The whale shark genome reveals how genomic and physiological properties scale with body size

Jessica A. Weber[a,1], Seung Gu Park[b,c,1], Victor Luria[d,1], Sungwon Jeon[b,c], Hak-Min Kim[b,c], Yeonsu Jeon[b,c], Youngjune Bhak[b,c], Je Hun Jun[e], Sang Wha Kim[f,g], Won Hee Hong[h], Semin Lee[b,c], Yun Sung Cho[e], Amir Karger[i], John W. Cain[j], Andrea Manica[k], Soonok Kim[l], Jae-Hoon Kim[m], Jeremy S. Edwards[n,2,3], Jong Bhak[b,c,e,2,3], and George M. Church[a,2,3]

[a]Department of Genetics, Harvard Medical School, Boston, MA 02115; [b]Korean Genomics Center, Ulsan National Institute of Science and Technology, 44919 Ulsan, Republic of Korea; [c]Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology, 44919 Ulsan, Republic of Korea; [d]Department of Systems Biology, Harvard Medical School, Boston, MA 02115; [e]Clinomics Inc., 44919 Ulsan, Republic of Korea; [f]Laboratory of Aquatic Biomedicine, College of Veterinary Medicine, Seoul National University, 08826 Seoul, Republic of Korea; [g]Research Institute for Veterinary Science, College of Veterinary Medicine, Seoul National University, 08826 Seoul, Republic of Korea; [h]Hanwha Marine Biology Research Center, 63642 Jeju, Republic of Korea; [i]IT–Research Computing, Harvard Medical School, Boston, MA 02115; [j]Department of Mathematics, Harvard University, Cambridge, MA 02138; [k]Department of Zoology, University of Cambridge, CB2 3EJ Cambridge, United Kingdom; [l]National Institute of Biological Resources, 37242 Incheon, Republic of Korea; [m]College of Veterinary Medicine and Veterinary Medical Research Institute, Jeju National University, 63243 Jeju, Republic of Korea; and [n]Department of Chemistry and Chemical Biology, UNM Comprehensive Cancer Center, University of New Mexico, Albuquerque, NM 87131

Contributed by George M. Church, June 8, 2020 (sent for review December 24, 2019; reviewed by Manuel Corpas and Xiaohua Huang)

The endangered whale shark (*Rhincodon typus*) is the largest fish on Earth and a long-lived member of the ancient Elasmobranchii clade. To characterize the relationship between genome features and biological traits, we sequenced and assembled the genome of the whale shark and compared its genomic and physiological features to those of 83 animals and yeast. We examined the scaling relationships between body size, temperature, metabolic rates, and genomic features and found both general correlations across the animal kingdom and features specific to the whale shark genome. Among animals, increased lifespan is positively correlated to body size and metabolic rate. Several genomic traits also significantly correlated with body size, including intron and gene length. Our large-scale comparative genomic analysis uncovered general features of metazoan genome architecture: Guanine and cytosine (GC) content and codon adaptation index are negatively correlated, and neural connectivity genes are longer than average genes in most genomes. Focusing on the whale shark genome, we identified multiple features that significantly correlate with lifespan. Among these were very long gene length, due to introns being highly enriched in repetitive elements such as CR1-like long interspersed nuclear elements, and considerably longer neural genes of several types, including connectivity, activity, and neurodegeneration genes. The whale shark genome also has the second slowest evolutionary rate observed in vertebrates to date. Our comparative genomics approach uncovered multiple genetic features associated with body size, metabolic rate, and lifespan and showed that the whale shark is a promising model for studies of neural architecture and lifespan.

whale shark | lifespan | body size | metabolic rate | neural genes

The relationships between body mass, longevity, and basal metabolic rate (BMR) across diverse habitats and taxa have been researched extensively over the last century and have led to generalized rules and scaling relationships that explain many physiological and genetic trends observed across the tree of life. While the largest extant animals on the planet are aquatic, the impact of marine habitats on body size and other physiological and genetic characteristics is only beginning to be discovered (1). In an effort to better understand the selective pressures imposed on body size in marine environments, studies of endothermic aquatic mammals have shown that selection for larger body sizes has been driven by the minimization of heat loss (2). In ectothermic vertebrates, however, the relationship between environmental temperature and body size is more complex. In these species, metabolic rate is directly dependent on temperature, and decreased temperatures are correlated with decreased BMRs, decreased growth rates, longer generational times, and increased body sizes (3, 4).

The whale shark (*Rhincodon typus*) is the largest extant fish, reaches lengths of 20 m (5) and 42 tons in mass (6) and has a maximum lifespan estimated at 80 y (6). Worldwide populations have been declining, and the whale shark has been classified as an endangered species by the International Union for Conservation of Nature. Whale sharks are one of three species of filter-feeding sharks that use modified gill rakers to sieve plankton and small nektonic prey from the water column in a method

Significance

We sequenced and analyzed the genome of the endangered whale shark, the largest fish on Earth, and compared it to the genomes of 84 other species ranging from yeast to humans. We found strong scaling relationships between genomic and physiological features. We posit that these scaling relationships, some of which were remarkably general, mold the genome to integrate metabolic constraints pertaining to body size and ecological variables such as temperature and depth. Unexpectedly, we also found that the size of neural genes is strongly correlated with lifespan in most animals. In the whale shark, large gene size and large neural gene size strongly correlate with lifespan and body mass, suggesting longer gene lengths are linked to longer lifespans.

Author contributions: J.A.W., V.L., Y.S.C., J.B., and G.M.C. designed research; H.-M.K., S.W.K., W.H.H., Y.S.C., S.K., and J.-H.K. performed research; J.A.W., S.G.P., V.L., S.J., H.-M.K., Y.J., Y.B., J.H.J., S.L., A.K., and J.W.C. analyzed data; and J.A.W., S.G.P., V.L., S.J., A.M., J.S.E., J.B., and G.M.C. wrote the paper.

Reviewers: M.C., Cambridge Precision Medicine Limited; and X.H., University of California San Diego.

The authors declare no competing interest.

Data deposition: The whale shark whole-genome project data have been deposited at INSDC: International Nucleotide Sequence Database Collaboration (accession no. QPMN00000000). The version described in this paper is version QPMN01000000. DNA sequencing reads have been uploaded to the National Center for Biotechnology Information Sequence Read Archive (SRP155581). The C++ code used for the Markov Cluster (MCL) algorithm was uploaded to the GitHub repository (https://github.com/jsungwon/MCL-clustering).

[1]J.A.W., S.G.P., and V.L. contributed equally to this work.

[2]J.S.E., J.B., and G.M.C. contributed equally to this work.

[3]To whom correspondence may be addressed. Email: jsedwards@salud.unm.edu, jongbhak@genomics.org, or gchurch@genetics.med.harvard.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1922576117/-/DCSupplemental.

First published August 4, 2020.

convergent with that of the baleen whales (1). Unlike the two smaller filter-feeding shark species (*Cetorhinus maximus*, *Megachasma pelagios*) that inhabit colder temperate waters with increased prey availability, whale sharks have a cosmopolitan tropical and warm subtropical distribution and have rarely been sighted in areas with surface temperatures less than 21 °C (7–9). However, recent global positioning system (GPS) tagging studies have revealed that they routinely dive to mesopelagic (200 to 1,000 m) and bathypelagic (1,000 to 4,000 m) zones to feed, facing water temperatures less than 4 °C (10). Observations of increased surface occupation following deeper dives have led to the suggestion that thermoregulation is a primary driver for their occupation of the warmer surface waters (7, 11). Since larger body masses retain heat for longer periods of time, the large body mass of whale sharks may slow their cooling upon diving and maximize their dive times to cold depths, where food is abundant, and could thus play a key role in metabolic regulation.

