


Comparative Analysis of Mammal Genomes Unveils Key Genomic Variability for Human Life Span

Xavier Farré,¹ Ruben Molina,² Fabio Barteri,¹ Paul R.H.J. Timmers,^{3,4} Peter K. Joshi,⁴ Baldomero Oliva ², Sandra Acosta,¹ Borja Esteve-Altava,¹ Arcadi Navarro,^{*} ^{1,5,6,7} and Gerard Muntané^{*} ^{1,8}

¹Department of Experimental and Health Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona, Spain

²Structural Bioinformatics Lab, Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain

³MRC Human Genetics Unit, MRC Institute of Genetics & Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom

⁴Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom

⁵Barcelonaβeta Brain Research Center (BBRC), Pasqual Maragall Foundation, Barcelona, Spain

⁶Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, Barcelona, Spain

⁷Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

⁸Hospital Universitari Institut Pere Mata, IISPV, Universitat Rovira i Virgili, Biomedical Network Research Centre on Mental Health (CIBERSAM), Reus, Spain

Institution at which research was done: Department of Experimental and Health Sciences, Institut Biologia Evolutiva, Universitat Pompeu Fabra—CSIC, 08003 Barcelona, Spain.

***Corresponding authors:** E-mails: arcadi.navarro@upf.edu; gerard.muntane@upf.edu, muntaneg@peremata.com.

Associate editor: Katja Nowick

Abstract

The enormous mammal's lifespan variation is the result of each species' adaptations to their own biological trade-offs and ecological conditions. Comparative genomics have demonstrated that genomic factors underlying both, species lifespans and longevity of individuals, are in part shared across the tree of life. Here, we compared protein-coding regions across the mammalian phylogeny to detect individual amino acid (AA) changes shared by the most long-lived mammals and genes whose rates of protein evolution correlate with longevity. We discovered a total of 2,737 AA in 2,004 genes that distinguish long- and short-lived mammals, significantly more than expected by chance ($P = 0.003$). These genes belong to pathways involved in regulating lifespan, such as inflammatory response and hemostasis. Among them, a total 1,157 AA showed a significant association with maximum lifespan in a phylogenetic test. Interestingly, most of the detected AA positions do not vary in extant human populations (81.2%) or have allele frequencies below 1% (99.78%). Consequently, almost none of these putatively important variants could have been detected by genome-wide association studies. Additionally, we identified four more genes whose rate of protein evolution correlated with longevity in mammals. Crucially, SNPs located in the detected genes explain a larger fraction of human lifespan heritability than expected, successfully demonstrating for the first time that comparative genomics can be used to enhance interpretation of human genome-wide association studies. Finally, we show that the human longevity-associated proteins are significantly more stable than the orthologous proteins from short-lived mammals, strongly suggesting that general protein stability is linked to increased lifespan.

Key words: comparative genomics, aging, maximum lifespan, genetics, GWAS, convergent evolution.

Introduction

Why some individuals within a species live longer than others is intimately related to the broader question of why some species live longer than others (de Magalhães and Toussaint 2002). Maximum lifespan (MLS) is a species-specific trait: flies, dogs, and humans all have different but consistent lifespans that are adapted to their ecology and biology. From an evolutionary standpoint, the main cause of these differences is lineage-specific ecological adaptations that modify the rates of extrinsic mortality. For instance, lifespan is usually lengthened by

arborealism (Shattuck and Williams 2010), flight (Pomeroy 1990), subterranean life (Buffenstein 2005), and body mass (Austad 2005; de Magalhães et al. 2007) since all these adaptations reduce mortality by predation. Most differences and similarities in MLS across species are explained by common physiological, biochemical, and genetic factors (Ma and Gladyshev 2017).

Mammals show a 100-fold variation in MLS, ranging from short-lived species like forest shrews (~2 years) to long-lived

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access

species like bowhead whales (~200 years, [Tacutu et al. 2018](#)), and so they are an ideal lineage to study the genomics of lifespan and to unveil genes and pathways that may be relevant for humans. Numerous studies have been devoted to examine mammal lifespan focusing on individual species, such as the bowhead whale ([Keane et al. 2015](#)) or the naked mole-rat ([Kim et al. 2011](#); [Ruby, Smith, and Buffenstein 2018](#)), or highly related species groups, such as bats ([Seim et al. 2013](#); [Huang et al. 2019](#); [Wang et al. 2020](#)). Although single-species studies have yielded some credible candidate genes associated with increased lifespan, it is difficult to obtain generalizations on universal mechanisms of lifespan regulation from them. Therefore, knowledge about lifespan evolution in mammals is still limited. Our mammalian ancestors, although diverse ([Pickrell 2019](#)), were small ([O'Leary et al. 2013](#)), and probably short-lived creatures ([Berkel and Cacan 2021](#)). The known long-term directional bias towards increasing size in mammals ([Baker et al. 2015](#); [Lyson et al. 2019](#)) could have driven a parallel trend towards increased lifespan, but there is only limited evidence in favor of that hypothesis ([Berkel and Cacan 2021](#)).

Other comparative genomics studies have focused on identifying rapid evolutionary changes in genomes or transcriptomes that correlate with changes in longevity ([Kim et al. 2011](#); [Munané et al. 2018](#); [Kowalczyk et al. 2020](#)); or have assessed the relationship between lifespan and other adaptations with life-history traits in different taxa ([Montgomery and Mundy 2012](#); [Zhang et al. 2014](#); [Foote et al. 2015](#); [Chikina et al. 2016](#); [Boddy et al. 2017](#); [Wang et al. 2020](#)). These studies have identified longevity pathways that are conserved across species, such as the insulin/IGF-1 pathway, telomere maintenance, DNA repair, coagulation and wound healing, proteostasis, and TOR signaling. The existence of such common pathways and mechanisms is consistent with the fact that long-lived animals show convergent phenotypes, including increased stress resistance, altered metabolism, and delayed reproduction and development ([Hekimi and Guarente 2003](#)).

A key mechanism that may contribute to differences in lifespan is the maintenance of the proteostasis network. Protein stability or proteostasis refers to the capacity to protect protein structures and functions against environmental stressors, including aging. In fact, dysfunction of the protein quality control mechanisms is a hallmark of aging ([López-Otín et al. 2013](#); [Santra et al. 2019](#)) and there is substantial evidence linking proteostasis and longevity (reviewed in [Tian et al. \[2017\]](#)). For instance, improved protein stability is determinant for longevity in exceptionally long-lived mollusks ([Treaster et al. 2014](#)) and in the naked mole-rat, the longest-living rodent ([Pérez et al. 2009](#)). In addition, interventions that enhance proteome stability can improve health or increase lifespan in model organisms ([Fontana and Partridge 2015](#)), such as pharmacological chaperones that have been investigated as potential therapeutic targets to reduce the adverse effects of misfolding of aging-related proteins ([Bullock et al. 1997](#); [Powers et al. 2009](#)).

Despite all the evidence outlined above, a mammalian-wide study of the genomic underpinnings of lifespan has

never been carried out with the combined goals of identifying individual mutations linked to longevity; analyzing the functional properties of their genes and the pathways in which they take part; and studying how the stability of proteins coded by these genes may differentiate long- and short-lived species. In fact, the largest scale studies conducted on mammalian lifespan have focused only on humans, with somewhat limited results ([Timmers et al. 2019](#)). Part of these limitations may be due to the low heritability of lifespan in humans. Twin studies already provide estimates of only 0.2–0.3 heritability ([Herskind et al. 1996](#); [Sebastiani and Perls 2012](#)); and, more recently, the analysis of family trees produced an estimation of around 0.1, suggesting that previous estimates were inflated due to assortative mating ([Kaplanis et al. 2018](#); [Ruby et al. 2018](#)). Several genome-wide association studies (GWAS) on human lifespan have been carried out using different indirect measures and strategies, such as parental lifespan ([Timmers et al. 2019](#)), extremely long-lived individuals ([Deelen et al. 2019](#)), or health span ([Zenin et al. 2019](#)). Although these studies have identified a set of genetic variants that are associated with an individual's lifespan, only a small fraction of the heritability—around 5%—has been explained by GWAS. In summary, not only we are missing important contributions to extant human variation on lifespan, probably due to genetic variants with small effects ([Munané et al. 2018](#); [de Magalhães and Wang 2019](#)); but also, and given the relatively small genetic variation in lifespan in our species, we still lack a map of the full landscape of factors underlying the genomic architecture of human longevity.

Here, we performed the largest phylogeny-based genome-phenotype analysis to date, focusing on the detection of individual mutations and genes that underlie the enormous variation of lifespan in mammals. We report the discovery of more than 2,000 longevity-related genes and show that, overall, they present a trend towards increased protein stability in long-lived organisms. In addition, we successfully show that our findings enhance the interpretation of the results of longevity GWAS that have been carried out in humans. Altogether, our results pave the way for the use of comparative genomics studies to shed light on human traits, particularly those of potential medical interest.

