{-2}	🟨	🟩	🟥	🟥	🟥	🟦	🟥	🟧	🟩		🟥		🟩
intracellular non-membrane-bounded organelle	GO:0043232	1.807×10^{-2}	🟨	🟩	🟥	🟥	🟥	🟦	🟥	🟧	🟩		🟥		🟩

1 to 8 of 8 |< < Page 1 of 1 > >|

REAC		stats													
Term name	Term ID	p_{adj}	0 $-\log_{10}(p_{adj})$ ≤16	ACTN3	ANK1	MRPL18	MYL1	PAXIP1	PPP1CA	SCN3B	SDCBP	SKIV2L	TUBG1	TYROBP	WIPF1
Neurofascin interactions	REAC:R-HSA-44...	6.899×10^{-3}	🟨		⬛						⬛				

1 to 1 of 1 |< < Page 1 of 1 > >|

WP		stats													
Term name	Term ID	p_{adj}	0 $-\log_{10}(p_{adj})$ ≤16	ACTN3	ANK1	MRPL18	MYL1	PAXIP1	PPP1CA	SCN3B	SDCBP	SKIV2L	TUBG1	TYROBP	WIPF1
Striated muscle contraction pathway	WP:WP383	4.502×10^{-2}	🟨	⬛			⬛								

1 to 1 of 1 |< < Page 1 of 1 > >|

Figure 3. Gene set enrichment analysis of cross-species hub genes for health(span) with g:profiler. Input are genes from Table 1 that are observed in healthspan-associated modules of multiple species. Terms with a low coverage of genes are not suitable to describe the selection as a whole but may still direct the interpretation of parts of the network where these genes are connected. Color codes are used for the *P* value (4th column) and color represents the source for the gene assignment (5th column), cf. the g:Profiler documentation, https://biit.cs.ut.ee/gprofiler/page/docs.

DISCUSSION

Method

The basis of this investigation were all experiments in GEO/ArrayExpress that mention ‘healthspan’ in their description (or ‘health’ or ‘healthspan’ in case of worm). For each experiment, from the descriptions that are provided for the samples in the database, a ‘health phenotype score’ was derived. A gene expression correlation analysis with WGCNA yielded a gene coexpression network for each experiment as a set of modules of genes that correlate with the health(span) phenotype. We were interested in genes that are most connected, i.e. hub genes, for each module, and in their interactions as described by the WGCNA network. The correlation of genes with the module eigengene (Table 2), to predict a positive or negative association with health in the molecular context of that module, was only of secondary interest to us.

In this analysis, we focussed on common observations across two or more species and a variety of health-related phenotypes, including the reaction to drugs that extend healthspan (Supplementary Table S1). The first steps of our analysis with WGCNA identified modules directly from the expression data, i.e. without inspecting a phenotype; the selection of health(span)-associated modules was performed in a later step. The WGCNA protocol was directly derived from the WGCNA tutorial.

The selection of genes, based on strong connectivity, from modules selected in such a way shall hence be considered robust even if the mapping of the multi-factorial sample descriptions to a single factor, that is, the health phenotype score describing the health-effect observed in samples, may allow for plausible alternatives. This is another reason, besides the need for abstraction to compare experiments, why we consider it advantageous to compare the module's genes against the module's eigengene, which is derived solely by an inspection of the expression data, and not against the health phenotype score (as done in Table 2). The manual intervention to derive the health phenotype score was solely needed to filter for health(span) associated modules (Supplementary Table S2).

To filter for gene interactions, we decided to filter for the strongest 5% of adjacencies from each experiment, further constrained to modules that are associated with the health(span) phenotype score; see the Methods section for details. This experiment-dependent threshold reflects that experiments differ in the number of samples and subgroups and hence in the contrasts to separate genes by their correlations.

The authors of WGCNA suggested that their software can be used to perform network meta-studies from multiple microarray experiments in a single WGCNA setup (74). But they clearly stated that the same module needs to be robust across experiments to directly perform WGCNA on a single joint matrix based on all expression data. For the very diverse set of experiments contributing to our analysis and their polygenic phenotype this is not necessarily expected to be the case, i.e. experiments may have their true healthspan-associated module in different sections of the transcriptome. Indeed, we did not observe any interactions to have orthologs across species. The setup presented here is pragmatic and robust, i.e. individual experiments can be removed without affecting the gene interactions determined for another experiment. Of major concern for us was that hub genes are expected to show a measurable effect on health(span) only under the conditions of those ArrayExpress/GEO experiments in which they are differentially expressed. To follow this work up with wet lab confirmations, it is hence essential to provide provenance information on how the change to the hub gene's expression was induced, i.e. a pointer to the ArrayExpress/GEO experiment. In a joint matrix across many experiments this information would be more difficult to retrieve, which suggests not to conduct the integration of experiments directly within a single WGCNA analysis.

