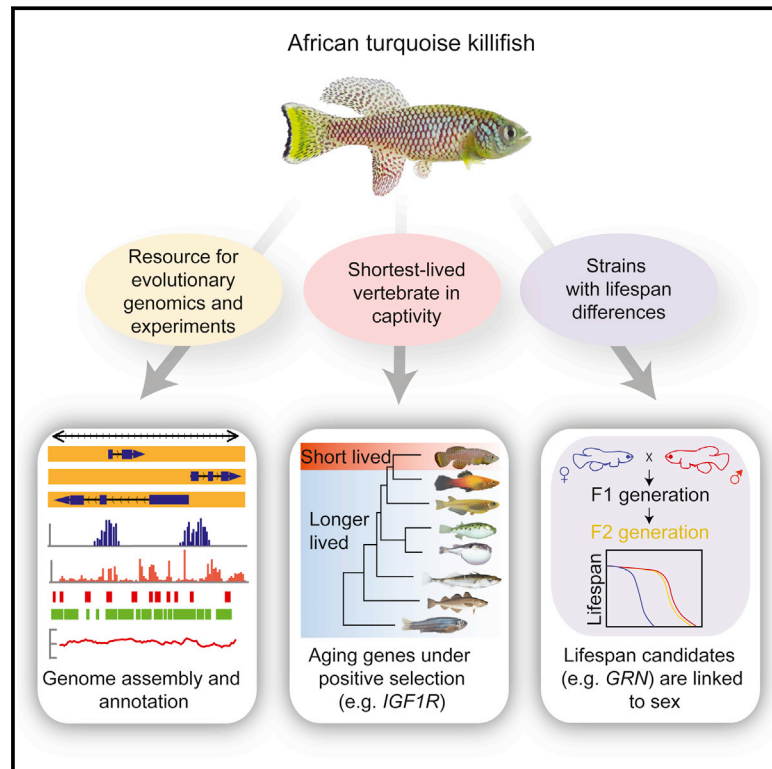


# The African Turquoise Killifish Genome Provides Insights into Evolution and Genetic Architecture of Lifespan

## Graphical Abstract



## Authors

Dario Riccardo Valenzano,  
B  r  nice A. Benayoun,  
Param Priya Singh, ..., Andreas Beyer,  
Eric A. Johnson, Anne Brunet

## Correspondence

dario.valenzano@age.mpg.de (D.R.V.),  
anne.brunet@stanford.edu (A.B.)

## In Brief

The genome of the African turquoise killifish, an exceptionally short-lived fish, is a useful resource to explore the genetic principles and the evolution of unique traits in lifespan and embryonic diapause. Linkage analysis suggests that short lifespan could have co-evolved with sex determination.

## Highlights

- De novo genome assembly and annotation of the African turquoise killifish
- Key aging genes are under positive selection in the turquoise killifish
- Differences in lifespan between killifish strains are genetically linked to sex
- A resource for comparative genomics and experimental aging studies



# The African Turquoise Killifish Genome Provides Insights into Evolution and Genetic Architecture of Lifespan

Dario Riccardo Valenzano,<sup>1,7,8,\*</sup> Bérénice A. Benayoun,<sup>1,7</sup> Param Priya Singh,<sup>1,7</sup> Elisa Zhang,<sup>1</sup> Paul D. Etter,<sup>2</sup> Chi-Kuo Hu,<sup>1</sup> Mathieu Clément-Ziza,<sup>3</sup> David Willemsen,<sup>4</sup> Rongfeng Cui,<sup>4</sup> Itamar Harel,<sup>1</sup> Ben E. Machado,<sup>1</sup> Muh-Ching Yee,<sup>1,9</sup> Sabrina C. Sharp,<sup>1</sup> Carlos D. Bustamante,<sup>1</sup> Andreas Beyer,<sup>5</sup> Eric A. Johnson,<sup>2</sup> and Anne Brunet<sup>1,6,\*</sup>

<sup>1</sup>Department of Genetics, Stanford University, California 94305, USA

<sup>2</sup>Institute of Molecular Biology, University of Oregon, Oregon 97403, USA

<sup>3</sup>CMMC, University of Cologne, Cologne, 50931, Germany

<sup>4</sup>Max Planck Institute for Biology of Ageing, Cologne, 50931, Germany

<sup>5</sup>Cellular Networks and Systems Biology, CECAD, University of Cologne, Cologne, 50931, Germany

<sup>6</sup>Glenn Laboratories for the Biology of Aging, Stanford University, California 94305, USA

<sup>7</sup>Co-first author

<sup>8</sup>Present address: Max Planck Institute for Biology of Ageing, Cologne, 50931, Germany

<sup>9</sup>Present address: Carnegie Institution for Science, Stanford, California 94305, USA

\*Correspondence: [dario.valenzano@age.mpg.de](mailto:dario.valenzano@age.mpg.de) (D.R.V.), [anne.brunet@stanford.edu](mailto:anne.brunet@stanford.edu) (A.B.)

<http://dx.doi.org/10.1016/j.cell.2015.11.008>

## SUMMARY

Lifespan is a remarkably diverse trait ranging from a few days to several hundred years in nature, but the mechanisms underlying the evolution of lifespan differences remain elusive. Here we de novo assemble a reference genome for the naturally short-lived African turquoise killifish, providing a unique resource for comparative and experimental genomics. The identification of genes under positive selection in this fish reveals potential candidates to explain its compressed lifespan. Several aging genes are under positive selection in this short-lived fish and long-lived species, raising the intriguing possibility that the same gene could underlie evolution of both compressed and extended lifespans. Comparative genomics and linkage analysis identify candidate genes associated with lifespan differences between various turquoise killifish strains. Remarkably, these genes are clustered on the sex chromosome, suggesting that short lifespan might have co-evolved with sex determination. Our study provides insights into the evolutionary forces that shape lifespan in nature.

## INTRODUCTION

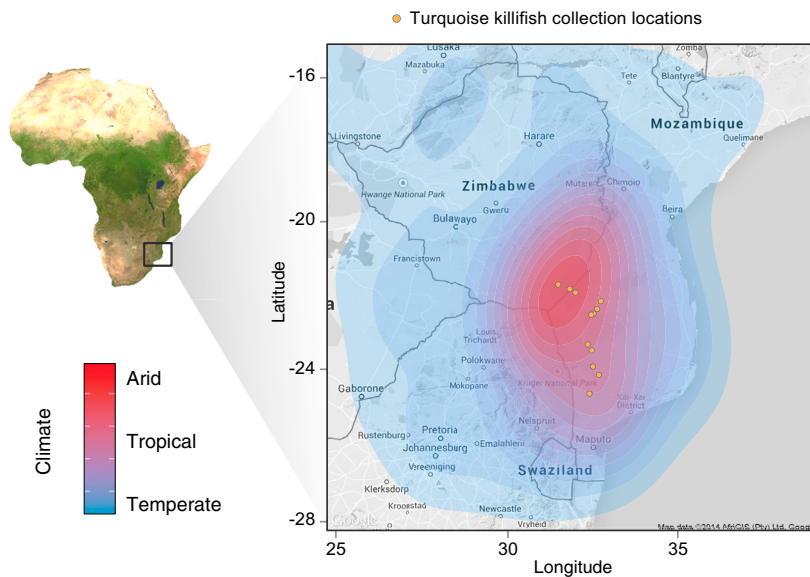
Nature offers an amazing diversity in the lifespan of species with a 150,000-fold difference between the shortest-lived and longest-lived species (Austad, 2010). Short-lived species have long been recognized as essential for experimental studies, and central pathways that regulate aging—notably the insulin-FOXO and mTOR pathways—have been discovered in short-lived models such as yeast, worms, and flies (Kaeberlein and Kennedy, 2011; Kapahi et al., 2010; Kenyon, 2010). These path-

ways have turned out to modulate aging all the way to humans (Flachsbart et al., 2009; Johnson et al., 2013). Many short-lived species tend to occupy unique ecological niches where the environment is harsh. Thus, these species can be used experimentally for the identification of conserved genes important for lifespan and can also provide insight into the evolutionary forces that shape lifespan strategies in nature.

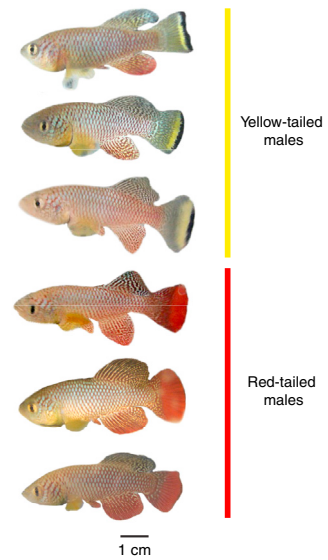
At the other end of the lifespan spectrum, exceptionally long-lived species, such as the naked mole rat (~30 years), the Brandt's bat (~40 years), and the bowhead whale (~200 years) have recently been used in comparative genomic studies to identify genes and residues that are uniquely changed or under positive selection in these organisms (Keane et al., 2015; Kim et al., 2011; Seim et al., 2013). Interestingly, several genes involved in aging and metabolic pathways are uniquely changed in the Brandt's bat (e.g., insulin-like growth factor 1 receptor [*IGF1R*] and growth hormone receptor) (Seim et al., 2013) and in the naked mole rat (e.g., the uncoupling protein *UCP1*) (Kim et al., 2011). Despite these advances, much remains to be learned about the genes that drive natural differences in lifespan and their evolutionary selection in the wild.

Here, we examine the evolution and genetic architecture of lifespan in a naturally short-lived vertebrate, the African turquoise killifish (*Nothobranchius furzeri*), which has been proposed as a vertebrate model for aging (Valdesalici and Cellarino, 2003). We report the de novo assembly and annotation of a reference genome for the turquoise killifish. Evolutionary genomics reveals genes that are positively selected in this short-lived fish and may underlie unique traits in this species, including its short life cycle. By performing a linkage analysis between two strains with experimental lifespan differences, we also identify candidate genes that are likely associated with lifespan differences between strains. Our study provides a resource for experimental and comparative genomics and gives insights into the evolution of naturally short lifespans in nature.

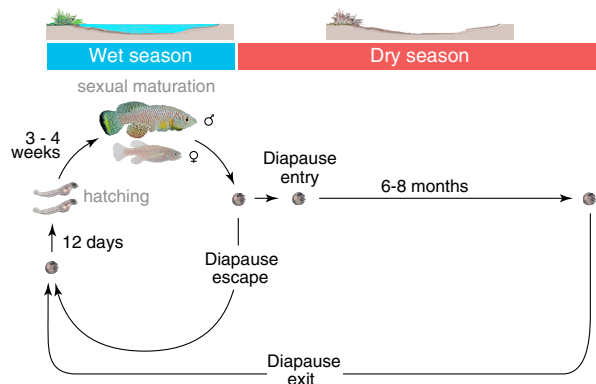
**A** Geographical location and climate of the turquoise killifish habitat



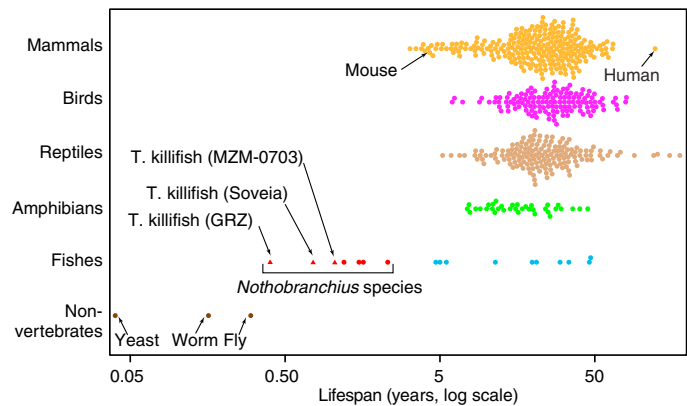
**B** Turquoise killifish morphologies



**C** Turquoise killifish life cycle



**D** Lifespans of different organisms



**Figure 1. The African Turquoise Killifish Is a Naturally Short-Lived Vertebrate with Multiple Strains**

(A) Geographical distribution of turquoise killifish in the wild (orange dots). The climate of the area measured by the Koeppen-Geiger index (concentric contours). (B) Examples of different morphological types in turquoise killifish males: yellow-tailed and red-tailed. (C) Life cycle of the turquoise killifish. The turquoise killifish achieves sexual maturation and reproduces during the wet season. Diapausing embryos can survive through the dry season, when the ponds are desiccated. Diapause can be skipped in the laboratory, resulting in a short generation time. (D) *Nothobranchius* species are among the shortest-lived vertebrates. Lifespan data for turquoise killifish strains are from our experimental data (age of the 10<sup>th</sup> percentile survivors) and are depicted by triangles. Lifespan of the strains can vary depending on housing conditions. Maximum lifespan data for the other organisms are from the AnAge database.

**RESULTS**

**The African Turquoise Killifish: a Vertebrate with a Naturally Compressed Lifespan**

The African turquoise killifish provides a unique system to test the evolution and genetics of longevity (Cellerino et al., 2015). This fish species comprises distinct populations that populate ephemeral ponds in arid regions of Zimbabwe and Mozambique (Figures 1A and 1B). The ponds are only present for 4–6 months during the wet season, and the turquoise killifish has evolved a

state of embryonic diapause—“suspended animation”—to survive through the dry season (Figure 1C). In the laboratory, diapause can be skipped, and this fish has a captive lifespan of 4–6 months (Figure 1C). Thus far, the turquoise killifish is the shortest-lived vertebrate that can be bred in captivity (Figure 1D). Interestingly, strains derived from the various populations of turquoise killifish from different regions of Zimbabwe and Mozambique can exhibit different experimental lifespans under similar conditions (Terzibasi et al., 2008) (Figure 1D). The extreme diversity in lifespan between the turquoise killifish (4–6 months)

and other species, as well as the presence of strains with different captive lifespans, offers an unprecedented paradigm to explore the evolution of lifespan.

### De Novo Assembly and Annotation of the Turquoise Killifish Genome

To gain insight into the evolution and genetic architecture of lifespan in the turquoise killifish, we de novo sequenced and assembled its genome. The GRZ strain, which originates from Zimbabwe, was selected to build the reference genome because it is inbred and has a low percentage of heterozygosity (Kirschner et al., 2012; Valenzano et al., 2009). By sequencing 10 paired-end or mate-pair Illumina libraries with a range of insert sizes, we obtained a 1.02 Gb genome assembly with a contig N50 of 9.3 kb and a scaffold N50 of 118 kb (Figure 2A). Using paired-end RNA-seq libraries and our high-density linkage map (see Figure 6C), we improved the contiguity of the genome resulting in a scaffold N50 of 247 kb (Figure 2A). The computational genome size estimate ranges from 1.3 Gb to 2.2 Gb (Figure 2A). The assembly statistics for our turquoise killifish genome are in the range of Illumina-based genome assemblies, although the assembly is still fragmented likely due to the high number of repeats (over 45% (Reichwald et al., 2009)) (Figure S1A). We next assessed the completeness and quality of the turquoise killifish genome assembly. Computational assessment indicates that the assembly contains 96.8% of core eukaryotic genes (Figure 2A). Importantly, assembled transcripts, paired-end RNA-seq reads, and contigs obtained by Sanger shotgun sequencing (Reichwald et al., 2009) properly mapped to the genome (Figures 2A, S1B, S1C and S1D). Thus, the reference turquoise killifish genome, while fragmented, contains most genes and is well assembled.

We next annotated the reference turquoise killifish genome. Using de novo gene prediction and sequence homology to 19 animal species, we identified 28,494 protein-coding gene models (Figure 2B and S1G). There was maximal enrichment of RNA-seq reads over protein-coding gene bodies (Figure S1E) and maximal enrichment of trimethylated lysine 4 on histone H3 (H3K4me3), a chromatin mark associated with promoters, at predicted transcription start sites (Figure S1F). The majority of the predicted protein-coding genes in the turquoise killifish have an ortholog in other vertebrate genomes (Table S2A). There was a good paralog correspondence and large blocks of syntenic genes between turquoise killifish and medaka, a fish with a high-quality Sanger-sequenced genome (Figures S2A–S2D). Together, these results indicate that protein-coding genes, including paralogs, are well annotated in the turquoise killifish assembly, although some split genes may be misannotated as paralogs.

We also identified 5,859 high-confidence long non-coding (lnc) RNA genes and predicted other classes of non-coding RNA genes, including miRNAs (Figures 2B, S2E and S2F, and Table S1A). Finally, we identified several families of transposable elements (Figure 2C and Table S1C), at least some of which were actively transcribed (Tables S1D and S1E). To facilitate the use of the genome by the community, we have deployed a fully functional genome browser website (<http://africanurquoisekillifishbrowser.org>) (Figure 2D).

### Evolutionary Analysis of the African Turquoise Killifish Genome

To gain insight into unique features of the African turquoise killifish, we analyzed its genome from an evolutionary perspective. The reconstituted phylogeny based on high-confidence one-to-one orthologs in 20 species (Tables S2B and S2C, Figure 3A) is consistent with fossil records and previous estimates (Reichwald et al., 2009), confirming the quality of the turquoise killifish genome assembly and annotation. To understand the evolution of features unique to this short-lived fish, we identified sites under positive selection in protein-coding genes in the turquoise killifish compared to seven longer-lived fish species (Figure 3A). We used a maximum likelihood approach to determine positive selection at the codon level (Yang, 2007) (Figures 3B and S3A). After multiple hypothesis correction, we identified 497 genes with at least one site under positive selection in the turquoise killifish genome (Table S3A). Of these, 249 genes had one or more sites with very high probability of being under positive selection, and represent the highest-confidence list of genes under positive selection (Table S3B). In the remainder of the manuscript, we used this highest-confidence list, unless otherwise indicated.

A gene ontology enrichment analysis revealed that the genes under positive selection in the turquoise killifish are enriched for components of signal transduction pathways, metabolism, development, proteostasis, and immunity (Figure 3C and Table S3C). Interestingly, several of these processes are relevant to the modulation of aging (Kennedy et al., 2014; López-Otín et al., 2013), although these categories are broad and could underlie other biological processes. Positively selected genes include the helicase *RTEL1*, which is involved in telomere elongation and has been associated with dyskeratosis congenita, a human syndrome with premature aging characteristics (Ballew et al., 2013). Another intriguing positively selected gene is *PSMD11*, which is involved in proteostasis and whose ortholog is important for lifespan in *C. elegans* (Vilchez et al., 2012). Using in silico prediction tools originally designed for human disease variants, we found that 340 of the 2,009 positively selected residues (corresponding to 121 proteins) had predicted functional effects on the protein by two independent methods (Figures 3D and S3B, and Table S3D). The genes under positive selection in the turquoise killifish tend to display gene expression changes throughout life ( $p = 0.045$  in Fisher's exact test at FDR 5%; Figures S3D and S3E and Table S3E). Thus, a subset of the positively selected genes in the turquoise killifish could be important for the evolution of a naturally short life trajectory. However, these genes could also underlie the evolution of other traits in this fish (e.g., diapause, resistance to high temperature, morphology, or sensitivity to microbial communities or pathogens).

### Aging and Longevity Genes Are under Positive Selection in the Turquoise Killifish

To focus on genes with potential roles in the turquoise killifish compressed lifespan trajectory, we intersected the set of 497 genes under positive selection in the turquoise killifish with known aging-related genes from GenAge (mouse and human combined) and LongevityMap (Table S4A) (Budovsky et al., 2013; de Magalhães et al., 2009). While the overlap was not statistically significant, 22 previously known aging-related

**A** Assembly statistics (NotFur1)

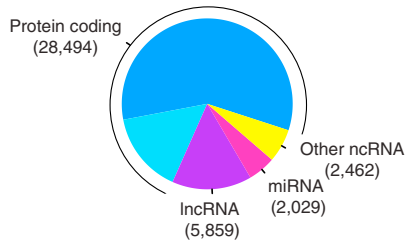
Metric	Contigs	Scaffolds	Scaffolds	Scaffolds
			(with RNA-seq)	(with RNA-seq and linkage map)
Number	230,961	46,729	45,505	42,796
Total bases (bp)	944,909,766	1,023,205,147 <sup>a</sup>	1,079,706,050 <sup>a</sup>	1,079,976,950 <sup>a</sup>
Maximum (kb)	82.7	667	1,081	27,998
Average (kb)	4.1	21.9	23.7	25.2
N50 length (kb) <sup>b</sup>	9.3	118	142	247
% N	0.04%	6.59%	7.66%	7.68%
Mean number of scaffolded contigs		4.9	5.3	5.6
% of total genome size estimate	43.0-72.7%	46.5-78.7%	49.1-83.1%	49.1-83.1%
			Total	Quality
CEGs (248 genes)			96.8%	89.9% (complete)
% assembled transcript mapping	Petzold et al, 2013		98.5%	75.2% (high-quality)
	This study		98.5%	77.1% (high-quality)
% RNA-seq mapping (average of 4 tissues)			81.5%	66.0% (properly paired)

<sup>a</sup> Length estimate includes gaps

<sup>b</sup> N50 length: such as 50% of assembly is of equal or longer length

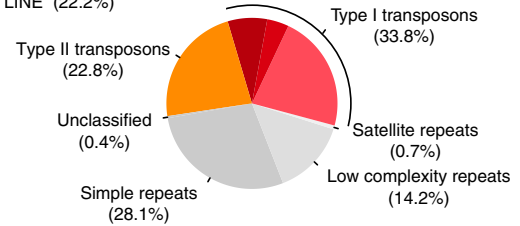
**B** Annotated coding and non-coding genes

- High-confidence protein coding (22,521)
- Other protein coding (5,973)

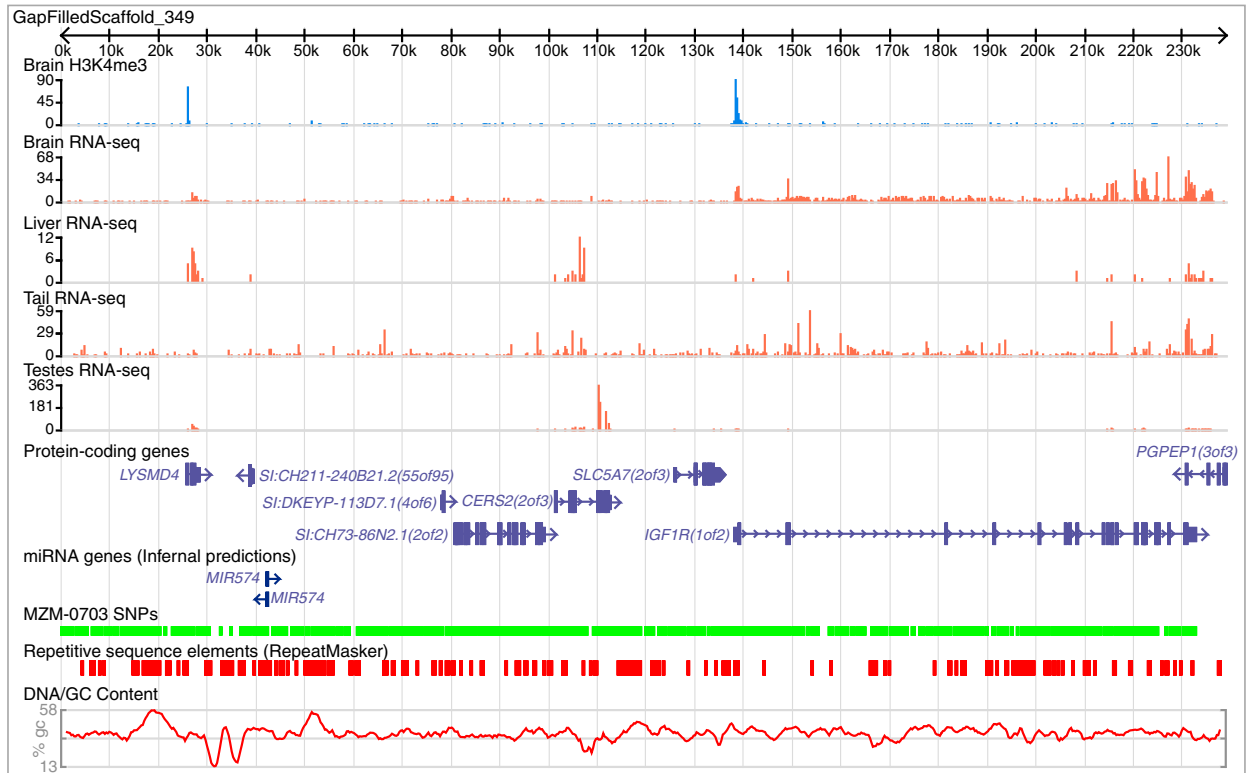


**C** Repetitive element composition

- SINE (7.5%)
- LTR (4.1%)
- LINE (22.2%)



**D** African turquoise killifish genome browser (Example)



(legend on next page)

genes were under positive selection in the turquoise killifish (Tables S4B and S4C). These genes include the insulin receptor A (*INSRA*, Figure 4A) and *IGF1* receptor (*IGF1R(1of2)*, also known as *IGF1RA*, Figure 4B). Reduced function in the insulin/*IGF1* receptors extends lifespan from worms to mice (Blüher et al., 2003; Holzenberger et al., 2003; Kenyon et al., 1993), and variants in *IGF1* receptor are associated with exceptional longevity in humans (Suh et al., 2008) (Figure 4C). Aging genes under positive selection also include *LMNA3* (*LMNA(3of3)*, Figure 4E), which encodes a nuclear Lamin-A/C. Mutations in human *LMNA* can give rise to Hutchinson-Gilford Progeria syndrome (Eriksson et al., 2003), although other variants are associated with exceptional longevity in humans (Conneely et al., 2012) (Figure 4C). Finally, another gene under positive selection is the DNA repair gene *XRCC5* (also known as *KU80*, Figure S4A). *XRCC5* deficiency in mice leads to premature aging (Vogel et al., 1999), though specific variants are also associated with human longevity (Figure 4C) (Soerensen et al., 2012).