Results

Convergent Amino Acid Substitutions: Discovery and Validation

To detect convergent amino acid (AA) substitutions (CAAS) shared between long-lived mammals, we split species into two groups selecting those in the extreme deciles of the longevity quotient (LQ) distribution (rounding up to six species in each group, [supplementary fig. 1, Supplementary Material online](#)). Throughout the manuscript we used the term “Convergent” rather than “Parallel” AA Substitutions to acknowledge that we cannot guarantee that each substitution had appeared independently in each species. In a first phase, the Discovery phase, we counted all AA changes in which the same AA was present in the reference genomes of the long-

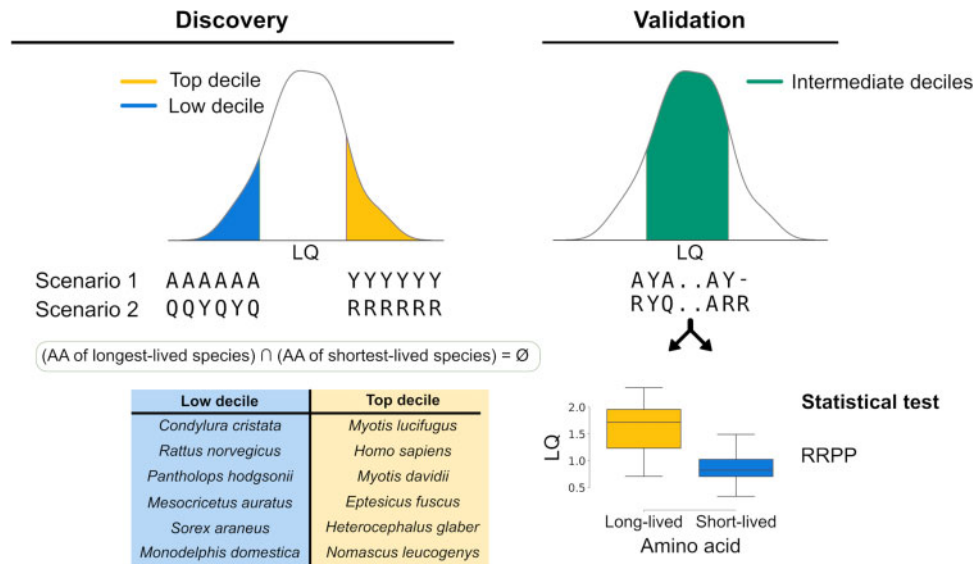


Fig. 1. Workflow used in this study for the detection of convergent amino acid substitutions (CAAS). In the Discovery phase, we identified AA substitutions that were exclusive from species in the top (yellow) and low (blue) deciles. In the Validation phase, we classified the species from intermediate deciles (green) in two groups, the species having the “long-lived” and the “short-lived” AA version. Finally, we ran a RRPP phylogenetic ANOVA to validate each discovered AA, keeping as significant only those for which we validated the direction of the effect (FDR < 0.05).

lived species, whereas the short-lived species presented either 1) a different fixed AA (Scenario 1) or 2) variable AAs that were different from the one in the long-lived group (Scenario 2). Positions where the short-lived group showed a fixed AA, and where segregating nonintersecting variation was observed in the long-lived group were also considered (Scenario 3). The species included in the Discovery phase were three Chiroptera (*Myotis lucifugus*, *Myotis davidii*, and *Eptesicus fuscus*), one Rodentia (*Heterocephalus glaber*), and two Primates (*Homo sapiens* and *Nomascus leucogenys*) in the long-lived group, and two Soricomorpha (*Condylura cristata* and *Sorex araneus*), two Rodentia (*Rattus norvegicus* and *Mesocricetus auratus*), one Didelphimorphia (*Monodelphis domestica*), and one Artiodactyla (*Pantholops hodgsonii*) in the short-lived group (see [fig. 1](#) and [supplementary table 1](#), [Supplementary Material](#) online). Since, to ensure fixation and differentiation, gaps were not accepted in any of the sequences, the number of genes that were screened for the CAAS discovery was reduced to 13,035. We scanned all the aligned positions finding a total of 2,737 CAAS in 2,004 genes: 284 belonged to Scenario 1, and 2,453 to Scenario 2. We also identified 533 CAAS belonging to Scenario 3 ([table 1](#)). It is worth noting that CAAS discovered in Scenario 2 were 4.6 times more frequent than CAAS in Scenario 3. These results might be biased by differences in the rates of protein evolution in the sets of species used for Discovery. For instance, the fact that we detect more AA changes corresponding to Scenario 2 (fixed in long-lived) than in Scenario 3 (fixed in short-lived) could be due to short-lived species presenting more AA diversity due to their higher rates of protein evolution. To control for this potential bias, we computed dN/dS ratios in long- and short-lived species and estimated their proportions. We observed higher evolutionary rates in the short-lived species, with a proportion of 1.19:1 relative to the long-lived species. If AA changes occur at random, these

expectations are significantly different from our observations (binomial test, $P = 1.65e-117$), which constitutes evidence for an evolutionary trend to increased lifespan in the mammalian lineage.

We performed two different resampling tests to evaluate if the number of detected CAAS was higher than expected by chance: either randomizing species independently of their phylogenetic relation or randomizing only within specific mammalian orders (“random” and “guided” resampling, see [Materials and Methods](#)). The probability of randomly obtaining a number of CAAS equal or higher than the observed one was 0.003 in both resampling tests ([supplementary fig. 2](#), [Supplementary Material](#) online). In addition, 2,087 CAAS (76.5%) sites were never recovered again in the random resampling and only 19 CAAS (0.7%) were found three times out of 1,000 permutations, being the maximum number of occurrences obtained. All these showing that our set of genes contains a statistical excess of CAAS and that, thus, it is enriched with AA substitutions and genes linked to mammalian longevity after correcting by body mass. In our data set, the *Didelphimorphia* and *Soricomorpha* orders only contained two and one species, respectively, and so the same species (i.e., *Monodelphis domestica*, *Condylura cristata*, and

Table 1. Lists of Discovered and Validated CAAS and Genes.

	Discovered		Validated RRPP	
	CAAS	Genes	CAAS	Genes
Scenario 1	284	273	131 (46.1%)	128
Scenario 2	2,453	1,822	1,025 (41.8%)	891
Scenario 3	533	495	185 (33.5%)	182
Scenarios 1 + 2	2,737	2,004	1,157 (42.3%)	996

NOTE.—Numbers in parentheses represent the percentage of phylogenetically validated positions.

Sorex araneus) were always included in the “guided” resampling, which resulted in a conservative test.

In the Validation phase, we confirmed the discovered CAAS using only the species in the intermediate deciles of the LQ distribution (that is, in the Validation phase, we do not use species in the extreme deciles). First, we intermediate species into two groups, those with the same AA as the long-lived species and those with the same AA combination as the short-lived species (fig. 1). Second, we tested whether the group presenting the long-lived AA showed significantly higher LQ using the phylogenetic ANOVA implemented in RRPP. Out of the 2,737 CAAS from Scenarios 1 and 2, we validated a total 1,157 that belong to 996 genes (table 1 and supplementary fig. 3, Supplementary Material online). This resulted in a 42.1% validation out of the discovered CAAS (for a discussion on discovery and validation at other thresholds see the supplementary note, Supplementary Material online). In addition, out of the 533 CAAS in Scenario 3, we validated 185 (34.7%) that belong to 182 genes. The list of all discovered and validated genes can be found in supplementary table 2, Supplementary Material online. We should point out that for some cases the validation test was underpowered, as for some comparisons there were too few species in one of the two groups, which makes it even more striking that such high percentages of AAs are validated.

Ancestral state reconstructions of the CAAS from Scenario 1 showed that, out of the 88 instances that were predicted with >80% probability (supplementary table 3, Supplementary Material online), only in four cases the long-lived AA was the ancestral state, whereas in the remaining 84 cases, the ancestral AA was the short-lived one (e.g., see supplementary fig. 4, Supplementary Material online). To estimate the number of parallel AA changes in each gene from Scenario 1, we simulated 100 stochastic character maps for each AA substitution in the tree (a total of 28,400 simulations). The results showed that the short-lived AA was at the root of the tree, and that the vast majority of CAAS appeared in parallel towards long-lived (supplementary fig. 5, Supplementary Material online).

We found that amongst the 2,004 discovered genes there was an enrichment of genes upregulated with age (false discovery rate [FDR] = 9.43e-04, enrichment ratio [ER] = 1.43) and a depletion of age-downregulated genes (depletion FDR = 9.99e-09, ER = 0.51), loss of proteostasis (FDR = 1.05e-07, ER = 0.26), essential genes (FDR = 2.76e-06, ER = 0.72), and genes with pLI > 0.9 (FDR = 5.77e-19, ER = 0.63). In contrast, there was no enrichment of genes previously associated with longevity from the GenAge database (supplementary table 4, Supplementary Material online). The significant enrichments were conserved in the subset of 996 genes phylogenetically validated (supplementary note, Supplementary Material online). Additionally, we studied functional enrichments in both the discovered and the validated gene sets with WebGestalt (for a complete list of processes enrichments see supplementary table 5, Supplementary Material online). The discovered genes were enriched in GO categories such as acute inflammatory response (FDR = 1.99e-03), leukocyte migration (FDR = 2.75e-

02), and cytokine binding (FDR = 2.96e-05), in pathways such as *Staphylococcus aureus* infection (FDR = 6.34e-04), and complement and coagulation cascades (FDR = 2.72e-03), and human diseases such as Gram-negative bacterial infections (FDR = 2.09e-07), autoimmune diseases (FDR = 7.14e-06), systemic inflammatory response syndrome (FDR = 1.17e-04), and Werner Syndrome (FDR = 1.85e-03), among many others. In addition, we found enrichment in hallmark gene sets (Liberzon et al. 2015) such as IL6 STAT3 signaling during acute phase response (FDR = 3.70e-05) and blood coagulation cascade (FDR = 9.50e-05).