Furthermore, for integrating interaction data from multiple experiments, the authors of WGCNA suggested to weigh the interactions from each experiment to derive a single joint adjacency matrix and they suggested to apply a threshold on that single matrix to derive a network. Because of the heterogeneity of our experiments, we cannot tell which experiment would be more informative for health(span), compared to another, and thus could not adjust weights accordingly. By treating all experiments individually, with the null hypothesis that all experiments have the same fraction of true interactions that shall be identified by the respective highest adjacency values, we could use an experiment-tailored threshold for filtering the interactions. Therefore, we used the 95th percentile of correlation values in the adjacency matrix, for each experiment, to adapt the selection of the interactions to be forwarded to describe a meta-study consensus (see Figures 1 and 2). These gene interactions may be trusted and they thus could be reassembled into a larger integrated meta-study network to reflect the molecular neighborhoods of hub genes, which we presented as Figure 2 (cross-species) and Supplement Figures 2a-d (for multiple modules of the respective same species). The comparison of findings across species further strengthens the confidence in the WGCNA results. Thus, we identified conserved candidate regulators of health(span).

An important technical concern lies with the interpretation of gene expression correlation data for RNA-seq experiments, which have an intrinsic high noise-level for low-abundant genes. We have shown (75) that even for array data (that are less noisy for low-abundant genes), the low-abundant genes have a measurable effect on a ranking of genes by Pearson correlation, and this is likely also the case for module calculations as performed here. This concern has to be borne in mind in the following interpretation of the modules in terms of biological functionality.

Cross-species hub genes and their interactions

Most of the hub genes identified by our analysis (Table 1) have been described in a health(span)-context before. The gene set enrichment analysis with g:profiler describes the molecular roles of the cross-species hub genes (Table 1) as specifically associated with a) features of the muscle and b) actin filament-based organelles and movement (Figure 3). The worm is a model species also for muscle development because of striking similarities of its muscles to mammalian muscle tissue (76), and movement (locomotion) is an important phenotype in all species towards operationalizing health by quantification (17). For human, rat and mouse in Supplementary Table S1, there are experiments for which samples were selectively taken from muscle tissue, but not so for the worm, which is routinely sequenced as a whole. Upon closer inspection of the enrichment results of Figure 3, we found that 'actin filament-based movement' refers to a wide spectrum of processes, i.e. genes that support actin polymerisation (WIPF1), the motor protein myosin (MYL1) or the transition of endosomes into exosomes for intercellular communication (SDCBP).

The number of experiments of vertebrates and invertebrates is balanced. Apart from a lack of tissue specificity, the experiments for the worm differ from rodents and humans, in that worm experiments may comprise samples from different larval stages. This may ease the task of finding strong correlations between genes, but specificity for aging-associated processes is likely reduced.

Inspecting the distribution of hub genes by species, we found no more than five of the 12 hub genes in worm, cf. Supplement Figure S2b, and four in human, cf. Supplement Figure S2a. The only conserved direct interaction between consensus hub genes was observed between MYL1 and ACTN3 (Figure 2). However, interactions were found multiple times for experiments of the same species, namely ABRA with VRK2, AQP11 with GSTA2 and CYLD with PCNX2 for the worm. These three interactions are shown in Supplement Figure S2b and the VRK2 gene remains directly connected with the PPP1CA hub gene also after the minimum-spanning-tree-based edge removal. VRK2 is described to have downstream effects on the consensus hub gene PPP1CA (77) via GSK3beta (78). Its genetic variants are associated with a series of neurological diseases and viral infection, but also with healthspan associated sleep patterns (79). The interactions conserved in multiple species are not confirmed in STRING for humans, however in worms, the consensus hub gene PPP1CA (C06A1.3) links to VRK2 (tag-191).

By interpreting the enrichments in Supplement Figure S3 we can gain more insight into how the genes we identified may be involved in health. An example is the enrichment referring to the TYROBP pathway described in wikipathways and to the GO term Leukocyte activation (Supplement Figure S3a). Genes connecting MYL1 and SKIV2L are involved in muscular structures (Supplement Figure S3b). Tubulins (e.g. TUBG1) are known to bind to PP1, of which PPP1CA is a subunit and together these proteins regulate histone acetylation (80), which is reflected by the genes connecting PPP1CA and TUBG1 (Supplement Figure S3c). Further, enhanced histone acetylation is associated with extended health and lifespan in worm (81).

The highly connected genes selected in this study differ from the list we recently published (18). This WGCNA-based study does not refer to prior knowledge about genetic contributions and does not perform a factor analysis to directly associate genes with a health(span) phenotype. Instead, our focus here is the network-centric interpretation of correlations within gene co-expression clusters, i.e. WCGNA modules. It is the module as a whole that correlates in its expression with health, not necessarily the individual genes. Table 2 lists genes within the clusters that are most representative for the features/characteristics of the cluster in question, i.e. that have the highest degree of *module membership* by WGCNA definition, and in the table, there are marks (by boldface) for the subset of genes that are also hub genes. Of the cross-species hub genes in Table 1, six are also listed in Table 2. Others are 'near misses', e.g. Table 2 does not list the consensus hub gene MRPL18 but MRPL19. And besides the consensus hub gene PPP1CA, other PP1 subunits like PPP1CB and PPP2R3C are found in two species (Table 2). The PP1 subunit PPP1R8 was found as a hub gene only for the worm (Supplement Figure S2b).