We next mapped the variants under positive selection on the protein structures or domains of several aging-related proteins, highlighting variants with predicted functional effects (Tables S4E–S4G). The majority of residues with functional effect under positive selection in the insulin/*IGF1* receptors (*INSRA* and *IGF1R1A*) are located in the extracellular domains implicated in ligand binding (Figures 4A, 4B, and 4D). In *LMNA3*, the functional residues under positive selection are located in the filamentous domain involved in the interaction with chromatin (Figure 4E). Finally, in *XRCC5*, they are located in the DNA binding domain (Figure S4A). Thus, positively selected residues with predicted functional effects are located in important domains in aging-related proteins (see also Figures S4B–S4D) and might have consequences on the role of these proteins in lifespan regulation.

The ability to undergo diapause—a developmental stage associated to drought-resistance—is likely under intense selective pressure given the transient nature of the ponds in which these fish live. We asked if the 497 genes under positive selection in the turquoise killifish overlapped with genes involved in diapause in other species, such as the insulin/*IGF1*-FOXO and the TGF $\beta$  pathways (Gottlieb and Ruvkun, 1994; Patterson et al., 1997). Only the genes in the insulin/*IGF1*-FOXO pathway (*INSRA*, *IGF1RA*, *FOXO1B(2of2)*), which are also involved in lifespan, were positively selected in the turquoise killifish (Figure S4E and Table S4D). Thus, positively selected genes in the insulin/*IGF1* pathway might play a role both in diapause and compressed life cycle in the turquoise killifish, perhaps depending on external conditions.

### Comparison of Aging Genes in the Turquoise Killifish and Other Species or Groups with Exceptional Longevity

Intriguingly, the genes under positive selection in the short-lived turquoise killifish are also under positive selection or uniquely changed in species with exceptional longevity (naked mole rat, Brandt's bat and bowhead whale). Indeed, *IGF1R(1of2)* was found to be uniquely changed in the long-lived Brandt's bat (Seim et al., 2013) (Figures 4C and 4D) and under positive selection in the short-lived marmoset (The Marmoset Genome Sequencing and Analysis Consortium, 2014). More generally, 11 other genes are under positive selection or uniquely changed in both turquoise killifish and “extreme longevity” species or groups of individuals (Figure 4C). These genes include a carboxyl ester lipase *CEL(7of7)*, which is involved in cholesterol metabolism and diabetes in humans (Raeder et al., 2006), and the complement system component *C3(3of3)*, which is implicated in age-related degenerative pathologies and Alzheimer's disease (Proitsi et al., 2012). These observations raise the intriguing possibility that the same genes can be under positive selection in both extremely short-lived and long-lived species.

Are the residues in proteins that are positively selected in short-lived and long-lived species similar or different? We mapped residues from the short-lived turquoise killifish, long-lived Brandt's bat, and humans onto the well-studied *IGF1* receptor (Figure 4D) and *LMNA* (Figure 4E). Many of the residues under positive selection in the turquoise killifish and the Brandt's bat are in proximity on the *IGF1* receptor sequence, but differ (Figure 4D). Furthermore, the residues under positive selection in the turquoise killifish and those associated with longevity in human are both located in the predicted *IGF1R* ligand-binding domains but are different (Figure 4D). These residues also differ from *C. elegans* longevity mutations in the insulin/*IGF1* receptor (*DAF-2*, Figure S4E). Similarly, the *LMNA3* residues under positive selection in the turquoise killifish also differ from variants in human centenarians or Hutchinson Gilford Progeria syndrome (Figure 4E). More generally, for the same protein, the residues under selection in the turquoise killifish differ from those uniquely changed in the long-lived bowhead whale (Table S4G, and mapping for *CEL(7of7)* in Figure S4F). Thus, proteins that act as central nodes could have been selected to underlie both compressed and extended life trajectories, depending on the residues. Alternatively, the same proteins could have been selected because both the turquoise killifish and long-lived species exhibit resistance to stress—during diapause for the turquoise killifish and throughout life for long-lived species.

### Figure 2. De Novo Sequencing and Assembly of the Reference Genome of the African Turquoise Killifish

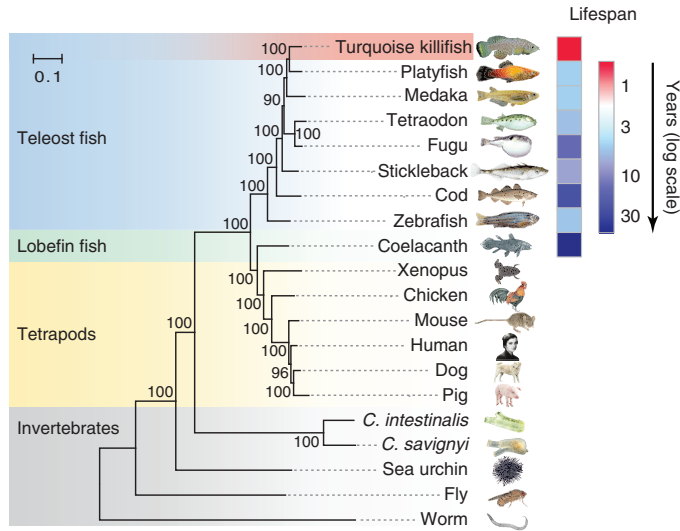
(A) Assembly statistics for draft genome (NotFur1) for the reference turquoise killifish strain (GRZ). Scaffolds that are not captured by the linkage map remain unplaced. CEGs: core eukaryotic genes. See also Figures 6C, S1B, S1C.

(B) Number of annotated coding and non-coding genes in the reference turquoise killifish genome. High-confidence protein coding genes are genes with homologs in at least 10 species. lncRNA: long non-coding RNA; ncRNA: non-coding RNA; miRNA: microRNA. See also Figures S2E and S2F and Tables S1A and S1B.

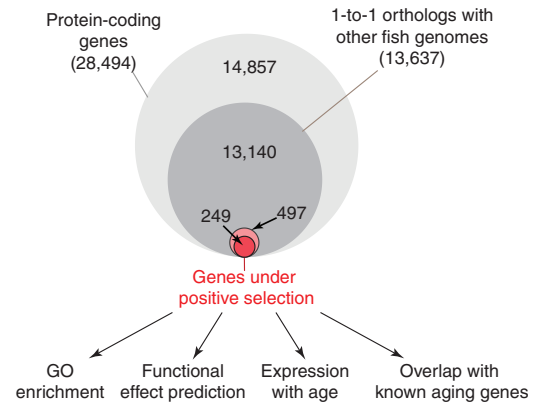
(C) Repetitive element composition in the reference turquoise killifish genome. LINE: long interspersed nuclear element; SINE: short interspersed nuclear element; LTR: long terminal repeat. Type I transposons are RNA-mediated. Type II transposons are DNA-mediated. See also Tables S1C–S1E.

(D) Example of a genomic region containing insulin-like growth factor 1 receptor (*IGF1RA*).

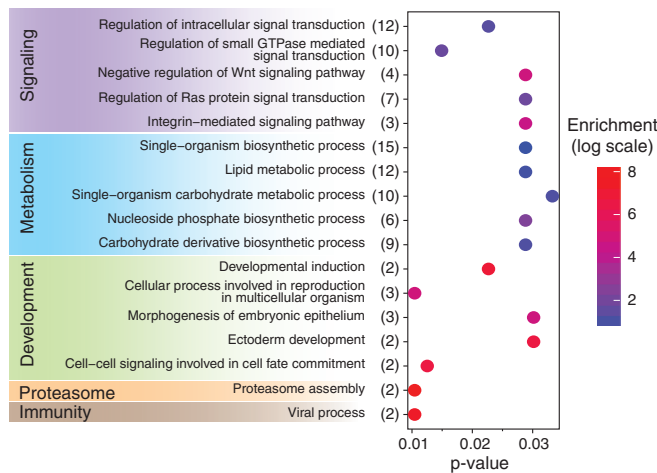
**A** Phylogenetic tree and lifespan



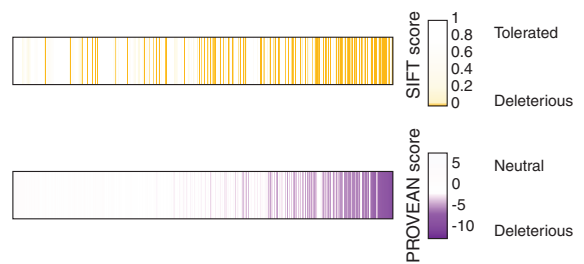
**B** Genes under positive selection in the turquoise killifish



**C** Selected GO term enrichment for the genes under positive selection



**D** Functional effect prediction for the sites under positive selection



**Figure 3. Evolutionary Analysis of the Turquoise Killifish Genome**

(A) Phylogenetic tree of 20 animal species, including the turquoise killifish, based on 619 one-to-one orthologs (Table S2C). Number on nodes: level of confidence (% bootstrap support). Scale bar: evolutionary distance (substitution per site). Maximum lifespan data are from our experimental data (turquoise killifish) or from the AnAge database (other fish species), and represented as a heat map.

(B) Proportion and analysis of the genes under positive selection in the turquoise killifish compared to 7 other fish species after multiple hypothesis correction (FDR < 5%). See also Figure S3A.

(C) Selected GO term enrichment for the genes under positive selection in the turquoise killifish. The number of genes associated with each category is indicated in brackets after the term description, and enrichment values are indicated in colored scale. See also Table S3C.

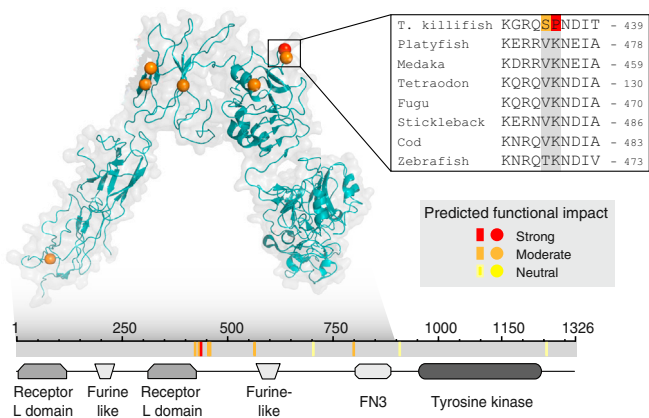
(D) Predicted functional effect on the protein of residues under positive selection in the turquoise killifish have based on SIFT (top row) and PROVEAN (bottom row). Residues are ordered from left to right based on the rank-product of the SIFT and PROVEAN scores. Only sites scored by both methods are displayed. See also Figure S3B and Tables S3D, and S4G.

**Sequencing Individuals from Additional Turquoise Killifish Strains Reveals Variants in Aging-Related Genes**

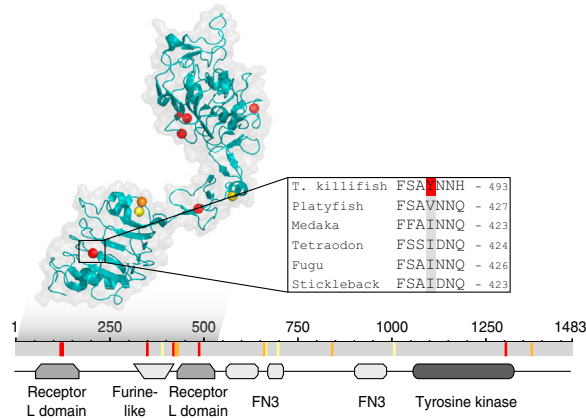
Within the turquoise killifish species, there exist several strains with reported differences in lifespan in specific laboratory environments (Kirschner et al., 2012; Terzibasi et al., 2008) (Figures 5A, S5A, and 6B), and these differences could be leveraged

to understand the genetic architecture of lifespan. To assess the genetic differences among turquoise killifish strains, we sequenced at lower coverage individuals from two additional strains that were captured in Mozambique in 2004 and 2007 (MZM-0403 and MZM-0703, respectively) and from a control GRZ individual (Figure 5A). This analysis uncovered over three million single nucleotide polymorphisms (SNPs) between

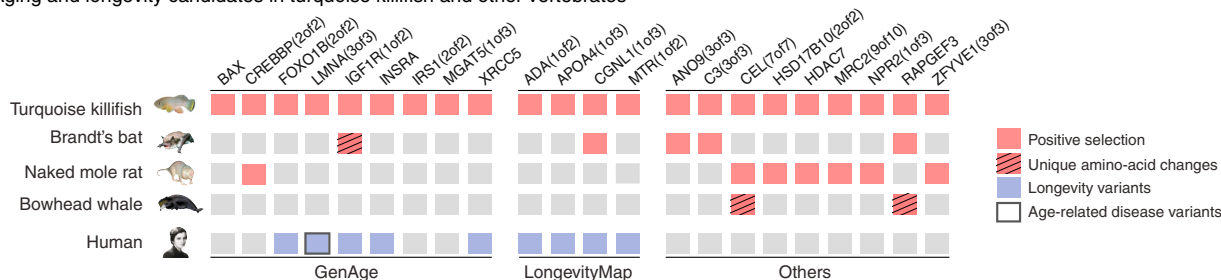
**A** Residues under positive selection in INSRA



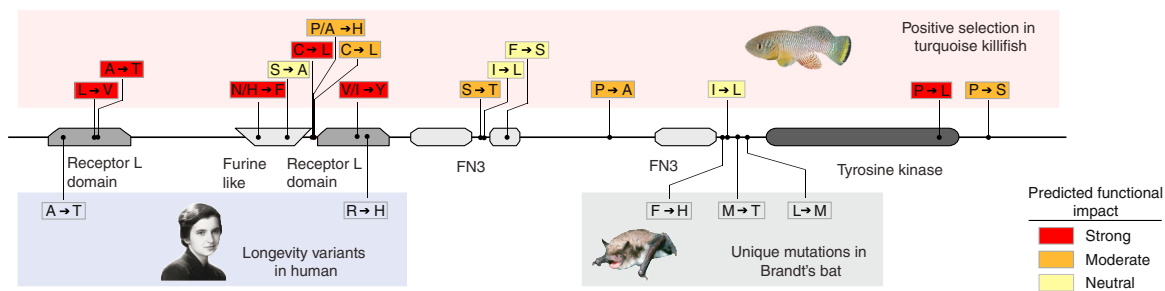
**B** Residues under positive selection in IGF1R(1of2)



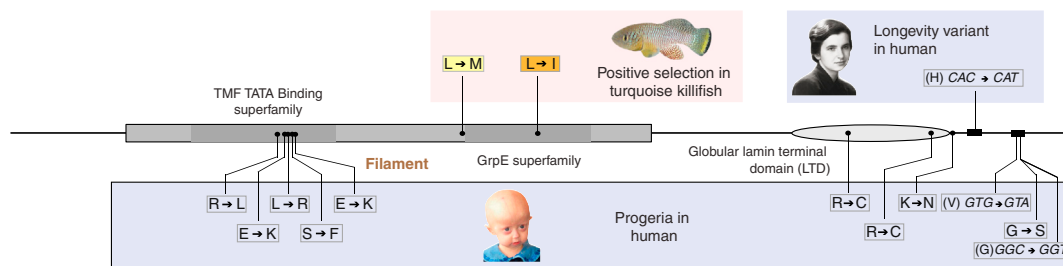
**C** Aging and longevity candidates in turquoise killifish and other vertebrates



**D** Residues and variants in IGF1R(1of2) in turquoise killifish and other organisms



**E** Residues and variants in LMNA(3of3) in turquoise killifish and human



**Figure 4. Aging and Longevity Genes under Positive Selection in the Turquoise Killifish and with Variants in Long-Lived Species or in Humans** (A and B) Location of residues under positive selection and with putative functional consequences in insulin receptor A (INSRA) (A), IGF1R(1of2) (B) in the turquoise killifish. Top: crystal structure of human orthologs. Color represents the strength of functional impact. Grey shadow: region of the protein with available crystal structure. Insert: alignment of an example residue with strong functional effect in the turquoise killifish and other fish. Bottom: schematic of the residues mapped on the turquoise killifish protein sequence (gray). Colored bars: residues under positive selection with different functional impacts. The conserved protein domains and functional sites are also indicated. FN3: Fibronectin type-III repeats.

(legend continued on next page)



MZM-0403 or MZM-0703 and the reference GRZ genome (Figures 5A, 5B, and S5B). As expected, there were fewer SNPs between the GRZ individual and the reference GRZ genome (Figures 5A and S5B).

To identify potential genetic differences associated with phenotypic diversity (e.g., lifespan, color, etc.) among strains, we focused on SNPs that are shared between the longer-lived red-tailed MZM strains, but not by the shorter-lived yellow-tailed GRZ reference strain (Figure 5C). We identified 22,389 non-synonymous SNPs in 10,638 genes, and 139 of these genes overlapped with known aging-related genes that encompassed all 9 “hallmarks of aging” (López-Otín et al., 2013) (Figure 5D and Table S5A). A number of these SNPs are predicted to have functional impact on the protein (Figure 5D and Table S5B). Genes with functional variants between turquoise killifish strains include insulin/IGF signaling pathway genes (*GHR(1of4)* and *FOXO4* transcription factor) and inter-cellular communication genes, such as progranulin (*GRN*, also known as *PGRN*) (Figure 5D). While this analysis does not identify the specific variants that explain lifespan differences between strains, it provides a resource for the study of genetic variation in this species.

### Identification of a Genomic Region Significantly Associated with Differences in Lifespan between Turquoise Killifish Strains

To identify genomic regions that are important for lifespan differences between turquoise killifish strains, we performed a genetic linkage analysis in a cross between the shorter-lived and longer-lived strains (Figure 6A). We used the shorter-lived reference strain captured in Zimbabwe in 1968 (GRZ) and longer-lived strains derived from expeditions in Mozambique in 2007 and 2006 (MZM-0703 and Soveia) (Figure S5A). These strains exhibit differences in lifespan in the laboratory conditions used in this study (Figure 6B and Table S6A), although the number of animals was low for the Soveia strain and differences between strains could have been accentuated by the laboratory conditions used. We crossed a female from the GRZ strain to a male from the MZM-0703 strain (“cross GxM,” Figure 6A), as well as a female from the Soveia strain with a male from the GRZ strain (“cross SxG,” Table S6A). F1 fish were interbred and the lifespan of as many F2 fish as possible (430 for cross GxM, and 130 for cross SxG) was scored. F1 and F2 fish live longer than the short-lived GRZ parental strain in both crosses (Figure 6B, and Table S6A), suggesting that the long lifespan trait is dominant over the short lifespan trait. Similar lifespan results were obtained with cross GxM (in which the female is from the shorter-

lived strain) and cross SxG (in which the male is from the shorter-lived strains), suggesting that there is no major maternal contribution to lifespan. Surprisingly, while males and females had a similar lifespan in the parental strains, males were significantly longer-lived than females in the F1 and F2 generation of cross GxM (Figures S6A–S6D), implying a possible interplay between lifespan and sex.

To identify the genomic regions significantly associated with lifespan differences, we conducted a genome-wide linkage analysis using cross GxM because of the higher number of F2 individuals in this cross. Using restriction-site associated DNA sequencing (RAD-seq) to genotype P0 grandparents (2), F1 fish (16), and a large subset of F2 fish (207), we identified 8,399 high-confidence genomic markers that displayed SNPs between the grandparents. These markers allowed us to build a high-resolution linkage map with 19 linkage groups (LGs), providing an anchor to our reference genome (Figure 6C). Importantly, the number of LGs is the same as the haploid number of chromosomes in this species (19) (Reichwald et al., 2009).

We then mapped individual lifespan as a phenotype using a method based on Random Forest. The most significant genetic region associated with lifespan is on linkage group 3 (LG-3) (Figure 6C). This region was found to be associated with lifespan differences by two independent statistical methods, Random Forest and log-rank test, and it reached genome-wide significance in the Random Forest QTL analysis (FDR < 5%). Importantly, the lifespan QTL on LG-3 was “non-transgressive,” i.e., the individuals with the longer-lived grandparents genotype at this locus (Figure 6D, red) were the longest-lived. The fact that individuals with a heterozygous genotype at this locus do not live longer than individuals with either grandparent genotype (Figure 6D) argues against hybrid vigor, consistent with the absence of global enrichment in heterozygous fish in longest-lived F2 individuals (Figure S6E). Individuals with one or both alleles from the longer-lived grandparent at the QTL peak marker exhibited a ~30% increase in lifespan ( $p = 1.6 \times 10^{-3}$  and  $p = 1.8 \times 10^{-3}$ , respectively, in log-rank tests) compared to individuals with both alleles from the shorter-lived grandparent (Figures 6E and S6F). Thus, this genetic region on LG-3 has a significant contribution to the lifespan difference in the F2 generation of this cross.

### The Genomic Region Associated with Lifespan Differences between Turquoise Killifish Strains Is on the Sex Chromosome





Interestingly, the genetic region associated with differences in lifespan among turquoise killifish strains is close to the

(C) Aging and longevity candidates under positive selection in the short-lived turquoise killifish and their variation in long-lived animal species and in humans. Left: aging-related genes from the GenAge database (human and mouse combined, Table S4A). Middle: genes identified in association studies in humans from the LongevityMap database (Table S4A). Right: other genes that are also under positive selection or uniquely changed in other species with extreme longevity phenotypes (naked mole rat, Brandt’s bat, bowhead whale).

(D) Location and variants of residues under positive selection in the turquoise killifish for IGF1R(1of2), and their location and variants in long-lived species and in human centenarians. Top: turquoise killifish variants, with the changed amino acid on the right. Color represents the strength of functional impact. Bottom: variants associated with centenarians in humans or residues with unique amino acid changes in long-lived Brandt’s bat mapped on the turquoise killifish sequence. The variants correspond to the amino acids on the right.

(E) Location and variants of residues under positive selection in the turquoise killifish for LMNA(3of3), and their location and variants in progeria and human centenarians. Top left: turquoise killifish variants, with the changed amino acid on the right. Top right: variants associated with centenarians in humans mapped onto the turquoise killifish sequence. Bottom: Variants in Hutchinson-Gilford Progeria Syndrome mapped onto the turquoise killifish sequence. These variants are the amino acids on the right. For the turquoise killifish, color represents the strength of functional impact.

**A** Resequencing of turquoise killifish individuals from various strains

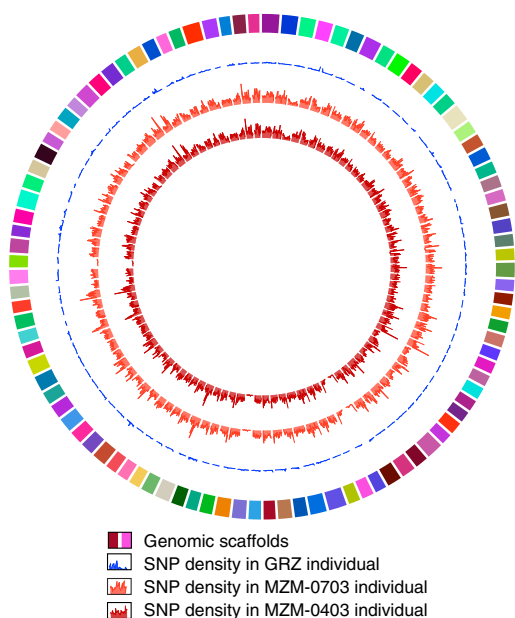
Sequencing	Strain		Captive lifespan <sup>c</sup>	Year / country of capture	# of individuals	Coverage	SNPs
Reference genome <sup>a</sup>	GRZ		Shorter-lived	1968 / Zimbabwe	9	80X	-
Genome resequencing <sup>b</sup>	GRZ		Shorter-lived	1968 / Zimbabwe	1	19X	284,386
	MZM-0703		Longer-lived	2007 / Mozambique	1	54X	4,671,729
	MZM-0403		Longer-lived	2004 / Mozambique	1	6X	2,754,094

<sup>a</sup> Data used for reference assembly

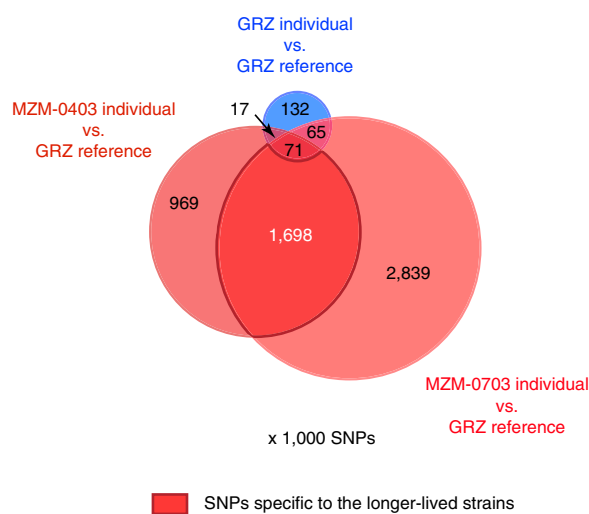
<sup>b</sup> Data from individual fish not used for the reference assembly

<sup>c</sup> This study; Terzibasi et al. 2008; Kirschner et al. 2012.



















**B** SNP density over the 100 longest scaffolds



**C** Unique and shared SNPs between resequenced individuals



**D** Example genes with non-synonymous SNPs in individuals from longer-lived strains

Hallmark of aging	Genes with SNPs specific to the longer-lived strains <sup>a</sup>	Example genes that have variants with predicted functional effect
 Altered intercellular communication	45 	<b>GRN, TNFB(3of3), PDGFRA</b>
 Genomic instability	39 	<b>BRCA1, TP53BP1, ERCC6</b>
 Deregulated nutrient sensing	22 	<b>GHR(1of4), IRS4, FOXO4</b>
 Cellular senescence	13 	<b>MYC(2of2), EGR1(1of5)</b>
 Epigenetic alterations	8 	<b>MED1(1of2), NCOR1(1of2)</b>
 Mitochondrial dysfunction	8 	<b>POLG, GSR</b>
 Stem cell exhaustion	8 	<b>MGAT5(1of3)</b>
 Telomere attrition	8 	<b>TERT</b>
 Loss of proteostasis	7 	<b>HSF1, HSPA9</b>

<sup>a</sup> One gene may belong to more than one hallmark

(legend on next page)

sex-determination region (Figure 7A). A previous study identified a lifespan QTL in the linkage group linked to sex but at the time this QTL could not be resolved from the sex-determination region (Kirschner et al., 2012). Using our high-density linkage map, we found that the lifespan QTL and the sex-determining region are on the same linkage group, but 36–38 cM apart (Figure 7B, top), although distance estimates in a sex-linked region cannot be fully accurate due to suppressed recombination near the sex-determining region. Individuals with both alleles from the long-lived grandparent at the lifespan QTL peak marker exhibited significant increase in lifespan (Figure 7B, bottom). The lifespan QTL also remained significant even after regressing for sex (Figures 6D and 7C). Importantly, the marker with the highest significance for lifespan was associated with longevity in both males and females (Figure S6F), ruling out the possibility that significance is due to a potential bias coming from the males. Together, these results strongly support that the lifespan QTL and the sex-determining region are linked, but distinct.