Body size, environmental temperature, metabolic rate, and generation time are all correlated with variations in evolutionary rates (12, 13). Since many of these factors are interconnected, modeling studies have shown that observed evolutionary rate heterogeneity can be predicted by accounting for the impact of body size and temperature on metabolic rate (14), suggesting that these factors together drive the rate of evolution through their effects on metabolism. Consistent with these results, brownbanded bamboo shark, cloudy catshark, and elephant fish have the slowest evolutionary rates reported to date (15, 16). Moreover, genome size and intron size have also been linked to metabolic rate in multiple clades. Intron length varies between species and plays an important role in gene regulation and splice-site recognition. In an analysis of amniote genomes, intron size was reduced in species with metabolically demanding powered flight and correlated with overall reductions in genome size (17, 18). However, since most previous studies were limited by poor taxonomic sampling and absence of genome data for the deepest branches of the vertebrate tree, comprehensive comparative genomic analyses across gnathostomes are necessary to gain a deeper understanding of the evolutionary significance of the correlations between genome size, intron size, and metabolic demands.

Here we sequenced, assembled, and analyzed the genome of the whale shark and compared its genome and biological traits to those of 84 eukaryotic species with a focus on gnathostomes such as fishes, birds, and mammals. In particular, we identified scaling relationships between body size, temperature, metabolic rates, and genomic features and found general genetic and physiological correlations that span the animal kingdom. We also examined characteristics unique to the whale shark and its slow-evolving, large genome.

Results

The Whale Shark Genome. The DNA of an *R. typus* individual was sequenced to a depth of 164× using a combination of Illumina short-insert, mate-pair, and TruSeq Synthetic Long Read (TSLR) libraries (*SI Appendix*, Tables S1, S2, S12, and S13), resulting in a 3.2-Gb genome (*SI Appendix*, Fig. S1 and Table S4) with a scaffold N50 of 2.56 Mb (*SI Appendix*, Tables S2, S5, and S6). A sliding-window approach was used to calculate guanine and cytosine (GC) content and resulted in a genome-wide average of 42%, which is similar to the coelacanth and elephant fish (*SI Appendix*, Fig. S2). Roughly 50% of the whale shark genome is composed of transposable elements (TEs), which were identified using both homology-based and ab initio approaches (19, 20). Of these, long interspersed nuclear elements (LINEs) made up 27% of the total TEs identified (*SI Appendix*, Table S7). A combination of homology-based and ab initio genome annotation methods (21, 22) resulted in a total of 28,483 predicted protein-coding genes (*SI Appendix*, Tables S8–S11).

Correlation of Physiological Characteristics with Genome Features across 85 Taxa. Body mass is intrinsically linked to physiological traits such as lifespan and BMR (23). To better understand how genomic traits correlate with physiological and ecological parameters such as body weight, lifespan, temperature, and metabolic rate, we compared the whale shark to 83 animals and yeast (*SI Appendix*, Tables S15 and S16) using physiological and genomic data (Fig. 1 and *SI Appendix*, Figs. S3–S6 and Table S16). Across the 85 species examined, we find a strong positive correlation with significant P values between the log-transformed values for body weight and maximum lifespan (Spearman's correlation coefficient $\rho = 0.787$, Fig. 2*A* and *SI Appendix*, Table S17*A*) and BMR (*SI Appendix*, Table S17*A* and Fig. S9*A*, $\rho = 0.962$, exponent B = 0.688; *SI Appendix*, Fig. S24; $n = 84$ species, yeast is excluded), consistent with previous reports (23). Comparisons of physiological traits and genome characteristics across these 84 animals and yeast revealed several genetic features that also scaled with body weight. Among these, total gene length, intron length, and genome size all show a moderate statistical correlation with body mass, lifespan, and BMR ($\rho = 0.4$ to 0.7) (Fig. 2 *B–E* and *SI Appendix*, Table S17*A*). These results are consistent with previous findings of decreased intron size correlating with increased metabolic rates. Furthermore, genome size and relative intron size are strongly correlated ($\rho = 0.72$) (Fig. 2*B* and *SI Appendix*, Table S17*A*), with the three sharks and the pika being notable outliers. Moreover, genome size, measured as golden path length, scales with gene size, measured as the summed length of exons and introns per gene (power law exponent B = 1.31, *SI Appendix*, Fig. S25). Additionally, we find that, unlike in bacteria (24) and crustaceans (25), genome size in Chordates scales positively with temperature (*SI Appendix*, Fig. S9*D*; B = 0.97, *SI Appendix*, Fig. S26).

Our comparisons of genome features revealed that exon length is remarkably constant across animals, regardless of genome size or intron length (Fig. 1*C*). Early observations of this phenomenon across small numbers of taxa led to the suggestion that the splicing machinery imposes a minimum exon size while exon skipping begins to predominate when exons exceed ~500 nucleotides in length (26). Interestingly, we also find a tight correlation ($\rho = 0.975$) between the overall GC content and GC3, the GC content of the third codon position (*SI Appendix*, Fig. S9*B* and Table S17), while both features are negatively correlated with the codon adaptation index (CAI) ($\rho = -0.804$ and $\rho = -0.837$, respectively; Fig. 2 *G* and *H* and *SI Appendix*, Table S17) in Eukaryota and negatively correlated with the genome size in Mammalia ($\rho = -0.440$ and $\rho = -0.456$, respectively) (*SI Appendix*, Table S17). These results are partially supported by previous research, which showed that GC3 is negatively correlated with body mass, genome size, and species longevity within 1,138 placental mammal orthologs (27). However, our results using whole-genome data do not support the GC3 correlation with body mass and longevity ($\rho = 0.074$ and $\rho = -0.267$; *SI Appendix*, Table S17). Thus, exon and intron length may affect body mass and longevity through a strong association between GC content and coding sequence length (28). Additionally, CAI and intron size are moderately positively correlated ($\rho = 0.493$; Fig. 2*I* and *SI Appendix*, Table S17). Since the CAI and codon usage bias have an inverse relationship, this is consistent with the negative correlation between intron length and codon usage bias in multicellular organisms (29).

Whale Shark Longevity and Genome Characteristics. The allometric scaling relationships between longevity, mass, temperature, and metabolic rate are well established (23), and the long lifespan of the whale shark can be explained by its large mass and the extremely low mass- and temperature-adjusted BMR (Fig. 1 *H* and *L*). There has been considerable debate in the literature over the evolutionary causes and consequences of genome size,