In Table 2, we report SQSTM1 as the only gene that is associated with health in three species. That gene was long suggested to be aging- and health-related (82,83), also for human, even though it was only found to be health(span) associated in the animal experiments of this study. Its transcript is negatively correlated with health, but SQSTM1 overexpression is known to extend healthspan in worm (84), which may be suggestive for a protective upregulation effect.

We performed the permutation test as a means for an internal validation of the analysis. The re-analysis on a differently normalized dataset means that any consensus is given with extra confidence in the association of PPP1CA with healthspan (this association was found in all runs with only 30 genes exported, for interactions and as a hub gene). Extending the quantile normalization's export to 60 genes also confirmed CEBPB, PAXIP1, and SQSTM1 with healthspan. If exporting 60 instead of 30 genes per module for the original data (data not shown) and for quantile normalization (Supplement Tables S3 and S4), the consensus grows to ABCD4, ANXA2, BAP1, GDF15, PPP1CA, SDCBP and SQSTM1, for which g:profiler finds an enrichment for 'exosome formation'.

In a review by Bartha and coworkers (85) a gene is defined as essential when its dysfunction has a strong effect on an individual's viability/fitness. Healthspan is also influenced by resilience, which without respective environmental stimuli will not show its effect on fitness and hence is less likely to be identified by knock-out/down experiments in vitro. Nevertheless, we provide references to the OGEE database for all genes in Tables 1 and 2. Most genes are found essential in a fraction of cell cultures, both for the selected hub genes and the genes that highly correlate with the healthspan phenotype. The OGEE database will keep growing, but if no indication of essentiality has yet been reported and a gene is positively correlated with the health score then this may be an early indication for a gene to be simply repairing or otherwise strengthening an individual's resilience. We also cross-checked with the GTEx database for genes that appear equally present across tissues and gender, which would be the expected characteristics for housekeeping genes. To represent the common notion of housekeeping genes, we combined multiple scores including the proportion of samples that express a certain gene as well as the gene's mean expression and its standard deviation (86). This finds the genes from both Table 1 and Table 2 to be enriched (Wilcoxon rank sum test $P_{\text{Table 1}} = 0.0001$ and $P_{\text{Table 2}} < 0.0001$) for their ubiquity across tissues. The most ubiquitous genes of Table 1 are PPP1CA, SDCBP, and MRPL1, of Table 2 these are RPL29, PPP1CA and PPP1CB.

We could not increase the number of health-associated modules (or genes within modules) by lowering the minimal module size that was defaulting to 30. Nevertheless, the low P values from the permutation tests give extra confidence in our findings. It should however be noted that the WGCNA tutorial's limit on 30 genes and the threshold imposed on the interaction score were key to achieve these significance levels. With more genes exported per module this increases the likelihood to find a match in another species by chance. With the quantile normalization applied we could double the export per module to 60 genes while remaining significant in the permutation test with 67 hub genes in the consensus across species ($P < 0.001$, 50% quantile at 44, Supplement Table S3).

Overall, our meta-analysis of a highly diverse set of transcriptomics experiments successfully identified genes which, for the most part, were already established to be closely associated with health(span), and together they have a strong and meaningful GO term enrichment. The enrichment of muscle-related genes can be credited to our focus on health(span) experiments, and our study found many 'actin filament-based movement' genes (Figure 3) that provide the cellular infrastructure not just for movement, but also for signaling and cell division, which may be triggered/blocked whenever cells start to feel unwell. If so, then it may be possible to detect many healthspan genes solely by inspecting cellular data. This hypothesis may be confirmed by an extension of our setup to a larger set of cellular transcriptomics data sets for which samples vary in their genetic or environmental exposure to stress factors.

This study provided a cross-species meta-study of gene interactions for health(span)-related datasets in ArrayExpress/GEO. It focused on a series of co-expression network analyses and subsequently on derived hub genes, instead of a focus on those genes that correlate the most with the 'healthspan phenotype score'. Tissues, technologies and experimental setups differ between the experiments but are homogeneous for each WGCNA analysis performed. This approach allows for an abstraction from the experiment at hand and permits a search for common mediators of an effect. The proposed consensus hub genes were plausible in their implication into health(span). Their interactions could be confirmed in STRING, or were found consistent with gene set enrichment analyses and they may support the interpretation of joint or epistatic effects between pairs of haplotypes in healthspan GWAS or linkage analyses. The protocol as provided with WGCNA is very transparent so that findings can be traced back to the experiments that are backing them, to serve as a template for further investigations in the wet lab.