### Genes in the Region Associated with Lifespan Differences

We next used the turquoise killifish genome to identify the specific genes in the region associated with lifespan differences between strains. The lifespan QTL region captures genomic scaffolds encompassing 31 protein-coding genes, 6 long non-coding RNA genes, and 2 small nucleolar RNA genes (snoRNA) (Figure 7C and Tables S7A–S7E). Among the 31 protein-coding genes, 7 had already been linked to the regulation of aging or lifespan in humans or model organisms (Figure 7C and Table S7A). These include the gene encoding progranulin (*GRN*), which has been implicated in neurodegenerative diseases (Wang et al., 2010) and lifespan regulation in mice (Ahmed et al., 2010) (Figure 7C). Another interesting candidate is *NUDT1*, which is involved in the hydrolysis of 8-oxo-dGTP, a deleterious nucleoside that increases with aging in mitochondrial DNA (Souza-Pinto et al., 1999) (Figure 7C). *NUDT1* overexpression extends lifespan in mice (De Luca et al., 2013). Yet another intriguing candidate is *GSTT1A*, which encodes a glutathione S-transferase, a class of redox homeostasis enzymes that regulate lifespan in worms and mice (Ayyadevara et al., 2007; Pesch et al., 2004) (Figure 7C). Finally, this region comprises genes encoding two transcription factors (*STAT3*, *STAT5.1(2of6)*) that have been implicated in regulating “inflammaging” (De-Fraja et al., 2000) (Figure 7C).

Of the 31 genes underlying the lifespan QTL, 15 harbored non-synonymous coding differences between the P0 individ-

uals (Figures 7C and S7A, and Table S7A). For 6 of these 15 genes (*ATXN7L1*, *GRN*, *HIPK2(11of26)*, *IFI35*, *TTYH3A* and *ZNF800A*), the coding differences occur in otherwise well-conserved residues, are also observed in the resequenced MZM-0403 individual, and the variant in the longer-lived P0 corresponds to the consensus from other species (Figures S7B–S7D; Table S7A). The presence of several of these variants in cross GxM P0 founders was confirmed by Sanger sequencing (Table S7E). Interestingly, one of the variants in the turquoise killifish *GRN* (W449 in the shorter-lived strain and C449 in the longer-lived strain) is within a motif that plays a key role in protein folding and that is mutated in frontotemporal dementia (FTD) (Wang et al., 2010) (Figure 7D). Consistently, this turquoise killifish variant is predicted to have functional consequences (Figure 7D and Tables S7G and S7H). This variant is also found in wild fish captured during our expedition to Mozambique and Zimbabwe in 2010 (Figure S7E and Table S7F), indicating that it is not a spurious mutation that arose in the laboratory or from the bottleneck of a rare allele. Thus, coding variations in *GRN* or other candidates may underlie differences in lifespan between strains of this species, although we cannot exclude that non-coding variants are responsible for these phenotypic difference (Figure S7A).

We wondered if the number of aging genes within the lifespan QTL was higher than expected by chance. Statistical analysis showed that the lifespan QTL region is significantly enriched for known aging-related genes (GenAge, mouse and human combined, Table S4A) compared to the rest of LG-3 ( $p = 6.4 \times 10^{-4}$ , Fisher’s exact test) or to all linkage groups ( $p = 2.1 \times 10^{-4}$ , Fisher’s exact test) (Figure 7E). This significant enrichment for known aging and longevity genes in the region underlying the lifespan QTL suggests that a haplotype block containing a cluster of genes, rather than a single gene, might be involved in the observed lifespan difference. Sex-determining regions are usually regions of suppressed recombination, and indeed the recombination frequencies are the lowest in the region associated with sex determination (Figures S7F and S7G). Intriguingly, the recombination frequencies are lower than expected throughout the entire LG-3, including the lifespan QTL region (Figures 7E, S7F, and S7G), compared to the rest of the genome. This haplotype block might have formed because of the suppressed recombination in this region, due to its proximity with the sex-determining region (Figure 7E). While this may be fortuitous, the presence of the lifespan-determining region on the sex chromosome might have evolved to couple strategies of fast reproduction with genes involved in overall fitness.

### Figure 5. Genetic Variation in Individuals from Different Strains of the Turquoise Killifish

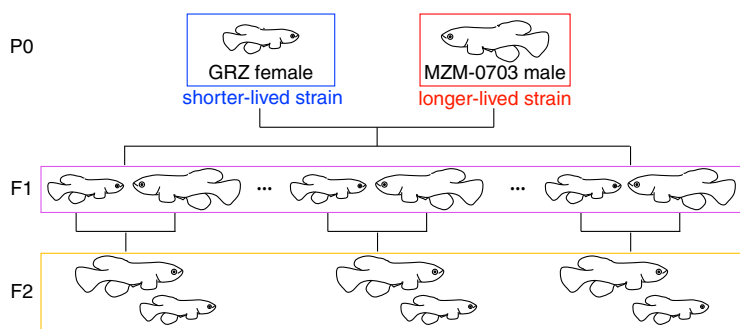
(A) Resequencing of individuals from different turquoise killifish strains (GRZ, MZM-0703, MZM-0403) with different reported lifespans in specific laboratory environments. SNPs: Single Nucleotide Polymorphisms.

(B) Circos plot of the 100 longest scaffolds showing SNP density in resequenced individuals from different turquoise killifish strains (GRZ, MZM-0703, MZM-0403) versus the reference GRZ assembly.

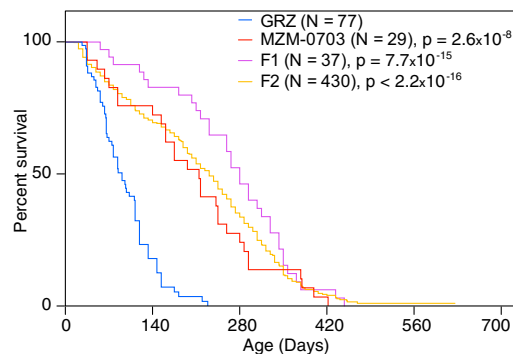
(C) Unique and shared SNPs between resequenced individuals from reported shorter-lived (GRZ) or longer-lived strains (MZM-0703 and MZM-0403) versus the reference GRZ assembly. Values in the Venn diagram should be multiplied by 1000 and are rounded for concision.

(D) Non-synonymous SNPs specific to individuals from the longer-lived strains (MZM-0403 and MZM-0703) in genes encompassing the hallmarks of aging. Aging-related genes were obtained from the GenAge database (human and mouse combined, Table S4A). All presented aging-related genes had at least one variant with predicted functional effect by both SIFT and PROVEAN (bolded genes) or by SIFT or PROVEAN (non bolded genes) (see also Table S5B).

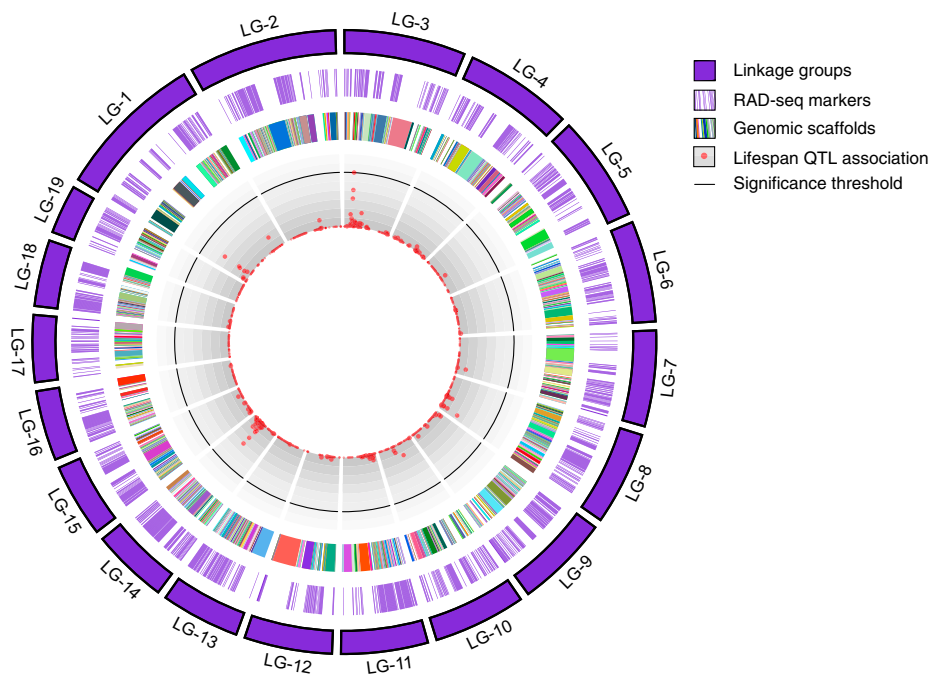
**A** Cross GxM scheme



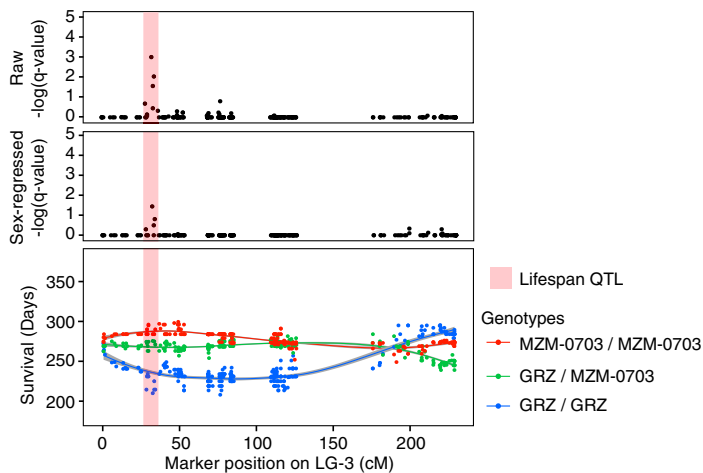
**B** Lifespan



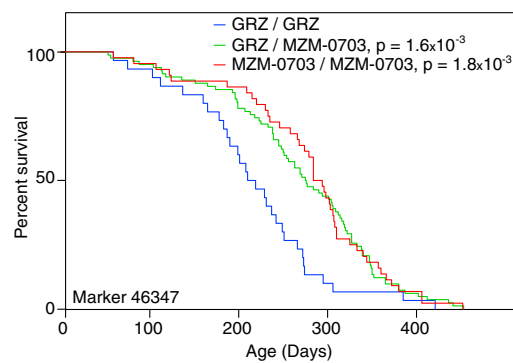
**C** Mapping of lifespan QTL



**D** Lifespan stratified by genotype



**E** Lifespan stratified by genotype at QTL peak marker



(legend on next page)

## DISCUSSION

### The Genome of One of the Shortest-Lived Vertebrate Species: A Resource for Comparative, Experimental, and Evolutionary Genomics

The African turquoise killifish reference genome and transcriptome, high-density genetic linkage map, and a comprehensive genome browser represent great resources for the exploration of the genetic principles and the evolution underlying several unique traits in this species, such as a compressed life cycle and embryonic diapause. These resources, combined with the rapid experimental capacity of this fish, provide an unprecedented setting for evolution and genetic studies in vertebrates.

### Evolution of the Lifespan Differences between Species

The forces that shape differences in lifespan in nature are still largely unknown. Extrinsic mortality imposes a strong constraint for rapid reproduction, which may ultimately result in the evolution of shorter adult lifespan (Chen and Maklakov, 2012). The complete desiccation of the turquoise killifish ponds during the dry season represents a potent extrinsic mortality constraint that may have resulted in the evolution of accelerated sexual maturation. Short lifespan might have evolved as the pleiotropic byproduct of rapid sexual maturation or could have resulted from relaxed negative selection on several genes. The abundance of transposable elements in the turquoise killifish genome might have played a role in the evolution of its compressed lifespan.

Intriguingly, genes implicated in insulin/IGF signaling and genome maintenance are under positive selection or uniquely changed in exceptionally long-lived mammals such as the naked mole rat (Kim et al., 2011), Brandt's bat (Seim et al., 2013), and bowhead whale (Keane et al., 2015). *IGF1R* has also been shown to be under positive selection in the genome of a short-lived primate, the marmoset (The Marmoset Genome Sequencing and Analysis Consortium, 2014), raising the exciting possibility that the same gene may be modified to evolve compressed or extended life trajectories under different environmental constraints. *IGF1R*, which controls organismal growth, sexual maturity, fertility, and fitness, may be particularly "tunable" for the evolution of different life strategies.

### Genetic Architecture of Lifespan between Turquoise Killifish Strains

A region associated to lifespan differences between turquoise killifish strains contains a cluster of known aging and longevity genes. The gene encoding progranulin (*GRN*) is particularly interesting as it influences lifespan in mice (Ahmed et al., 2010) and contributes to age-related diseases in humans, in particular frontotemporal dementia (Baker et al., 2006). It will be important to determine the functional consequences of the turquoise killifish *GRN* variants, for example in stress response (Judy et al., 2013). Variants in non-coding genes or regulatory regions, such as enhancers, may also mediate the observed lifespan differences between fish strains. Finally, the differences in lifespan between turquoise killifish strains may be accentuated in the specific laboratory setting used in this study, and the genetic region identified may represent the interaction between genetic determinants in this particular environment.

### A Cluster of Aging and Longevity Genes Linked to the Sex-Determining Region

The most important region associated to differences in lifespan between turquoise killifish strains is on the sex chromosome, consistent with a previous finding (Kirschner et al., 2012). Our high-resolution genetic linkage map enabled us to show that the lifespan QTL is linked to, but distinct from the sex-determining region, within a region of suppressed recombination. The association between the sex-determining region and the lifespan QTL raises the intriguing possibility that these regions might have co-evolved. Alternatively, conditions of partially suppressed recombination might have allowed this trait to hitchhike on the sex-determining region without any direct fitness benefit.

In conclusion, our analyses reveal a possible association between lifespan and sex determination in the turquoise killifish and identify several potential candidates that may underlie lifespan differences across taxa. The African turquoise killifish genomic and genetic resources should help further explore the genetic basis of longevity and the evolutionary forces that shape vertebrate life history traits in nature.

## EXPERIMENTAL PROCEDURES

Additional details are provided in the [Supplemental Experimental Procedures](#).

### Figure 6. Genetic Architecture of Lifespan Using a Cross between Shorter-Lived and Longer-Lived Strains of the Turquoise Killifish

(A) Scheme of cross GxM. A female from the shorter-lived GRZ strain was crossed with a male from the longer-lived MZM-0703 strain (P0) to generate F1 progeny. F1 individuals were mated to generate F2 progeny.

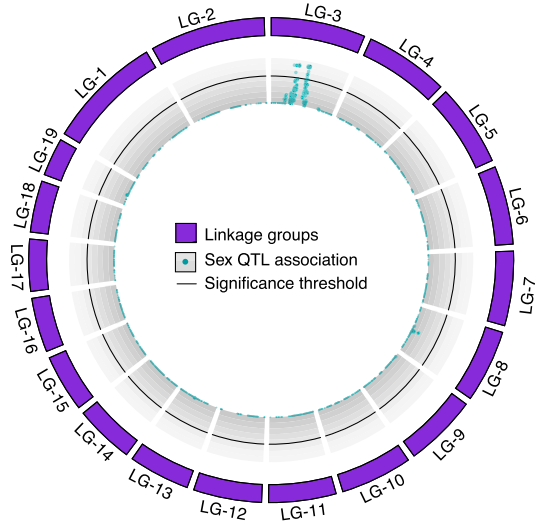
(B) Lifespan of the parental strains, F1, and F2 progeny of cross GxM in the captive conditions used in this study (pooled males and females). p values for differential survival compared to GRZ individuals in log-rank tests are indicated. See [Table S6A](#) for complete statistics.

(C) Circos plot representing the linkage map of cross GxM and association of markers with lifespan by quantitative trait locus (QTL) analysis. The linkage map is composed of 19 linkage groups (LG). The association of each RAD-seq marker with differences in lifespan is represented as  $-\log_{10}$  of the Random Forest Analysis q-value (red dots). The 5% FDR significance threshold is denoted by a black line. There is one marker above this threshold in LG-3 (lifespan QTL). Genomic scaffold length is scaled to genetic distance (cM) and not physical distance (bp).

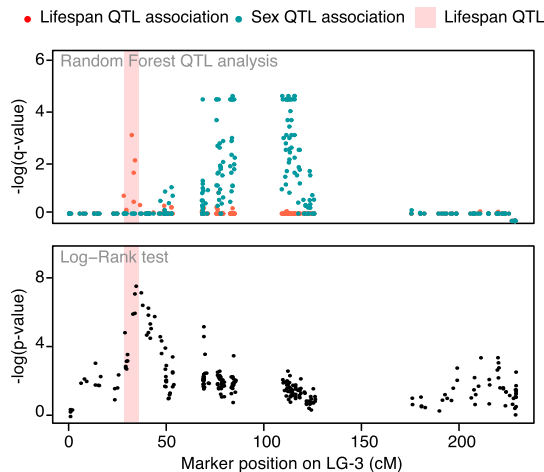
(D) The lifespan QTL is non-transgressive. Upper: Raw  $-\log_{10}$  (Random Forest Analysis q-value) for association of markers to individual fish lifespan. Middle:  $-\log_{10}$  of the Random Forest Analysis q-values for association of markers to lifespan after sex regression to account for the possible effect of sex as a confounding variable. Bottom: Survival stratified by genotype associated with each marker on LG-3. Homozygotes with alleles coming from the long-lived MZM-0703 grandparent (red) exhibit highest survival at the  $\sim 35$  cM position on LG-3, whereas homozygotes with alleles coming from the short-lived GRZ grandparent (blue) exhibit lowest survival. Light red rectangle: lifespan QTL region.

(E) Lifespan of fish with different genotypes at the marker that is most significantly associated to the lifespan QTL (RAD-seq marker 46347). p values for differential survival compared to individuals with the GRZ/GRZ genotype in log-rank tests are indicated. See also [Table S6B](#) for complete statistics.

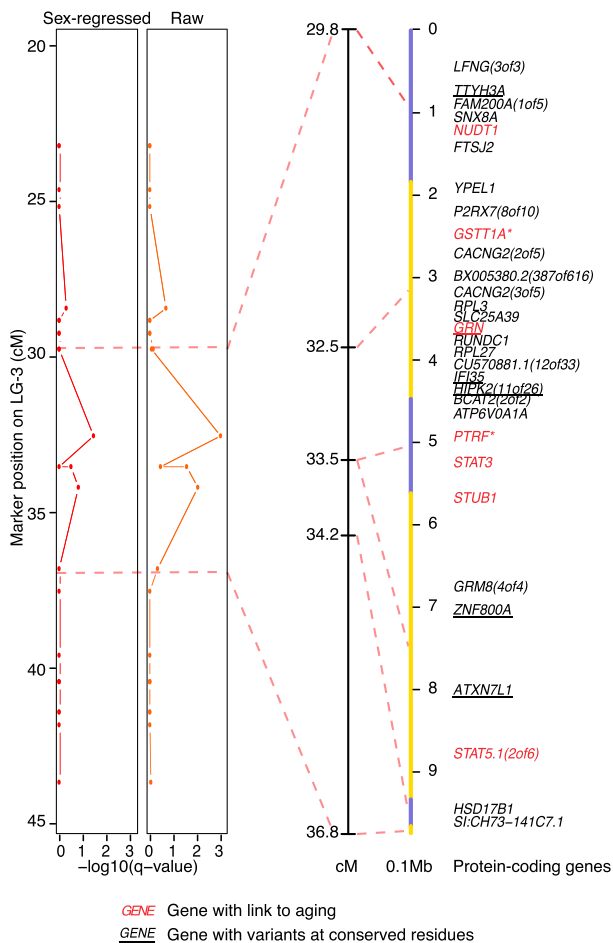
**A Mapping of sex-determining region**



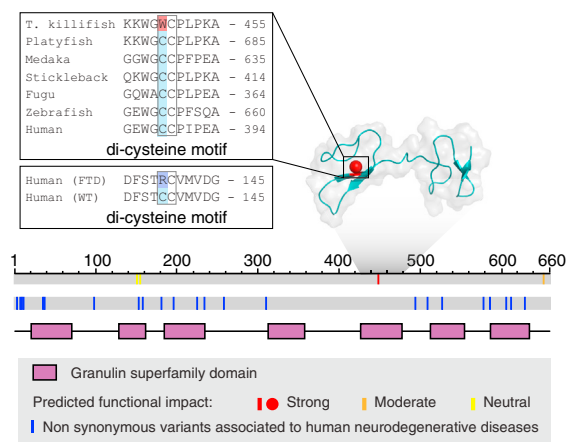
**B Marker association with lifespan or sex**



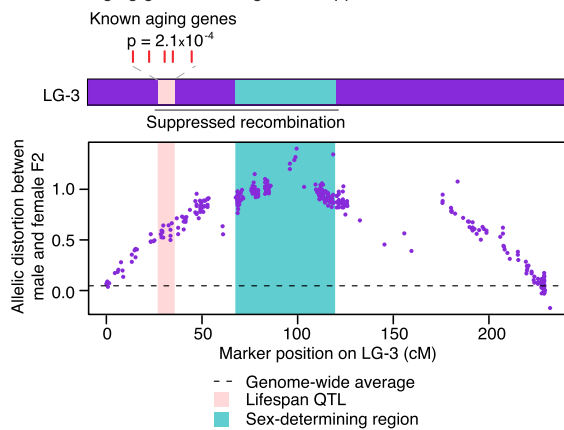
**C Genes underlying the lifespan QTL**



**D Impact of turquoise killifish GRN non-synonymous variants**



**E Cluster of aging genes in a region of suppressed recombination**



**Figure 7. The Lifespan QTL Is Linked to the Sex-Determining Region and Contains a Cluster of Aging Genes**

(A) Circos plot representing the association of markers with sex by QTL analysis in cross GxM. The association of each RAD-seq marker with sex is represented as  $-\log_{10}$  of the Random Forest Analysis q-value (turquoise dots). The 5% FDR significance threshold is denoted by a black line. There is a cluster of markers above this threshold in LG-3 (sex-determining region).

(legend continued on next page)

### Assembly of the Turquoise Killifish Genome and Identification of Genetic Variants between Different Strains

Genomic DNA was isolated from African turquoise killifish individuals from the inbred strain GRZ, and 10 libraries with varying insert sizes were constructed for Illumina sequencing on HiSeq2000 instruments. Assembly was performed using SGA (Simpson and Durbin, 2012) and SOAPdenovo (Luo et al., 2012), scaffolding using SSPACE Basic (Boetzer et al., 2011) and Gap-filling using GapFiller (Nadalin et al., 2012). The SGA and SOAP genome assemblies were reconciled using GARM (Soto-Jimenez et al., 2014). Higher-order scaffolding was performed using raw reads from Illumina paired-end RNA-seq libraries and the linkage map from cross GxM (see below). To identify genetic variants among different strains of the turquoise killifish, we resequenced the founders of cross GxM (GRZ female and MZM-0703 male) and a male individual from another wild-derived strain (MZM-0403). We used the GATK genotyping pipeline (McKenna et al., 2010) to call variants between turquoise killifish strains. The turquoise killifish genome browser is at <http://africanturquoisekillifishbrowser.org>.

### De Novo Prediction and Annotation of Protein Coding Gene Models

The MAKER2 pipeline was used to generate putative protein coding gene predictions (Holt and Yandell, 2011) supported by two ab initio gene prediction approaches. To annotate protein-coding genes from predictions, a homology-based approach using 19 other genomes (Table S2B) was implemented, leading to a final set of 28,494 turquoise killifish putative protein coding genes.

### Identification of Turquoise Killifish Genes under Positive Selection and Prediction of Their Functional Impact

The genes under positive selection were identified using a branch-site model implemented in PAML (Yang, 2007). Single ortholog protein families were identified using the eight teleost fish genomes in our analysis. Proteins from each family were aligned using PRANK (Löytynoja and Goldman, 2005) and the resulting alignments were filtered in GUIDANCE (Penn et al., 2010). CODEML was then used to predict the genes and individual sites under positive selection in the turquoise killifish lineage after stringent filtering and multiple hypothesis correction. Functional impact for the residues under positive selection or the non-synonymous variants between killifish strains was assessed using two prediction algorithms: PROVEAN (Choi et al., 2012) and SIFT (Kumar et al., 2009).

### Linkage Map and QTL Mapping

One female from the GRZ strain was crossed with one male from the MZM-0703 strain (cross GxM), and one female from the Soveia strain was crossed with a male from the GRZ strain (cross SxG). F1 fish were interbred in families to generate F2 fish and lifespan was scored as the age at death for all individuals (raised in cohorts of mixed sexes). RAD-seq libraries for 225 samples from the cross GxM and for 86 from the SxG cross were prepared as described (Eter et al., 2011). This resulted in 8,399 polymorphic markers for cross GxM. We built a linkage map for cross GxM with R/qtl using RAD-seq markers that had

homozygous haplotypes in the grandparents. After stringent filtering of markers and individuals, the final linkage map comprised 193 F2 individuals and 5,757 RAD-seq markers. QTL detection was conducted using a Random Forest-based method (Clément-Ziza et al., 2014), which uses genetic markers as predictors to model the traits and where population structure is modeled as covariates.

### ACCESSION NUMBERS

The accession number for all new GRZ genome and transcriptome libraries and RAD-seq libraries reported in this paper is SRA: SRP041421. The accession number for the H3K4me3 ChIP-seq data in the brain reported in this paper is SRA: SRP045718. The accession number for the draft genome reported in this paper is GenBank: JNBZ000000000.

### SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and seven tables and can be found with this article online at <http://dx.doi.org/10.1016/j.cell.2015.11.008>.