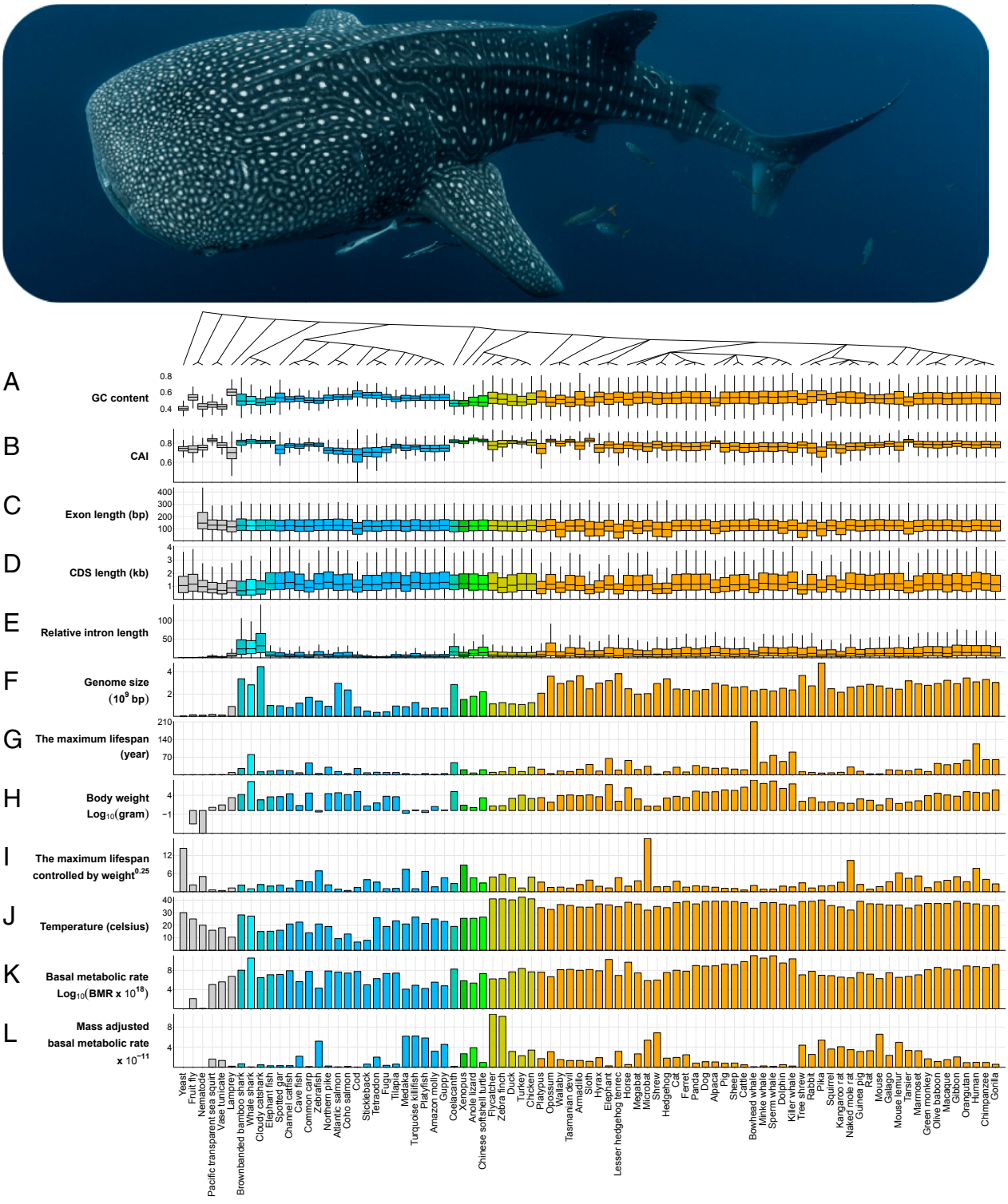

Fig. 1. Comparative genomic analysis across 85 species reveals traits linked to lifespan and bodyweight. (*Top*) Image of a whale shark. (*Bottom*) The phylogenetic tree was constructed using the NCBI common tree (https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi) without divergence times. The rows show the following values in 85 species: five genomic parameters (*A*–*E*), golden path length (*F*), maximum lifespan (*G*), body weight (*H*), maximum lifespan controlled by weight$^{0.25}$ (*I*), body temperature (optimal temperatures for cold-blooded animals) (*J*), basal metabolic rate (yeast is excluded) (*K*), and basal metabolic rate adjusted by weight (yeast is excluded) (*L*). The exon length (*C*) shows length of exons in the coding region. The exon lengths of yeast (median length = 1,032 bp) and fruit fly (median length = 217 bp) and the weight of yeast (6E-14 kg), fruit fly (4.43E-10 kg), and nematode (1.2E-09 kg) are not shown here as they are too extreme to fit in each chart. The relative intron length (*E*) was calculated by dividing the total intron length between the first coding exon and the last coding exon by the CDS length. The nine colors of boxes and bars indicate biological classification (gray: Hyperoartia, Ascidiacea, Chromadorea, Insecta, and Saccharomycetes; turquoise: Chondrichthyes (the cyan color indicates whale shark); light blue: Actinopterygii; aquamarine: Sarcopterygii; dark green: Amphibia; light green: Reptilia; dark yellow: Aves; orange: Mammalia).

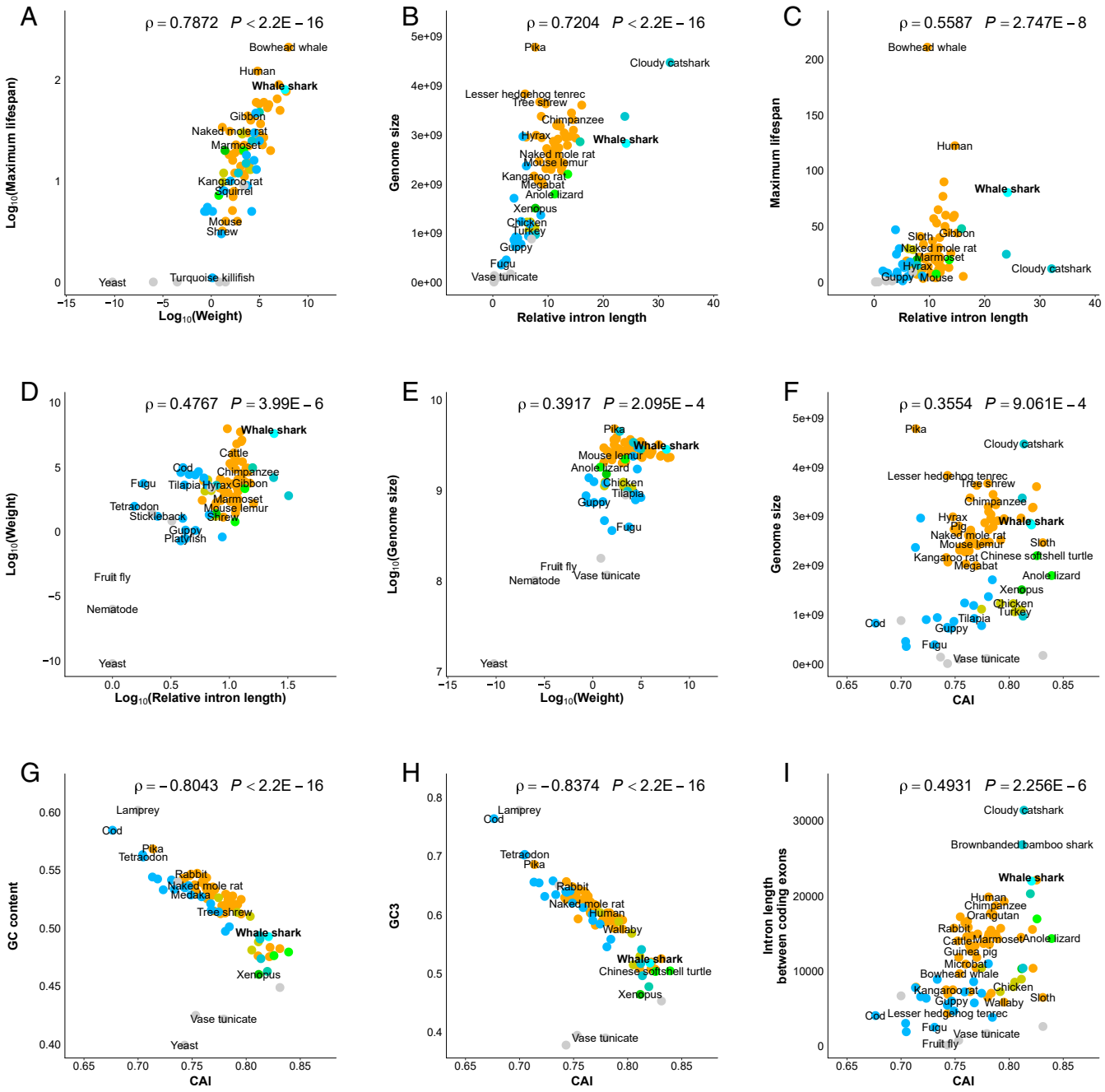

Fig. 2. Scaling relationships between genomic and physiologic properties across 85 species. For each plot, the properties on the *x* axis and *y* axis were used to calculate the 481 Spearman's rank correlation coefficient: (*A*) Log_{10}(Maximum lifespan) and Log_{10}(Weight); (*B*) genome size and relative intron length; (*C*) maximum lifespan and relative intron length; (*D*) Log_{10}(Weight) and Log_{10}(Relative intron length); (*E*) Log_{10}(Genome Size) and Log_{10}(Weight); (*F*) genome size and CAI; (*G*) GC content and CAI; (*H*) GC3 and CAI; and (*I*) intron length between coding exons and CAI. All *P* values and rho (ρ) values are shown at the top of each plot. Overlapping species names in the same layer were not plotted. The nine dot colors indicate biological classification (gray: Hyperoartia, Ascidiacea, Chromadorea, Insecta, and Saccharomycetes; turquoise: Chondrichthyes [the cyan color indicates whale shark]; light blue: Actinopterygii; aquamarine: Sarcopterygii; dark green: Amphibia; light green: 484 Reptilia; dark yellow: Aves; orange: Mammalia).