Among the 2,004 genes harboring CAAS we found eight genes that have been previously linked with longevity. One example is the *WRN* gene, which plays a critical role in repairing damaged DNA, showing two mutations that differ between long- and short-lived mammals. One was from Scenario 1, with two AA clearly differentiating long- and short-lived mammals (F1018L) and another from Scenario 2 (N1055S/R/K/I/T). Both mutations were in the RQC domain of the protein, which makes these two mutations good candidates for follow-up studies and experimental validation. Another example is *CASP10*, a gene involved in the activation cascade of caspases responsible for apoptosis execution, showed six CAAS, all of them located in the caspase domain and validated with the phylogenetic test (supplementary fig. 6, Supplementary Material online). A final example, *ZC3HC1*, which has been recently identified in the GWAS of parental lifespan (Timmers et al. 2019), contained a validated Scenario 2 substitution (T366S/A).

Human Variation in CAAS

Translating CAAS nucleotide changes from short-lived mammals to humans using TransVar, we found 2,704 out of the 2,737 CAAS mapped to a single nucleotide substitution. We excluded the remaining 33 CAAS because they needed more than one nucleotide substitution. We identified human genetic variation in only 516 out of the 2,704 CAAS (19%), but only in six cases (0.22%) the minor allele frequency (MAF) was higher than 1% (supplementary table 6, Supplementary Material online). This observation is much lower than that expected by randomly selecting 100 subsets of 2,737 AA substitutions among the substituted positions between human and rat, and between human and green monkey (empirical $P < 0.01$ in both). In the randomization we observed a mean percentage of 22.4% and 32.85% positions with human genetic variation in the 100 subsets, respectively, and in all simulations that percentage was higher than the observed 19%. Moreover, the number of nucleotide substitutions with a MAF higher than 1% exhibited a mean of 0.60% and 2.36% and only in one out of the 100 randomizations between human and rat, the mean was lower than the 0.22% of the observation (supplementary fig. 7, Supplementary Material online). We also evaluated conservation at these sites by checking PhastCons and PhyloP scores and found that the CAAS set was less conserved than the random subsets between human and rat ($P < 0.01$), and between human and green monkey ($P < 0.01$). All this suggests that the vast majority of the identified CAAS sites (6 of 2,704; 99.87%) exhibit

no or little variation in human populations and, thus, that they may be evolutionarily conserved in our lineage, perhaps because they correspond to genomic factors contributing to human species-specific lifespan or related traits. Since they do not segregate in the population, such potentially critical changes are invisible to analyses that exploit variation in current human populations (e.g., GWAS).

We obtained SIFT and PolyPhen information for 2,175 out of the 2,704 CAAS that mapped to a single nucleotide substitution. A total of 2,134 and 2,112 of the substitutions were considered benign or tolerated in humans according to PolyPhen and SIFT scores, respectively, with 2,079 being benign according to both metrics. CADD scores evaluated the 516 variants showing human variation as likely benign (summarized in [supplementary table 7, Supplementary Material](#) online). This represented, for all the scores, an enrichment of tolerated substitutions compared with 100 random samplings (empirical $P < 0.01$, [supplementary note, Supplementary Material](#) online).

Protein Models

We compared the FoldX changes of total energy between modeled structures of human and rat sequences in 40 protein pairs with validated CAAS. The difference of energy showed that the genes harboring CAAS code for proteins that are more stable in long-lived mammals (represented by human) than in short-lived organisms (represented by rat). To test whether this trend is general or is a property of the longevity-related proteins detected in this study, we analyzed the energies of all the sequences of *R. norvegicus* with known structure and without validated CAAS that had similar human sequences whose structure was either known or could be modeled. The permutation test proved the over stability of human sequences with a P value of $6.3e-04$ ([supplementary fig. 8A, Supplementary Material](#) online). This trend was further validated in a larger background set of about 500 structural models of human and rat sequences, and the significance was preserved with P value of $4.5e-04$ ([supplementary fig. 8B, Supplementary Material](#) online). Increased stability of the proteins coded by these genes was validated ($P = 0.03$) using the naked mole-rat as long-lived model (see [Supplementary Material](#) online). In short: the accumulation of longevity-related differences in AA residues between short- and long-lived mammals has resulted in increased stability in these proteins in long-lived organisms. The specific role of these AA residues is unclear, as the variability of local energies of FoldX is not remarkable for any specific partial energy (with the only exception of Van Der Waals clashes).

Gene–Phenotype Coevolution

To identify genes with rates of protein evolution associated with changes in LQ across the mammalian phylogeny, we computed the root-to-tip ω for each gene and species and evaluated its association with LQ using phylogenetic generalized least squares (PGLS). Among the 18,266 gene alignments, 297 were removed because for more than half of the species we were not able to compute a root-to-tip dN/dS, finally evaluating 17,969 protein-coding sequences.

After FDR correction, in the PGLS analysis, four genes showed a significant association between gene root-to-tip ω and species LQ ([fig. 2](#)): *SPAG16* ($P = 3.58e-7$, slope = 3.14), *TOR2A* ($P = 2.26e-7$, slope = -2.44), *ADCY7* ($P = 1.63e-06$, slope = -2.79), and *CDK12* ($P = 7.81e-06$, slope = 3.92). Among the four significant genes, two (*SPAG16* and *CDK12*) showed a positive association between rate of protein evolution and LQ, and the other two showed a negative association (*TOR2A* and *ADCY7*). These associations between the root-to-tip ω and the LQ values in the four genes were strong, since even after applying the P value conservative method (see Materials and Methods), they were still the top four genes in the analysis ([supplementary table 8, Supplementary Material](#) online). Moreover, 705 genes showed a nominal significant association between rate of protein evolution and LQ ($P_{\text{cons}} < 0.05$).

Human Life Span GWAS Signal Enrichment

Finally, we evaluated whether the gene sets obtained in our analyses were enriched in current human lifespan heritability as estimated from GWAS data. We used data from the largest study on human parental lifespan GWAS to date, based on data from the UK Biobank ([Timmers et al. 2019](#)). We partitioned heritability on the genic fraction of the SNPs using LD-score regression (LDSC) and observed a 3.4-fold enrichment of explained heritability in the set of genes with CAAS compared with the set of screened genes ($P = 3.46e-04$). The enrichment was a 2.6-fold for the phylogenetically validated set ($P = 0.06$). Although genes with a nominal significant association ($P_{\text{cons}} < 0.05$) between rates of protein evolution and LQ showed a 4.2-fold enrichment in GWAS heritability ($P = 0.04$, [supplementary table 9, Supplementary Material](#) online). [Figure 3](#) shows these enrichments using a stratified Q–Q plot, in which a leftward deflection from the null expectation of the subset of SNPs of interest implies an enrichment in GWAS signal. SNPs in the genes that were screened by the CAAS method did not significantly deviate from the expected P values since it remained close to the line for all the SNPs from the GWAS. On the other hand, the discovered and validated genes, as well as genes that were nominally significant in the PGLS approach deviate from the null expectation and showed an enrichment on GWAS significant P values. This enrichment on heritability from the parental lifespan GWAS was also confirmed using Pathway Scoring Algorithm (PASCAL) ([Lamparter et al. 2016](#)). A chi-squared P value of $3.27e-05$ was obtained for the gene set comprising discovered genes with CAAS ([supplementary table 10, Supplementary Material](#) online). The gene set created with the genes resulting from the phylogenetic validation also showed a significant enrichment (chi-squared $P = 0.039$). Finally, the set of genes that were significant ($P_{\text{cons}} < 0.05$) after a PGLS between LQ and root-to-tip ω 's also showed significant enrichment (chi-squared $P = 8.66e-04$).

Discussion

The largest scale studies trying to unveil the genomic architecture of lifespan variation, including human GWAS, have focused on single species. As a consequence, these studies