DATA AVAILABILITY

Our implementation is available online at https://bitbucket.org/ibima/healthspantranscriptomicsnetworks/.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

The authors thank Joao Pedro de Magalhães for his pointer to gene essentiality. An anonymous reviewer stimulated the investigation if the reported hub genes are enriched for their number of interactions in protein-protein interaction databases and the investigation of housekeeping characteristics with GTex. Daniel Palmer contributed a series of valuable improvements to the manuscript. Steve Horvath commented on the robustness of WGCNA to linear transformations.

FUNDING

European Union's Horizon 2020 research and innovation programme [633589 (Aging with Elegans)]; NRDI Fund based on the charter of bolster issued by the NRDI Office under the auspices of the Ministry for Innovation and Technology [BME NC TKP2020, BME IE-BIO TKP2020]; Artificial Intelligence National Laboratory Programme; BMBF (Validierung des technologischen und gesellschaftlichen Innovationspotenzials wissenschaftlicher Forschung – VIP+, FKZ) [03VP06230]; A. Gézsi was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences; computations were performed at the Uni Rostock ITMZ cluster with support by the European Union (EFRE, 'European Fund for Regional Development') [GHS-15-0019, 2016-2021].

Conflict of interest statement. None declared.

REFERENCES

1. Tacutu,R., Thornton,D., Johnson,E., Budovsky,A., Barardo,D., Craig,T., Diana,E., Lehmann,G., Toren,D., Wang,J. *et al.* (2018) Human ageing genomic resources: new and updated databases. *Nucleic Acids Res.*, **46**, D1083–D1090.
2. Zahn,J.M., Poosala,S., Owen,A.B., Ingram,D.K., Lustig,A., Carter,A., Weeraratna,A.T., Taub,D.D., Gorospe,M., Mazan-Mamczarz,K. *et al.* (2007) AGEMAP: a gene expression database for aging in mice. *PLoS Genet.*, **3**, e201.
3. Consortium,AgingAtlas, Liu,G.-H., Bao,Y., Qu,J., Zhang,W., Zhang,T., Kang,W., Yang,F., Ji,Q., Jiang,X *et al.* (2021) Aging atlas: a multi-omics database for aging biology. *Nucleic Acids Res.*, **49**, D825–D830.
4. Dönertaş,H.M., Valenzuela,Fuentealba, Partridge,M. and Thornton,J.M. (2018) Gene expression-based drug repurposing to target aging. *Aging Cell*, **17**, e12819.
5. de Magalhães,J.P., Curado,J. and Church,G.M. (2009) Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics*, **25**, 875–881.
6. Tacutu,R., Craig,T., Budovsky,A., Wuttke,D., Lehmann,G., Taranukha,D., Costa,J., Fraifeld,V.E. and de Magalhães,J.P. (2012) Human ageing genomic resources: integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Res.*, **41**, D1027–D1033.
7. van Dam,S., Cordeiro,R., Craig,T., van Dam,J., Wood,S.H. and de Magalhães,J. (2012) GeneFriends: an online co-expression analysis tool to identify novel gene targets for aging and complex diseases. *BMC Genomics*, **13**, 535.
8. Kotlyar,M., Pastrello,C., Malik,Z. and Jurisica,I. (2019) IID 2018 update: context-specific physical protein–protein interactions in human, model organisms and domesticated species. *Nucleic Acids Res.*, **47**, D581–D589.
9. Leiserson,M.D., Eldridge,J.V., Ramachandran,S. and Raphael,B.J. (2013) Network analysis of GWAS data. *Curr. Opin. Genet. Dev.*, **23**, 602–610.
10. Cowen,L., Ideker,T., Raphael,B.J. and Sharan,R. (2017) Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.*, **18**, 551–562.
11. Lee,J., Shah,M., Ballouz,S., Crow,M. and Gillis,J. (2020) CoCoCoNet: conserved and comparative co-expression across a diverse set of species. *Nucleic Acids Res.*, **48**, W566–W571.
12. Avelar,R.A., Ortega,J.G., Tacutu,R., Tyler,E.J., Bennett,D., Binetti,P., Budovsky,A., Chatsirisupachai,K., Johnson,E., Murray,A. *et al.* (2020) A multidimensional systems biology analysis of cellular senescence in aging and disease. *Genome Biol.*, **21**, 91.
13. Obayashi,T., Kagaya,Y., Aoki,Y., Tadaka,S. and Kinoshita,K. (2019) COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. *Nucleic Acids Res.*, **47**, D55–D62.
14. Faisal,F.E. and Milenković,T. (2014) Dynamic networks reveal key players in aging. *Bioinformatics*, **30**, 1721–1729.
15. van Dam,S., Craig,T. and de Magalhães,J.P. (2015) GeneFriends: a human RNA-seq-based gene and transcript co-expression database. *Nucleic Acids Res.*, **43**, D1124–D1132.
16. Adler,P., Kolde,R., Kull,M., Tkachenko,A., Peterson,H., Reimand,J. and Vilo,J. (2009) Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. *Genome Biol.*, **10**, R139.
17. Fuellen,G., Jansen,L., Cohen,A.A., Luyten,W., Gogol,M., Simm,A., Saul,N., Cirulli,F., Berry,A., Antal,P. *et al.* (2019) Health and aging: unifying concepts, scores, biomarkers and pathways. *Aging Dis.*, **10**, 883.
18. Möller,S., Saul,N., Cohen,A.A., Köhling,R., Sender,S., Murua Escobar,H., Junghanss,C., Cirulli,F., Berry,A., Antal,P. *et al.* (2020) Healthspan pathway maps in c. elegans and humans highlight transcription, proliferation/biosynthesis and lipids. *Aging*, **12**, 12534–12581.
19. Franz,M., Rodriguez,H., Lopes,C., Zuberi,K., Montojo,J., Bader,G.D. and Morris,Q. (2018) GeneMANIA update 2018. *Nucleic Acids Res.*, **46**, W60–W64.