particularly as it relates to BMR. At 3.2 Gb, the whale shark genome is significantly larger than the elephant fish genome, although both exon number and size are comparable. Similar to the brownbanded bamboo shark and cloudy catshark, the whale shark is notable among fish for its long introns (Fig. 1*E* and *SI Appendix*, Figs. S3*E* and S4*E*). Analyses of single-copy orthologous gene (SCOG) clusters did not reveal any large intron gains or losses in whale shark, brownbanded bamboo shark, or cloudy catshark (*SI Appendix*, Fig. S10), although retrotransposon analysis revealed a significant expansion of CR1-like LINEs and Penelope-like elements within introns (Fig. 3*A* and *SI Appendix*, Figs. S11–S15). The CR1-like LINEs are the dominant family of TEs in nonavian reptiles and birds (30). In these three sharks, the proportion of CR1-like LINE elements accounts for more than 10% of the total intron length (*SI Appendix*, Fig. S13*C*), which is higher than that of the anole lizard, a species known for expanded CR1-like LINEs (30). The total length of intronic repetitive elements is as great as in the opossum genome, known to

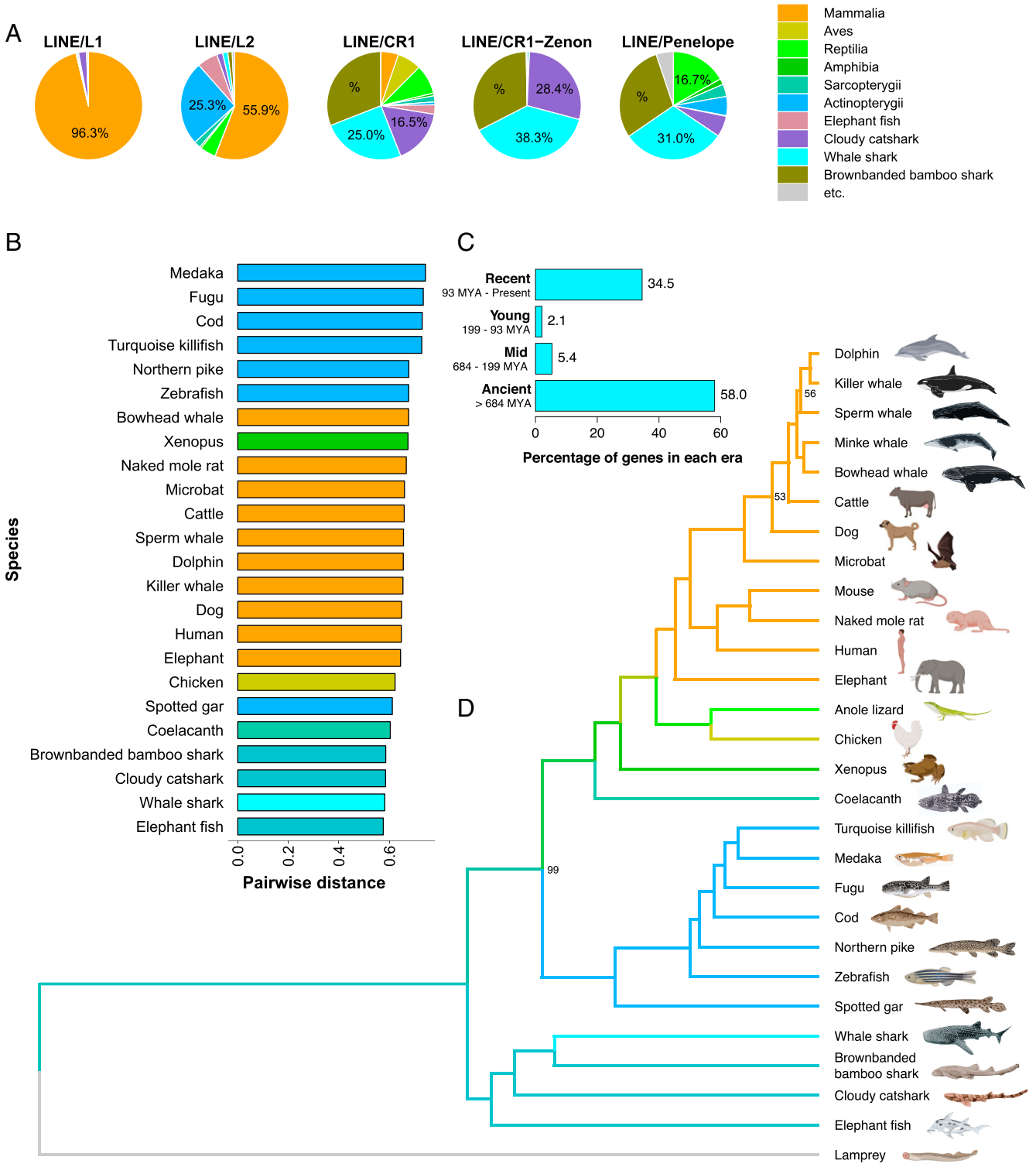

Fig. 3. Repetitive elements, evolutionary rate model, and flow of genes in the whale shark genome. (*A*) Each pie chart summarizes the lengths of predicted intronic repetitive elements (labeled at the top of pies). Values from 84 animals were averaged across six classes (Mammalia, Aves, Reptilia, Amphibia, Sarcopterygii, Actinopterygii). The whale shark and the elephant fish are listed separately. Yeast was excluded from these analyses. (*B*) All pairwise distances from sea lamprey were calculated using the R-package "ape" (32). The species were ordered by the pairwise distances. The eight bar colors indicate biological classification (turquoise: Chondrichthyes (the cyan color indicates whale shark); light blue: Actinopterygii; aquamarine: Sarcopterygii; dark green: Amphibia; light green: Reptilia; dark yellow: Aves; orange: Mammalia). (*C*) While most genes (~58%) in the whale shark genome are ancient, some (~5.4%) are of intermediate age, a few (~2%) are young, and a significant fraction (~34.6%) are new. (*D*) Maximum-likelihood phylogenetic tree of 28 species (for orders with more than one member represented in our 85-species dataset, one species was randomly selected). Bootstrap support values are 100 unless otherwise marked at the nodes. Terminal branches are colored according to the biological classification (gray: Hyperoartia, Ascidiacea, Chromadorea, Insecta, and Saccharomycetes; turquoise: Chondrichthyes (the cyan color indicates whale shark); light blue: Actinopterygii; aquamarine: Sarcopterygii; dark green: Amphibia; light green: Reptilia; dark yellow: Aves; orange: Mammalia).

be rich in repetitive elements (31) (*SI Appendix*, Fig. S11*A*). Although the whale shark has the fourth longest repetitive elements (*SI Appendix*, Fig. S11*A*), it has the highest proportion of LINEs (*SI Appendix*, Fig. S12*B*), particularly CR1-like LINEs and CR1-Zenon like LINEs (*SI Appendix*, Fig. S13 *C* and *D*). In the whale shark genome, 38% of the CR1-like LINEs, 39% of the CR1-Zenon like LINEs, and 30% of the Penelope-like elements are located in intronic regions (*SI Appendix*, Fig. S14). Strikingly, most genes (more than 88%) in the whale shark genome have CR1-like LINE elements within their introns (*SI Appendix*, Fig. S15), a proportion higher than in other Chondrichthians. Moreover, 56% of whale shark genes also have LINE1 elements (*SI Appendix*, Fig. S15). Thus, the whale shark has a relatively large genome and long introns due to an expansion of multiple types of repetitive elements.

Codon usage and the evolutionary age of genes are associated in metazoans (33). Interestingly, two principal component analyses (PCA) of relative synonymous codon usage (RSCU) from 85 and 79 species (6 species having distant codon usage patterns were excluded), respectively, revealed that the whale shark pattern of RSCU is most similar to that of the coelacanth, with well-separated patterns of RSCU for each class (*SI Appendix*, Fig. S16). While the whale shark genome has a relatively short exon length (smaller than those of 25 species; *SI Appendix*, Table S11), notably, it has a smaller number of exons per gene than all but 3 species (yeast, fruit fly, and brownbanded bamboo shark) (*SI Appendix*, Figs. S3*B* and S4*G* and Table S11). Thus, the whale shark coding sequence (CDS) length is shorter than the CDS length of all species except the brownbanded bamboo shark and the vase tunicate (Fig. 1*D* and *SI Appendix*, Fig. S4*D*).