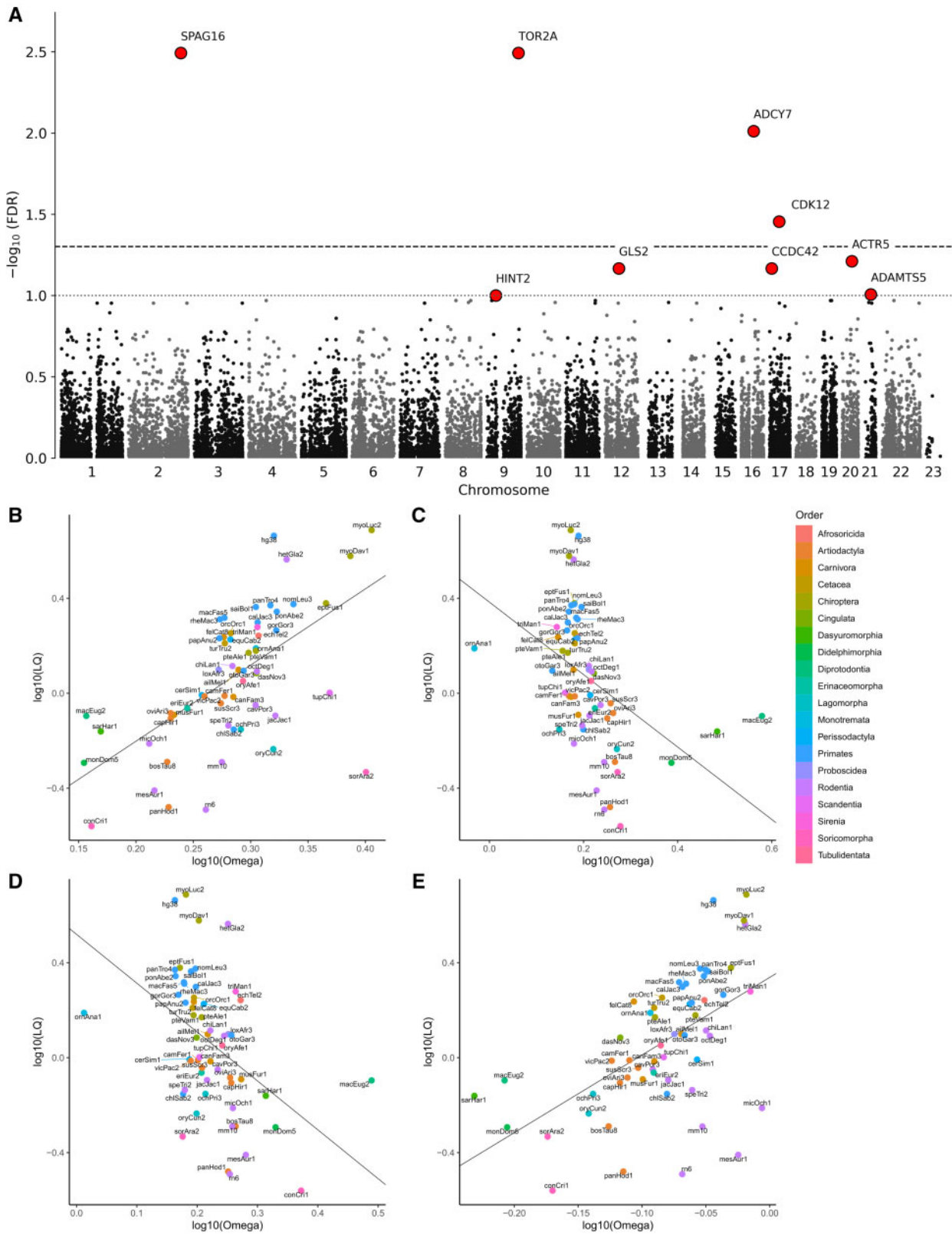


FIG. 2. (A) Manhattan plot of gene-based association results of the phylogenetic-controlled PGLS regressions for LQ. Each dot represents a gene, those depicted in red represent $FDR < 0.1$. The negative logarithm of the FDR P value for each gene tested is reported on the y axis. P value cutoffs corresponding to the Benjamini–Hochberg threshold $FDR = 0.05$ and $FDR = 0.1$, based on the 17,969 genes tested, are denoted by the dashed and dotted lines respectively. Phylogenetically controlled regression (PGLS) between \log_{10} root-to-tip ω for the significant genes (B) SPAG16, (C) TOR2A, (D) ADCY7, and (E) CDK12 are displayed against \log_{10} LQ. Black lines represent the linear regression line. UCSC version names were used for species labeling. Correspondence to the species names can be found in [supplementary table 1, Supplementary Material](#) online.

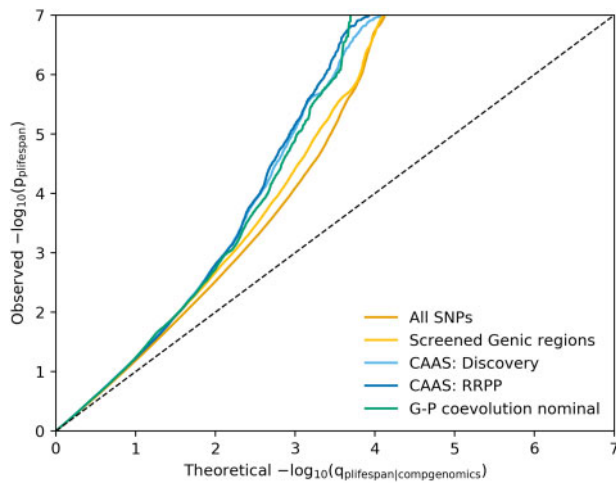


Fig. 3. Stratified Q–Q plot for human longevity shows consistent enrichment across several assessed gene sets. Annotation categories were: 1) all SNPs in the GWAS (orange); 2) SNPs in genic regions of genes screened by the CAAS method (yellow); 3) SNPs in the discovered genes (light blue); 4) SNPs in genes validated with RRPP (dark blue); and 5) SNPs in genes nominally significant ($P_{\text{cons}} < 0.05$) in the PGLS regression (green). All genic regions were defined by gene boundaries plus 5 kb. In summary, the genes we validated in the study were enriched in human longevity signal.

cannot detect variation that, while fundamental to define lifespan-related phenotypes, may have been fixed in the lineage of a species and may contribute crucially to differences in longevity across species. Comparative genome–phenome analysis, therefore, is essential to obtain a complete view of the genetic architecture of lifespan, to unveil important longevity-related genes and genomic features, and to understand the evolution of long-lived species. Here, we leverage longevity variation across mammalian species to explore cross-species variation and identify mutations and genes linked to the evolution of lifespan. The genes detected belong to pathways potentially involved in longevity, have an increased protein stability in long-lived species, and capture a significant part of the variance in the lifespan of current human populations explained by GWAS.

Genetic Architecture of Longevity across Mammals

Our comparative analysis discovered 3,270 CAAS in 2,314 genes using species in the extreme values of LQ distribution. Among them, 2,737 CAAS in 2,004 genes were from cases in which all reference genomes of long-lived species present the same reference AA, whereas short-lived species always present a different AA. This was a statistical excess of discoveries, emphasizing that our gene set is enriched with lifespan-related genetic variation. Out of the 2,737 discoveries, 1,157 CAAS (a 42.3%) in 996 genes were validated using the species in the intermediate deciles with a phylogenetic ANOVA test. The observed 42.3% is much higher than the 5% validation expected by chance, showing again, that our approach can unveil true longevity signals. Furthermore, it should be noted that in the Validation phase, some of the CAAS could not be validated as there was insufficient statistical power due to the small number of species in one of the two groups.

Our results strongly support the use of a comparative genetics approach to inform and complement the interpretation of human lifespan GWAS. First, we observed the enrichment of GWAS signals in stratified QQ-plots of human lifespan. Second, focusing on genes that contain discovered and validated AAs, we evaluated whether the proportion of lifespan heritability that they explain was larger than expected by chance. Third, we analyzed a custom pathway created with the obtained gene sets for enrichment in GWAS by using PASCAL. All resulted in significant enrichments for the lists including genes from the convergent substitutions and the gene–phenotype coevolution analysis. Moreover, comparative genomics is the only way to pinpoint most of the genes we report here, since most of the detected AA changes were fixed or almost fixed in current human populations, with only five out of 2,230 substitutions (0.22%) segregating at $\text{MAF} > 1\%$, despite being more conserved positions across mammals. This is significantly less than what is expected by selecting random SNP across the genome and highlights the fact that variation associated with longevity in mammals is almost all fixed in humans. In sum, we demonstrated that a cross-species comparative genomics approach can complement the analysis of the genetic architecture of complex traits like LQ.

A number of different biologically significant phenomena (e.g., point mutations and posttranslational modifications) can change the folding stability of a protein. Here, we found that the proteins harboring AA changes linked to increased lifespan show increased stability in humans compared with short-lived mammals, represented by rats. The exact cause of increased protein stability cannot be determined from the data collected in this work. Some trends suggest that over stability may be due to the contacts in the hydrophobic core, but results were not significant, with the exception of a reduction in van der Waals clashes. Still, an overall explanation for our findings may be that these proteins have accumulated AA changes resulting in increased resistance to the general proteome destabilization that comes with age. In fact, we also observed a significant depletion of genes linked to loss of proteostasis among the discovered gene set. These observations are consistent with evidence showing that long-lived animals have improved protein stability mechanisms (Pérez et al. 2009; Kim et al. 2011; Treaster et al. 2014). Taken together, this is compelling evidence that a common cross-mammalian mechanism to increase lifespan involves proteome stabilization.

In line with this observation, the identified longevity-related set was depleted in essential genes and genes intolerant of loss-of-function variation ($\text{pLI} \geq 0.9$). Moreover, the deleteriousness scores of the mutations we discovered were enriched in tolerated substitutions compared with random mutations. This could be explained if lifespan evolves through genes and pathways that are not essential to the organisms, which would be fitting with the considerable evolutionary plasticity of longevity in mammals (Ratikainen and Kokko 2019).

It is of note that there were only 533 cases in which the long-lived species showed variable AA and the reference

genomes in the short-lived species presented the same AA (Scenario 3). Assuming randomness, this represents a significant reduction from the 2,453 AAs in Scenario 2 (same AA in all long-lived species, different AAs in the short-lived ones). To the best of our knowledge, this is the first genomic evidence of a trend to increased lifespan in mammals, probably driven by positive selection throughout the mammal clade, which is consistent with the fact that stem mammals were small (O'Leary et al. 2013) and with a recent and detailed study of the distribution of maximum ages in Chordata (Berkel and Cacan 2021). This observation was validated by the fact that only 4.5% of the long-lived variants from Scenario 1 were estimated to be present at the root of the mammalian phylogeny, whereas in the remaining 95.5%, the mammalian ancestor was assigned the short-lived variants. In many cases, for instance, substitutions were present in a long-lived ancestor and are shared by sister species, even if there are no sister species in the top and bottom deciles used in the Discovery phase. However, for most AA changes in Scenario 1 we have been able to validate that long-lived species incorporated the AA mutation multiple times in parallel. These cases are examples of independent mutations that appeared in parallel with lifespan shifts across mammals, as described for other biological adaptations, such as echolocation (Liu et al. 2010) or adaptation to aquatic environments (Foote et al. 2015).