20. Kucera,M., Isserlin,R., Arkhangorodsky,A. and Bader,G.D. (2016) AutoAnnotate: a cytoscape app for summarizing networks with semantic annotations. *F1000Research*, **5**, 1717.
21. Magger,O., Waldman,Y.Y., Ruppin,E. and Sharan,R. (2012) Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLoS Comput. Biol.*, **8**, e1002690.
22. Li,M.-L., Wu,S.-H., Zhang,J.-J., Tian,H.-Y., Shao,Y., Wang,Z.-B., Irwin,D.M., Li,J.-L., Hu,X.-T. and Wu,D.-D. (2019) 547 transcriptomes from 44 brain areas reveal features of the aging brain in non-human primates. *Genome Biol.*, **20**, 258.
23. van Dam,S., Võsa,U., van der Graaf,A., Franke,L. and de Magalhães,J.P. (2018) Gene co-expression analysis for functional classification and gene–disease predictions. *Brief. Bioinform.*, **19**, 575–592.
24. Clough,E. and Barrett,T. (2016) The gene expression omnibus database. In: Mathé,E. and Davis,S. (eds). *Statistical Genomics, Methods in Molecular Biology*. Springer New York, NY, Vol. **1418**, pp. 93–110.
25. Athar,A., Füllgrabe,A., George,N., Iqbal,H., Huerta,L., Ali,A., Snow,C., Fonseca,N.A., Petryszak,R., Papatheodorou,I. *et al.* (2019) ArrayExpress update – from bulk to single-cell expression data. *Nucleic Acids Res.*, **47**, D711–D715.
26. Zhao,Y., Wong,L. and Goh,W.W.B. (2020) How to do quantile normalization correctly for gene expression data analyses. *Sci. Rep.*, **10**, 15534.
27. Gurumayum,S., Jiang,P., Hao,X., Campos,T.L., Young,N.D., Korhonen,P.K., Gasser,R.B., Bork,P., Zhao,X.-M., He,L. *et al.* (2021) OGEE v3: online GEne essentiality database with increased coverage of organisms and human cell lines. *Nucleic Acids Res.*, **49**, D998–D1003.
28. Leinonen,R., Akhtar,R., Birney,E., Bower,L., Cerdeno-Tarraga,A., Cheng,Y., Cleland,I., Faruque,N., Goodgame,N., Gibson,R. *et al.* (2011) The European nucleotide archive. *Nucleic Acids Res.*, **39**, D28–D31.
29. Bushnell,B., Rood,J. and Singer,E. (2017) BBMerge – accurate paired shotgun read merging via overlap. *PLoS One*, **12**, e0185056.
30. Yates,A.D., Achuthan,P., Akanni,W., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R. *et al.* (2020) Ensembl 2020. *Nucleic Acids Res.*, **48**, D682–D688.
31. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
32. Li,B., Ruotti,V., Stewart,R.M., Thomson,J.A. and Dewey,C.N. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**, 493–500.
33. Langfelder,P. and Horvath,S. (2008) WGCNA: an r package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.
34. Zhao,W., Langfelder,P., Fuller,T., Dong,J., Li,A. and Hovarth,S. (2010) Weighted gene coexpression network analysis: state of the art. *J. Biopharm. Stat.*, **20**, 281–300.
35. Szklarczyk,D., Gable,A.L., Nastou,K.C., Lyon,D., Kirsch,R., Pyysalo,S., Doncheva,N.T., Legeay,M., Fang,T., Bork,P. *et al.* (2021) The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.*, **49**, D605–D612.
36. Pickering,C. and Kiely,J. (2018) ACTN3, morbidity, and healthy aging. *Front. Genet.*, **9**, 15.
37. Spracklen,C.N., Horikoshi,M., Kim,Y.J., Lin,K., Bragg,F., Moon,S., Suzuki,K., Tam,C.H.T., Tabara,Y., Kwak,S.-H. *et al.* (2020) Identification of type 2 diabetes loci in 433,540 east asian individuals. *Nature*, **582**, 240–245.
38. Qin,L., Nie,Y., Zhang,H., Chen,L., Zhang,D., Lin,Y. and Ru,K. (2020) Identification of new mutations in patients with hereditary spherocytosis by next-generation sequencing. *J. Hum. Genet.*, **65**, 427–434.
39. Morris,G., Berk,M., Maes,M. and Puri,B.K. (2019) Could alzheimer's disease originate in the periphery and if so how so?*Mol. Neurobiol.*, **56**, 406–434.
40. Zhang,X., Gao,X., Coots,R.A., Conn,C.S., Liu,B. and Qian,S.-B. (2015) Translational control of the cytosolic stress response by mitochondrial ribosomal protein L18. *Nat. Struct. Mol. Biol.*, **22**, 404–410.
41. Stuart,C.A., Stone,W.L., Howell,M.E.A., Brannon,M.F., Hall,H.K., Gibson,A.L. and Stone,M.H. (2016) Myosin content of individual human muscle fibers isolated by laser capture microdissection. *Am. J. Physiol.-Cell Physiol.*, **310**, C381–C389.