Evolutionary Rate and Historical Demography of the Whale Shark. Analyses of the whale shark genome show that it is the second slowest evolving vertebrate yet characterized. A relative rate test and two cluster analyses revealed that the whale shark has a slower evolutionary rate than those of brownbanded bamboo shark, of cloudy catshark, and of all other bony vertebrates examined, including the coelacanth (16) (Fig. 3*B* and *SI Appendix*, Fig. S17 and Tables S18–S20). These results support previous work predicting a slow evolutionary rate in ectothermic, large-bodied species with relatively low body temperatures (compared to similarly sized warm-blooded vertebrates) (14). They also are consistent with previous studies of nucleotide substitution rates in elasmobranchs, which are significantly lower than those of mammals (34, 35).

A phylogenetic analysis of the 175 SCOG clusters from the whale shark and 27 other animal genomes (Fig. 3*D*) showed a divergence of the Elasmobranchii (sharks) and Holocephali (chimaeras) roughly 333 million years ago (MYA) and of the Chondrichthyes from the bony vertebrates about 358 MYA (Fig. 3*D*), consistent with previous estimates. To better understand the evolutionary history of the genes within the whale shark genome, we evaluated the age of the whale shark protein-coding genes based on protein sequence similarity (36). Grouping the whale shark genes into four broad evolutionary eras, we observed that, while the majority (58%) of genes are ancient (older than 684 MYA), a few (~5.4%) are middle age (684 to 199 MYA), fewer (~2%) are young (199 to 93 MYA), and many (34.6%) are new (93 MYA to present) (Fig. 3*C*). Normalizing the number of genes by evolutionary time suggests that gene turnover is highest near the present time (*SI Appendix*, Fig. S18). Examining the age of genes shows that many genes are ancient and also that many genes appear very young (*SI Appendix*, Fig. S19). These results highlight both the conservation of a large part of the coding genome and the innovative potential of the whale shark genome, since many new genes have appeared within the last 93 million years.

Length of Neural Genes and Correlation with Physiological Features. Gene length has recently emerged as an important feature of neural genes, as long genes are preferentially expressed in neural tissues and their expression is under tight transcriptional and epigenetic control (37). Within 84 animals and yeast, we compared the dimensions of average genes with those of 10 categories of neural genes (neuronal connectivity, cell adhesion, olfactory receptors, ion channels, unfolded protein-response–associated genes, neuronal activity and memory, neuropeptides, homeobox genes, synaptic genes, and neurodegeneration) (Fig. 4*A* and *SI Appendix*, Figs. S20 and S21). Interestingly, we found that neuronal connectivity genes are longer than average genes in most vertebrates, with the length increase being significant in whale shark and most mammals, as well as in coelacanth and platypus (Fig. 4*A* and *SI Appendix*, Fig. S21*A*). Surprisingly, we found that neural genes are scaled to average genes with an exponent greater than 1 (B = 1.038, *SI Appendix*, Fig. S27*A*), with the whale shark, brownbanded bamboo shark, and cloudy catshark showing an extreme lengthening of neural genes (*SI Appendix*, Figs. S20, S21, and S27). Moreover, we found that cell adhesion, ion channels, unfolded protein-response–related genes, and neurodegeneration genes are increased in length in the whale shark and two other shark species (Fig. 4*B* and *SI Appendix*, Fig. S28), suggesting that this may be a general feature of sharks. Finally, neuronal functions are enriched in long genes in more than 60 species (Fig. 4*C* and *SI Appendix*, Table S21 and Dataset S1).

To determine which genes are linked with three physiological traits (maximum lifespan, body weight, and BMR), we examined the correlation of gene size and three physiological traits in SCOG families. We found 172 SCOG families in which gene lengths significantly correlated to three physiological traits (*SI Appendix*, Tables S22–S24). Gene Ontology (GO) analyses of the 172 SCOG families showed statistical enrichment of biological processes such as regulation of telomere maintenance via telomerase (GO: 0032210—*CCT3*, *CDK7*, *MAPKAPK5*, *NAT10*, and *XRCC5*) and tRNA metabolic processes (GO: 0006399—*DDX1*, *EEF1E1*, *KARS*, *LARS2*, *NARS2*, *NAT10*, *PUS10*, *RARS2*, *RTCB*, and *ZBTB8OS*; Fig. 4*D* and *SI Appendix*, Tables S22 and S23), both of which are associated with longevity and cancer (38, 39). Furthermore, the size of *DLD*, a gene with proteolytic activity and metabolic function (40, 41), significantly correlates with BMR (ρ = 0.67) (*SI Appendix*, Fig. S23). These results suggest that an evolutionary relationship exists between gene size and several physiological traits size such as body size, metabolic rate, and lifespan. This holds particularly for genes whose functions are essential for living long lives, such as telomere maintenance and metabolic activity.

Discussion

We sequenced and assembled the genome of the whale shark (*R. typus*), an endangered species that is the largest extant fish on Earth. We compared it to the genomes of 84 eukaryotic species, and related genomic traits to physiological traits and environmental variables such as temperature to understand how ecological constraints shape genomes. Several major findings emerged from our comparative evolutionary analyses. First, at 3.2 Gb, the relatively large genome of the whale shark is the second slowest evolving vertebrate genome found to date and has a striking number of CR1-like LINE transposable elements. Second, in most genomes, we found that major genomic traits, including intron length and gene length, scale with body size, temperature, and lifespan. Some genomic traits are correlated to metabolic rate, which scales with both animal mass and temperature, thus reflecting both physiology and environment. These results suggest that ecological variables mold both genomes and morphology. Third, we found that GC content and codon adaptation index are negatively correlated. Furthermore,