Relevant Genes and Pathways

Discovered genes were enriched in processes involving immune and inflammatory response, cytokine binding, and hemostasis, all of them pathways with well-known relationships with lifespan (Maynard et al. 2015). Coagulation, which has an important role in the maintenance of hemostasis, is known to increase with age and contributes to the higher incidence of cardiovascular diseases in the elderly (Franchini 2006). These pathways overlap with recent findings from Kowalczyk et al. (2020), where they used the number of AA substitutions on a phylogenetic branch to infer shifts associated with lifespan. They found that pathways such as inflammation, DNA repair, cell death, the *IGF1* pathway, and immunity were under increased evolutionary constraint in large and long-lived mammals.

We did not find a specific enrichment in aging-associated lists among the uncovered genes. This might be explained by the fact that our analysis is intended to identify genes and pathways that are shared across mammals, whereas most of the aging-related genes discovered so far are mainly the result of single-species approaches that capture genes driving current variation within species rather than crucial changes fixed along the phylogeny. However, many genes have been previously associated with longevity. A good example is two AA changes in *WRN* gene that were validated in our study at positions 1018 and 1055, both in the RQC domain, which is crucial for DNA binding and for many protein interactions (Tadokoro et al. 2012). *WRN* codifies for the Werner protein, which plays a critical role in maintaining the structure and integrity of the DNA. More than 60 mutations in the *WRN* gene are known to cause Werner's syndrome, which is characterized by dramatically early appearance of features

associated with aging. The two positions we identified (g.31141514T>C and g.31141706A>G) have not been reported before, because they show no variation in humans, suggesting a potential role of these specific positions in the evolution of lifespan in mammals.

Four genes showed a significant relationship between LQ and root-to-tip ω (PGLS at FDR < 0.05): *TOR2A*, *ADCY7*, *CDK12*, and *SPAG16*. Since they have coevolved with longevity patterns across mammals, these are very good candidates for future aging studies. Some have been previously involved in regulating longevity-related pathways. For example, *TOR2A* is a gene involved in cardiovascular diseases (Sun et al. 2021), which is included in a CNV region in chromosome 9 correlating with longevity (Zhao et al. 2018). Interestingly, *TOR2A* gene codifies for the protein Prosalusin, which is cleaved into two gene products (salusin-alpha and salusin-beta), both important vasoactive peptides that play a pivotal role in the maintenance of blood pressure homeostasis regulating cell proliferation and apoptosis (Shichiri et al. 2003; Fujie et al. 2020). *ADCY7* is an adenylate cyclase that catalyzes the formation of cyclic AMP from ATP. Adenylate cyclase genes are key genes in the longevity regulating pathway (Yan et al. 2012) and *ADCY7* was recently associated with cancer mortality in dog breeds (Doherty et al. 2020). *CDK12* is a cyclin-dependent kinase that regulates genome stability via stimulating the transcription of key DNA repair genes (Dubbury et al. 2018). Moreover, the inhibition of *CDK12* emerged as a promising anticancer strategy (Pilarova et al. 2020). In contrast, *SPAG16* has not been previously associated with longevity. *SPAG16* has a main role in sperm motility. However, its expression is not restricted to sperm cells and has been found elevated in various tumors (Siliņa et al. 2011; Knevel et al. 2014). Overall, using both approaches, we provide a list of genes that align with previous observations, but there are also new longevity-associated genes and mutations that will need to be validated experimentally.

Limitations of the Study

We should acknowledge some limitations in our study. First, we analyzed a reduced number of species and genes. Given the number of species and quality of gene alignments, we could get genetic and phenotypic data for 57 mammals. Out of 19,170 genes, we evaluated 13,035 good-quality genes. Very recently, the genomes of 250 mammalian species, including 132 assemblies, have been released (Zoonomia Consortium 2020), that resource can eventually be used for analyses similar to the one presented here. Second, our power to validate CAAS depends on the numbers and phylogenetic distribution of nonextreme species showing either the long- or the short-lived AA residue. In some cases, the long/short-lived AA identified was not present in any other mammal, or only in a few, which made validation impossible. These cases are of course of interest, but beyond the scope of the present work. Third, because the decile strategy was decided a priori, it is likely that we have not explored the full set of AA changes that were involved in changes in mammalian lifespan. However, supplementary figure 3, Supplementary Material online, suggests that using more extreme longevity values

to classify CAAS may provide even more information of the genomic architecture of lifespan. Fourth, in this study we used the short-lived mammal with the largest available data on protein folding (rat and human), to demonstrate the level of stabilization of the proteins coded by aging-related genes detected in this study. However, the comparison was performed upon the reduced set of proteins with available folding data, which may not be representative of the whole proteome. Finally, in the CAAS method we identified point mutations that are ultimately assigned to genes. To study the heritability explained by these genes we assigned to each gene all the SNPs lying in its coding region (plus a 5-kb window). This can introduce a bias because of complicated LD patterns and because many parts of the gene might not be relevant even if a GWAS association is lying in the gene. However, such bias would of course be conservative, even in cases where there is a GWAS signal lying in a gene.

Conclusions

Our findings provide evidence on the genes and cellular mechanisms that may play a role in regulating mammalian lifespan, strongly suggesting that protein stability is linked to increased longevity and supporting an evolutionary trend towards longer lifespan in mammals. Moreover, our study is the first to showcase how comparative-genomic studies can illuminate the genetic architecture of human traits, including clinical and medical phenotypes, and supports the use of comparative genomics studies to understand complex human traits.

Materials and Methods

Genomic and Phenotypic Data

Amino acid and nucleotide alignments for 39,178 orthologous coding sequences were retrieved from the Multiz alignment of 100 vertebrate genomes (Human 100-way) together with the mammalian phylogenetic tree, which was also downloaded from UCSC (<https://genome.ucsc.edu/>, last accessed August, 2019). Amongst the 100 vertebrate species, we kept the 62 species belonging to the class Mammalia ([supplementary fig. 1, Supplementary Material](#) online). For each gene, only the longest transcript was kept and protein alignments with an overall number of gaps >50% or in human alternate contigs were excluded ($n = 905$). After this filtering, a total of 18,266 protein transcripts were included in the analyses.

Variation in MLS across species correlates with many life-history traits, including body mass, growth rate, age at sexual maturity, and body temperature, which can bias comparative studies of lifespan ([Speakman 2005](#)). The most relevant and studied confounding factor is body mass, so longevity is usually corrected by it using the LQ, which indicates whether a species has an average lifespan or is unusually long- or short-lived relative to its body size. LQs is computed as the ratio of a species MLS to the expected MLS given its body mass ([Austad and Fischer 1991](#)). MLS and adult body mass were obtained from the AnAge database ([Tacutu et al. 2018](#), build 14) and missing information was complemented, when available,

using data from the Animal Diversity Web ([Myers et al. 2019](#), last accessed August 2020). The LQ of each species was calculated using the allometric equation for mammals ([de Magalhães et al. 2007](#)). After filtering out species for those we were unable to obtain LQ data, we kept a final number of 57 mammalian species for subsequent analyses ([supplementary fig. 1 and table 1, Supplementary Material](#) online).

Convergent AA Substitutions: Discovery and Validation

Convergent AA substitutions are AA changes that have occurred independently at least twice across the phylogeny. For the purposes of this work, we focused on CAAS that coincide with extreme lifespan values in the set of mammalian species under study. We designed a two-phase procedure to identify such instances of CAAS. First, in the “Discovery phase,” we selected the species in the top and low deciles of the LQ distribution, which we named long- and short-lived, respectively, for a total of 12 extremely lived species (six top and six low). Subsequently, an in-house script was used to detect specific protein positions in which the reference genomes of the long-lived species had the same AA and the short-lived group presented either another AA (Scenario 1) or a set of segregating AAs that were different from the reference AAs in the long-lived group (Scenario 2). Positions where the short-lived group showed a fixed AA, and where segregating, nonintersecting variation was observed in the long-lived group were also considered (Scenario 3). For the purposes of this work, we only focused on Scenarios 1 and 2, representing the AA substitutions converging in the mammal long-lived species (discussed in [supplementary note, Supplementary Material](#) online). We required full information from all species in the extremes, so AA positions for which one or more of the species had a gap were excluded from the analysis. The filter is conservative, since it tends to exclude rapidly evolving genes which may tend to have more phenotypic associations but allows guaranteeing AA sharing and differentiation in all extreme species. The final set is composed by 13,035 genes evaluated using the CAAS procedure ([fig. 1](#)).

To ascertain whether the number of CAAS identified as linked to extremely lived species groups was different than random expectations, we performed two resampling tests. In both tests, two groups of six species were randomly taken from the phylogeny 1,000 times and the procedure to identify CAAS was repeated. The P value was the empirical probability of getting a number of CAAS equal or larger than the original observation. The two resampling procedures differed in their consideration of the phylogeny. The first resampling was independent of the phylogeny; the species were selected completely at random (“random” resampling). The second resampling method was designed to maintain the same proportion of species in each order as those in the observed data, that is: three chiroptera, two primates, and one rodentia species in one group and two rodentia, one didelphimorphia, two soricomorpha, and one artiodactyla in the other group (“guided” resampling). For mammalian orders where there

were no other species to resample, we were conservative and always included the same species.