42. Cobley,J., Malik,Ab., Morton,Z., Close,J., Edwards,G. and Burniston,J. (2016) Age- and Activity-Related differences in the abundance of myosin essential and regulatory light chains in human muscle. *Proteomes*, **4**, 15.
43. De Gregoriis,G., Ramos,J.A., Fernandes,P.V., Vignal,G.M., Brianese,R.C., Carraro,D.M., Monteiro,A.N., Struchiner,C.J., Suarez-Kurtz,G., Vianna-Jorge,R. *et al.* (2017) DNA repair genes *PAXIP1* and *TP53BP1* expression is associated with breast cancer prognosis. *Cancer Biol. Ther.*, **18**, 439–449.
44. Ceulemans,H. and Bollen,M. (2004) Functional diversity of protein phosphatase-1, a cellular economizer and reset button. *Physiol. Rev.*, **84**, 1–39.
45. Castro,M.E., Ferrer,I., Cascon,A., Guijarro,M.V., Lleonart,M., Cajal,S.R., Leal,J.F.M., Robledo,M. and Carnero,A. (2007) PPP1CA contributes to the senescence program induced by oncogenic ras. *Carcinogenesis*, **29**, 491–499.
46. Sun,H., Ou,B., Zhao,S., Liu,X., Song,L., Liu,X., Wang,R. and Peng,Z. (2019) USP11 promotes growth and metastasis of colorectal cancer via PPP1CA-mediated activation of ERK/MAPK signaling pathway. *EBioMedicine*, **48**, 236–247.
47. Korrodi-Gregório,L., Silva,J.V., Santos-Sousa,L., Freitas,M.J., Felgueiras,J. and Fardilha,M. (2014) TGF-β cascade regulation by PPP1 and its interactors -impact on prostate cancer development and therapy. *J. Cell. Mol. Med.*, **18**, 555–567.
48. Ruiz,L., Traskine,M., Ferrer,I., Castro,E., Leal,J.F.M., Kaufman,M. and Carnero,A. (2008) Characterization of the p53 response to oncogene-induced senescence. *PLoS One*, **3**, e3230.
49. Silva,J.V., Freitas,M.J. and Fardilha,M. (2014) Phosphoprotein phosphatase 1 complexes in spermatogenesis. *Curr. Mol. Pharmacol.*, **7**, 136–146.
50. Banzhaf-Strathmann,J., Benito,E., May,S., Arzberger,T., Tahirovic,S., Kretzschmar,H., Fischer,A. and Edbauer,D. (2014) Micro RNA -125b induces tau hyperphosphorylation and cognitive deficits in alzheimer's disease. *EMBO J.*, **33**, 1667–1680.
51. Ishikawa,T., Takahashi,N., Ohno,S., Sakurada,H., Nakamura,K., On,Y.K., Park,J.E., Makiyama,T., Horie,M., Arimura,T. *et al.* (2013) Novel SCN3B mutation associated with brugada syndrome affects intracellular trafficking and function of nav1.5. *Circ. J. Off. J. Jpn. Circ. Soc.*, **77**, 959–967.
52. Li,S., Han,J., Guo,G., Sun,Y., Zhang,T., Zhao,M., Xu,Y., Cui,Y., Liu,Y. and Zhang,J. (2020) Voltage-gated sodium channels β3 subunit promotes tumorigenesis in hepatocellular carcinoma by facilitating p53 degradation. *FEBS Lett.*, **594**, 497–508.
53. Rajesh,S., Bago,R., Odintsova,E., Muratov,G., Baldwin,G., Sridhar,P., Rajesh,S., Overduin,M. and Berditchevski,F. (2011) Binding to syntenin-1 protein defines a new mode of ubiquitin-based interactions regulated by phosphorylation. *J. Biol. Chem.*, **286**, 39606–39614.
54. Baietti,M.F., Zhang,Z., Mortier,E., Melchior,A., Degeest,G., Geeraerts,A., Ivarsson,Y., Depoortere,F., Coomans,C., Vermeiren,E. *et al.* (2012) Syndecan-syntenin-ALIX regulates the biogenesis of exosomes. *Nat. Cell Biol.*, **14**, 677–685.
55. Fares,J., Kashyap,R. and Zimmermann,P. (2017) Syntenin: key player in cancer exosome biogenesis and uptake?*Cell Adhes. Migr.*, **11**, 124–126.
56. Herrera-Moyano,E., Porreca,R.M., Ranjha,L., Gonzalez-Franco,R., Skourti,E., Sun,Y., Stylianakis,E., Montoya,A., Kramer,H. and Vannier,J.-B. (2020) Human SKI is a telomere-associated complex involved in DNA-RNA hybrid control and telomere stability molecular biology. bioRxiv doi: https://doi.org/10.1101/2020.05.20.107144, 22 May 2020, preprint: not peer reviewed.
57. Vardi,I., Barel,O., Sperber,M., Schvimer,M., Nunberg,M., Field,M., Ouahed,J., Marek-Yagel,D., Werner,L., Haberman,Y. *et al.* (2018) Genetic and structural analysis of a SKIV2L mutation causing Tricho-hepato-enteric syndrome. *Dig. Dis. Sci.*, **63**, 1192–1199.
58. Shuai,P., Ye,Z., Liu,Y., Qu,C., Liu,X., Luo,H., Feng,X., Li,X., Shi,Y. and Gong,B. (2017) Association between *SKIV2L* polymorphism rs429608 and age-related macular degeneration: a meta-analysis. *Ophthalmic Genet.*, **38**, 245–251.