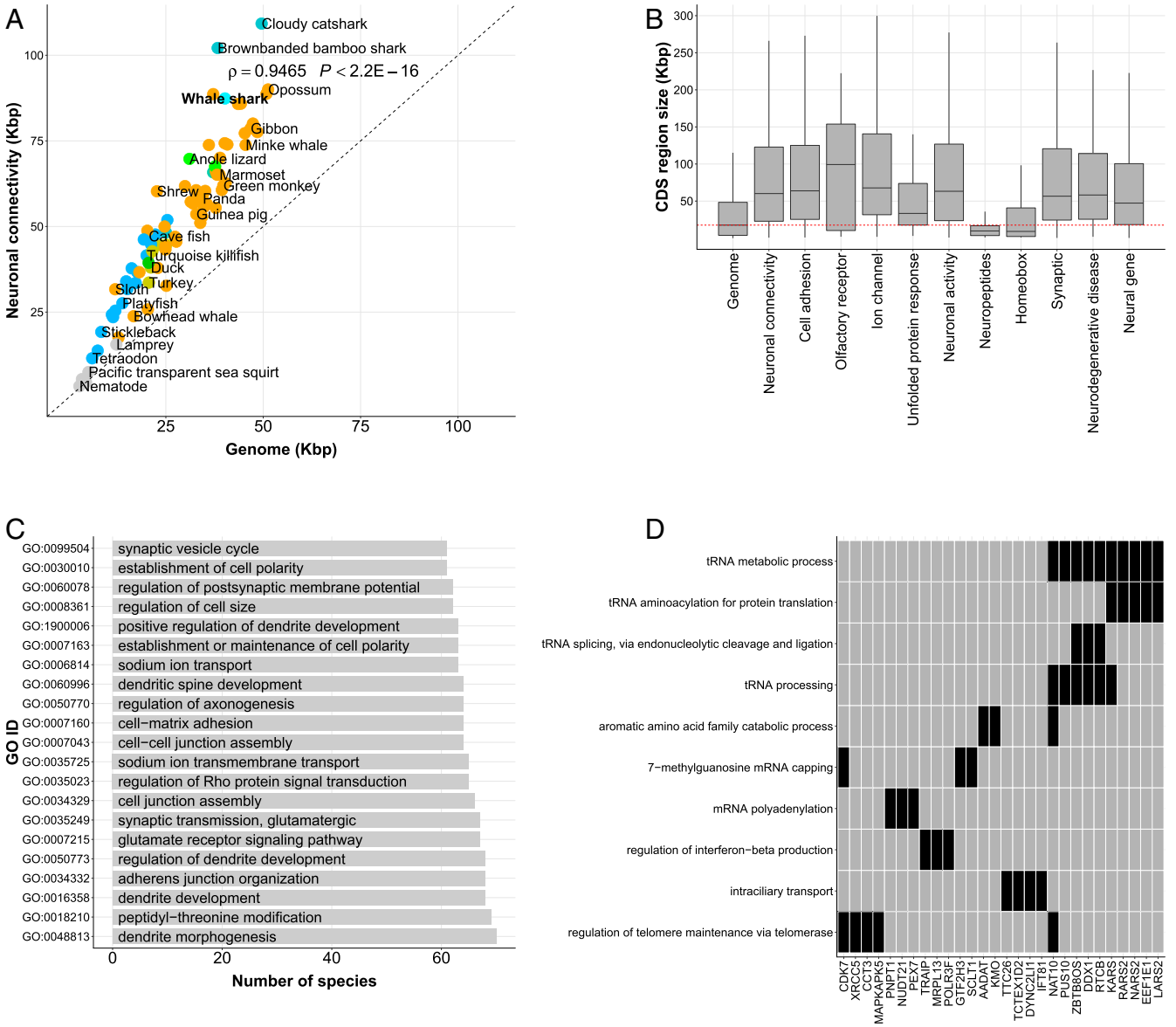

Fig. 4. The relationship between gene length and neural genes and SCOG families with correlations between gene length and maximum lifespan, weight, and BMR. (*A*) Neuronal connectivity genes are longer than average genes in 84 animals. The *x* and *y* axes show the average gene length and the gene length of neuronal connectivity-related genes, respectively. The dashed diagonal line represents "$y = x$." Spearman's rho correlation coefficient and *P* value are shown in the top right corner of the plot. (*B*) Of the 12 categories of neural genes that we analyzed in the whale shark genome, several are longer than average whale shark genes. (*C*) Most common GO terms are relevant to neural function. GO terms, shown based on the number of species in which they were found, were computed with Gene Set Enrichment Analysis. (*D*) Enriched GO functions in SCOG families in which relative intron length positively correlates with maximum lifespan. For each GO term, black boxes indicate human gene symbols representative of the family.

while the correlation of GC3 content and overall genomic GC content had previously been established (42), we extended the validity of this correlation to a wide range of species, densely sampling chordate genomes at genome scale. Fourth, unexpectedly, we found that neural connectivity genes are substantially longer than average genes. While it has previously been observed that neural genes are longer than average genes in the human genome, our comparative analysis has dramatically extended the range of this observation to more than 80 species. Interestingly, we found that introns are longer in the shark genomes than in most other species due to the high proportion of repetitive elements. Finally, we found that neural genes of several types, including neurodegeneration genes, are much longer than average genes in species with long lifespans.

As a general approach, studying whether distinct quantitative traits are correlated at vastly different spatial and temporal scales is an important discovery tool. First, for pairs of traits the correlation of which was not anticipated, such as intron size and body weight, the quantification of scaling enables the generation of mechanistic hypotheses. Second, examining the relationships between quantitative traits on a large evolutionary scale, as we have done here in a group of 85 Eukaryotic species centered on Chordates, enables the identification of the mathematical functions that best describe the relationships between traits. For some pairs of traits, these functions can be expressed as power law equations that may be succinctly summarized as scaling exponents. It should also be noted that, for many of the strongly correlated traits, there are notable outliers, such as the bowhead whale, in comparisons of longevity. When traits such as genome

size and lifespan correlate, large-scale evolutionary comparisons can be used to identify the outlier species that are most suited for addressing particular research questions. Together, these results show the power of the comparative evolutionary approach and of mathematical modeling to uncover both general and specific relationships that reveal how genome architecture is shaped by size and ecology.

Methods

Sample Preparation and Sequencing. Genomic DNA was isolated from heart tissue acquired from a 7-y-old, 4.5-m deceased male whale shark from the Hanwha Aquarium, Jeju, Korea. DNA libraries were constructed using a TruSeq DNA library kit for the short-read libraries and a Nextera Mate Pair sample prep kit for the mate-pair libraries. Sequencing was performed using the Illumina HiSeq2500 platform. Libraries were sequenced to a combined depth of 164× (*SI Appendix*, Tables S1 and S2).

Genome Assembly and Annotation. Reads were quality-filtered (*SI Appendix*, Table S3), and the error-corrected reads from the short insert size libraries (<1 kb) and mate-pair libraries (>1 kb) were used to assemble the whale shark genome using SOAPdenovo2 (43). As the quality of the assembled genome can be affected by the *K*-mer size, we used multi-K-mer values (minimum 45 to maximum 63) with the "all" command in the SOAPdenovo2 package (43). The gaps between the scaffolds were closed in two iterations with the short insert libraries (<1 kb) using the GapCloser program in the SOAPdenovo2 package (43). We then aligned the short insert size reads to the scaffolds using BWA-MEM (44) with default options. Variants were identified using SAMtools (45) and scaffolds were error corrected by substituting the short insert read allele. For heterozygous mapped alleles, the first variant was substituted. Finally, we mapped the Illumina TruSeq TSLRs to the assembly, corrected the gaps covered by the synthetic long reads to reduce erroneous gap regions in the assembly (*SI Appendix*, Tables S5 and S13), and assessed the genome assembly and genome completeness using the BUSCO approach (*SI Appendix*, Table S14) (46).

The GC distribution of the whale shark genome was calculated using a sliding-window approach. We employed 10-kb sliding windows to scan the genome and calculate the GC content. Tandem repeats were predicted using the Tandem Repeats Finder program (version 4.07) (47). TEs were identified using both homology-based and ab initio approaches. The Repbase database (version 19.02) (48) and RepeatMasker (version 4.0.5) (19) were used for the homology-based approach, and RepeatModeler (version 1.0.7) (20) was used for the ab initio approach. All predicted repetitive elements were merged using in-house Perl scripts. Two candidate gene sets were built to predict the protein-coding genes in the whale shark genome using AUGUSTUS (22) and Evidence Modeler (21), respectively (*SI Appendix*, 1.7 Annotation of protein-coding genes).

Genomic Context Calculations. From 85 species (*SI Appendix*, Table S15), we computed the following genomic factors: GC3 (GC content at third codon position), CAI, number and length of coding exon(s), and relative intron length between the first and last exon (or coding exon). CDS sequences with premature stop codons and lengths that were not multiples of three were excluded. The relative intron length was calculated by dividing the total intron length between first and last exon (or coding exon) by the CDS length (or messenger RNA length). GC3 was computed from concatenated third codon nucleotides (49). We measured RSCU using the method from Sharp et al. (50) and the CAI in a CDS using Sharp and Li's method (51) for each of the 85 species. The PCA on RSCU was performed using the R packages (version 3.3.0) (52) ggplot2 (53) and ggfortify (54).

Orthologous Gene Family Clustering and Phylogeny Construction. To identify orthologous gene families among the whale shark and the other 85 species, we downloaded all pair-wise reciprocal BLASTP results using the "peptide align feature" in the Ensembl genome database project (release 86) (55). To generate pair-wise orthologues that were not available in the Ensembl resources, we performed reciprocal BLASTP (56) with the "-evalue 1e-05 -seg no -max_hsps_per_subject 1 -use_sw_tback" options. From the pair-wise reciprocal BLASTP results among the 85 species, we generated similarity matrices by connecting possible orthologous pairs. To constrain the computational load, we did not join additional nodes when the number of nodes was larger than 1,500. The normalized weights for the similarity matrix were calculated using the OrthoMCL approach (57). We identified orthologous gene families using an in-house C++ script based on the Markov Cluster (MCL) algorithm (58) with inflation index 1.3. A total of 1,556,795 genes were assigned to 245,314 clusters including 209,992 singletons, and 175 single-copy gene families were extracted from 28 species. Multiple sequence alignments were performed using MUSCLE 3.8.31 (59) and were concatenated without gap regions. The phylogenetic tree was constructed using RAxML 8.2 (60) with maximum likelihood (1,000 bootstraps), using the PROTCATLG amino acid substitution model (Fig. 3*D*).