### AUTHOR CONTRIBUTIONS

D.R.V. and A. Brunet designed and initiated the study. D.R.V. performed all the fish experiments, conducted the 2010 expedition in Mozambique, and performed the RAD-seq analysis, linkage, and trait mapping. B.A.B. performed the de novo assemblies, annotations, re-sequencing analyses, RNA-seq analysis and deployed the genome browser portal. P.P.S. identified orthologs, filtered annotations, performed all phylogenetic and evolutionary analyses, and helped with the genome browser. E.Z. assisted in fish experiments and linkage analyses. P.D.E. and E.A.J. prepared and sequenced RAD-seq libraries. C.K.H. built the brain and tail RNA-seq libraries. M.C.Z. ran the random forest model. D.W. measured recombination suppression, performed synteny analysis, and generated gene trees. R.C. assisted with genome size estimate. I.H. performed the H3K4me3 ChIP-seq experiment. M.-C.Y. made initial genome and RNA-seq libraries. B.E.M. performed the Sanger re-sequencing of individuals from the wild. S.C.S. helped with large-scale fish cohorts. C.D.B. provided intellectual contributions. A. Beyer established the random forest-based QTL mapping. D.R.V., B.A.B., P.P.S. and A. Brunet wrote the paper.

### ACKNOWLEDGMENTS

We thank Robert Piskol, Duygu Ucar, Erik Lenhart, Joanna Kelley, and Julian Catchen for their help with genome analysis. We thank Jonathan Pritchard and Yang Li for helping with evolutionary analyses, statistics, and the manuscript. We thank Dmitri Petrov and David Enard for their help with the

(B) The lifespan QTL is distinct from the sex-determining region. Top: Random Forest analysis for marker association with lifespan (red) or sex (turquoise). Lower: log-rank survival analysis in the F2 generation of cross GxM between homozygotes with alleles coming from GRZ grandparent versus the MZM-0703 grandparent at each marker. Light red rectangle: Lifespan QTL region.

(C) Identification of the genes underlying the lifespan QTL on LG-3. Left: Sex-regressed or raw  $-\log_{10}$  of the Random Forest Analysis q-value for association with lifespan. Dashed lines delimit the lifespan QTL. Right: lifespan QTL region on LG3 and corresponding anchored genomic scaffolds in alternating yellow and slate blue colors. Markers are linked to the mid-points of the scaffolds. Genes in red have been previously linked to aging in the GenAge database (human and mouse combined, Table S4A) or manually curated from the literature (asterisks, Table S7A). Underlined genes have non-synonymous variants at evolutionary-conserved residues. See also Figures S7B–S7G.

(D) Location of residues with putative functional consequences and associated to human neurodegenerative diseases on GRN in the turquoise killifish. Color represents the strength of functional impact. Top: NMR structure of human orthologous domain. Grey shadow: GRN domain with available NMR structure. Top insert: alignment of the residue with strong functional effect W449 in the turquoise killifish with other species (see also Figure S7D and Table S7H). Bottom insert: region surrounding a mutation found in human frontotemporal dementia (FTD) patients, involving an analogous di-cysteine motif residue. Bottom: schematic of the residues mapped on the turquoise killifish protein sequence (gray).

(E) Cluster of known aging genes in the lifespan QTL region, in a region of suppressed recombination. Top: schematic of the enrichment for known aging-related genes (from GeneAge, human and mouse combined) in the lifespan QTL region ( $p = 2.1 \times 10^{-4}$ , in Fisher's exact test, compared to rest of the genome,  $p = 6.4 \times 10^{-4}$  in Fisher's exact test, compared to the rest of LG-3). Bottom: measure of suppressed recombination by allelic distortion between the male and female F2 progeny at each marker on LG-3. See also Figures S7H and S7I. Dash line indicates the genome-wide average for allelic distortion. Light red rectangle: lifespan QTL region. Turquoise rectangle: sex-determining region.

evolutionary analysis. We thank Art Owen for his help with statistics. We thank David Kingsley, Felicity Jones, and Frank Chan for helping with QTL analysis and the manuscript. We thank members of the Brunet lab for help on the manuscript. We thank Aaron Daugherty, Ben Dulken, Katja Hebestreit, Andrew McKay, and Robin Yeo for helping with independent code verification. We thank Aimee Kao for helpful discussion about *GRN*. This work was supported by NIH DP1AG044848 (A. Brunet), the Glenn Laboratories for the Biology of Aging (A. Brunet), the Max Planck Society and the Max Planck Institute for Biology of Ageing (D.R.V., D.W. and R.C.), the Dean's fellowship at Stanford and NIH K99AG049934 (B.A.B.), the Stanford Center for Computational Evolutionary and Human Genomics fellowship (P.P.S.), the Life Sciences Research Foundation fellowship (C.K.H.), the Damon Runyon, Rothschild, and HFSP fellowships (I.H.), and the German Federal Ministry of Education and Research (A. Beyer., M.C.Z., Grant: Sybacol).

Received: March 16, 2015

Revised: September 7, 2015

Accepted: November 2, 2015

Published: December 3, 2015

## REFERENCES

- Ahmed, Z., Sheng, H., Xu, Y.F., Lin, W.L., Innes, A.E., Gass, J., Yu, X., Wuertzer, C.A., Hou, H., Chiba, S., et al. (2010). Accelerated lipofuscinosis and ubiquitination in granulin knockout mice suggest a role for progranulin in successful aging. *Am. J. Pathol.* **177**, 311–324.
- Austad, S.N. (2010). Methuselah's Zoo: how nature provides us with clues for extending human health span. *J. Comp. Pathol.* **142** (Suppl 1), S10–S21.
- Ayyadevara, S., Dandapat, A., Singh, S.P., Siegel, E.R., Shmookler Reis, R.J., Zimniak, L., and Zimniak, P. (2007). Life span and stress resistance of *Caenorhabditis elegans* are differentially affected by glutathione transferases metabolizing 4-hydroxynon-2-enal. *Mech. Ageing Dev.* **128**, 196–205.
- Baker, M., Mackenzie, I.R., Pickering-Brown, S.M., Gass, J., Rademakers, R., Lindholm, C., Snowden, J., Adamson, J., Sadovnick, A.D., Rollinson, S., et al. (2006). Mutations in progranulin cause tau-negative frontotemporal dementia linked to chromosome 17. *Nature* **442**, 916–919.
- Ballew, B.J., Yeager, M., Jacobs, K., Giri, N., Bolland, J., Burdett, L., Alter, B.P., and Savage, S.A. (2013). Germline mutations of regulator of telomere elongation helicase 1, RTEL1, in Dyskeratosis congenita. *Hum. Genet.* **132**, 473–480.
- Blüher, M., Kahn, B.B., and Kahn, C.R. (2003). Extended longevity in mice lacking the insulin receptor in adipose tissue. *Science* **299**, 572–574.
- Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579.
- Budovsky, A., Craig, T., Wang, J., Tacutu, R., Csordas, A., Lourenço, J., Fraielfeld, V.E., and de Magalhães, J.P. (2013). LongevityMap: a database of human genetic variants associated with longevity. *Trends Genet.* **29**, 559–560.
- Cellerino, A., Valenzano, D.R., and Reichard, M. (2015). From the bush to the bench: the annual *Nothobranchius* fishes as a new model system in biology. *Biol. Rev. Camb. Philos. Soc.* <http://dx.doi.org/10.1111/brv.12183>.
- Chen, H.Y., and Maklakov, A.A. (2012). Longer life span evolves under high rates of condition-dependent mortality. *Curr. Biol.* **22**, 2140–2143.
- Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., and Chan, A.P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* **7**, e46688.
- Clément-Ziza, M., Marsellach, F.X., Codlin, S., Papadakis, M.A., Reinhardt, S., Rodríguez-López, M., Martin, S., Marguerat, S., Schmidt, A., Lee, E., et al. (2014). Natural genetic variation impacts expression levels of coding, non-coding, and antisense transcripts in fission yeast. *Mol. Syst. Biol.* **10**, 764.
- Conneely, K.N., Capell, B.C., Erdos, M.R., Sebastiani, P., Solovieff, N., Swift, A.J., Baldwin, C.T., Budagov, T., Barzilai, N., Atzmon, G., et al. (2012). Human longevity and common variations in the LMNA gene: a meta-analysis. *Aging Cell* **11**, 475–481.
- De-Fraja, C., Conti, L., Govoni, S., Battaini, F., and Cattaneo, E. (2000). STAT signalling in the mature and aging brain. *Int. J. Dev. Neurosci.* **18**, 439–446.
- De Luca, G., Ventura, I., Sanghez, V., Russo, M.T., Ajmone-Cat, M.A., Cacci, E., Martire, A., Popoli, P., Falcone, G., Michelini, F., et al. (2013). Prolonged lifespan with enhanced exploratory behavior in mice overexpressing the oxidized nucleoside triphosphatase hMTH1. *Aging Cell* **12**, 695–705.
- de Magalhães, J.P., Budovsky, A., Lehmann, G., Costa, J., Li, Y., Fraielfeld, V., and Church, G.M. (2009). The Human Ageing Genomic Resources: online databases and tools for biogerontologists. *Aging Cell* **8**, 65–72.
- Eriksson, M., Brown, W.T., Gordon, L.B., Glynn, M.W., Singer, J., Scott, L., Erdos, M.R., Robbins, C.M., Moses, T.Y., Berglund, P., et al. (2003). Recurrent *de novo* point mutations in lamin A cause Hutchinson-Gilford progeria syndrome. *Nature* **423**, 293–298.
- Etter, P.D., Bassham, S., Hohenlohe, P.A., Johnson, E.A., and Cresko, W.A. (2011). SNP discovery and genotyping for evolutionary genetics using RAD sequencing. *Methods Mol. Biol.* **772**, 157–178.
- Flachsbarf, F., Caliebe, A., Kleindorfer, R., Blanché, H., von Eller-Eberstein, H., Nikolaus, S., Schreiber, S., and Nebel, A. (2009). Association of FOXO3A variation with human longevity confirmed in German centenarians. *Proc. Natl. Acad. Sci. USA* **106**, 2700–2705.
- Gottlieb, S., and Ruvkun, G. (1994). *daf-2*, *daf-16* and *daf-23*: genetically interacting genes controlling Dauer formation in *Caenorhabditis elegans*. *Genetics* **137**, 107–120.
- Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491.
- Holzenberger, M., Dupont, J., Ducos, B., Leneuve, P., Géloën, A., Even, P.C., Cervera, P., and Le Bouc, Y. (2003). IGF-1 receptor regulates lifespan and resistance to oxidative stress in mice. *Nature* **421**, 182–187.
- Johnson, S.C., Rabinovitch, P.S., and Kaeberlein, M. (2013). mTOR is a key modulator of ageing and age-related disease. *Nature* **493**, 338–345.
- Judy, M.E., Nakamura, A., Huang, A., Grant, H., McCurdy, H., Weiberth, K.F., Gao, F., Coppola, G., Kenyon, C., and Kao, A.W. (2013). A shift to organismal stress resistance in programmed cell death mutants. *PLoS Genet.* **9**, e1003714.
- Kaeberlein, M., and Kennedy, B.K. (2011). Hot topics in aging research: protein translation and TOR signaling, 2010. *Aging Cell* **10**, 185–190.
- Kapahi, P., Chen, D., Rogers, A.N., Katewa, S.D., Li, P.W., Thomas, E.L., and Kockel, L. (2010). With TOR, less is more: a key role for the conserved nutrient-sensing TOR pathway in aging. *Cell Metab.* **11**, 453–465.
- Keane, M., Semeiks, J., Webb, A.E., Li, Y.I., Quesada, V., Craig, T., Madsen, L.B., van Dam, S., Brawand, D., Marques, P.I., et al. (2015). Insights into the evolution of longevity from the bowhead whale genome. *Cell Rep.* **10**, 112–122.
- Kennedy, B.K., Berger, S.L., Brunet, A., Campisi, J., Cuervo, A.M., Epel, E.S., Franceschi, C., Lithgow, G.J., Morimoto, R.I., Pessin, J.E., et al. (2014). Geroscience: linking aging to chronic disease. *Cell* **159**, 709–713.
- Kenyon, C.J. (2010). The genetics of ageing. *Nature* **464**, 504–512.
- Kenyon, C., Chang, J., Gensch, E., Rudner, A., and Tabtiang, R. (1993). A *C. elegans* mutant that lives twice as long as wild type. *Nature* **366**, 461–464.
- Kim, E.B., Fang, X., Fushan, A.A., Huang, Z., Lobanov, A.V., Han, L., Marino, S.M., Sun, X., Turanov, A.A., Yang, P., et al. (2011). Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature* **479**, 223–227.
- Kirschner, J., Weber, D., Neuschl, C., Franke, A., Böttger, M., Zielke, L., Powalsky, E., Groth, M., Shagin, D., Petzold, A., et al. (2012). Mapping of quantitative trait loci controlling lifespan in the short-lived fish *Nothobranchius furzeri*—a new vertebrate model for age research. *Aging Cell* **11**, 252–261.
- Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081.

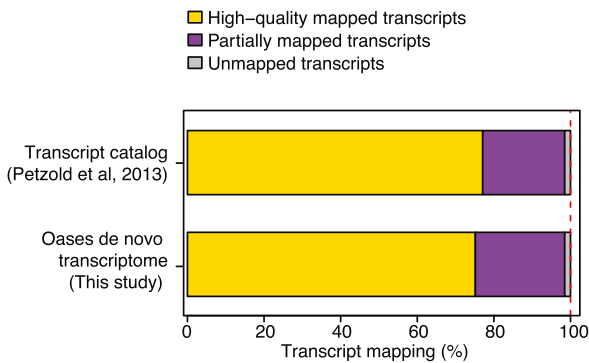


- López-Otín, C., Blasco, M.A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. *Cell* *153*, 1194–1217.
- Löytynoja, A., and Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. USA* *102*, 10557–10562.
- Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* *1*, 18.
- Marmoset Genome Sequencing and Analysis Consortium (2014). The common marmoset genome provides insight into primate biology and evolution. *Nat. Genet.* *46*, 850–857.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytzky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* *20*, 1297–1303.
- Nadalín, F., Vezzi, F., and Policriti, A. (2012). GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics* *13* (Suppl 14), S8.
- Patterson, G.I., Koweeck, A., Wong, A., Liu, Y., and Ruvkun, G. (1997). The DAF-3 Smad protein antagonizes TGF-beta-related receptor signaling in the *Caenorhabditis elegans* dauer pathway. *Genes Dev.* *11*, 2679–2690.
- Penn, O., Privman, E., Landan, G., Graur, D., and Pupko, T. (2010). An alignment confidence score capturing robustness to guide tree uncertainty. *Mol. Biol. Evol.* *27*, 1759–1767.
- Pesch, B., Düsing, R., Rabstein, S., Harth, V., Grentrup, D., Brüning, T., Landt, O., Vetter, H., and Ko, Y.D. (2004). Polymorphic metabolic susceptibility genes and longevity: a study in octogonarians. *Toxicol. Lett.* *151*, 283–290.
- Proitsi, P., Lupton, M.K., Dudbridge, F., Tsolaki, M., Hamilton, G., Daniilidou, M., Pritchard, M., Lord, K., Martin, B.M., Johnson, J., et al. (2012). Alzheimer's disease and age-related macular degeneration have different genetic models for complement gene variation. *Neurobiol. Aging* *33*, 1843.e9–1843.e17.
- Raeder, H., Johansson, S., Holm, P.I., Haldorsen, I.S., Mas, E., Sbarra, V., Neramo, I., Eide, S.A., Grevle, L., Bjørkhaug, L., et al. (2006). Mutations in the *CEL* VNTR cause a syndrome of diabetes and pancreatic exocrine dysfunction. *Nat. Genet.* *38*, 54–62.
- Reichwald, K., Lauber, C., Nanda, I., Kirschner, J., Hartmann, N., Schories, S., Gausmann, U., Taudien, S., Schilhabel, M.B., Szafranski, K., et al. (2009). High tandem repeat content in the genome of the short-lived annual fish *Nothobranchius furzeri*: a new vertebrate model for aging research. *Genome Biol.* *10*, R16.
- Seim, I., Fang, X., Xiong, Z., Lobanov, A.V., Huang, Z., Ma, S., Feng, Y., Turanov, A.A., Zhu, Y., Lenz, T.L., et al. (2013). Genome analysis reveals insights into physiology and longevity of the Brandt's bat *Myotis brandtii*. *Nat. Commun.* *4*, 2212.
- Simpson, J.T., and Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome Res.* *22*, 549–556.
- Soerensen, M., Dato, S., Tan, Q., Thinggaard, M., Kleindorp, R., Beekman, M., Jacobsen, R., Suchiman, H.E., de Craen, A.J., Westendorp, R.G., et al. (2012). Human longevity and variation in GH/IGF-1/insulin signaling, DNA damage signaling and repair and pro/antioxidant pathway genes: cross sectional and longitudinal studies. *Exp. Gerontol.* *47*, 379–387.
- Soto-Jimenez, L.M., Estrada, K., and Sanchez-Flores, A. (2014). GARM: genome assembly, reconciliation and merging pipeline. *Curr. Top. Med. Chem.* *14*, 418–424.
- Souza-Pinto, N.C., Croteau, D.L., Hudson, E.K., Hansford, R.G., and Bohr, V.A. (1999). Age-associated increase in 8-oxo-deoxyguanosine glycosylase/AP lyase activity in rat mitochondria. *Nucleic Acids Res.* *27*, 1935–1942.
- Suh, Y., Atzmon, G., Cho, M.O., Hwang, D., Liu, B., Leahy, D.J., Barzilai, N., and Cohen, P. (2008). Functionally significant insulin-like growth factor I receptor mutations in centenarians. *Proc. Natl. Acad. Sci. USA* *105*, 3438–3442.
- Terzibas, E., Valenzano, D.R., Benedetti, M., Roncaglia, P., Cattaneo, A., Domenici, L., and Cellerino, A. (2008). Large differences in aging phenotype between strains of the short-lived annual fish *Nothobranchius furzeri*. *PLoS ONE* *3*, e3866.
- Valdesalici, S., and Cellerino, A. (2003). Extremely short lifespan in the annual fish *Nothobranchius furzeri*. *Proc. Biol. Sci.* *270* (Suppl 2), S189–S191.
- Valenzano, D.R., Kirschner, J., Kamber, R.A., Zhang, E., Weber, D., Cellerino, A., Englert, C., Platzer, M., Reichwald, K., and Brunet, A. (2009). Mapping loci associated with tail color and sex determination in the short-lived fish *Nothobranchius furzeri*. *Genetics* *183*, 1385–1395.
- Vilchez, D., Morante, I., Liu, Z., Douglas, P.M., Merkwirth, C., Rodrigues, A.P., Manning, G., and Dillin, A. (2012). RPN-6 determines *C. elegans* longevity under proteotoxic stress conditions. *Nature* *489*, 263–268.
- Vogel, H., Lim, D.S., Karsenty, G., Finegold, M., and Hasty, P. (1999). Deletion of *Ku86* causes early onset of senescence in mice. *Proc. Natl. Acad. Sci. USA* *96*, 10770–10775.
- Wang, J., Van Damme, P., Cruchaga, C., Gitcho, M.A., Vidal, J.M., Seijo-Martínez, M., Wang, L., Wu, J.Y., Robberecht, W., and Goate, A. (2010). Pathogenic cysteine mutations affect progranulin function and production of mature granulins. *J. Neurochem.* *112*, 1305–1315.
- Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* *24*, 1586–1591.

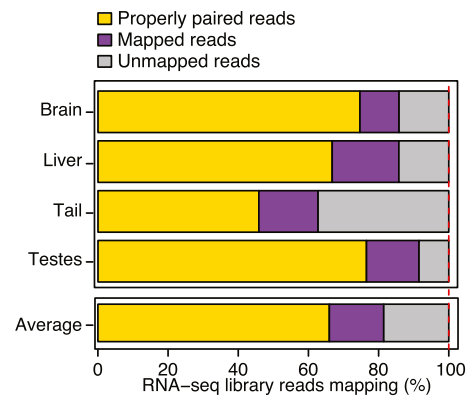
## A Regions with excess coverage in the turquoise killifish genome assembly

Regions with excess coverage	Length (Mb)	Portion of assembly (%)	Proportion of reads (%)	Over-representation
> 2 times the expected coverage	455	44.5	70.1	1.6X
> 3 times the expected coverage	47	4.6	26.0	5.7X
> 4 times the expected coverage	21	2.0	20.1	10.4X
> 10 times the expected coverage	10	1.0	17.4	17.4X

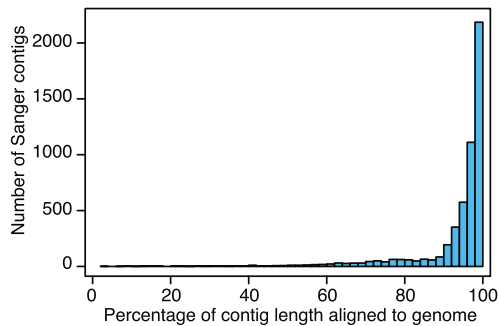
## B Assembled transcript mapping



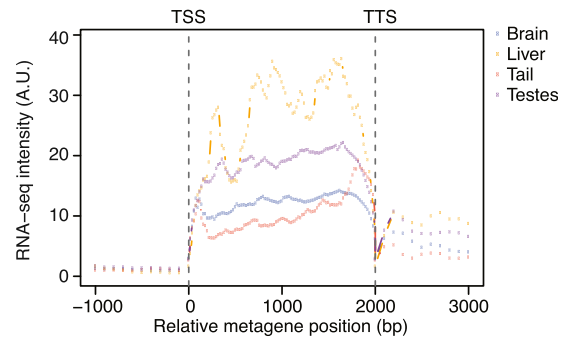
## C Paired-end RNA-seq mapping



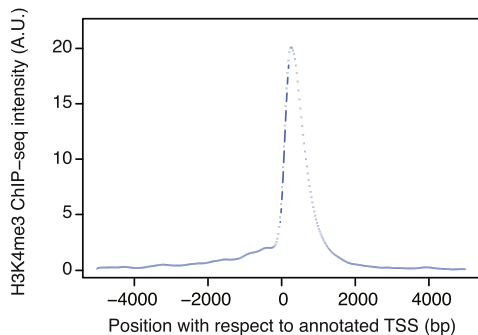
## D Alignment of Sanger-sequenced genome shotgun contigs



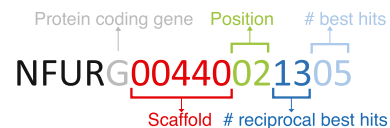
## E Gene body RNA-seq support



## F Promoter annotation quality



## G Nomenclature for protein-coding genes



**Figure S1. Quality Controls for the De Novo Genome Assembly and Annotation of Protein-Coding Genes in the African Turquoise Killifish, Related to Figure 2**

(A) Sequencing coverage of some regions is considerably higher than expected (see [Experimental Procedures](#)), suggesting the presence of contigs assembled as one copy but repeated many times in the genome. A conservative genome size estimate of 2 Gb was used for these calculations.

(legend continued on next page)

---

(B) Mapping of previously published turquoise killifish transcript catalog (Petzold et al., 2013) or our de novo-assembled Oases transcriptome (this study) to the turquoise killifish genome assembly. High-quality mapped transcripts: alignment by exonerate filtered by MAKER2. Partially mapped transcripts: alignment by BLAT (transcripts not mapped by exonerate) (see [Experimental Procedures](#)).

(C) Mapping of turquoise killifish RNA-seq paired-end libraries to the turquoise killifish genome assembly. Alignment was performed with Tophat2. Note that the majority of mapped reads are properly paired, indicative of a good genome assembly quality.

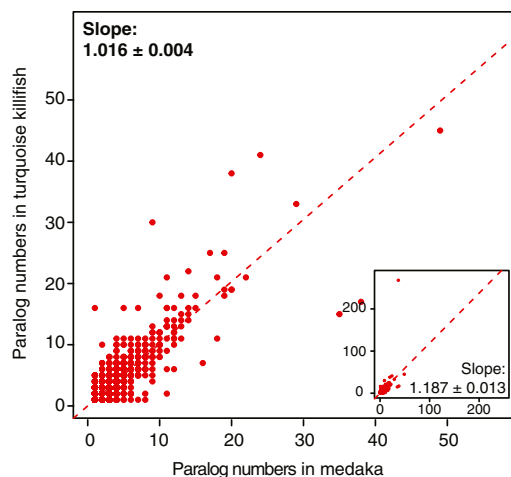
(D) Distribution histogram of previously published Sanger-sequenced shotgun contigs ([Reichwald et al., 2009](#)) mapping to our turquoise killifish genome assembly. Alignments were performed using BLAT.

(E) Meta-gene analysis using RNA-seq data from four adult tissues. Annotated protein-coding genes were used. The RNA-seq signal is maximal over gene bodies, consistent with correct gene prediction and annotation.

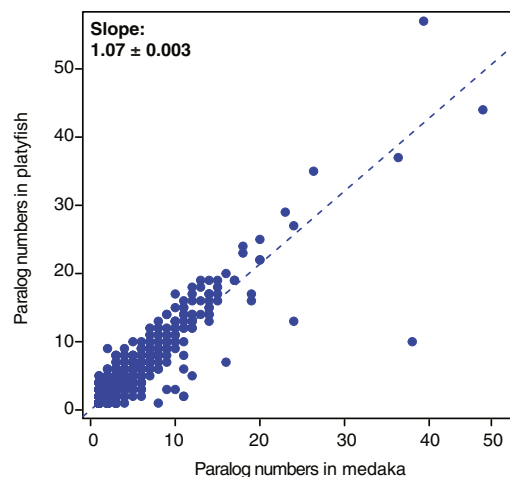
(F) Meta-promoter analysis using H3K4me3 ChIP-seq data from brain tissues. Annotated protein-coding genes were used. The H3K4me3 ChIP-seq signal is maximal at the transcriptional start site (TSS), consistent with correct promoter prediction.

(G) Naming convention for putative protein-coding genes that do not have a clear name from orthology. Gene ID is obtained using the scaffold number, gene position on the scaffold, number of best reciprocal blast hits between turquoise killifish and 19 other species, and number of BLASTp hits among all 20 species.