The second phase, a “Validation phase,” was applied to each AA pin-pointed in the “Discovery phase.” It consisted in validating whether the species in the intermediate deciles (middle 80% of the LQ distribution, a total of 45 species) that had the same AA as the long-lived species also had a higher LQ than those species having the same AA change/s as the short-lived species. When short-lived species displayed more than one AA (Scenario 2), all the short-lived AA combinations were included in the Validation phase. However, AA present among the species of the intermediate deciles but that were not observed in the long- or the short-lived group, were discarded. For Validation, we used a phylogenetic ANOVA test as implemented in the RRPP package in R, using 10,000 iterations for significance testing (Collyer and Adams 2018, fig. 1). Finally, to further validate the longevity signal recovered in the gene set we also performed an external validation with another mammal set of species (supplementary note, Supplementary Material online).

Annotation of CAAS

We analyzed the functional effects of the nucleotide changes leading to CAAS, their population frequency in humans and their association to complex diseases. The most likely nucleotide substitutions corresponding to AA substitutions associated with high LQ in mammals were ascertained using the *panno* option from TransVar (Zhou et al. 2015) and visualized in the protein context (supplementary note, Supplementary Material online). The frequency of each genetic variant in current human populations was obtained from the GnomAD v3 variant database (Karczewski et al. 2020). In those positions showing variable AA in the short-lived species (Scenario 2), we selected the more conservative option to avoid duplicated sites. That is, we assessed all possible combinations and kept the alternative with the highest allele frequency in humans. In addition, those cases in which the most plausible variation leading to the AA mutation implied the change of more than one nucleotide of a codon were excluded from the variation analysis. The same procedure was repeated for 100 random sets of AA substitutions to test whether our observations on genetic variation in the discovered positions fitted the random expectations (supplementary note, Supplementary Material online).

The functional prediction of the genetic variants, as well as the SIFT and PolyPhen2 scores were obtained from the Variant Effect Predictor (McLaren et al. 2016). SIFT and PolyPhen2 predict the functional impact of an AA substitution, the first by leveraging the sequence homology and physical properties of the AA (Ng and Henikoff 2003), and the later by using physical and comparative models based on evolutionary conservation and structure (Adzhubei et al. 2010). CADD scores were obtained from the CADD project website (<https://cadd.gs.washington.edu/>, Rentzsch et al. 2019, last accessed May 2021), and used to assess the deleteriousness of genetic variants, by classifying those with a Phred score higher than 30 as likely deleterious variants.

For the identified positions in Scenario 1, ancestral states were reconstructed to assess the likelihood of the last common ancestor of mammals harboring the putatively long- or short-lived AA. Simulation of the ancestral AA was performed using an empirical Bayes method as implemented in the R package *phytools* (Revell 2012). To avoid cases in which the ancestral AA was uncertain, we only kept those in which the AA in the root of the tree had a probability higher than 0.8. We then quantified the cases in which the ancestral reconstructed AA was the one present in the short- or long-lived mammals. Additionally, for Scenario 1 substitutions, we simulated 100 stochastic character maps using a fixed transition matrix that assumes the same rate of change for any AA transition to estimate the number of AA changes of each type, in order to quantify the number of changes across the phylogeny from any AA to the long-lived or to the short-lived AA (Huelsenbeck et al. 2003).

Protein Models

We aimed to compare protein energy variations between short- and long-lived mammals. Since *R. norvegicus* proteins have been the object of numerous studies, we selected that species as the representative of short-lived group of mammals in our analysis. Humans were selected as long-lived representatives. Not all proteins had known structures for both organisms. Out of 104 protein sequences in which we have validated CAAS and that have known structures, we only found 40 protein sequence pairs from human and rat. This is, sequences with validated CAAS and similar human and rat sequences aligning with a relevant percentage of identical residues, ranging from 50% to a maximum of 99%. Then, we used MODELLER (Webb and Sali 2016) to model both structures of the pair, using the structures of the known templates and the sequence alignments obtained with “matcher” (from EMBOSS package) (Rice et al. 2000). The modeled structures were optimized with the repair-pdb protocol from FoldX (Buß et al. 2018). We used the optimized structures to calculate the differences of Fold X energies (ΔE) between the human and rat protein sequences.

We selected all the sequences of *R. norvegicus* available in Uniprot with known 3D structure ($n = 667$) as a background set to compare the ΔE distributions of proteins coded by genes harboring CAAs with the distributions of ΔE s of pairs of sequences from genes without CAAS. After removing the ones in the set with CAAS described in the previous paragraph, we obtained 337 structural models of both human and rat paired sequences (with alignments ranging from 50% to 99% of identical residues) from which the distributions of ΔE s were calculated. Outliers were removed from both distributions by generating an interquartile range (IQR) with a weight of 4: $IQR = 4 \times (\text{upper quartile} - \text{lower quartile})$. The distribution was normalized by the maximum value to have comparable ranges between 0 and 1. We compared the distribution of the pairs with validated CAAS with the background distribution of nonvalidated CAAS using a permutation two-sample test (<https://statlab.github.io/permute/user/two-sample.html>).

We further tested the robustness of the results by increasing the size of the background. We increased the background with protein pairs of related human and rat sequences without validated CAAS, selected randomly. After parsing around 4,000 sequences, we obtained 500 pairs of rat–human sequences whose structure could be modeled for both species. We used the same protocol for modeling and optimization of the structures, and the distribution of ΔE s was similarly normalized and analyzed, using a permutation two-sample test for the comparison.

Gene–Phenotype Coevolution across Mammals

The nucleotide alignments that underwent previous quality control (<50% of gaps) were used for studying the coevolution of genes and phenotype. We estimated root-to-tip rates of protein evolution (the dN/dS ratio or ω) using the free-ratio model from PAML 4.9a (Yang 1997). The root-to-tip ω is a property of the species tip rather than of the terminal branch, thus being more inclusive of the evolutionary history of a locus and, therefore, it is more suitable for regressions against phenotypic data from extant species (Montgomery and Mundy 2012). Briefly, for each gene and species we computed the root-to-tip dNs and dSs and the ratio between these values to obtain the root-to-tip ω , as previously described in Muntané et al. (2018). To avoid numerical problems with the log transformation and unrealistic substitution rates, the species for which the root-to-tip ω was 0 were discarded, with 877 genes having at least one species removed. Genes for which we could not estimate a root-to-tip value for, at least, half of the species were also removed from further analyses, resulting in a final set of 17,969 genes.

For each gene, we studied the association between its rate of protein evolution and longevity regressing root-to-tip ω and LQ by means of PGLS as implemented in the *caper* library in R (Orme 2018). PGLS allows to incorporate the phylogenetic relationship among species in the error term of a generalized least squares model, thus controlling for the phylogenetic inertia (close species may have more similar phenotypes than distant species). Pagel's lambda (λ) was estimated through maximum likelihood in each case. Pagel's λ values equal or close to 1 indicate that a character is evolving stochastically (Brownian motion) along the tree, whereas $\lambda < 1$ indicates that a character evolution is independent of the phylogeny. In 295 genes, estimations of λ resulted in a value of 0 and the log-likelihood plots showed a flat likelihood surface (e.g., see supplementary fig 9, Supplementary Material online), which is most likely due to reduced sample size (DeCasien et al. 2017). Consequently, for these cases we set the value of λ to the genome-wide median λ value. To control for the effect of effective population size covarying with LQ, median genome-wide root-to-tip ω 's was included in the PGLS models as covariate (Boddy et al. 2017). In addition, species with studentized residuals greater than ± 3 were considered outliers and, thus, removed from the regression and PGLS was fitted again. Moreover, to control that no single species was biasing the PGLS models, we performed an additional step and repeated regressions: by removing one species at a time and keeping the maximum P value (P value

conservative). In all regressions, both the LQ and the root-to-tip ω 's were \log_{10} transformed. Finally, we applied a Benjamini–Hochberg FDR with an FDR of 5% for multiple test corrections. To avoid associations due to one single species, when we refer to genes that are nominally significant in the PGLS analysis, we always refer to those that were nominally significant for the P value conservative regressions ($P_{\text{cons}} < 0.05$).

Functional Enrichments

To study whether there was an over- or underrepresentation of genes previously related to aging in our gene sets we used hypergeometric tests. We included lists of genes that had been previously associated with aging (supplementary note, Supplementary Material online). Biological mechanisms underlying CAAS were evaluated with WebGestalt, that allows checking for pathway overrepresentation of specific GO terms, pathways, and disease-associated genes from GLAD4U (Liao et al. 2019). For each evaluated gene set, FDR was controlled using the Benjamini–Hochberg procedure and the set of evaluated genes ($n = 13,035$ genes) was used as the background. PGLS results were filtered keeping only those genes that were nominally significant ($P_{\text{cons}} < 0.05$) and then ranked by the t-statistic obtained in the PGLS analysis, subsequently gene set enrichment analysis, from WebGestalt, was carried out to study enriched categories in genes with both a positive and a negative association between root-to-tip ω and LQ.