Gene Age Estimation. Phylostratigraphy uses BLASTP-scored sequence similarity to estimate the minimal age of every protein-coding gene. The National Center for Biotechnology Information (NCBI) nonredundant database was queried with a protein sequence to detect the most distant species in which a sufficiently similar sequence is present and then posit that the gene is at least as old as the age of the common ancestor (36). Using NCBI taxonomy for every species, the timing of lineage divergence events was estimated with TimeTree (61). To facilitate detection of protein sequence similarity, we used the e-value threshold of 10^{-3}. We evaluated the minimal evolutionary age of all protein-coding genes the protein sequence lengths of which are between 40 amino acids and 4,000 amino acids. First, we counted the number of genes in each phylostratum (PS), from the most ancient (PS 1, cellular organisms) to the most recent (PS 20, *R. typus*). It should be noted that, within the Rhincodontidae family (PS 18) and the *Rhincodon* genus (PS 19), the whale shark is currently the only species with a sequenced genome. Therefore, the large number of genes that appeared species-specific (7,647 genes in phylostratum *R. typus*, *SI Appendix*, Fig. S19) may include marginally older genes that are restricted to the genus *Rhincodon* (PS 19) or to the family Rhincodontidae (PS 18), but cannot presently be assigned to these two phylostrata until additional high-quality genomes are sequenced and assembled for species in these clades. To evaluate broad evolutionary patterns, we aggregated the counts from several phylostrata into four broad evolutionary eras: ancient (PS 1 to 7, cellular organisms to Deuterostomia, 4,204 to 684 MYA), middle (PS 8 to 14, Chordata to Selachii, 684 to 199 MYA), young (PS 15 to 17, Galeomorpha to Orectolobiformes, 199 to 93.2 MYA), and newest (PS 18 to 20, Rhincodontidae to *R. typus*, 93.2 MYA to present). To estimate the gene flow per time unit, we normalized the number of genes in an era by the age and the duration of that evolutionary era.

Correlation Tests in Orthologous Gene Families. From these 85 species, we selected 9,180 SCOG gene families found in at least 40 species to calculate the correlation between gene length and three physiological properties (the maximum lifespan, body weight, and BMR). We identified gene families that had significant correlations between gene length and maximum lifespan (3,521 genes), body mass (2,620 genes), and BMR (3,267 genes). The statistical significance of correlations was evaluated by calculating Spearman's rho (ρ) correlation coefficient and applying the Benjamini–Hochberg adjustment (adjusted P value ≤ 0.05). All of these gene families were subject to alignment filtering criterion that included more than 50% of conserved exon–exon boundaries (intron position) in their CDS alignments. This step reduces the effect of gene length changes due to intron gain or loss and increases the accuracy of multiple sequence alignments (*SI Appendix*, Fig. S23). Finally, we acquired four sets of gene families where we observed correlations between gene length and three properties: 1) 18 gene families in which gene length correlated with the maximum lifespan only (*SI Appendix*, Table S24), 2) 3 gene families correlated with the body weight only (*SI Appendix*, Table S24), 3) 7 gene families correlated with the BMR only, and 4) 148 gene families correlated with all three physiological properties (*SI Appendix*, Table S23).

Statistical Analysis. For all pairs of median values of physiological and genomic features assessed in this analysis of 85 species, the Spearman's rank correlation coefficient rho (ρ) values were calculated using the cor.test package in R with the following options: method = "spearman," exact = "true," and were plotted using the ggplot2 package. If the resulting P values were lower than 2.2e-16, the smallest value output using this package, $P <$ 2.2e-16, was listed, rather than an exact value. For selected pairs of values, the plots were displayed using median values or $\log_{10}$-transfomed values, as appropriate. Second, for the nine pairs of genomic and physiological features the scaling correlations of which were evaluated in Fig. 2, the robustness of the Spearman's correlation coefficients was calculated using a leave-one-out jackknifing procedure (62), performed using the StatPlus program (n = 85 species, 85 iterations). The results are reported in *SI Appendix*, Table S17*B*, where the Spearman's correlation coefficient values (ρ) are shown along with measures of ρ variation (minimum, maximum, and SD). Third, the variability, skewness, and kurtosis were evaluated for all

median values of physiological and genomic features and were also evaluated for the nine Spearman's correlation coefficient distributions generated by jackknifing. Fourth, for every physiological and genomic feature assessed in this analysis, pairwise comparisons between the 85 species were done as two-sided Wilcoxon rank-sum tests and displayed as correlation matrices. All *P* values were adjusted using the Benjamini–Hochberg procedure, log-transformed, and displayed in a color scale ranging from 0.000 to 0.01; values higher than 0.01 are shown in gray.

Scaling Analysis. The adjustment of the basal metabolic rate to mass is based on Gillooly's Eq. 1 (14) relating the mass-adjusted basal metabolic rate to mass and temperature $B = b_0 M^{-1/4} e^{-E/kT}$, where B = basal metabolic rate, b_0 is a coefficient independent of body size and temperature, M = organism mass, E_i = average activation energy for enzyme-catalyzed biochemical reactions of metabolism (~0.65 eV), T = absolute temperature (for poikilo-terms, the environmental temperature at which the organism lives; for homeotherms, of the organism itself), and $e^{-E/kT}$ = Arrhenius or Boltzmann factor, which includes $k = 8.62 * 10^{-5}$ eV·K^{-1} (Boltzmann constant). Furthermore, we compared BMR values that were measured experimentally to BMR values calculated with Gillooly's Eq. 1 (14). We found a very strong correlation between measured and calculated BMR values (*SI Appendix*, Fig. S7*A*, Spearman's $\rho = 0.954$, $n = 24$ species; *SI Appendix*, Fig. S7*B*, $\rho = 0.935$, $n = 21$ species excluding cattle, pig, and human, which have very high BMRs). Since the whale shark routinely dives to cold, deep depths, we also calculated the whale shark BMR across its temperature range (*SI Appendix*, Fig. S8).

Data and Materials Availability. The whale shark whole-genome project has been deposited in the INSDC: International Nucleotide Sequence Database Collaboration under accession no. QPMN00000000. The version described in this paper is version QPMN01000000. DNA sequencing reads have been uploaded to the NCBI Sequence Read Archive (SRP155581). The C++ code used for the MCL algorithm was uploaded to the GitHub repository (https://github.com/jsungwon/MCL-clustering).

ACKNOWLEDGMENTS. We thank Dr. Mark Erdmann for generously providing the whale shark photograph used in Fig. 1. Portions of Fig. 3 were created with https://biorender.com/. V.L. thanks Marc W. Kirschner (M.W.K.) and Anne O'Donnell-Luria (A.O.L.) for discussions and support. This work was supported by the Genome Korea Project in Ulsan (800 genome sequencing) Research Fund (1.180017.01) of the Ulsan National Institute of Science and Technology (UNIST); the Genome Korea Project in Ulsan (200 genome sequencing) Research Fund (1.180024.01) of UNIST; NIH Grants R01 HD073104 and R01 HD091846 (to M.W.K.); and the William Randolph Hearst Fund Award and a Boston Children's Hospital Career Development Fellowship (to A.O.L.).