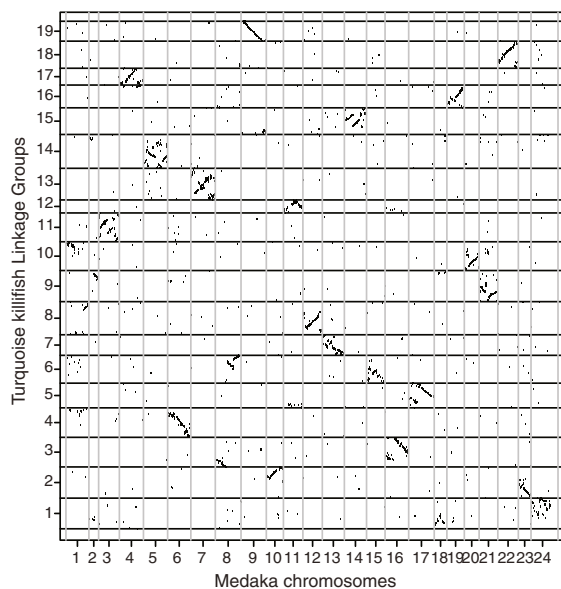
**A** Paralog numbers in turquoise killifish vs. medaka



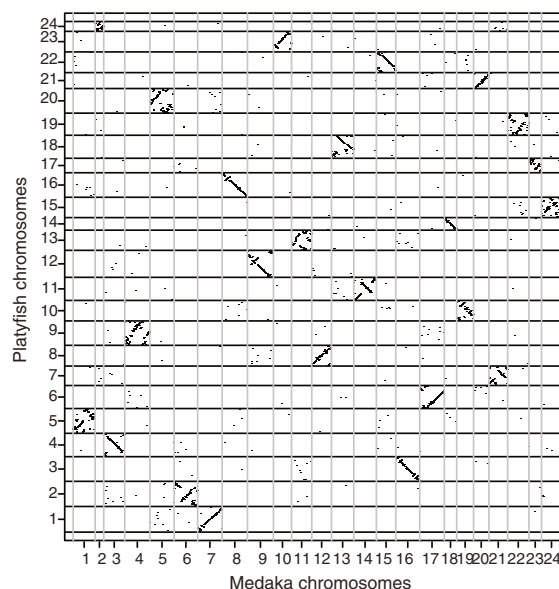
**B** Paralog numbers in platyfish vs. medaka



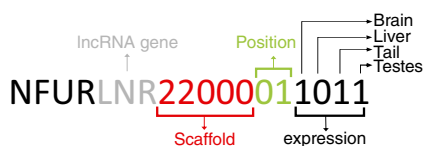
**C** Synteny between turquoise killifish and medaka



**D** Synteny between platyfish and medaka



**E** Nomenclature for lncRNA genes



**F** ncRNA gene predictions

ncRNA gene type	Number	Prediction algorithm
putative rRNA genes	200	RNAmer*
	320	Infernal
	504	RepeatMasker
putative tRNA genes	1,841	tRNA-scan SE*
	4,344	tRNA-scan SE (pseudogenes)
	1,760	Infernal
	2,172	RepeatMasker
putative snRNA genes	230	Infernal*
	223	RepeatMasker
putative snoRNA genes	191	Infernal*
putative miRNA genes	2,029	Infernal*

\*Higher specificity predictions

(legend on next page)

---

**Figure S2. Quality Controls for Correct Paralog Assembly and Non-coding RNA Predictions in the Genome of the Turquoise Killifish, Related to Figure 2**

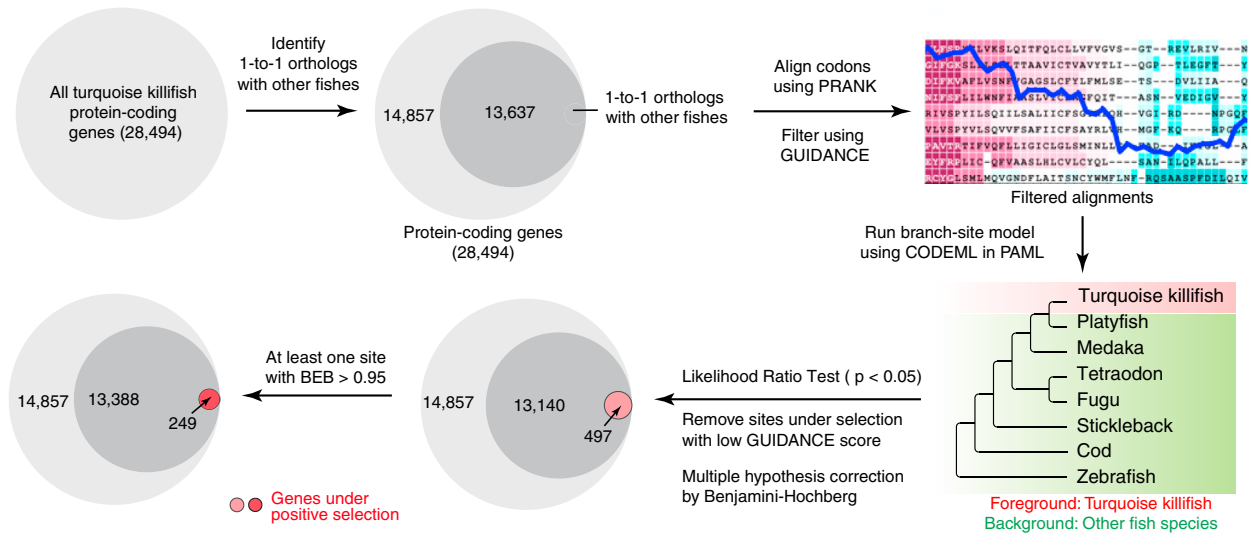
(A and B) Correlation of paralog numbers within gene families in turquoise killifish versus medaka (A) and in platyfish versus medaka for comparison (B). Medaka was chosen because it is the closest fish to turquoise killifish and platyfish with a Sanger-sequenced genome. The slope of the linear least square regression model without an intercept term and its standard error are reported. Gene families with at least one gene from each species are displayed. The results shown are for 7,593 gene families (A) and 8,303 families (B), using a similarity threshold of 70. Similar results were obtained when using other similarity thresholds (see [Experimental Procedures](#)). The main plot in (A) corresponds to the analysis after removing the largest gene family in the turquoise killifish (268 members, including multiple copies/isoforms of the *ZNF235* gene) whereas the inset corresponds to the entire analysis.

(C and D) Oxford Grid plots showing synteny between turquoise killifish linkage groups and medaka chromosomes (C), and between platyfish and medaka chromosomes for comparison (D). Note the presence of large blocks of syntenic genes, indicating a good quality of assembly, including for paralogs (see [Experimental Procedures](#)).

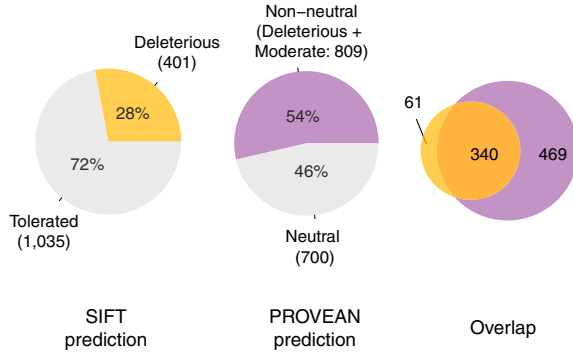
(E) Naming convention for lncRNA genes. Gene ID is obtained using the scaffold number, gene position on the scaffold, and 0/1 encoding to reflect the tissues in which RNA level evidence was found.

(F) ncRNA gene predictions in the turquoise killifish. Predictions from different tools for a single class of genes (e.g., tRNA genes) may be overlapping. \*predictions used to compute the numbers of ncRNA genes reported in [Figure 2B](#).

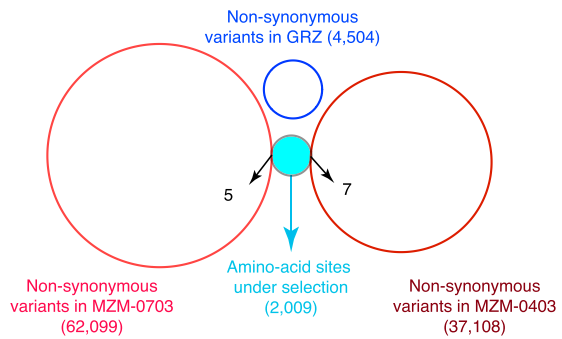
**A Identification of turquoise killifish genes under positive selection**



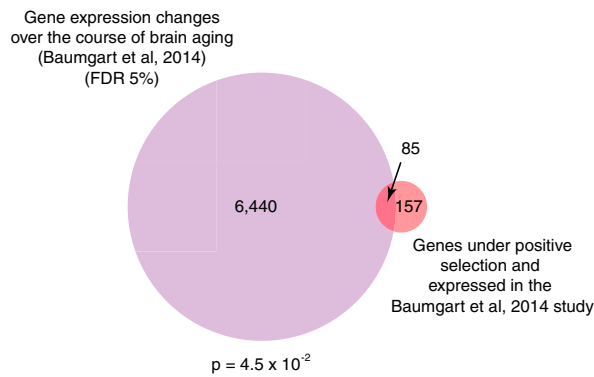
**B Functional effect prediction for the sites under positive selection**



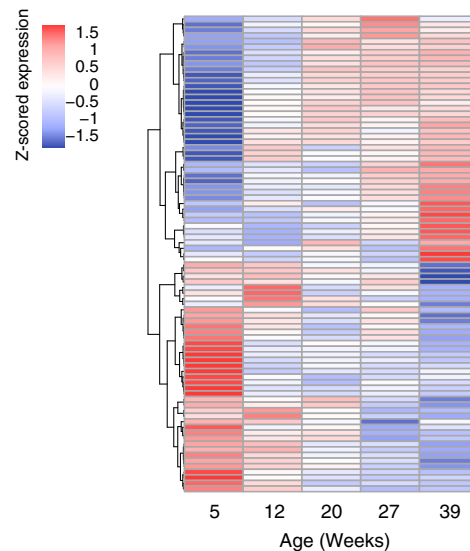
**C Comparison of residues under positive selection and inter-individual/strain variants**



**D Genes under positive selection and deregulated during brain aging (MZM-0410; FDR 5%)**



**E Expression profile of genes under positive selection and deregulated during brain aging (MZM-0410; FDR 5%)**



(legend on next page)

---

**Figure S3. Evolutionary Analysis of the Turquoise Killifish Genome and Prediction of Functional Effects of Positively Selected Variants, Related to Figure 3**

(A) Flowchart of the analysis of genes under positive selection in the turquoise killifish genome. Only genes with 1-to-1 orthologs with other bony fish species were included for alignment. The phylogenetic tree of the fish species used to analyze positive selection with PAML is indicated. Multiple hypothesis correction using Benjamini-Hochberg was applied, leading to 497 genes with significant p value from Likelihood Ratio Test (Table S3A). A subset of 249 genes had at least one site with BEB probability > 0.95 ("highest confidence list," Table S3B).

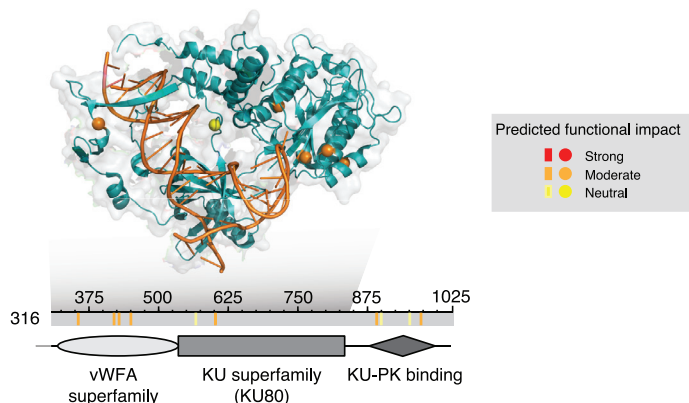
(B) Fraction of residues under positive selection with functional effect using SIFT (left) or PROVEAN (middle) as well as overlap between both methods (right). See also Table S3D.

(C) The vast majority of sites under positive selection in the GRZ reference genome do not overlap with inter-individual or inter-strain non-synonymous coding variants. None of the sites under positive selection in turquoise killifish using the GRZ reference genome overlap with non-synonymous variants in another GRZ individual (see also Figures 5 and S5B). Only 5 and 7 sites under positive selection in turquoise killifish using the GRZ reference genome overlap with non-synonymous variants identified in individuals from MZM-0703 and MZM-0403, respectively (see also Figures 5 and S5B). The fact that most of the genes under positive selection do not overlap with inter-individual or inter-strain variation is compatible with the identification of species-level selective constraints.

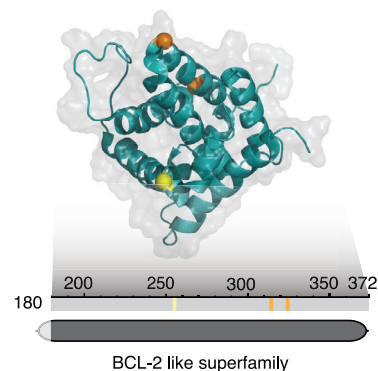
(D) Venn diagram for the overlap of genes under positive selection and genes showing a significant change in expression (FDR threshold of 5%) over the course of aging in brain tissues in the MZM-0410 strain (Baumgart et al., 2014) (see Experimental Procedures). p value is from the significance of overlap for enrichment in Fisher's exact test.

(E) Heatmap of expression over lifespan for genes under positive selection that show significant change in expression over the course of aging in brains of individuals from the MZM-0410 strain. Displayed genes in the heatmap correspond to the 85 gene overlap from Figure S3D and are listed in Table S3E.

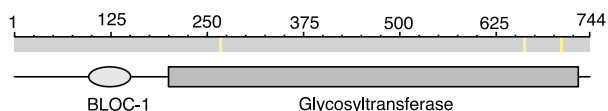
### A Residues under positive selection in XRCC5



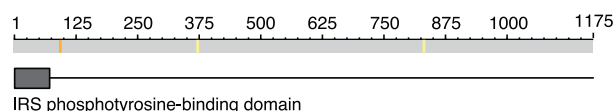
### B Residues under positive selection in BAX



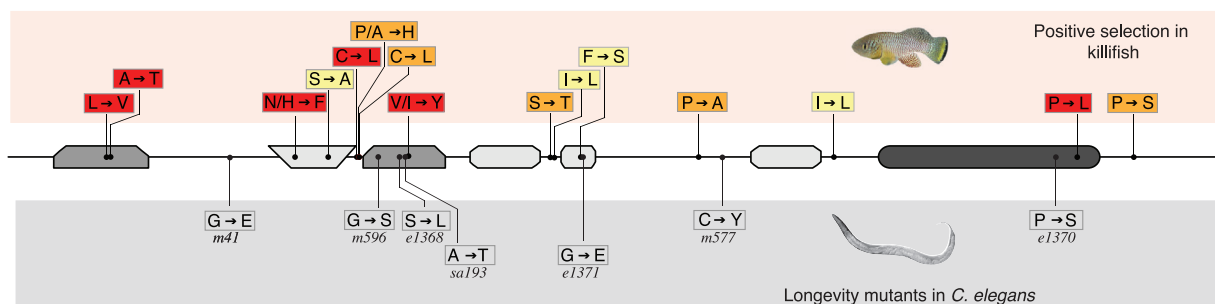
### C Residues under positive selection in MGAT5(1of3)



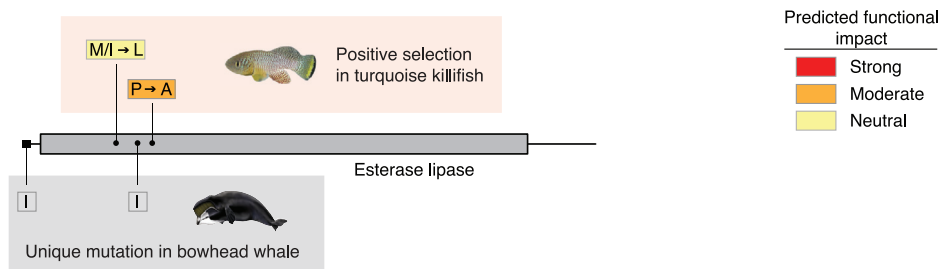
### D Residues under positive selection in IRS1(2of2)



### E Residues and variants in IGF1R(1of2) in turquoise killifish and DAF-2 in *C. elegans*



### F Residues and variants in CEL(7of7) in the turquoise killifish and bowhead whale



### Figure S4. Analysis of Aging-Related Genes under Positive Selection in the Turquoise Killifish, Related to Figure 4

(A and B) Location and functional effect of residues under positive selection in XRCC5 (A) and BAX (B) in the turquoise killifish. Red: changes with strong functional impact (identified by SIFT, PROVEAN, and structural analysis). Orange: changes with moderate functional impact (identified by at least SIFT, PROVEAN, or structural analysis). Yellow: changes with neutral functional impact. Top: crystal structure of human ortholog. Residues and their functional impacts are denoted by colored dots. Grey shadow: region of the protein with available crystal structure. Bottom: schematic of the residues mapped on the turquoise killifish protein sequence (gray). Residues and their functional impacts are denoted by colored bars. The conserved protein domains and functional sites predicted from the NCBI conserved domain database (Marchler-Bauer et al., 2015) are indicated. The indicated start codon is based on manually curated evidence. vWFA: Von Willebrand factor type A.

(C and D) Location of residues under positive selection in MGAT5(1of3) (C) and IRS1(2of2) (D) in the turquoise killifish. Schematic of the residues mapped on the turquoise killifish protein sequence (gray). Residues and their functional impacts are denoted by colored bars. The strength of the functional effect is color-coded

(legend continued on next page)



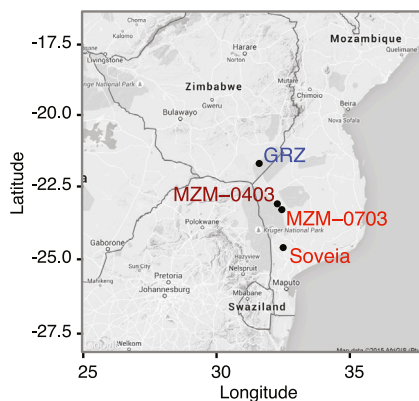
---

in the same way as in [Figure 4A](#). The conserved protein domains and functional sites predicted from the NCBI conserved domain database (Marchler-Bauer et al., 2015) are indicated. BLOC-1: biogenesis of lysosome-related organelles complex-1.

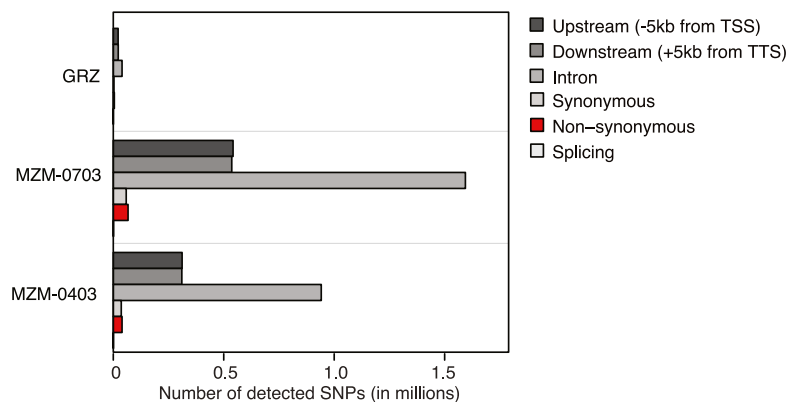
(E) Location and variants of residues under positive selection in the turquoise killifish for IGF1R(1of2), and the location and variants of residues linked to a longevity or dauer diapause phenotype in *C. elegans* insulin/IGF1 receptor (DAF-2). Top: turquoise killifish variants, with the changed amino acid on the right. The strength of the functional effect is color-coded in the same way as in [Figure 4A](#). Bottom: mutations in DAF-2 that are associated with an “extended lifespan” phenotype in *C. elegans* (from WormBase). Amino-acid changes in DAF-2 are from (Patel et al., 2008) and are mapped to turquoise killifish IGF1R(1of2). These mutants are on the right. Note that some of these mutants can also lead to dauer diapause (Gems et al., 1998). Conserved protein domains and functional sites predicted from the NCBI conserved domain database (Marchler-Bauer et al., 2015).

(F) Location and variants of residues under positive selection in the turquoise killifish for CEL(7of7), and their location and variants in the long-lived bowhead whale. Top: turquoise killifish variants, with the changed amino acid on the right. Bottom: uniquely mutated residues in the bowhead whale (Keane et al., 2015) mapped onto the turquoise killifish sequence. For the turquoise killifish, the strength of the functional effect is color-coded in the same manner as in [Figure S4A](#) (note that the structural analysis could not be performed for that protein).

## A Geographic origin of strains



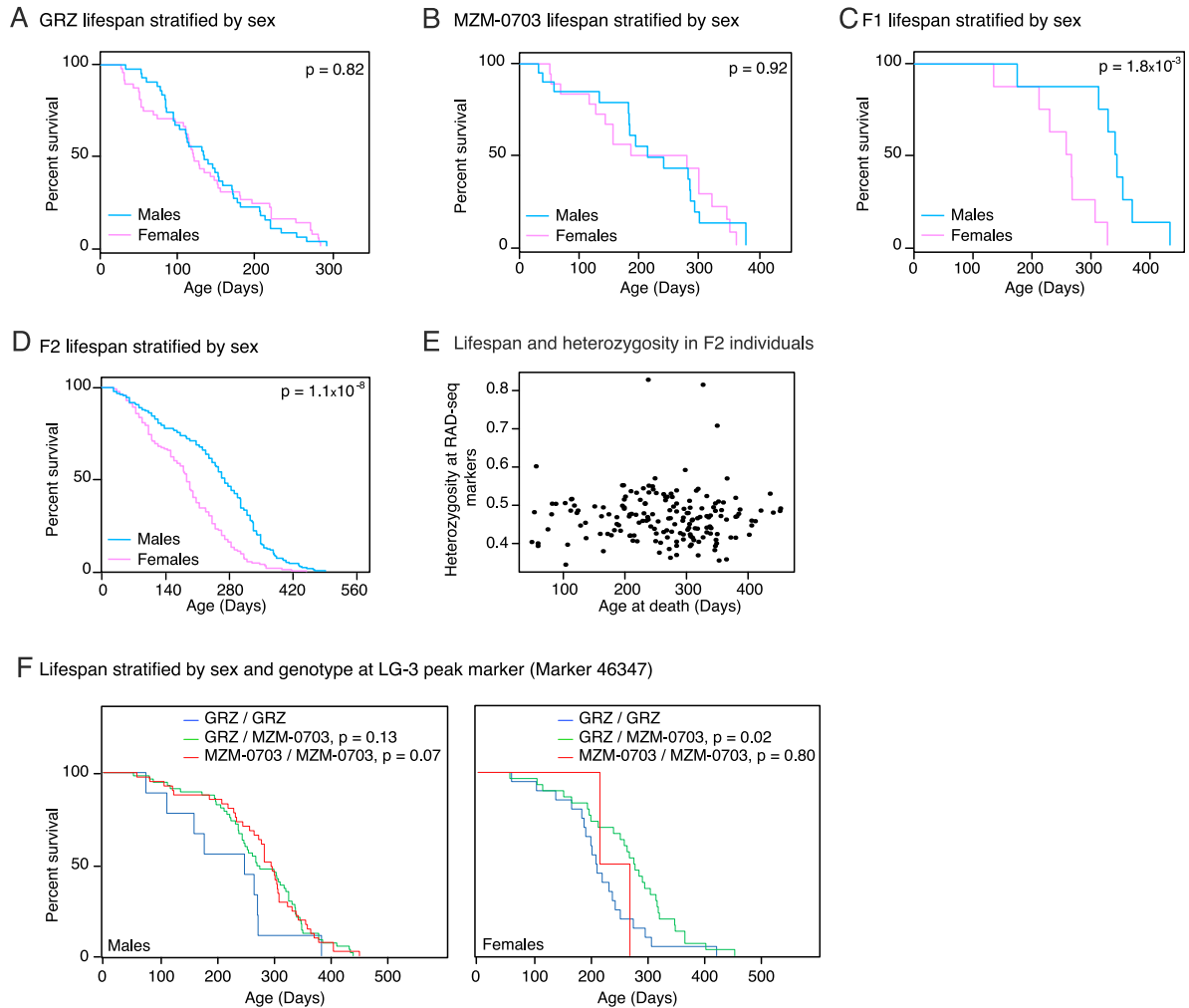
## B Summary of predicted effect of annotated SNPs



**Figure S5. Geographic Location of Different Strains of Turquoise Killfish and Genetic Variation in Individuals from Different Strains, Related to Figures 5 and 6**

(A) Geographic origin of turquoise killfish strains used for resequencing and genetic variation (Figure 5) and QTL mapping (Figure 6). Blue: shorter-lived strain in specific captive conditions. Red: longer-lived strains in specific captive conditions.

(B) Barplot with predicted functional effect of detected SNPs in resequenced individuals from the GRZ, MZM-0703, and MZM-0403 strains. As expected, most of the variation is in regulatory regions or in introns, with only a minority leading to non-synonymous coding changes (in red).



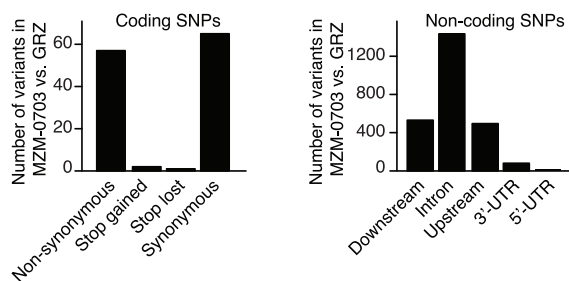
**Figure S6. Genetic Architecture of Lifespan Using Crosses between Shorter-Lived and Longer-Lived Strains of the African Turquoise Killifish, Related to Figure 6**

(A–D) Lifespan stratified by sex for GRZ parental strain (A), parental MZM-0703 strain (B), cross GxM F1 (C), and cross GxM F2 progeny (D).  $p$  values for differential survival between males and females in log-rank tests are indicated.