Human Life Span GWAS Heritability Enrichment

To test whether the genes obtained from our comparative analyses could explain genetic variation in human lifespan, LDSC was used (Finucane et al. 2015). Specifically, we tested whether both, the genes with CAAS (discovered and validated) and those nominally significant after the PGLS regression, explained a larger fraction of SNP heritability in a human lifespan GWAS than would be expected by chance. Briefly, SNPs on the GWAS of parental lifespan (Timmers et al. 2019) were assigned to genes by annotating and keeping only SNPs in genic regions plus a window of 5 kb around each gene. Subsequently, the custom annotation file for LDSC was prepared for five categories: 1) all genic SNPs, 2) SNPs that map into the genes that were evaluated, 3) SNPs located in the genes containing discovered CAAS, and 4) SNPs in the genes harboring phylogenetically validated CAAS and 5) SNPs in the genes that were significant ($P_{\text{cons}} < 0.05$) in the PGLS genome–phenotype analysis. Enrichment in human lifespan heritability was evaluated in the detected genes (Categories 3–5) compared heritability explained by the genes evaluated (Category 2). To help in visualizing the results, stratified QQ-plots were also performed with the SNPs in the five categories.

PASCAL is a software that allows testing if a given gene set (or pathway) is enriched in GWAS signal (Lamparter et al. 2016). With this aim, we computed gene scores by aggregating using the sum of chi-squared option (SOCS), SNP P values from the GWAS on parental lifespan, while correcting for linkage disequilibrium data. These computed gene-based scores were then aggregated across sets of related genes

with the pathway analysis tools in PASCAL to obtain a pathway score. Pathway enrichment was evaluated using the chi-squared method. With the aim of testing whether our identified genes were enriched in human lifespan GWAS signal, we built custom pathways using the genes resulting from our analyses (the five categories aforementioned) and computed pathway scores for them. For all of them, we kept only genetic regions including a window of 5 kb around each gene.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank T. Marquès-Bonet and Salvador Macip for their insights discussing the analyses. We also would like to thank the reviewers for their thoughtful observations and efforts toward improving the manuscript. This work was supported by AEI-PGC2018-101927-BI00 (FEDER/UE), the Spanish National Institute of Bioinformatics of the Instituto de Salud Carlos III (PT17/0009/0020), FEDER (Fondo Europeo de Desarrollo Regional)/FSE (Fondo Social Europeo), “Unidad de Excelencia María de Maeztu,” funded by the AEI (CEX2018-000792-M) and Secretaria d’Universitats i Recerca and CERCA Programme del Departament d’Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 880).

Author Contributions

G.M. and A.N. conceived, planned, and supervised the study. X.F. processed the data and performed the analyses. R.M. and B.O. performed the computations for protein modeling. P.R.H.J.T. and P.K.J. helped in supervising the GWAS heritability analyses. S.A., F.B., and B.E. contributed to the interpretation of results. G.M. and A.N. wrote the manuscript with the input from all authors.

Data Availability

All relevant data are within the paper and its [Supplementary Material](#) online. The alignment data that support the findings of this study are openly available from UCSC database <http://hgdownload.cse.ucsc.edu/goldenpath/hg19/multiz100way/alignments/>. Maximum lifespan data for mammal species are available at AnAge database <https://genomics.senescence.info/species/>. GWAS summary statistics data can be obtained at <https://datashare.ed.ac.uk/handle/10283/3209>. The in-house code for discovery and validation analyses associated with the current submission is available on demand.

References

- Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods*. 7(4):248–249.
- Austad SN. 2005. Diverse aging rates in metazoans: targets for functional genomics. *Mech Ageing Dev, Funct Genomics Ageing II*. 126(1):43–49.
- Austad SN, Fischer KE. 1991. Mammalian aging, metabolism, and ecology: evidence from the bats and marsupials. *J Gerontol*. 46(2):B47–B53.
- Baker J, Meade A, Pagel M, Venditti C. 2015. Adaptive evolution toward larger size in mammals. *Proc Natl Acad Sci U S A*. 112(16):5093–5098.
- Berkel C, Cacan E. 2021. Analysis of longevity in chordata identifies species with exceptional longevity among taxa and points to the evolution of longer lifespans. *Biogerontology* 22(3):329–343.
- Boddy AM, Harrison PW, Montgomery SH, Caravas JA, Raghanti MA, Phillips KA, Mundy NI, Wildman DE. 2017. Evidence of a conserved molecular response to selection for increased brain size in primates. *Genome Biol Evol*. 9(3):700–713.
- Buffenstein R. 2005. The naked mole-rat: a new long-living model for human aging research. *J Gerontol A Biol Sci Med Sci*. 60(11):1369–1377.
- Bullock AN, Henckel J, DeDecker BS, Johnson CM, Nikolova PV, Proctor MR, Lane DP, Fersht AR. 1997. Thermodynamic stability of wild-type and mutant P53 core domain. *Proc Natl Acad Sci U S A*. 94(26):14338–14342.
- Buř O, Rudat J, Ochsenreither K. 2018. FoldX as protein engineering tool: better than random based approaches? *Comput Struct Biotechnol J*. 16:25–33.
- Chikina M, Robinson JD, Clark NL. 2016. Hundreds of genes experienced convergent shifts in selective pressure in marine mammals. *Mol Biol Evol*. 33(9):2182–2192.
- Collyer ML, Adams DC. 2018. RRPP: An R package for fitting linear models to high-dimensional data using residual randomization. *Methods Ecol Evol*. 9(7):1772–1779.
- DeCasien AR, Williams SA, Higham JP. 2017. Primate brain size is predicted by diet but not sociality. *Nat Ecol Evol*. 1(5):0112.
- Deelen J, Evans DS, Arking DE, Tesi N, Nygaard M, Liu X, Wojczynski MK, Biggs ML, van der Spek A, Atzmon G, et al. 2019. A meta-analysis of genome-wide association studies identifies multiple longevity genes. *Nat Commun*. 10(1):3669–3674.
- de Magalhães JP, Costa J, Church GM. 2007. An analysis of the relationship between metabolism, developmental schedules, and longevity using phylogenetic independent contrasts. *J Gerontol A Biol Sci Med Sci*. 62(2):149–160.
- de Magalhães JP, Toussaint O. 2002. The evolution of mammalian aging. *Exp Gerontol*. 37(6):769–775.
- de Magalhães JP, Wang J. 2019. The fog of genetics: what is known, unknown and unknowable in the genetics of complex traits and diseases. *EMBO Rep*. 20(11):e48054.
- Doherty A, Lopes I, Ford CT, Monaco G, Guest P, de Magalhães JP. 2020. A scan for genes associated with cancer mortality and longevity in pedigree dog breeds. *Mamm Genome*. 31(7–8):215–227.
- Dubbury SJ, Boutz PL, Sharp PA. 2018. CDK12 regulates DNA repair genes by suppressing intronic polyadenylation. *Nature* 564(7734):141–145.
- Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, Anttila V, Xu H, Zang C, Farh K, et al. 2015. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet*. 47(11):1228–1235.
- Fontana L, Partridge L. 2015. Promoting health and longevity through diet: from model organisms to humans. *Cell* 161(1):106–118.
- Foot AD, Liu Y, Thomas GWC, Vinar T, Alföldi J, Deng J, Dugan S, van Elk CE, Hunter ME, Joshi V, et al. 2015. Convergent evolution of the genomes of marine mammals. *Nat Genet*. 47(3):272–275.
- Franchini M. 2006. Hemostasis and aging. *Crit Rev Oncol Hematol*. 60(2):144–151.
- Fujie S, Hasegawa N, Sanada K, Hamaoka T, Maeda S, Padilla J, Martinez-Lemus LA, Iemitsu M. 2020. Increased serum salusin- α by aerobic exercise training correlates with improvements in arterial stiffness in middle-aged and older adults. *Ageing* 12(2):1201–1212.
- Hekimi S, Guarente L. 2003. Genetics and the specificity of the aging process. *Science* 299(5611):1351–1354.
- Herskind AM, McGue M, Holm NV, Sørensen TI, Harvald B, Vaupel JW. 1996. The heritability of human longevity: a population-based study of 2872 Danish twin pairs born 1870–1900. *Hum Genet*. 97(3):319–323.
- Huang Z, Whelan CV, Foley NM, Jebb D, Touzalin F, Petit EJ, Puechmaile SJ, Teeling EC. 2019. Longitudinal comparative transcriptomics