1. J. A. Goldbogen *et al*., Why whales are big but not bigger: Physiological drivers and ecological limits in the age of ocean giants. *Science* **366**, 1367–1372 (2019).
2. W. Gearty, C. R. McClain, J. L. Payne, Energetic tradeoffs control the size distribution of aquatic mammals. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 4194–4199 (2018).
3. D. Atkinson, Temperature and organism size: A biological law for ectotherms? *Adv. Ecol. Res.* **25**, 1–58 (1994).
4. D. Atkinson, B. J. Ciotti, D. J. Montagnes, Protists decrease in size linearly with temperature: Ca. 2.5% degrees C(-1). *Proc. Biol. Sci.* **270**, 2605–2611 (2003).
5. C.-T. Chen, "Preliminary report on Taiwan's whale shark fishery" in *Elasmobranch Biodiversity, Conservation and Management: Proceedings of the International Seminar and Workshop, Sabah, Malaysia, July 1997*, S. Fowler, T. M. Reed, F. A. Dipper, Eds. (IUCN, Gland, Switzerland, 1997), pp. 162–167.
6. H. H. Hsu, S. J. Joung, R. E. Hueter, K. M. Liu, Age and growth of the whale shark (Rhincodon typus) in the north-western Pacific. *Mar. Freshw. Res.* **65**, 1145–1154 (2014).
7. J. G. Colman, A review of the biology and ecology of the whale shark. *J. Fish Biol.* **51**, 1219–1234 (1997).
8. D. Rowat, K. S. Brooks, A review of the biology, fisheries and conservation of the whale shark Rhincodon typus. *J. Fish Biol.* **80**, 1019–1056 (2012).
9. A. M. Sequeira, C. Mellin, L. Floch, P. G. Williams, C. J. Bradshaw, Inter-ocean asynchrony in whale shark occurrence patterns. *J. Exp. Mar. Biol. Ecol.* **450**, 21–29 (2014).
10. J. P. Tyminski, R. de la Parra-Venegas, J. González Cano, R. E. Hueter, Vertical movements and patterns in diving behavior of whale sharks as revealed by pop-up satellite tags in the eastern Gulf of Mexico. *PLoS One* **10**, e0142156 (2015).
11. M. Thums, M. Meekan, J. Stevens, S. Wilson, J. Polovina, Evidence for behavioural thermoregulation by the world's largest fish. *J. R. Soc. Interface* **10**, 20120477 (2013).
12. A. P. Martin, S. R. Palumbi, Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl. Acad. Sci. U.S.A.* **90**, 4087–4091 (1993).
13. C. D. Laird, B. L. McConaughy, B. J. McCarthy, Rate of fixation of nucleotide substitutions in evolution. *Nature* **224**, 149–154 (1969).
14. J. F. Gillooly, A. P. Allen, G. B. West, J. H. Brown, The rate of DNA evolution: Effects of body size and temperature on the molecular clock. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 140–145 (2005).
15. B. Venkatesh *et al*., Elephant shark genome provides unique insights into gnathostome evolution. *Nature* **505**, 174–179 (2014).
16. C. T. Amemiya *et al*., The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496**, 311–316 (2013).
17. Q. Zhang, S. V. Edwards, The evolution of intron size in amniotes: A role for powered flight? *Genome Biol. Evol.* **4**, 1033–1043 (2012).
18. A. Kapusta, A. Suh, C. Feschotte, Dynamics of genome size evolution in birds and mammals. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E1460–E1469 (2017).
19. J. A. Bedell, I. Korf, W. Gish, MaskerAid: A performance enhancement to RepeatMasker. *Bioinformatics* **16**, 1040–1041 (2000).
20. G. Abrusán, N. Grundmann, L. DeMester, W. Makalowski, TEclass: A tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**, 1329–1330 (2009).
21. B. J. Haas *et al*., Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
22. M. Stanke, B. Morgenstern, AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465-7 (2005).
23. G. B. West, J. H. Brown, B. J. Enquist, A general model for the origin of allometric scaling laws in biology. *Science* **276**, 122–126 (1997).
24. N. Sabath, E. Ferrada, A. Barve, A. Wagner, Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. *Genome Biol. Evol.* **5**, 966–977 (2013).
25. K. Alfsnes, H. P. Leinaas, D. O. Hessen, Genome size in arthropods: Different roles of phylogeny, habitat and life history in insects and crustaceans. *Ecol. Evol.* **7**, 5939–5947 (2017).
26. D. A. Sterner, T. Carlo, S. M. Berget, Architectural limits on split genes. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 15081–15085 (1996).
27. J. Romiguier, V. Ranwez, E. J. Douzery, N. Galtier, Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. *Genome Res.* **20**, 1001–1009 (2010).
28. J. L. Oliver, A. Marín, A relationship between GC content and coding-sequence length. *J. Mol. Evol.* **43**, 216–223 (1996).
29. A. E. Vinogradov, Intron-genome size relationship on a large evolutionary scale. *J. Mol. Evol.* **49**, 376–384 (1999).
30. A. Suh *et al*., Multiple lineages of ancient CR1 retroposons shaped the early genome evolution of amniotes. *Genome Biol. Evol.* **7**, 205–217 (2014).
31. A. J. Gentles *et al*., Evolutionary dynamics of transposable elements in the short-tailed opossum Monodelphis domestica. *Genome Res.* **17**, 992–1004 (2007).
32. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
33. Y. Prat, M. Fromer, N. Linial, M. Linial, Codon usage is associated with the evolutionary age of genes in metazoan genomes. *BMC Evol. Biol.* **9**, 285 (2009).
34. A. P. Martin, G. J. Naylor, S. R. Palumbi, Rates of mitochondrial DNA evolution in sharks are slow compared with mammals. *Nature* **357**, 153–155 (1992).
35. A. P. Martin, Substitution rates of organelle and nuclear genes in sharks: Implicating metabolic rate (again). *Mol. Biol. Evol.* **16**, 996–1002 (1999).
36. T. Domazet-Loso, J. Brajković, D. Tautz, A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* **23**, 533–539 (2007).
37. H. W. Gabel *et al*., Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature* **522**, 89–93 (2015).
38. K. L. Rudolph *et al*., Longevity, stress response, and cancer in aging telomerase-deficient mice. *Cell* **96**, 701–712 (1999).
39. Z. Ke *et al*., Translation fidelity coevolves with longevity. *Aging Cell* **16**, 988–993 (2017).
40. N. E. Babady, Y. P. Pang, O. Elpeleg, G. Isaya, Cryptic proteolytic activity of dihydrolipoamide dehydrogenase. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 6158–6163 (2007).
41. M. H. Odièvre *et al*., A novel mutation in the dihydrolipoamide dehydrogenase E3 subunit gene (DLD) resulting in an atypical form of α-ketoglutarate dehydrogenase deficiency. *Hum. Mutat.* **25**, 323–324 (2005).
42. J. P. Thiery, G. Macaya, G. Bernardi, An analysis of eukaryotic genomes by density gradient centrifugation. *J. Mol. Biol.* **108**, 219–235 (1976).
43. R. Luo *et al*., SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012).
44. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*:1303.3997 (16 March 2013).
45. H. Li *et al*.; 1000 Genome Project Data Processing Subgroup, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
46. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
47. G. Benson, Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
48. J. Jurka *et al*., Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
49. T. Tatarinova, E. Elhaik, M. Pellegrini, Cross-species analysis of genic GC_3 content and DNA methylation patterns. *Genome Biol. Evol.* **5**, 1443–1456 (2013).

50. P. M. Sharp, T. M. Tuohy, K. R. Mosurski, Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.* **14**, 5125–5143 (1986).
51. P. M. Sharp, W. H. Li, The codon Adaptation Index: A measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295 (1987).
52. R. C. Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2014), 2013.
53. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*, (Springer, 2016).
54. M. Horikoshi, Y. Tang, W. Li, ggfortify: Unified interface to visualize statistical results of popular R packages. *R Journal* **8**, 478–489 (2016).
55. J. Herrero *et al.*, Ensembl comparative genomics resources. *Database (Oxford)* **2016**, bav096 (2016).
56. C. Camacho *et al.*, BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
57. L. Li, C. J. Stoeckert Jr., D. S. Roos, OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
58. A. J. Enright, S. Van Dongen, C. A. Ouzounis, An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
59. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
60. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
61. S. Kumar, G. Stecher, M. Suleski, S. B. Hedges, TimeTree: A resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
62. A. J. Bishara, J. B. Hittner, Confidence intervals for correlations when data are not normal. *Behav. Res. Methods* **49**, 294–309 (2017).