(E) Scatterplot of lifespan versus marker heterozygosity in cross GxM F2 progeny. Note that there is no global correlation between heterozygosity and lifespan, compatible with the absence of widespread hybrid vigor effect (where heterozygosity would be correlated to extended longevity).

(F) Lifespan stratified by sex and genotype in cross GxM F2 at the marker that is most significantly associated to the lifespan QTL (RAD-seq marker 46347).  $p$  values for differential survival compared to individuals with the GRZ/GRZ genotype in log-rank tests are indicated. See also Table S6B for complete statistical analyses.

**A** SNPs in MZM-0703 vs. GRZ in the region underlying the lifespan QTL (cross GxM P0)



**B** Candidate genes in lifespan QTL region with non-synonymous SNPs at conserved sites (Illumina sequencing)

Gene	GRZ aa	MZM-0703 aa	MZM-0403 aa	Consensus aa
<i>ATXN7L1</i>	G589	R589	R589	R/K
<i>GRN</i>	Q151, W449	H151, C449	H151, C449	H/F, C
<i>HIPK2(11of26)</i>	H218, K321	Y218, T321	Y218, T321	Y, T
<i>IFI35</i>	M196	L196	L196	L
<i>TTYH3A</i>	S423	T423	T423	T
<i>ZNF800A</i>	N489	T489	T489	T/S

**C** GRN mutiple protein alignment (turquoise killifish residue 151)

```

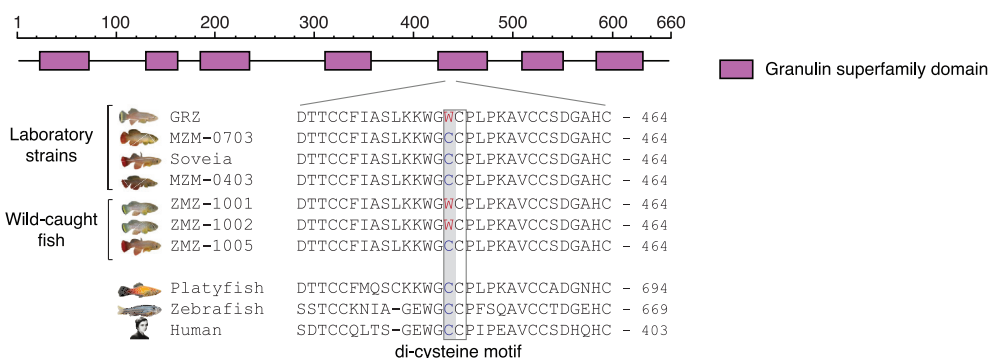
T. killifish CPDGKHCPEGQRCSNACHSCKK - 163
Platyfish CSDGKHCPEGHHCSDRSRSCIKK - 163
Medaka CPDGKHCPEGHQCSLDRSVCVK - 163
Fugu CPDSKTM----- - 107
Stickleback CSDGKHCPEGHQCSADRCRSCIKQ - 117
Zebrafish CSDGKHCPCNDHECSDSLSLCVKR - 162
Human CGDGHHCCPRGFHCSDAGRSCFQR - 115
    
```

**D** GRN mutiple protein alignment (turquoise killifish residue 449)

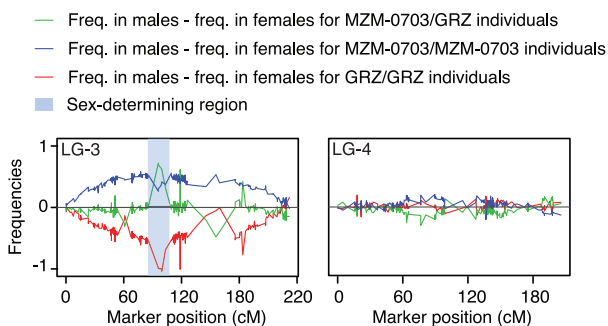
```

T. killifish DTTCCFIASLKKWGCPLPKAVCCSDGAHC - 464
Platyfish DTTCCFMQSKKWGCCPLPKAVCCADGNHC - 694
Medaka QTTCCKTQE-GGWGCCPFPEAVCCADGEHC - 644
Fugu GSTCCKMAS-GQWACCPLPEAVCCEDGDHC - 373
Stickleback SNTCCFMAESQKWGCCPLPKAVCCSDGNHC - 423
Zebrafish SSTCCKNIA-GEWGCCPFQAVCCTDGEHC - 669
Human SDTCCQLTS-GEWGCCPIPEAVCCSDHQHC - 403
    
```

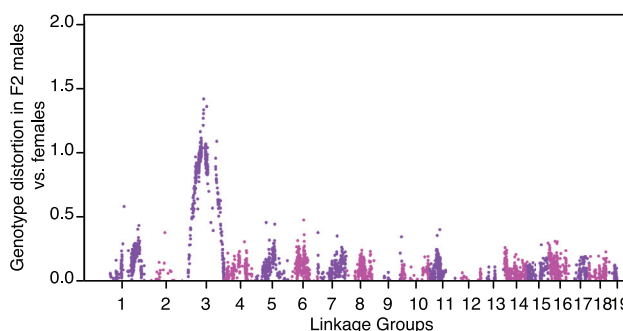
**E** GRN protein sequence in turquoise killifish strains, wild-caught fish and other species (turquoise killifish residue 449)



**F** Genotype frequency distortion and recombination



**G** Genotype frequency distortion in F2 males vs. females



**Figure S7. The Lifespan QTL Region Contains Genetic Variation between the GRZ and MZM-0703 Founders and Is Linked to the Sex-Determining Region, Related to Figure 7**

(A) Barplot depicting the predicted effect of SNPs in the region underlying the lifespan QTL. Note that there are both coding and non-coding variants between the GRZ and MZM-0703 founders (P0). The majority of annotated variants in the lifespan QTL region is not expected to impact protein sequence, but may have regulatory impact.

(B) Genes in the region underlying the lifespan QTL in LG-3 that have variants between the GRZ and MZM-0703 cross founders at conserved sites. Amino-acid variants are from next-generation sequencing genotyping in GRZ, MZM-0703, and MZM-0403 individuals. Only positions with similar variants in the MZM-0703 cross founder and in the resequenced MZM-0403 individual were considered. The line for the *GRN* gene is highlighted in red, because of *GRN* implication in aging and age-related diseases. See also Table S7A.

(legend continued on next page)

---

(C and D) Multiple alignment around residues with variants of interest at GRN residues 151 (C) and 449 (D). Orthologs in selected fish species and human are displayed. The residues of interest are shaded.

(E) Multiple alignment of GRN around the 449<sup>th</sup> amino-acid of the turquoise killifish protein sequence in laboratory strains, wild-caught fish, and selected vertebrate species. The alignments for the turquoise killifish, platyfish, zebrafish and human protein sequences are the same as in [Figure S7D](#). The turquoise killifish 449 amino-acid position is shaded in gray. The conserved amino-acid in the alignment is a cysteine (C), but GRZ carries a tryptophan (W). For wild-caught fish, the amino-acid corresponding to the major allele is reported (out of the five sequenced individuals from each geographical location). Full genotypes of individuals determined by Sanger sequencing are reported in [Table S7F](#).

(F) Measure of suppressed recombination by allelic distortion between the male and female F2 progeny at each marker on selected linkage groups. Genotype distortion is observed on LG-3, but not on other LGs. Values are reported for the frequency of males minus frequency of females carrying both alleles from the longer-lived grandparent (blue), one allele from the longer-lived and one from the shorter-lived grandparent (green) or both alleles from the shorter-lived grandparent (red). LG-4 is shown as a representative non-sex linked control linkage group.

(G) Manhattan plot of measure of suppressed recombination by allelic distortion between the male and female F2 progeny at each marker on the whole linkage map. Markers on different linkage groups are represented by random alternating colors. The frequency of males minus frequency of females carrying both alleles from the shorter-lived grandparent (e.g., corresponding to the red line in [Figure S7F](#)) is subtracted from the frequency of males minus frequency of females carrying both alleles from the longer-lived grandparent (e.g., corresponding to the blue line in [Figure S7F](#)).

Cell

Supplemental Information

**The African Turquoise Killifish Genome  
Provides Insights into Evolution  
and Genetic Architecture of Lifespan**

Dario Riccardo Valenzano, Bérénice A. Benayoun, Param Priya Singh, Elisa Zhang, Paul D. Etter, Chi-Kuo Hu, Mathieu Clément-Ziza, David Willemsen, Rongfeng Cui, Itamar Harel, Ben E. Machado, Muh-Ching Yee, Sabrina C. Sharp, Carlos D. Bustamante, Andreas Beyer, Eric A. Johnson, and Anne Brunet

## Supplemental Experimental Procedures

### Fish housing and husbandry

Fish were raised at 25°C in a centralized filtration water system at a density of up to 1 fish per 1.4L in 2.8L and 9L tanks. Fish were fed freshly hatched *Artemia* nauplii until 3 weeks of age and then dried bloodworm (*Chironomous sp.*) twice a day during the week and once a day during weekends. Adult fish spawned on a sand substrate in 2.8L and 9L tanks within the centralized filtration system. Dead fish were removed daily from the tanks, weighted, and stored in 50 mL of 100% ethanol. Embryos were collected on a weekly basis and plated on sterile dry peat moss until they were ready to hatch. Once ready to hatch, indicated by the presence of a distinct yellow iris in the eye and by continuous body twitching within the chorion, embryos were immersed in a 4°C Yamamoto embryo solution (17 mM NaCl, 2.7 mM KCl, 2.5 mM CaCl<sub>2</sub>, 0.02 mM NaHCO<sub>3</sub> pH 7.3) (Rembold et al., 2006) supplemented with peat moss extract and oxygen tablets. Hatched fry were placed in 0.2-gallon tanks at the density of 5 fry per tank and fed with brine shrimp nauplii.

### Koepfen-Geiger analysis of climate

Data on annual precipitation and temperature were collected in 102 meteorological stations from <http://en.climate-data.org/> and were used to compute a Koepfen-Geiger climate classification in southeastern Africa according to the classification by Peel (Peel et al., 2007).

### Next-generation sequencing of the turquoise killifish genome

Genomic DNA was isolated from tissues (muscle or tail) of 9 African turquoise killifish individuals (8 males and one female) from the inbred reference strain GRZ. The predominance of male individuals allows a more comprehensive survey of the genome of this species, because males are heterogametic in this species (Kirschner et al., 2012; Valenzano et al., 2009). Ten libraries with varying insert sizes were constructed for Illumina sequencing as indicated in the table below from 9 independent GRZ fish (one fish was used to build two libraries). Genomic sequences for de novo assembly were generated on Illumina HiSeq2000 instruments (Beijing Genome Institute, University of Oregon, and Stanford Center for Genomics and Personalized Medicine). Paired-end libraries were obtained as 2x101bp raw reads, and mate-pair libraries were obtained as 2x50bp raw reads. All genomic sequencing libraries have been deposited to SRA (SRP041421).

Library	Insert size (bp)	Strategy	Total QC length (bp)	Coverage (X) <sup>a</sup>	Center	Sex
BGI170	170	Paired-end	24,461,613,765	12.23	BGI	M
BGI500	500	Paired-end	27,538,490,563	13.77	BGI	M
GRZ340	340	Paired-end	9,892,583,577	4.95	Stanford	M
GRZ540	540	Paired-end	3,856,857,526	1.93	Stanford	M
RADGP0	200	Paired-end	1,892,127,538	0.95	Oregon	F
RADAAP0	200	Paired-end	1,032,838,324	0.52	Oregon	M
GRZ300	300	Paired-end	22,661,204,008	11.33	Stanford	M
GRZ400	400	Paired-end	22,531,661,933	11.27	Stanford	M
BGI2K	2,000	Mate-pair	32,648,404,824	16.32	BGI	M
BGI5K	5,000	Mate-pair	12,742,324,364	6.37	BGI	M
<b>Total</b>			159,258,106,422	79.63		

<sup>a</sup> coverage estimate based on a conservative genome size estimate of 2Gb (see below). Note that the total coverage estimate ranges from 72.4X (for the maximal genome estimate of 2.2Gb) to 122.5X (for the minimal genome size estimate of 1.3Gb).

## **Read quality filtering and trimming for de novo genome and transcriptome assembly**

Sequencing reads in all Illumina libraries for de novo genome and transcriptome assembly were quality filtered and trimmed using the trim\_galore software ([http://www.bioinformatics.babraham.ac.uk/projects/trim\\_galore/](http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/)), with a Phred score threshold of 30 and a minimum remaining read length of 50bp in either read of the pair after trimming. Due to nucleotide composition biases at the beginning of sequencing reads, all reads were also further trimmed of their first four most 5' bases using fastx\_trimmer ([http://hannonlab.cshl.edu/fastx\\_toolkit](http://hannonlab.cshl.edu/fastx_toolkit)).

## **Genome size estimate from Illumina sequencing libraries**

Genome size was estimated using two independent methods. The first method was based on (Li et al., 2010). Briefly, 25-mer frequencies were counted using the Jellyfish software (Marcais and Kingsford, 2011) and the corresponding maximum k-mer frequency was graphically determined. Genome size was estimated using the count of base pairs in the used reads, the mean length of the quality-trimmed reads, and the maximum k-mer frequency. With this method, the turquoise killifish genome size was estimated to be 1.9-2.2Gb, taking into account three independent libraries (GRZ300, GRZ340 and GRZ400).

The second method used the preqc module in SGA (Simpson, 2014). This method also takes a k-mer frequency approach, but accounts for sequencing error rates, potential heterozygous sites, and effect of repeat sequences (Simpson, 2014). The preqc module (SGA v 0.10.13) was run on three independent libraries (BGI170, GRZ300 and GRZ340). The genome size was estimated to be 1.3-1.6Gb by this method. Thus, the computational estimate of the turquoise killifish genome size ranges from 1.3-2.2Gb. This is consistent with (Reichwald et al., 2009). Based on this range, the percentage of genome assembled is 83.1% to 49.1%, respectively. In the manuscript, we use a conservative genome size estimate of 2Gb (~54% of genome assembled, with ~80 fold coverage).

## **De novo genome assembly with SGA and SOAPdenovo**

The overlapping 170bp Illumina paired-end library was preprocessed to obtain bona fide longer sequence fragments using FLASH (Magoc and Salzberg, 2011). The length distribution of output fragments was compatible with the library specifications. SGA v0.9.19-10 (string graph assembler) (Simpson and Durbin, 2012) was applied to the filtered pre-processed Illumina paired-end libraries to obtain a high-quality master assembly. We chose SGA as our main contig assembler because of its lower reported rate of misassemblies compared to other assemblers (Salzberg et al., 2012). Final assembly parameters in SGA were a correction k-mer of 51, an overlap length of 65bp in the string-graph, and a merging overlap length of 75bp to generate the contig assembly. Parameters to the sga-assemble function were set to “-d 0.2 -g 0.1 -r 10” to account for potential indels and SNPs in the paired-end (PE) libraries (constructed from 7 independent GRZ individuals, 6 males and 1 female), in the range recommended in the software instructions.

A second assembly was obtained using the de Bruijn graph assembly SOAPdenovo V1.05 (Luo et al., 2012), with an assembly k-mer of 81 and using the error-corrected reads from the sga pipeline, requiring support of 5X coverage and minimum assembled length of 200bp for contig reporting.

## **Scaffolding, GapFilling and assembly reconciliation**

Scaffolding was performed on the SGA and SOAP assemblies using the SSPACE Basic v2.0 scaffolder (Boetzer et al., 2011). All paired-end and mate-pair libraries (10 libraries total) were inputted in hierarchical size order, requiring support of  $\geq 5$  independent pairs to create a scaffold link. Parameters were set to: no contig extension, a 0.7 maximum link ratio, only 1bp allowed gap for alignment, and reporting only contigs longer than 200bp. Gap-filling was conducted on the scaffolded assemblies using GapFiller v1.10 (Nadalin et al., 2012) inputting all paired-end and mate-pair libraries in hierarchical size order. Parameters were set to: a minimum overlap with the gap of  $\geq 31$ bp, a minimum of 5 supporting reads, a



minimum of 15bp overlap to merge sequences of a closing gap, a 15bp trimming, a 50bp gap-close difference, a base ratio of 0.7, all over 10 iterations.

Because of different biases associated with each assembly algorithms (Earl et al., 2011; Salzberg et al., 2012), assembly reconciliation was performed using the SGA assembly as ‘master’ assembly and the SOAP assembly as ‘slave’ using GARM v0.7 (Soto-Jimenez et al., 2014). Assembly reconciliation is known to improve upon individual assemblies by leveraging their different strengths (Soto-Jimenez et al., 2014; Yao et al., 2012). Gapfilling and scaffolding were run again on the output to refine unresolved regions. The reconciled assembly had increased contiguity over the starting SGA or SOAP assemblies, as measured by a higher N50 statistic (118kb vs. 66kb or 31kb, respectively), a lower number of scaffolds (46,729 vs. 565,630 or 203,468, respectively). The reconciled assembly also had a higher number of complete core eukaryotic genes according to CEGMA (Parra et al., 2007) (223 vs. 219 or 217, respectively; see below). The final assembly has been deposited in Genbank under accession number JNBZ00000000 (first version, referred to as NotFur1 assembly). The turquoise killifish genome browser is at <http://africanturquoisekillifishbrowser.org>.

### **Assessment of completeness of the draft genome using CEGMA**

The CEGMA algorithm (Parra et al., 2007) was applied to the turquoise killifish draft genome to gain an estimate of the completeness of the genome as well as initial gene models for these core eukaryotic genes (CEGs) (*i.e.* proportion of a conserved eukaryotic core genes present in the draft genome).

### **De novo transcriptome assembly using Oases**

Strand-specific RNAseq libraries from GRZ adult individuals were constructed from liver, brain, testes and tail tissues using the standard Illumina protocol. The liver and testes libraries were polyA-selected, and the brain and tail libraries were rRNA-depleted (Harel et al., 2015). Libraries were sequenced on Illumina HiSeq2000 machines, as paired-end 101bp reads. After read quality preprocessing (see above), libraries were run through FLASH to provide extended merged reads as well as unmerged remaining paired reads. De novo transcriptome assembly was performed using the Oases de Bruijn graph-based algorithm (oases v0.2.06, and velvet v1.2.03 dependency) (Schulz et al., 2012). Assemblies were generated using a k-mer range of 43 to 91 and a step of 4, and keeping only assembled sequences longer than 200bp with supporting evidence of  $\geq 5X$  coverage. Transcriptomes from the four different tissues were assembled separately. Resulting tissue-specific assemblies were then merged, and redundantly assembled transcripts were eliminated using uclust (Edgar, 2010) and cdhit-est cluster (Li and Godzik, 2006), asking for  $\geq 90\%$  reciprocal sequence homology. Resulting transcripts constituted our reference transcriptome assembly.

### **Higher-order scaffolding using RNA-seq data and RAD-seq linkage map**

The raw reads from our Illumina paired-end RNA-seq libraries were used for higher-order scaffolding over potentially unresolved introns. The tophat-fusion pipeline (Kim and Salzberg, 2011) was used to identify read pairs mapping over 2 different scaffolds. Mapping was run for each library as: ‘tophat2 --bowtie1 --fusion-search -m 1 -g 5 --fusion-multipairs 1 -o ./OUTPUT-DIR/ --solexa-quals -p 4 -r 100 --mate-std-dev 50 --no-coverage-search --no-mixed --segment-length 55 index\_name RNAseq\_file\_1.fq RNAseq\_file\_2.fq’. The ‘fusion.out’ output file was parsed to filter pairs to retain only putative high-confidence links between scaffolds (more than 20 supporting pairs and no contradicting pairs, or more than 50 supporting pairs and less than 1% contradicting pairs). Scaffolding was performed using high-confidence putative links between scaffolds and by stringently filtering out links that could be ambiguous or hard to resolve (e.g., links potentially resulting from alternative splicing). The orientation determined by Tophat for the fusion according to the RNA-seq mapping (forward-forward, reverse-reverse, forward-reverse, reverse-forward) was used to orient the scaffolds with respect to one another. An AGP format file ([https://www.ncbi.nlm.nih.gov/assembly/agg/AGP\\_Specification/](https://www.ncbi.nlm.nih.gov/assembly/agg/AGP_Specification/)) was generated to summarize the scaffold links and placement, and is available for download on the genome browser website.

To perform even higher-order scaffolding on the RNA-seq scaffolded assembly, the linkage map from cross GxM (see below) was used. First, RAD-seq markers were mapped in fasta format to the genome using bowtie1. The alignment file was parsed to extract information on captured genomic scaffolds and on their relationships with mapped markers to build a scaffolding roadmap. During parsing, markers that led to contradictory or ambiguous links that could not be resolved were discarded. A limit resolution confidence threshold of 5cM was used to resolve conflicts arising between assembled genomic contigs and mapping sequence: below 5cM, the genomic order was preferred; above 5cM, the genetic map was preferred. Captured scaffolds were then ordered according to the genetic linkage map using an AGP format output. The rest of the scaffolds were considered 'unplaced' in the assembly and annotated as such in the final AGP output file. This file is available for download on the genome browser website.

### **Estimate of unresolved repeats in draft assembly**

Genomes with a high-repeat content are known to yield relatively fragmented drafts when assembled using short-read sequencing technology (Treangen and Salzberg, 2012). A previous assessment estimated that the turquoise killifish genome contained at least 45% of repetitive sequences (Reichwald et al., 2009). We examined regions of the genome with an excess fold coverage compared to the expected coverage from the libraries (Figure S1A). This was done based on the conservative genome size estimate of 2Gb. A small portion of the assembly captured a large portion of the reads. For example, regions that have 10 times the expected coverage (i.e. mean coverage based on depth of sequencing and used genome size estimate) represent 1% of the genome assembly but capture 17.4% of reads (Figure S1A). Regions with excess coverage likely correspond to unresolved repeats that are present in many copies in the actual genome sequence (Treangen and Salzberg, 2012).

### **Analysis of repetitive sequences in draft assembly**

The annotation of repetitive elements present in the assembly was performed using RepeatMasker v3.3.0 (Smit et al., 1996-2004) and the RepBase repeat library (2012-04-18 version) (Jurka et al., 2005). We used RMBlast version-2.2.27 as an alignment engine, and restricted the similarity search to the library of elements from teleost fishes.

### **Mapping of Sanger shotgun sequence contigs to the draft assembly**

Previously published Sanger shotgun contigs from GRZ genomic DNA were obtained from GenBank (accession ABLO01) (Reichwald et al., 2009). Contigs were aligned to the draft assembly using BLAT (Kent, 2002). The longest alignment of the contig to the genome (the best match) was then used to compute the aligned fraction of each Sanger contig.

### **Annotation of protein-coding genes using the MAKER2 pipeline**

The MAKER2 pipeline was used to generate consensus gene predictions derived from ab initio predicted models, RNA-seq reads, de novo transcriptome assembly and EST/transcript data, and protein similarity (Holt and Yandell, 2011). MAKER2 predicts the most likely gene model and outputs a confidence score (Annotation Edit Distance, AED) to each prediction based on the degree of support by experimental evidence (EST/transcript mapping, RNA-seq mapping, protein homology, etc.). Several sources of transcriptomic sequences were generated to support gene predictions. These include: 1) published assembled transcript sequences from the turquoise killifish downloaded from Genbank (Petzold et al., 2013) and ESTs from another killifish, *H. fondulus* (Tingaud-Sequeira et al., 2013), 2) our de novo assembled transcriptome sequences, and 3) our paired-end Illumina RNA-seq data from brain, liver, testes and tails as well as previously published RNA-seq libraries from skin and whole fish (Petzold et al., 2013). RNA-seq data were aligned to the masked reference genome and exon-exon junctions were modeled using the tuxedo suite (Trapnell et al., 2012), as recommended in the MAKER2 manual. The full-length transcriptome data and splice junctions were used as transcript evidence in the MAKER pipeline. In addition to the transcript sequences, the complete reference proteomes of *Danio rerio*, *Oryzias latipes* and

*Takifugu rubripes* were downloaded from Uniprot (on 05-30-2013) to provide the MAKER pipeline with protein homology evidence. The MAKER pipeline also incorporates a repeat masking step before running gene predictors, and the *teleostei* repeat library from the RepBase database (2012-04-18 version) (Jurka et al., 2005) was used for this step (see above). The discovery of single-exon genes was enabled, though it is disabled by default in the pipeline, to allow for the discovery of potentially important mono-exonic genes. MAKER aligns transcript and protein evidence to the genome using the sensitive/specific splice site-aware alignment algorithm exonerate (v2.2.0) (Slater and Birney, 2005). Alignments retained by the MAKER pipeline have > 20 score, do not overlap low-complexity regions, and were used to estimate the proportion of transcripts mapping to the assembly with high quality.

The ab initio predictor SNAP (Semi-HMM-based Nucleic Acid Parser) was first trained specifically for the turquoise killifish using CEG models from the output of CEGMA (Parra et al., 2007). After a first run of the MAKER pipeline, gene models with an AED score of 0 (most supported by RNA or protein evidence) were then used to retrain SNAP for a second round to obtain a higher quality hidden Markov model. Ab initio predictor Augustus was then trained using AED=0 gene models from this second run, using the included zebrafish model parameters as a starting point to create a turquoise killifish-specific gene prediction model. We then performed a third and last run of the MAKER pipeline with all evidence support, enabling gene prediction from transcript sources (i.e. exonerate alignment of transcripts and Cufflinks junctions) and from trained ab initio gene predictors (i.e. Augustus and SNAP). In this final run, 61,418 putative protein-coding gene models were predicted by MAKER at any AED threshold (0-1). Gene model filtering steps and annotation of protein coding gene models by sequence orthology are described below.

### **Analysis of alignment of turquoise killifish transcripts and RNA-seq to the draft genome**

Long assembled transcripts from a published catalog (Petzold et al., 2013) and from our de novo assembled transcriptome were used to assess the quality of our genome assembly. Transcript alignments were performed and filtered based on quality using MAKER, which uses exonerate (v2.2.0), an alignment algorithm that is sensitive, specific, and splice site-aware (Slater and Birney, 2005). Alignments retained by MAKER have > 20 score and do not overlap low-complexity regions. They were used to assess the proportion of transcripts mapping with high-quality to the genome assembly. The remaining transcripts that were not part of this high-quality set were then mapped to the genome using BLAT (using the -trimHardA -trimT -extendThroughN options), and transcripts with at least one partial hit to the genome were quantified.