- reveals unique mechanisms underlying extended healthspan in bats. *Nat Ecol Evol.* 3(7):1110–1120.
- Huelsenbeck JP, Nielsen R, Bollback JP. 2003. Stochastic mapping of morphological characters. *Syst Biol.* 52(2):131–158.
- Kaplanis J, Gordon A, Shor T, Weissbrod O, Geiger D, Wahl M, Gershovits M, Markus B, Sheikh M, Gymrek M, et al. 2018. Quantitative analysis of population-scale family trees with millions of relatives. *Science* 360(6385):171–175.
- Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581(7809):434–443.
- Keane M, Semeiks J, Webb AE, Li YI, Quesada V, Craig T, Madsen LB, van Dam S, Brawand D, Marques PI, et al. 2015. Insights into the evolution of longevity from the bowhead whale genome. *Cell Rep.* 10(1):112–122.
- Kim EB, Fang X, Fushan AA, Huang Z, Lobanov AV, Han L, Marino SM, Sun X, Turanov AA, Yang P, et al. 2011. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature* 479(7372):223–227.
- Knevel R, Klein K, Somers K, Ospelt C, Houwing-Duistermaat JJ, van Nies JAB, de Rooy DPC, de Bock L, Kurreeman FAS, Schonkeren J, et al. 2014. Identification of a genetic variant for joint damage progression in autoantibody-positive rheumatoid arthritis. *Ann Rheum Dis.* 73(11):2038–2046.
- Kowalczyk A, Partha R, Clark NL, Chikina M. 2020. Pan-mammalian analysis of molecular constraints underlying extended lifespan. *ELife* 9(February):e51089.
- Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S. 2016. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS Comput Biol.* 12(1):e1004714.
- Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. 2019. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* 47(W1):W199–W205.
- Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov J, Tamayo P. 2015. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* 1(6):417–425.
- Liu Y, Cotton JA, Shen B, Han X, Rossiter SJ, Zhang S. 2010. Convergent sequence evolution between echolocating bats and dolphins. *Curr Biol.* 20(2):R53–R54.
- López-Otín C, Blasco MA, Partridge L, Serrano M, Guido K. 2013. The hallmarks of aging. *Adv Exp Med Biol.* 1002(6):153–1217.
- Lyson TR, Miller IM, Bercovici AD, Weissenburger K, Fuentes AJ, Clyde WC, Hagadorn JW, Butrim MJ, Johnson KR, Fleming RF, et al. 2019. Exceptional continental record of biotic recovery after the cretaceous–paleogene mass extinction. *Science* 366(6468):977–983.
- Ma S, Gladyshev VN. 2017. Molecular signatures of longevity: insights from cross-species comparative studies. *Semin Cell Dev Biol.* 70(October):190–203.
- Maynard S, Fei Fang E, Scheibye-Knudsen M, Croteau DL, Bohr VA. 2015. DNA damage, DNA repair, aging, and neurodegeneration. *Cold Spring Harb Perspect Med.* 5(10):1–18.
- McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl variant effect predictor. *Genome Biol.* 17(1):122.
- Montgomery SH, Mundy NI. 2012. Evolution of *Aspm* is associated with both increases and decreases in brain size in primates. *Evolution* 66(3):927–932.
- Muntané G, Farré X, Rodríguez JA, Pegueroles C, Hughes DA, de Magalhães JP, Gabaldón T, Navarro A. 2018. Biological processes modulating longevity across primates: a phylogenetic genome-phenome analysis. *Mol Biol Evol.* 35(8):1990–2004.
- Myers P, Espinosa R, Parr CS, Jones T, Hammond GS, A Dewey T. 2019. The animal diversity web (online) . Available from: <https://animal-diversity.org>.
- Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31(13):3812–3814.
- O’Leary MA, Bloch JJ, Flynn JJ, Gaudin TJ, Giallombardo A, Giannini NP, Goldberg SL, Kraatz BP, Luo Z-X, Meng J, et al. 2013. The placental mammal ancestor and the post-K-Pg radiation of placentals. *Science* 339(6120):662–667.
- Orme D. 2018. The Caper package: comparative analysis of phylogenetics and evolution in R. Available from: <https://cran.r-project.org/web/packages/caper/vignettes/caper.pdf>.
- Pérez VI, Buffenstein R, Masamsetti V, Leonard S, Salmon AB, Mele J, Andziak B, Yang T, Edrey Y, Friguet B, et al. 2009. Protein stability and resistance to oxidative stress are determinants of longevity in the longest-living rodent, the naked mole-rat. *Proc Natl Acad Sci U S A.* 106(9):3059–3064.
- Pickrell J. 2019. How the earliest mammals thrived alongside dinosaurs. *Nature* 574(7779):468–472.
- Pilarova K, Herudek J, Blazek D. 2020. CDK12: cellular functions and therapeutic potential of versatile player in cancer. *NAR Cancer.* 2(1):zcaa003.
- Pomeroy D. 1990. Why fly? The possible benefits for lower mortality. *Biol J Linn Soc.* 40(1):53–65.
- Powers ET, Morimoto RI, Dillin A, Kelly JW, Balch WE. 2009. Biological and chemical approaches to diseases of proteostasis deficiency. *Annu Rev Biochem.* 78:959–991.
- Ratikainen II, Kokko H. 2019. The coevolution of lifespan and reversible plasticity. *Nat Commun.* 10(1):538.
- Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. 2019. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 47(D1):D886–D894.
- Revell LJ. 2012. Phytools: an R package for phylogenetic comparative biology (and other things): *phytools: R Package. Methods Ecol Evol.* 3(2):217–223.
- Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* 16(6):276–277.
- Ruby JG, Smith M, Buffenstein R. 2018. Naked mole-rat mortality rates defy gompertzian laws by not increasing with age. *ELife* 7:e31157.
- Ruby JG, Wright KM, Rand KA, Kermany A, Noto K, Curtis D, Varner N, Garrigan D, Slinkov D, Dorfman I, et al. 2018. Estimates of the heritability of human longevity are substantially inflated due to assortative mating. *Genetics* 210(3):1109–1124.
- Santra M, Dill KA, de Graff AMR. 2019. Proteostasis collapse is a driver of cell aging and death. *Proc Natl Acad Sci U S A.* 116(44):22173–22178.
- Sebastiani P, Perls TT. 2012. The genetics of extreme longevity: lessons from the New England Centenarian study. *Front Genet.* 3:277.
- Seim I, Fang X, Xiong Z, Lobanov AV, Huang Z, Ma S, Feng Y, Turanov AA, Zhu Y, Lenz TL, et al. 2013. Genome analysis reveals insights into physiology and longevity of the Brandt’s bat *Myotis brandtii*. *Nat Commun.* 4(1):2212.
- Shattuck MR, Williams SA. 2010. Arboreality has allowed for the evolution of increased longevity in mammals. *Proc Natl Acad Sci U S A.* 107(10):4635–4639.
- Shichiri M, Ishimaru S, Ota T, Nishikawa T, Isogai T, Hirata Y. 2003. Salusins: newly identified bioactive peptides with hemodynamic and mitogenic activities. *Nat Med.* 9(9):1166–1172.
- Siliņa K, Zayakin P, Kalniņa Z, Ivanova L, Meistere I, Endzeliņš E, Abols A, Stengrēvics A, Leja M, Ducena K, et al. 2011. Sperm-associated antigens as targets for cancer immunotherapy: expression pattern and humoral immune response in cancer patients. *J Immunother.* 34(1):28–44.
- Speakman JR. 2005. Correlations between physiology and lifespan—two widely ignored problems with comparative studies. *Aging Cell.* 4(4):167–175.
- Sun S, Zhang F, Pan Y, Xu Y, Chen A, Wang J, Tang H, Han Y. 2021. A TOR2A gene product: salusin- β contributes to attenuated vasodilatation of spontaneously hypertensive rats. *Cardiovasc Drugs Ther.* 35(1):125–139.
- 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(D1):D1083–D1090.

- Tadokoro T, Kulikowicz T, Dawut L, Croteau DL, Bohr VA. 2012. DNA binding residues in the RQC domain of Werner protein are critical for its catalytic activities. *Aging* 4(6):417–429.
- Tian X, Seluanov A, Gorbunova V. 2017. Molecular mechanisms determining lifespan in short- and long-lived species. *Trends Endocrinol Metab.* 28(10):722–734.
- Timmers PR, Mounier N, Lall K, Fischer K, Ning Z, Feng X, Bretherick AD, Clark DW, Shen X, Esko T, et al. 2019. Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *Elife* 8:e39856.
- Treaster SB, Ridgway ID, Richardson CA, Gaspar MB, Chaudhuri AR, Austad SN. 2014. Superior proteome stability in the longest lived animal. *Age* 36(3):9597.
- Wang H, Zhao H, Sun K, Huang X, Jin L, Feng J. 2020. Evolutionary basis of high-frequency hearing in the cochleae of echolocators revealed by comparative genomics. *Genome Biol Evol.* 12(1):3740–3753.
- Webb B, Sali A. 2016. Comparative protein structure modeling using MODELLER. *Curr Protoc Bioinformatics.* 54:5.6.1–5.6.37.
- Yan L, Park JY, Dillinger J-G, De Lorenzo MS, Yuan C, Lai L, Wang C, Ho D, Tian B, Stanley WC, et al. 2012. Common mechanisms for calorie restriction and adenylyl cyclase type 5 knockout models of longevity. *Aging Cell.* 11(6):1110–1120.
- Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci.* 13(5):555–556.
- Zenin A, Tsepilov Y, Sharapov S, Getmantsev E, Menshikov LI, Fedichev PO, Aulchenko Y. 2019. Identification of 12 genetic loci associated with human healthspan. *Commun Biol.* 2:41.
- Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, et al. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* 346(6215):1311–1320.
- Zhao X, Liu X, Zhang A, Chen H, Huo Q, Li W, Ye R, Chen Z, Liang L, Liu QA, et al. 2018. The correlation of copy number variations with longevity in a genome-wide association study of Han Chinese. *Aging* 10(6):1206–1222.
- Zhou W, Chen T, Chong Z, Rohrdanz MA, Melott JM, Wakefield C, Zeng J, Weinstein JN, Meric-Bernstam F, Mills GB, et al. 2015. TransVar: a multilevel variant annotator for precision genomics. *Nat Methods.* 12(11):1002–1003.
- Zoonomia Consortium. 2020. A comparative genomics multitool for scientific discovery and conservation. *Nature* 587(7833):240–245.