59. Schott,G. and Garcia-Blanco,M.A. (2021) MHC class III RNA binding proteins and immunity. *RNA Biol.*, **18**, 640–646.
60. Eckard,S.C., Rice,G.I., Fabre,A., Badens,C., Gray,E.E., Hartley,J.L., Crow,Y.J. and Stetson,D.B. (2014) The SKIV2L RNA exosome limits activation of the RIG-I-like receptors. *Nat. Immunol.*, **15**, 839–845.
61. Alvarado-Kristensson,M. (2018) γ-tubulin as a signal-transducing molecule and meshwork with therapeutic potential. *Signal Transduct. Target. Ther.*, **3**, 24.
62. Blanco,I., Kuchenbaecker,K., Cuadras,D., Wang,X., Barrowdale,D., de Garibay,G.R., Librado,P., Sánchez-Gracia,A., Rozas,J., Bonifaci,N. *et al.* (2015) Assessing associations between the AURKA-HMMR-TPX2-TUBG1 functional module and breast cancer risk in BRCA1/2 mutation carriers. *PLoS One*, **10**, e0120020.
63. Maounis,N.F., Dráberová,E., Mahera,E., Chorti,M., Caracciolo,V., Sulimenko,T., Riga,D., Trakas,N., Emmanouilidou,A., Giordano,A. *et al.* (2012) Overexpression of γ-tubulin in non-small cell lung cancer. *Histol. Histopathol.*, **27**, 1183–1194.
64. Caracciolo,V., D'Agostino,L., Dráberová,E., Sládková,V., Crozier-Fitzgerald,C., Agamanolis,D.P., de Chadarévian,J.-P., Legido,A., Giordano,A., Dráber,P *et al.* (2010) Differential expression and cellular distribution of gamma-tubulin and betaIII-tubulin in medulloblastomas and human medulloblastoma cell lines. *J. Cell. Physiol.*, **223**, 519–529.
65. Ma,J., Jiang,T., Tan,L. and Yu,J.-T. (2015) TYROBP in alzheimer's disease. *Mol. Neurobiol.*, **51**, 820–826.
66. Pottier,C., Ravenscroft,T.A., Brown,P.H., Finch,N.A., Baker,M., Parsons,M., Asmann,Y.W., Ren,Y., Christopher,E., Levitch,D. *et al.* (2016) TYROBP genetic variants in early-onset alzheimer's disease. *Neurobiol. Aging*, **48**, 222.e9–222.e15.
67. Liu,H., Yang,Y., Ge,Y., Liu,J. and Zhao,Y. (2019) TERC promotes cellular inflammatory response independent of telomerase. *Nucleic Acids Res.*, **47**, 8084–8095.
68. Wu,P., Xiang,T., Wang,J., Lv,R. and Wu,G. (2020) *TYROBP* is a potential prognostic biomarker of clear cell renal cell carcinoma. *FEBS Open Bio.*, **10**, 2588–2604.
69. Jiang,J., Ding,Y., Wu,M., Lyu,X., Wang,H., Chen,Y., Wang,H. and Teng,L. (2020) Identification of TYROBP and C1QB as two novel key genes with prognostic value in gastric cancer by network analysis. *Front. Oncol.*, **10**, 1765.
70. Stranahan,A.M., Martin,B., Chadwick,W., Park,S.-S., Wang,L., Becker,K.G., Woodiii,W.H., Zhang,Y. and Maudsley,S. (2012) Metabolic context regulates distinct hypothalamic transcriptional responses to antiaging interventions. *Int. J. Endocrinol.*, **2012**, 732975.
71. Zhang,T., Shen,X., Liu,R., Zhu,G., Bishop,J. and Xing,M. (2017) Epigenetically upregulated WIPF1 plays a major role in BRAF V600E-promoted papillary thyroid cancer aggressiveness. *Oncotarget*, **8**, 900–914.
72. Pan,Y., Lu,F., Xiong,P., Pan,M., Zhang,Z., Lin,X., Pan,M. and Huang,H. (2018) WIPF1 antagonizes the tumor suppressive effect of miR-141/200c and is associated with poor survival in patients with PDAC. *J. Exp. Clin. Cancer Res.*, **37**, 167.