Illumina paired-end RNA-seq reads from adult turquoise killifish tissues (SRP041421) were aligned to our reference draft genome using bowtie 0.12.7 and TopHat2 v2.0.4 (Trapnell et al., 2012). Alignments were retained only if they mapped to at most three loci in the assembly (-g 3). Statistics of percentage of mapped reads and proper pair orientations were obtained from the resulting alignment files using samtools (v 0.1.17).

### **Alignment of transcripts to proper paralogs for specific genes**

For selected genes, we determined independently proper mapping of transcripts to different paralogs (Table S4I). In these cases, the corresponding annotated transcripts from the published catalog of assembled turquoise killifish transcripts (Petzold et al., 2013) were downloaded from the NFIN website (<http://nfintb.fli-leibniz.de/nfintb/>). Sequences were mapped using BLAT (-trimHardA -extendThroughN -fine options), and alignments were visualized using our genome browser (Table S4I).

### **Annotation of ribosomal RNA genes and short non-coding RNA genes**

RNAmmer (Lagesen et al., 2007) was used to annotate high-quality rRNA genes. tRNAscan-SE (Lowe and Eddy, 1997) was used to annotate high-quality tRNA genes. Putative tRNA with a non-conventional anticodon, or labeled as likely pseudogenes by the software, were discarded from the high-confidence tRNA gene predictions. Annotated tRNA gene type and distribution per anticodon are reported in Table S1A. Infernal (Nawrocki and Eddy, 2013) was used to annotate putative snRNA, snoRNA and miRNA

genes. All programs were run with default settings. RepeatMasker also provided secondary predictions of rRNA, tRNA and snRNA genes, and Infernal provided secondary predictions of tRNA genes.

### **Annotation of long non-coding RNA genes**

Illumina paired-end RNA-seq reads from adult turquoise killifish tissues (SRP041421) were aligned to the reference draft genome using bowtie 0.12.7 and TopHat2 v2.0.4 (Trapnell et al., 2012). Alignments were retained only if they mapped to at most three loci in the assembly (-g 3). De novo transcriptome assembly guided by the genome was performed using cufflinks2 v2.1.1, using the predicted protein-coding gene as prior information, and with a maximum pre-mRNA fraction parameter set at 0.1 (-j 0.1). Previously identified transcripts from predicted protein-coding genes were excluded from subsequent steps.

Next, the EMBOSS software suite was used to predict the longest ORF in each of the remaining predicted transcripts (Mullan and Bleasby, 2002). Transcripts longer than 150bp, with at least 5X sequencing coverage in one library, and whose longest predicted ORF was strictly shorter than 50 amino-acids, were retained for further processing. Finally, as a stringent cutoff, we filtered the putative lncRNA genes to retain only those with an H3K4me3 peak, indicative of promoter activity, in an H3K4me3 ChIP-seq dataset from turquoise killifish adult brain (SRP045718) (Harel et al., 2015).

### **Ortholog identification and gene model annotation**

To further annotate protein-coding genes from all the 61,418 MAKER gene predictions and exclude spurious predictions and untranslated sequences, we used homology-based evidence from 19 fully sequenced genomes (Table S2B), including seven additional fish genomes. Protein sequences for all the organisms were downloaded from Ensembl (release 75) (Cunningham et al., 2014) using BioMart (Kinsella et al., 2011). The sea urchin proteome was downloaded from Ensembl Metazoa. For genes with multiple protein products due to alternative splicing, only the longest protein was used.

An all-against-all BLASTp search was run using an e-value of  $10^{-5}$ . Both best and bidirectional best hits for every turquoise killifish protein in each of the analyzed genome were determined. In addition, a BLASTp search against NCBI nr (all non-redundant protein sequences) was performed using all the predicted turquoise killifish proteins. We discarded 32,924 predicted turquoise killifish genes whose protein product did not have either a match in at least two organisms or a match in NCBI nr with at least 30% query coverage, leading to a final set of 28,494 killifish protein coding genes. These genes were divided into three tiers of confidence levels based on the homology evidence from all of the other genomes. Turquoise killifish genes with bidirectional best hits in at least 10 analyzed organisms were considered Tier-1 (10,329 genes). Turquoise killifish genes with bidirectional best hits in less than 10 organisms, but with homologous sequences in at least 10 genomes, were considered Tier-2 (12,192 genes). Turquoise killifish genes with a BLASTp hit in less than 10 organisms, and a best hit with NCBI nr database, were considered Tier-3 (5,973 genes). Tier-1 and Tier-2 genes represent the high quality gene prediction in the current assembly, and are considered high-confidence protein coding genes (Figure 2B).

Turquoise killifish genes from all the three tiers were assigned a gene symbol based on the consensus symbol from all the genomes having a homologous sequence. In cases where a consensus gene symbol could not be reached, preference was given to the gene symbols supported only by the 7 teleost fish genomes. If still no consensus gene symbol could be identified, the gene symbol was chosen from organisms in the following order of priority: human, mouse, zebrafish, and medaka. For multiple genes with the same gene symbol (possible gene duplicates), a number was assigned randomly to each ('NofN'). Finally, an ID was assigned to each protein (Figure S1G) and long non-coding RNA genes (Figure S2E). This ID contains information about the scaffold, synteny, and number of organisms having a homolog of the gene, or for the lncRNA, the tissues where the gene is expressed.

### **Gene family analysis**

To construct gene families in turquoise killifish and medaka, an all-against-all BLASTp was run using all the filtered turquoise killifish and medaka proteins (e-value  $< 10^{-5}$ ). The hits were then clustered at 5

similarity thresholds (50 to 90) to generate gene families using TransClust (Wittkop et al., 2010). Using the resulting families with at least one gene from both the organisms, we performed linear regression without an intercept term in R. As a control, an identical analysis was performed between platyfish and medaka proteins. A good correlation genome-wide between gene families indicates proper assembly of paralogs. Differences in the paralog numbers of specific families between killifish and medaka could be due to evolutionary events such as lineage specific gene duplication or to loss misassembly/misannotation in one of the species.

### **Computation of codon usage**

All high-quality annotated protein-coding genes (Tiers 1, 2 and 3) were used to estimate codon-usage in the turquoise killifish. The R (<http://cran.r-project.org>) package 'seqinr' (Charif and Lobry, 2007) was used to compute codon usage from the corresponding coding sequences (Table S1B).

### **Metagene profiles for RNA-seq and metapromoter profile for H3K4me3 ChIP-seq**

For metagene analysis of RNA-seq, normalized aligned read counts were extracted around the annotated TSSs (-tss option), gene bodies (-rna), or TTSs (-tts) using the 'annotatePeaks.pl' script from the HOMER suite (Heinz et al., 2010) with '-hist' option and with the final annotation gff3 file for the turquoise killifish genome. Average values are used to plot the metagene profiles.

For metapromoter analysis of H3K4me3 ChIP-seq, normalized aligned read counts were extracted around the annotated TSSs (-tss option) using the same method as above. Average values are used to plot the metapromoter profile.

### **Analysis of RNA-seq data for transposase expression**

Our de novo assembled transcriptome was used to obtain transcript reference sequences derived from independent transposon insertions of the same family. We mapped reads to the transcriptome instead of the genome because genome-based quantification of RNA expressed from transposable elements is problematic due to ambiguous mapping of reads to the genome. Transposon-derived transcripts were annotated using BLASTp hit to the NCBI nr database or using predicted domains from the NCBI conserved domains database (<http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>). The RNA-seq reads were mapped to the reference transcriptome using Tophat2, supported by bowtie1 (Trapnell et al., 2012). Aligned reads over transcripts were counted using the bedtools suite 'coverageBed' command. Read counts were normalized by transcript length and library size to obtain final FPKM values.

### **Synteny analysis**

To build whole genome synteny maps, genomic scaffolds were assigned to specific RAD-seq markers along the linkage map using BLAT. Linkage groups for the turquoise killifish were then represented as a linear series of scaffolds matching the RAD-seq markers (not oriented). Coding sequences matching each mapped scaffold were then extracted using a GTF file of the turquoise killifish annotated genes, and a new file was built containing gene names along the mapped scaffolds. Medaka, stickleback, and platyfish GTF files were obtained from <ftp://ftp.ensembl.org/pub/>. The ordered linkage groups in the turquoise killifish were matched to the medaka and platyfish chromosomes based on reciprocal best BLASTp hits (turquoise killifish/medaka synteny), or on identical gene names (platyfish/medaka synteny). High quality protein genes from Tier 1 and Tier 2 were used for the turquoise killifish. Matching pairs of genes between species were plotted using either the turquoise killifish linkage map position or the platyfish chromosome position on the y-axis and the medaka chromosome position on the x-axis (OxGrid plot).

## Construction of the species tree

For the generation of the species tree, we selected all the 619 one-to-one orthologs from each of the organisms where the same killifish gene was the reciprocal best BLASTp hit in each of the 19 animal genomes. These genes represent the entire set of high confidence one-to-one orthologs for the 20 organisms, including turquoise killifish, in our analysis (Table S2A). Known aging-related genes are well represented in this list (see Table S2C). A single long sequence for each organism was constructed by concatenating these 619 proteins in the same order. The sequences were then aligned using MAFFT (Katoh and Standley, 2013). The conserved blocks were identified from the corresponding multiple sequence alignment to remove non-conserved or misaligned regions using Gblocks (Talavera and Castresana, 2007). ProtTest (Darriba et al., 2011) was used to identify the best model (LG+G+I) for construction of phylogeny. A maximum likelihood tree was then constructed using PhyML v3.1 with 100 bootstrap steps for statistical support (Guindon et al., 2010). We used discrete Gamma model with 4 categories (shape parameter: 0.821) and allowed PhyML to estimate the proportion of invariant sites (0.066). The resulting unrooted tree was rooted in MEGA-6 based on *C. elegans* as the outgroup. Phylogeny based on neighbor-joining (Tamura et al., 2013) produced identical topology.

## Phylogenetic trees for specific gene families

The predicted protein sequence corresponding to a specific turquoise killifish gene was used to identify best BLASTp hits in the proteomes of 7 fish species. Hits for medaka, stickleback, *Tetraodon*, *Takifugu*, cod and zebrafish were obtained from the UCSC/ENSEMBL portals if available; if not, the best BLASTp hits from NCBI nr database were used. The best hits for platyfish protein sequences were always obtained using BLASTp hits against NCBI nr.

Protein sequence alignment was performed using ClustalX v2.1. PHYLIP proml v3.695 was used to build trees, using zebrafish sequences as the outgroup. The plots were generated as rooted phylograms using Unipro UGENE v1.17.0.

## Identification of turquoise killifish genes under positive selection

To identify genes in the African turquoise killifish lineage that are under positive selection, we used PAML (Phylogenetic Analysis by Maximum Likelihood, version 4.8) (Yang, 2007; Yang and Nielsen, 2002), which implements a maximum likelihood framework to evaluate adaptive selection based on non-synonymous by synonymous substitution rate ratio ( $K_A/K_S$ ,  $D_N/D_S$  or  $\omega$  ratio). We used a branch-site model implemented in PAML to identify individual amino-acid sites targeted by positive selection in the turquoise killifish branch using 7 other long-lived fish species as a background (Figure S3A).

First, single ortholog gene families were selected from the eight teleost fish genomes (including the turquoise killifish) where a clear bidirectional best hit with the same turquoise killifish gene was present in at least four other fish genomes. Hence, we required at least 5 fish sequences including the turquoise killifish for further analysis, leading to a set of 13,637 ortholog families. These sequences were aligned using PRANK (Loytynoja and Goldman, 2005) (codon model incorporated in GUIDANCE (Penn et al., 2010)), an algorithm known to generate highly accurate alignments to detect positive selection (Fletcher and Yang, 2010; Jordan and Goldman, 2012). Since alignment quality is one of the critical steps in accurate detection of positive selection, we applied stringent filtering using GUIDANCE (Penn et al., 2010), which is known to be a highly accurate alignment quality-control package for such analysis (Jordan and Goldman, 2012). The following (local and global) stringent filters were applied on each alignment using GUIDANCE: mean residue pair score > 0.85 (indicative of the overall alignment quality), mean column score > 0.85 (indicative of the overall alignment quality), no individual sequences with score < 0.85 (indicative of the sequences that may lead to bad alignment quality) and no sequence-pairs with score < 0.85 (indicative of the sequence pairs leading to bad alignment).

For the alignments that passed our quality filters, we used the branch-site model implemented in CODEML that is designed to detect positive selection that affects only a few sites on the specified branches of a phylogeny. On the species tree of the eight fish genomes, the turquoise killifish lineage was marked as 'foreground' and the rest of the fish species as 'background' lineages (Yang, 2007; Yang and Nielsen,

2002). We then performed a likelihood ratio test between model M2a\_null (model = 2, NSsites = 2; fix\_omega = 0) and M2a\_selection (model = 2, NSsites = 2; fix\_omega = 1, omega = 1) as recommended (PAML User Guide: <http://abacus.gene.ucl.ac.uk/software/pamlDOC.pdf>).

A p-value to assess the significance of the likelihood ratio test was calculated using  $\chi^2$  test for twice the difference of likelihood from model M2a\_selection versus likelihood M2a\_null, with one degree of freedom. We used Bayes Empirical Bayes (BEB) probabilities provided by CODEML to identify individual sites under selection (Yang et al., 2005). However, we further excluded the selected sites if the GUIDANCE column score for the site under selection was less than 0.85 or if there was a gap in the alignment (in any sequence) within +/- 5 amino-acids from the selected site, even if the BEB probability value was significant. This stringent filter was introduced to ensure that the sites under selection are in the well-aligned regions. Finally, we subjected the p-value of the likelihood ratio test to multiple hypothesis correction in R by Benjamini–Hochberg method and identified the genes with significant overall p-value at 5% FDR ( $p < 0.05$ ). There were 497 genes under positive selection with a corrected overall p-value  $< 0.05$  (Table S3B). To generate an even more stringent list, we restricted the genes to those with at least one site with BEB probability of selection  $> 0.95$ , which led to 249 genes (Table S3A). The ‘highest confidence’ list of 249 genes was used for GO enrichment analysis, functional effect prediction and overlap with genes under selection in extremely long-lived vertebrates. The list of 497 genes was used to assess the overlap with aging genes in vertebrate model organisms from the GenAge and LongevityMap databases.

While we designed our analysis to identify the genes and residues under positive selection with high confidence, there may be additional genes/sites under selection that are missed or some false positives due to inherent limitations of this kind of analysis (e.g. misalignment of codons, choice of background species, missing exons due to alternative splicing, parameter sensitivity of likelihood ratio calculations, etc.). In general, the false positive rate of branch-site tests ranges from 4 to 6.4% (Wong et al., 2004; Yang and dos Reis, 2011).

### **GO enrichment analysis of genes under positive selection**

Gene Ontology (GO) terms for the predicted killifish genes were obtained by attributing the corresponding GO terms of zebrafish genes to their one-to-one orthologs in the turquoise killifish genome. GO enrichment analysis for the stringent list of 249 genes under positive selection was performed in R (version 3.1.1) using ‘GStats’ package (Falcon and Gentleman, 2007). We used all the filtered 13,637 genes as background and employed hyper-geometric test implemented in GStats to obtain the significantly enriched terms after Benjamini-Hochberg correction for multiple testing, and filtering out the terms that showed significant depletion from the results. Selected GO terms with p-values, number of genes, and enrichment values  $\geq 2$  fold from the top enriched GO terms are shown in Figure 3C.

### **Overlap between genes under positive selection in the turquoise killifish and genes that change in expression with age**

For analysis of the publicly available brain aging dataset (MZM-0410 strain) (Baumgart et al., 2014), the STAR ultrafast universal RNA-seq aligner (Dobin et al., 2013) was used. Reads over gene models were counted using the featureCounts feature of Subread package (Liao et al., 2014). Library size adjustment and dispersion normalizations were performed using the R ‘DEseq2’ package (Love et al., 2014). DESeq2 was used to model expression changes as a function of the age of the fish (5, 12, 20, 27 and 39 weeks). Differentially expressed genes according to that model were called at FDR threshold of 5%. The list of differentially expressed genes was overlapped with the list of 249 positively selected genes in the turquoise killifish. Enrichment was measured by a Fisher test on expressed genes. Expression levels of genes of interest were plotted as a heatmap using the ‘pheatmap’ package in R (<http://cran.r-project.org/web/packages/pheatmap/index.html>).

## Turquoise killifish orthologs of aging-related genes from GenAge and LongevityMap

Aging-related genes in mouse or human were downloaded from the GenAge database (Built 17) (de Magalhaes et al., 2009; de Magalhaes and Toussaint, 2004). Genes with longevity variants in human were obtained from the LongevityMap database (Budovsky et al., 2013) after excluding all the non-significant variants. We used the turquoise killifish orthologs from either the mouse or human genes in the all-against-all BLASTp analysis from all 20 organisms, as described above. These lists are included in Table S4A.

## Prediction of functional impact of turquoise killifish variants

To determine the residues under positive selection in the turquoise killifish or the non-synonymous variants between killifish strains that have a functional impact, we used two sequence-based prediction algorithms: PROVEAN (Protein Variation Effect Analyzer) (Choi et al., 2012) and SIFT (Sorting Intolerant from Tolerant) (Kumar et al., 2009). For the residues under positive selection in the 249 genes that had a medaka ortholog in our analysis, we used the medaka protein sequences as reference, and evaluated the potential impact of their substitution to the turquoise killifish specific residue. We then calculated the SIFT score using SwissProt database and PROVEAN score using the NCBI nr database. A replacement was classified as 'DELETERIOUS' by SIFT if the prediction score is  $< 0.05$ . For PROVEAN, a score  $< -1.3$  was considered to have a 'Moderate' effect and a score  $< -2.5$  was considered to have a 'Deleterious' effect. After removing potential false positives, we could obtain a functional prediction for 1509 residues under positive selection in 199 genes from SIFT or PROVEAN (Tables S3D and S4D).

To have a neutral reference during the prediction of functional effect using SIFT and PROVEAN, we also predicted the effect of residues under positive selection using the ancestral sequences for the 7 aging genes [BAX, IGF1R(1of2), INSRA, IRS1(2of2) XRCC5, LMNA(3of3) and MGAT5(1of3)]. The sequences of the common ancestors of medaka, platyfish and killifish were generated using maximum-likelihood approach in PhyloBot (<http://PhyloBot.com>) using zebrafish/cod/stickleback as outgroups. For LMNA(3of3) and MGAT5(1of3), medaka was not a part of our analysis because there was no bidirectional best hit to a medaka ortholog. Therefore, in these two cases, we used the sequence of the common ancestor of platyfish, killifish, fugu, and tetraodon. We then computed the functional impact of the killifish residues at the corresponding position using ancestral sequence from both SIFT and PROVEAN (Table S4E).

To determine the potential functional impact of non-synonymous variants in aging-related genes between killifish strains, we used either the GRZ variant or the variant that is common between MZM-0403 and MZM-0703 ('MZM') as reference points, with the other serving as alternative sequence. This yielded 4 predictions for each considered position: effect of the MZM residue in the GRZ protein sequence context according to SIFT (1) or PROVEAN (2), and effect of the GRZ residue in the MZM protein sequence context according to SIFT (3) or PROVEAN (4). We considered that there was a predicted functional impact if at least one of these predictions was significant. The impact was considered higher confidence if both SIFT and PROVEAN predicted an impact on protein function.

For a subset of aging and age-related genes (those for which we could map the residues under positive selection on the available protein structure, e.g. BAX, IGF1R(1of2), INSRA, XRCC5, and GRN), we used five different methods to assess protein folding/stability changes upon point mutations: ENCoM: <http://bcf.med.usherbrooke.ca/encom.php> (Frappier et al., 2015), PoPMuSiC 3.1: <http://dezyme.com/> (Dehouck et al., 2011), I-Mutant v3.0: <http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi> (Capriotti et al., 2006), DUET: <http://bleoberis.bioc.cam.ac.uk/duet/> (Pires et al., 2014), and CUPSAT: <http://cupsat.tu-bs.de/> (Parthiban et al., 2006). All these methods model point mutations on protein structure and determine if the free energy of folding ( $\Delta\Delta G$ ) is significantly different between the 'mutant' (in this case, turquoise killifish residue) and 'wild-type' (in this case, human or mouse residue in the structure). If the  $\Delta\Delta G$  value was greater than 1 between wild type and mutant, the mutation was considered to have a strong effect on the folding and stability of the protein. If the  $\Delta\Delta G$  was  $> 0.5$ , it was considered to have a moderate effect (Tables S4F and S7H).



## Mapping residues and variants on available protein structures or domains

Residues under positive selection in the turquoise killifish were mapped on the available three-dimensional protein structures for the orthologs of 4 aging genes: INSRA, IGF1R(1of2), XRCC5, and BAX. We first performed a structure-based sequence alignment using the corresponding chain in the available crystal structure in Protein Data Bank (PDB; see accession numbers in Table S4F). To this end, we used all the fish species that were part of our selection pipeline, including the ancestral sequences generated by PhyloBot (<http://PhyloBot.com>), and aligned them with the corresponding chain in PDB structure using PROMALS3D (Pei et al., 2008). To map the non-synonymous variant with functional effect in GRN (W449), we aligned only the corresponding GRN domain in the turquoise killifish with the PDB structure of the human ortholog (Table S7H) using PROMALS3D. Corresponding residues were then mapped and highlighted on the PDB structures using JalView (Waterhouse et al., 2009) and PyMOL: <https://www.pymol.org/> (see Tables S4F and S7H for details).

There were no protein structure available for the orthologs of MGAT5(1of3) and IRS1(2of2), and the available structure for LMNA did not encompass the orthologous residues under selection. Therefore, for these proteins, we mapped the residues under selection on the predicted domains from the NCBI Conserved Domain search (Marchler-Bauer et al., 2015) (CDD: <http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>).

To map human variants, we obtained the position of residues associated with exceptional human longevity from LongevityMap (Budovsky et al., 2013; Suh et al., 2008) (IGF1R), and (Conneely et al., 2012; Sebastiani et al., 2012) (LMNA). We aligned the human sequence with multiple fish sequences using PRANK and mapped the residues on the turquoise killifish ortholog. We obtained the residues mutated in LMNA in Hutchinson Gilford Progeria Syndrome via OMIM: <http://omim.org/entry/150330> and UniProt: <http://www.uniprot.org/uniprot/P02545>. We also mapped residues with unique amino-acid changes in 34 genes in the bowhead whale (Keane et al., 2015) (Table S4G) and residues with unique amino-acid changes in IGF1R in the Brandt's bat (Seim et al., 2013). *daf-2* alleles with the phenotype 'extended life span' in *C. elegans* were obtained from WormBase (<http://wormbase.org>, Release:WS249; Date: September 3, 2015). Amino-acid changes in DAF-2 were from (Patel et al., 2008) and were mapped to IGF1R(1of2) in the turquoise killifish.

We obtained GRN mutations associated to neurodegenerative diseases in humans from OMIM (<http://www.omim.org/entry/138945>) and from the 'Alzheimer Disease & Frontotemporal Dementia Mutation Database' (Cruts et al., 2012). After manual inspection of the multiple protein alignments of human and fish sequences, we mapped the diseases mutations to the NMR structure of GRN using alignment by MUSCLE. Multiple alignments for ZNF800A and IFI35 were also performed using MUSCLE.

## Targeted Sanger re-sequencing of selected residues from aging-related genes

Genomic DNA from fish tissues was extracted using 200  $\mu$ L of DirectPCR Tail (Viagen Biotech Inc) with 4  $\mu$ L of Proteinase K (Invitrogen Inc). Samples were incubated at 50°C overnight, boiled at 100°C for 10 min, and centrifuged at 8,000g for 5 min. The supernatant was directly used as a template for PCR reactions.

To confirm the sequence of genes of interest around the residues that were identified to be under positive selection according to the reference GRZ genome, small amplicons (300-400bp) encompassing the residues of interest were amplified by PCR, using the GoTaq Green Master Mix (Promega) with an annealing temperature of 58°C. Two independent GRZ individuals were used for each analysis.