73. Staub,E., Groene,J., Heinze,M., Mennerich,D., Roepcke,S., Klaman,I., Hinzmann,B., Castanos-Velez,E., Pilarsky,C., Mann,B. *et al.* (2009) An expression module of WIPF1-coexpressed genes identifies patients with favorable prognosis in three tumor types. *J. Mol. Med.*, **87**, 633–644.
74. Langfelder,P., Mischel,P.S. and Horvath,S. (2013) When is hub gene selection better than standard meta-analysis?*PLoS One*, **8**, e61505.
75. Struckmann,S., Ernst,M., Fischer,S., Mah,N., Fuellen,G. and Möller,S. (2021) Scoring functions for drug-effect similarity. *Brief. Bioinform.*, **22**, bbaa072.
76. Christian,C.J. and Benian,G.M. (2020) Animal models of sarcopenia. *Aging Cell*, **19**, e13223.
77. Cossa,G., Roeschert,I., Prinz,F., Baluapuri,A., Silveira Vidal,R., Schülein-Völk,C., Chang,Y.-C., Ade,C.P., Mastrobuoni,G., Girard,C. *et al.* (2020) Localized inhibition of protein phosphatase 1 by NUAK1 promotes spliceosome activity and reveals a MYC-Sensitive feedback control of transcription. *Mol. Cell*, **77**, 1322–1339.
78. Lee,E., Ryu,H.G., Kim,S., Lee,D., Jeong,Y.-H. and Kim,K.-T. (2016) Glycogen synthase kinase 3β suppresses polyglutamine aggregation by inhibiting Vaccinia-related kinase 2 activity. *Sci. Rep.*, **6**, 29097.
79. Dashti,H.S., Jones,S.E., Wood,A.R., Lane,J.M., van Hees,V.T., Wang,H., Rhodes,J.A., Song,Y., Patel,K., Anderson,S.G. *et al.* (2019) Genome-wide association study identifies genetic loci for self-reported habitual sleep duration supported by accelerometer-derived estimates. *Nat. Commun.*, **10**, 1100.
80. Ding,H., Dolan,P.J. and Johnson,G.V.W. (2008) Histone deacetylase 6 interacts with the microtubule-associated protein tau. *J. Neurochem.*, **106**, 2119–2130.
81. Zhang,M., Poplawski,M., Yen,K., Cheng,H., Bloss,E., Zhu,X., Patel,H. and Mobbs,C.V. (2009) Role of CBP and SATB-1 in aging, dietary restriction, and insulin-like signaling. *PLoS Biol.*, **7**, e1000245.
82. Bitto,A., Lerner,C.A., Nacarelli,T., Crowe,E., Torres,C. and Sell,C. (2014) p62/SQSTM1 at the interface of aging, autophagy, and disease. *AGE*, **36**, 9626.
83. Sánchez-Martín,P. and Komatsu,M. (2018) p62/SQSTM1 – steering the cell through health and disease. *J. Cell Sci.*, **131**, jcs222836.
84. Kumsta,C., Chang,J.T., Lee,R., Tan,E.P., Yang,Y., Loureiro,R., Choy,E.H., Lim,S.H.Y., Saez,I., Springhorn,A. *et al.* (2019) The autophagy receptor p62/SQST-1 promotes proteostasis and longevity in c. elegans by inducing autophagy. *Nat. Commun.*, **10**, 5648.
85. Bartha,I., di Iulio,J., Venter,J.C. and Telenti,A. (2018) Human gene essentiality. *Nat. Rev. Genet.*, **19**, 51–62.
86. Projahn,E. and Möller,S. (2022) Ubigen: Interactive Ranking and Enrichment Test for Ubiquitously Expressed Genes Bioinformatics. bioRxiv doi: https://doi.org/10.1101/2022.07.28.501880, 01 August 2022, preprint: not peer reviewed.