The following primers were used for PCR amplification of genomic sequences:

IGF1R_1of2_V121-T126_F	TTTGCCCTAACACATCTCCATTC
IGF1R_1of2_V121-T126_R	GTGATGTTCTCAGGTTGTACAG
IGF1R_1of2_F351_F	TTCTGAAACTGACCTCTTCACCT
IGF1R_1of2_F351_R	AGTGTACGTCCTTACCGATTAGT
IGF1R_1of2_A391_F	CGTTCCTCTGTCTACCTCAGTGT

IGF1R_1of2_A391_R	TGAAGTGTACGTCCTTACCGATT
IGF1R_1of2_L426-H428-L429_F	ACCACTGCTAGAATTTTCAGACG
IGF1R_1of2_L426-H428-L429_R	TCTTCTCCCAAACCTTCCAGAGA
IGF1R_1of2_Y489_F	CCATGTTTGTGACCATCTGATG
IGF1R_1of2_Y489_R	CGACACCATGAAATGAGAAAAGC
IGF1R_1of2_A843_F	CCTCGCTATATTTCACTGTTTGG
IGF1R_1of2_A843_R	AAGTTTGCACATTTCCATCAAGT
IGF1R_1of2_L1008_F	AATCTGAGCCCAGGAAACTATTC
IGF1R_1of2_L1008_R	CAGAAGCACTCACCTCTTTTTGT
IGF1R_1of2_L1305_F	GTTGTATTGTGGGAAATTGCCAC
IGF1R_1of2_L1305_R	CTCTATTGGCCAACACTGATGAG
IGF1R_1of2_S1374_F	CCCATTTCAAGGGAAGTAAGTTTC
IGF1R_1of2_S1374_R	ATTCATGTGTGCGTATGGTTGTA
INSRA_N425_F	AACACAAGCTAAGCATCCTGAAG
INSRA_N425_R	CAGCTTCATCACGTCTATGTTCA
INSRA_S434-P435_F	TGAAAGCGTGCTGATTCAAATTG
INSRA_S434-P435_R	TGCAACAAAGTCTAGCGATTTCT
INSRA_V457-V459_F	TTTCCCCTCCAGTTTATTTTCAT
INSRA_V457-V459_R	GCAAAGCAACAGTTTGGTCTAGT
INSRA_N566_F	CTGATGGAGACAAAGAGCGTATT
INSRA_N566_R	TCCGTGATTGACTTCCTGTTTAT
INSRA_A705_F	TAAATTCTCCCAAGTTCTGGACA
INSRA_A705_R	GGGTTTTCGCTTTTATTCAAGAT
INSRA_P801_F	AACCCCAGATGTTTTTGGTACT
INSRA_P801_R	TATAGGATGATGATGCCGTTAGG
INSRA_V910_F	TAATTGTACCTGAATGGGGTGAC
INSRA_V910_R	TACTCAGGGTTTGAAGAGGCATA
INSRA_T1258_F	TTGTTCTGGAGCTCTAACTCACC
INSRA_T1258_R	CAGTATTCGTCCATTGGTCTTTC
XRCC5_G888_F	ACCCCTTTAACTTTGTTCTCAGC
XRCC5_G888_R	ATACAGAAAGTGGTCCGTTTCAC
IRS1_2of2_L94_F	ATGATTCAGTTGTAGCCAGAAC
IRS1_2of2_L94_R	GCAATCCCATCCTCACCTTTC
LMNA3_M307_I358_F	ATTTGAGAGCAAACCTGGCAGA
LMNA3_M307_I358_R	AGTAAAAGCTGTCCGAGTACCTG

The presence of single PCR products was assessed on 1% TAE agarose gels, and PCR products were purified using the Qiagen PCR purification kit. Sequence of individuals was determined by direct Sanger sequencing of PCR products using the original amplification primers (MCLAB sequencing services). When direct sequencing of the PCR products did not work readily, the PCR products were cloned into the PCR4 TOPO-TA vector (Lifetechnologies), and at least two independent bacterial clones from each PCR were sequenced using the M13 reverse primer. Sequence analysis was done using pairwise BLAST alignments to the reference genomic sequence. All the tested PCR amplicons had sequences identical to the genomic reference around the residues of interest.

#### **Genetic variation in GRZ, MZM-0703, and MZM-0403 individuals**

To identify genetic variants among different strains of the turquoise killifish, we used next-generation sequencing to genotype the founders of cross GxM (GRZ female and MZM-0703 male). Ethanol-preserved

tissues were used for genomic DNA extraction, library construction (200bp inserts), and paired-end sequencing on Illumina HiSeq2000 instruments (BGI).

Library name	Strain	Insert size	Strategy	Total QC length (bp)	Coverage (X) <sup>a</sup>	Sex
POGFA	GRZ	200	Paired-end	38,540,978,200	19.27	F
POGMA	MZM-0703	200	Paired-end	107,816,807,600	53.91	M

<sup>a</sup> coverage estimate based on a conservative genome size estimate of 2Gb.

A library using genomic DNA from a male from another wild-derived strain, MZM-0403, was constructed at Stanford University and sequenced at the Stanford Genome Center.

Library name	Strain	Insert size	Strategy	Total QC length (bp)	Coverage (X) <sup>a</sup>	Sex
MZM0403	MZM-0403	300	Paired-end	12,817,270,468	6.41	M

<sup>a</sup> coverage estimate based on a conservative genome size estimate of 2Gb .

Quality filtered and trimmed paired-end sequencing reads were aligned to the reference genome using bwa v0.6.1-r104, which is a sensitive aligner recommended for variant calling from next-generation sequencing data (Liu et al., 2013). The GATK genotyping pipeline (McKenna et al., 2010), which was developed for human variant calling in the 1000 genome project, was used to call SNPs. GATK v1.6-13 was used with underlying support from picard-tools v1.55. Briefly, PCR-duplicates were filtered out, indels were fine-realigned, base phred scores were recalibrated, and the unified genotyper was run. To recalibrate SNP scores, we assembled a benchmark SNP database using turquoise killifish SNPs from dbSNP (Kirschner et al., 2012) mapped to our genome as well as high quality and a catalog of high-depth SNPs identified by RAD-seq. GATK filters used as parameters to “-A” were: AlleleBalance, DepthOfCoverage, HomopolymerRun, QualByDepth, and MappingQualityRankSumTest. To minimize the rate of false negative calls in the library from the GRZ individual (i.e. sites present but detected with lower depth), we did not apply a sequencing depth cutoff in the GRZ individual (i.e. DepthOfCoverage filter), but we applied sequencing depth cutoffs for the MZM-0703 individual (Depth  $\geq$  20) and MZM-0403 individual (Depth  $\geq$  5, to take into account the overall lower sequencing depth of that sample). Only variants with mapping quality  $\geq$  4 and depth  $\geq$  5 and low strand bias were selected, which eliminates variants that are hard to validate. Variants with non-straightforward allelic frequencies (i.e. not clearly homozygous or heterozygous) were also eliminated, as these might be artifacts of poorly resolved repetitive regions. Variants annotated as ‘LowQual’ by GATK were also eliminated. Variants called independently in more than one library are considered as higher quality variants.

To annotate the identified variants and predict their potential function, we used the SNPeff pipeline (Cingolani et al., 2012) on final vcf call files and the high quality predicted gene models (Tiers 1, 2 and 3) gff3 file. The SNPeff pipeline categorizes genetic variants based on their location with respect to genes (e.g. upstream (-5kb from the annotated TSS), downstream (+5kb from the annotated TTS), intronic, etc.), and their potential impact on the coding sequence (e.g. synonymous, non-synonymous, etc.).

### Genetic crosses

One female from the GRZ strain was crossed with one male from the MZM-0703 strain (cross GxM), and one female from the Soveia strain was crossed with a male from the GRZ strain (cross SxG). F1 fish were interbred in families to generate F2 fish.

	F1	F1 families	F2
Cross GxM	36	16	430
Cross SxG	9	4	130

Observed lifespan was scored as the age at death for all the fish. For this study, fish were raised in cohorts of mixed sexes, in the conditions described above. Note that both genetic and environmental factors affect the age of death. Casualties due to non-natural death causes, e.g. inter-individual aggression or occasional tank exclusion from the water recirculation system, were not scored as “observed lifespan” and these individuals were censored for the lifespan analysis. Survival analysis was done with the R package “survival” using Logrank test statistics.

### **RAD-seq library construction for Illumina sequencing**

RAD-seq libraries for 225 samples from the cross GxM and for 86 from the SxG cross were prepared as described (Etter and Johnson, 2012). For the libraries, either 125 or 200 ng of genomic DNA from each sample was digested for 60 min at 37°C in a 25 µL reaction volume containing 2.5 µL 10x Buffer 4 and 10 units (U) SbfI-HF (New England Biolabs [NEB]) in a 96-well PCR plate. Samples were heat-inactivated for 20 min at 65°C then allowed to cool at room temperature for 1 hour. 1.0 µL or 1.6 µL of 6bp barcoded SbfI-P1 Adapter (100 nM), a modified Illumina © adapter (2006 Illumina, Inc; top oligo: 5'-ACACTCTTTCCTACACGACGCTCTTCCGATCTxxxxxxTGC\*A-3'; bottom oligo: 5'-Phos-yyyyyyAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-3' [x and y denote barcode and reverse complement, respectively]), was added to each sample along with 1.4 µL rATP (25 mM, Epicentre), 0.4 µL 10x NEB Buffer 4, 0.25 µL (500 U) T4 DNA Ligase (high concentration, NEB), 1.95 µL H<sub>2</sub>O and incubated at room temperature for 30 min. Samples were heat-inactivated for 20 min at 65°C and cooled at room temperature for 30 min, then combined in sub-libraries of 24-30 F2 individuals (15 µL per F2 library) and processed as 4 parallel libraries. 180 µL of each pooled sample was randomly sheared (Bioruptor) to an average size of 500 bp. 30 µL sheared sample was run on a 1.25% agarose gel to determine size before a 1.0X AMPure XP bead size selection and purification was performed on the remaining sheared volume. The Quick Blunting Kit (NEB) was used to polish the ends of the DNA in a 25 µL reaction volume containing 2.5 µL 10x Blunting Buffer, 2.5 µL dNTP Mix and 1.0 µL Blunt Enzyme Mix incubated at room temperature for 30 min. The sample was purified with AMPure XP beads, including a 0.5x size-exclusion step prior to 1.0X purification, and incubated at 37°C for 20 min with 10 U Klenow Fragment (3'-5' exo- with 2.5 µL NEB Buffer 2 and 0.5 µL dATP (10 mM, Fermentas), to add 3' adenine overhangs to the DNA. The reaction was moved to room temperature for 30 min, purified (1.0X) and 1.0 µL of Paired-End-P2 Adapter (PE-P2; 10 µM), a divergent modified Illumina © adapter (2006 Illumina, Inc.; see (Etter et al., 2011)) was ligated to the DNA fragments. The sample was purified (1.0x), eluted in 50 µL, and quantified using the Qubit™. We used 2.1-4.2 ng equivalent per individual as template in a 50 µL PCR amplification with 25 µL Phusion Hot Start Flex 2X Master Mix and 2.0 µL modified Illumina © amplification primer mix (10 µM, 2006 Illumina, Inc.; long-P1-forward primer: 5'-AATGATACGGCGACCACCGAGATCTACTCTTTCCTACACGACGCTCTTCCGATC\*T-3', short-P2-reverse primer: 5'-CAAGCAGAAGACGGCATACG\*A-3'). PCR was carried out with an initial denaturing step at 98°C for 3 min, then 14 cycles of 40 sec at 98°C, 15 sec at 65°C, and 30 sec at 72°C followed after 14 cycles by 5 min at 72°C. RAD-PE GP libraries were prepared together from 225 ng of genomic DNA and 2.0 µL of P1 adapter each and processed as above as a sub-library except that a gel extraction was performed after PCR and bead cleanup to enrich for longer fragments as previously described (Etter et al., 2011). Libraries were purified, diluted to 10 nM and submitted to the University of Oregon Genomics Core Facility for qPCR quantification. Sub-libraries were mixed at equal molar quantities and sequenced on the HiSeq 2000 following Illumina protocols for 100bp single-end reads (paired-end in the case of the GP library). All sequences are available at the NCBI Short Read Archive (accession number: SRP041421).

### **RAD-seq analysis of genetic cross**

RAD-seq data from the GxM genetic cross was processed and analyzed using STACKS (Catchen et al., 2011). We identified a total of 65,773 RAD-tags. Tags containing only one SNP were selected (-F snp\_l=1; -F snp\_u=1), corresponding to 9,529 in cross GxM. RAD-tags represented in less than 25% of the genotyped subjects were excluded from the analysis, resulting in 8,399 markers.

## Linkage map generation based on the RAD-seq data

We built a linkage map with R/qtl using RAD-seq markers that had homozygous haplotypes in the grandparents. This map was used for genome scaffolding and identification of QTL. Individuals that had genotypes in <1400 markers were dropped. Markers were considered duplicates and removed if >80% of pairs of individuals had matching genotypes. Pairs of individuals with >80% matching genotypes were also removed. Markers with distorted segregation patterns ( $p < 10^{-10}$ ) were also dropped. Genotype frequencies were examined and confirmed to be 1:2:1 AA:AB:BB. The final cross GxM linkage map comprised 193 F2 individuals and 5,757 RAD-seq markers.

## Random Forest QTL mapping

RAD-seq markers with a minor allele frequency below 25% were removed, i.e. 2,672 markers, corresponding to 32% of the total markers. Additionally, individuals for which genotyping information was not available from more than 25% of the remaining markers were excluded for this analysis. This corresponded to 53 individuals, i.e. 24% of the total number of fishes. To test the effect of sex on longevity, QTLs were mapped in: i) all individuals, ii) males only, iii) females only, iv) the residual of the regression of the longevity data on sex; furthermore genetic association with v) weight, vi) weight in males, vii) weight in female, viii) the residual of the regression of weight on sex, as well as other discrete traits: ix) gender, x) tail color, and xi) presence of a black strip.

QTL detection was carried out using a method based on Random Forest (Michaelson et al., 2010), that we previously adapted to take missing values and population structure into account (Clement-Ziza et al., 2014). This method uses genetic markers as predictors to model the traits and population structure is modelled as covariates. First, we estimated the kinship matrix, which scores the relatedness between the strains. We removed markers with more than ten missing genotype values for the population structure estimation. Missing values were randomly replaced by random alleles with probabilities following the distribution of the alleles for the marker of interest. The procedure was repeated 2,000 times, each time building a new kinship and performing singular value decomposition. The average of all generated matrices was used as the final estimate of the kinship. As covariate, we selected the eigenvectors corresponding to the top five eigenvalues, which accounted for more than 75% of the genotype variance.

For the QTL mapping, forests of 14,400 trees were grown (120 forests of 120 trees) using the R implementation of the 'RandomForest' algorithm. The strategies described previously (Clement-Ziza et al., 2014) and above were used to handle missing genotype values and model the population structure. The *mtry* parameter was left to default (one third of the total number of predictors). The QTLs were then scored using the predictor selection frequency as previously proposed (Michaelson et al., 2010). To estimate the significance of the linkages, each trait was permuted 50,000 times and random forests were trained for each permutation. The correspondence between the covariates and the permuted traits was maintained in order to properly estimate the significance of the trait-marker linkages. We obtained null distributions of the selection frequencies for each trait and each marker. They were used to estimate p-values for the selection frequencies. We then reused the permutation results to also estimate p-values for each randomized trait and thus obtained a null distribution of p-values. We then considered the 11 mapped traits together to estimate the false discovery rates (FDR) for the entire analysis based of null distribution of the p-values.

## Identification of genomic scaffolds underlying the lifespan QTL

Bowtie 0.12.7 was used to map the 6 RAD-seq markers corresponding to the lifespan QTL peak on LG-3 to our assembled scaffolds. Scaffolds were considered to belong to the peak if at least one of these RAD-seq markers uniquely mapped to them. This analysis resulted in 6 scaffolds (Table S7). Of note, two of these scaffolds (GapFilledScaffold\_60 and GapFilledScaffold\_883) also contain RAD-seq markers that were attributed to other linkage groups by R/qtl (marker 42243 was attributed to LG-2 and markers 18875 and 24430 were attributed to LG-15), although the genes in GapFilledScaffold\_883 showed synteny with the equivalent region in medaka that is syntenic to LG-3. To be comprehensive, both scaffolds were kept in the analysis. Protein-coding genes and non-coding RNA genes contained in these scaffolds were then identified from our annotation files. Gene positions were plotted to scale ordered on the scaffolds using the 'grid' R

package. SNPs between the GxM cross founders falling on these scaffolds were extracted for further analysis. Circular linkage maps and scaffolds were plotted using Circos: <http://circos.ca/> (Krzyszowski et al., 2009).

### Analysis of enrichments in aging genes at the lifespan QTL

To compare the potential enrichment in aging genes at the lifespan QTL with the rest of the linkage map, we first extracted all the scaffolds that contained RAD-seq markers as described above. Genes belonging to these scaffolds were attributed to the corresponding linkage groups (LGs). There was a total of 13,242 genes anchored to the linkage map.

To test the enrichment of aging-related genes at the lifespan QTL, we used the list of combined human and mouse aging and longevity-related genes from the GenAge database (Table S4A). Out of the 462 genes in the mouse and aging complete list, 238 were anchored to our linkage map, including 20 in LG-3, and 5 at the QTL peak. Under the hypothesis of random distribution of genes, we applied Fisher's exact test to measure enrichment of aging-related genes at the QTL compared to the entire linkage map or just LG-3. We also used the list of mouse aging and longevity-related genes from GenAge (Table S4A). Out of the 142 genes in this mouse list, 70 were anchored to our linkage map, including 6 in LG-3, and 3 at the QTL peak.

### 2010 fish collection expedition to Mozambique

To identify genetic variants present in wild turquoise killifish populations, we conducted an expedition and collected wild specimens from 5 different localities along the Chefu river drainage in southern Mozambique in April-May 2010.

Strain name	GPS coordinates	Male coloration
ZMZ-1001	S21° 48.933' E031° 55.872'	Yellow
ZMZ-1002	S21° 55.011' E021° 05869'	Yellow
ZMZ-1003	S22° 08.803' E032° 49.465'	N.A.*
ZMZ-1004	S22° 28.924' E032° 49.465'	N.A.*
ZMZ-1005	S22° 30.497' E032° 33.055'	Yellow; Red
ZMZ-1006	S23° 27.548' E032° 33.855'	Red
ZMZ-1007	S24° 06.293' E032° 46.117'	Red

\*only females were identified, therefore male coloration could not be assessed

### Targeted Sanger re-sequencing of candidate genes in individuals from different strains or from the wild

Genomic DNA was extracted as described above. To genotype the *GRN*, *IFI35* and *ZNF800A* genes, small amplicons (300-600bp) encompassing the residues of interest were amplified by PCR as described above with the following primers:

GRN_W449_F1	TGTGAGGACAAGGAGCACTG
GRN_W449_R1	CAAACCTCCATGCAGAAAGAGC
GRN_W449_F2	CACTTACTGAAACTTCCTCCACTGT
GRN_W449_R2	GCTGCTAAACAATGAAATATTCCTG
GRN_Q151_F1	TCATTCCAGAGTTGATTTTCACA
GRN_Q151_R1	AAGGGCAGACGTTGTGTACC
IFI35_M196_F	CATCTCATTAGTGGCGAGCA
IFI35_M196_R	AGAGTCGATCTGTGGGATGG
ZNF800A_N489_F	CCGCTGTTAGACTCCTCGTC
ZNF800A_N489_R	CAGAGTGTCCCATGAAAGA

Genotype of individuals was determined by direct Sanger sequencing of PCR products (MCLAB sequencing services). Sequences were manually confirmed using the chromatogram profiles to detect variants at the homozygous or heterozygous state (Table S7E, F).

### Code sharing

The code generated for this work is available for download at <http://africanturquoisekillifishbrowser.org/downloads.html>.

### Supplemental References

- Baumgart, M., Groth, M., Priebe, S., Savino, A., Testa, G., Dix, A., Ripa, R., Spallotta, F., Gaetano, C., Ori, M., et al. (2014). RNA-seq of the aging brain in the short-lived fish *N. furzeri* - conserved pathways and novel genes associated with neurogenesis. *Aging Cell* *13*, 965-974.
- Capriotti, E., Calabrese, R., and Casadio, R. (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* *22*, 2729-2734.
- Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W., and Postlethwait, J.H. (2011). Stacks: building and genotyping Loci de novo from short-read sequences. *G3* *1*, 171-182.
- Charif, D., and Lobry, J.R. (2007). SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations (Biological and Medical Physics, Biomedical Engineering)*, U. Bastolla, M. Porto, H.E. Roman, and M. Vendruscolo, eds. (Berlin, Heidelberg, Germany: Springer Verlag), pp. 207-232.
- Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* *6*, 80-92.
- Cruts, M., Theuns, J., and Van Broeckhoven, C. (2012). Locus-specific mutation databases for neurodegenerative brain diseases. *Hum. Mutat.* *33*, 1340-1344.
- Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2014). Ensembl 2015. *Nucleic Acids Res.* *43*, D662-669.
- Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* *27*, 1164-1165.
- de Magalhaes, J.P., and Toussaint, O. (2004). GenAge: a genomic and proteomic network map of human ageing. *FEBS Lett.* *571*, 243-247.
- Dehouck, Y., Kwasigroch, J.M., Gilis, D., and Rooman, M. (2011). PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics* *12*, 151.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* *29*, 15-21.
- Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., Yu, H.O., Buffalo, V., Zerbino, D.R., Diekhans, M., et al. (2011). Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res.* *21*, 2224-2241.
- Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* *26*, 2460-2461.

- Etter, P.D., and Johnson, E. (2012). RAD paired-end sequencing for local de novo assembly and SNP discovery in non-model organisms. *Methods Mol. Biol.* 888, 135-151.
- Falcon, S., and Gentleman, R. (2007). Using GOSTATS to test gene lists for GO term association. *Bioinformatics* 23, 257-258.
- Fletcher, W., and Yang, Z. (2010). The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol. Biol. Evol.* 27, 2257-2267.
- Frappier, V., Chartier, M., and Najmanovich, R.J. (2015). ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability. *Nucleic Acids Res.* 43, W395-400.
- Gems, D., Sutton, A.J., Sundermeyer, M.L., Albert, P.S., King, K.V., Edgley, M.L., Larsen, P.L., and Riddle, D.L. (1998). Two pleiotropic classes of *daf-2* mutation affect larval arrest, adult behavior, reproduction and longevity in *Caenorhabditis elegans*. *Genetics* 150, 129-155.
- Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307-321.
- Harel, I., Benayoun, B.A., Machado, B., Singh, P.P., Hu, C.K., Pech, M.F., Valenzano, D.R., Zhang, E., Sharp, S.C., Artandi, S.E., et al. (2015). A platform for rapid exploration of aging and diseases in a naturally short-lived vertebrate. *Cell* 160, 1013-1026.
- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* 38, 576-589.
- Jordan, G., and Goldman, N. (2012). The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol. Biol. Evol.* 29, 1125-1139.
- Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 110, 462-467.
- Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772-780.
- Kent, W.J. (2002). BLAT - the BLAST-like alignment tool. *Genome Res.* 12, 656-664.
- Kim, D., and Salzberg, S.L. (2011). TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.* 12, R72.
- Kinsella, R.J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., et al. (2011). Ensembl BioMart: a hub for data retrieval across taxonomic space. *Database (Oxford)* 2011, bar030.
- Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639-1645.
- Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T., and Ussery, D.W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35, 3100-3108.
- Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., et al. (2010). The sequence and de novo assembly of the giant panda genome. *Nature* 463, 311-317.
- Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658-1659.
- Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923-930.
- Liu, X., Han, S., Wang, Z., Gelernter, J., and Yang, B.Z. (2013). Variant callers for next-generation sequencing data: a comparison study. *PLoS One* 8, e75619.



- Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* *15*, 550.
- Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* *25*, 955-964.
- Magoc, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* *27*, 2957-2963.
- Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* *27*, 764-770.
- Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., et al. (2015). CDD: NCBI's conserved domain database. *Nucleic Acids Res.* *43*, D222-226.
- Michaelson, J.J., Alberts, R., Schughart, K., and Beyer, A. (2010). Data-driven assessment of eQTL mapping methods. *BMC Genomics* *11*, 502.
- Mullan, L.J., and Bleasby, A.J. (2002). Short EMBOSS User Guide. European Molecular Biology Open Software Suite. *Brief. Bioinform.* *3*, 92-94.
- Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* *29*, 2933-2935.
- Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* *23*, 1061-1067.
- Parthiban, V., Gromiha, M.M., and Schomburg, D. (2006). CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res.* *34*, W239-242.
- Patel, D.S., Garza-Garcia, A., Nanji, M., McElwee, J.J., Ackerman, D., Driscoll, P.C., and Gems, D. (2008). Clustering of genetically defined allele classes in the *Caenorhabditis elegans* DAF-2 insulin/IGF-1 receptor. *Genetics* *178*, 931-946.
- Peel, M.C., Finlayson, B.L., and McMahon, T.A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrol. Earth Syst. Sci.* *11*, 1633-1644.
- Pei, J., Kim, B.H., and Grishin, N.V. (2008). PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* *36*, 2295-2300.
- Petzold, A., Reichwald, K., Groth, M., Taudien, S., Hartmann, N., Priebe, S., Shagin, D., Englert, C., and Platzer, M. (2013). The transcript catalogue of the short-lived fish *Nothobranchius furzeri* provides insights into age-dependent changes of mRNA levels. *BMC Genomics* *14*, 185.
- Pires, D.E., Ascher, D.B., and Blundell, T.L. (2014). DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.* *42*, W314-319.
- Rembold, M., Lahiri, K., Foulkes, N.S., and Wittbrodt, J. (2006). Transgenesis in fish: efficient selection of transgenic fish by co-injection with a fluorescent reporter construct. *Nat. Protoc.* *1*, 1133-1139.
- Salzberg, S.L., Phillippy, A.M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T.J., Schatz, M.C., Delcher, A.L., Roberts, M., et al. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* *22*, 557-567.
- Schulz, M.H., Zerbino, D.R., Vingron, M., and Birney, E. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* *28*, 1086-1092.
- Sebastiani, P., Solovieff, N., Dewan, A.T., Walsh, K.M., Puca, A., Hartley, S.W., Melista, E., Andersen, S., Dworkis, D.A., Wilk, J.B., et al. (2012). Genetic signatures of exceptional longevity in humans. *PLoS One* *7*, e29848.
- Simpson, J.T. (2014). Exploring genome characteristics and sequence quality without a reference. *Bioinformatics* *30*, 1228-1235.

- Slater, G.S., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6, 31.
- Smit, A.F.A., Hubley, R., and Green, P. (1996-2004). RepeatMasker Open-3.0. <http://www.repeatmasker.org>.
- Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* 56, 564-577.
- Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* 30, 2725-2729.
- Tingaud-Sequeira, A., Lozano, J.J., Zapater, C., Otero, D., Kube, M., Reinhardt, R., and Cerda, J. (2013). A rapid transcriptome response is associated with desiccation resistance in aerially-exposed killifish embryos. *PLoS One* 8, e64410.
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562-578.
- Treangen, T.J., and Salzberg, S.L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* 13, 36-46.
- Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189-1191.
- Wittkop, T., Emig, D., Lange, S., Rahmann, S., Albrecht, M., Morris, J.H., Bocker, S., Stoye, J., and Baumbach, J. (2010). Partitioning biological data with transitivity clustering. *Nat. Methods* 7, 419-420.
- Wong, W.S., Yang, Z., Goldman, N., and Nielsen, R. (2004). Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* 168, 1041-1051.
- Yang, Z., and dos Reis, M. (2011). Statistical properties of the branch-site test of positive selection. *Mol. Biol. Evol.* 28, 1217-1228.
- Yang, Z., and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19, 908-917.
- Yang, Z., Wong, W.S., and Nielsen, R. (2005). Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* 22, 1107-1118.
- Yao, G., Ye, L., Gao, H., Minx, P., Warren, W.C., and Weinstock, G.M. (2012). Graph concordance of next-generation sequence assemblies. *Bioinformatics* 28, 13-16.