# Intra-Species Differences in Population Size shape Life History and Genome Evolution

**Authors:** David Willemsen[1], Rongfeng Cui[1], Martin Reichard[2], Dario Riccardo Valenzano[1,3*]

**Affiliations:**

[1]Max Planck Institute for Biology of Ageing, Cologne, Germany.

[2]The Czech Academy of Sciences, Institute of Vertebrate Biology, Brno, Czech Republic.

[3]CECAD, University of Cologne, Cologne, Germany.

*Correspondence to: dvalenzano@age.mpg.de

Key words: life history, evolution, genome, population genetics, killifish, *Nothobranchius furzeri*, lifespan, sex chromosome, selection, genetic drift

## Abstract

The evolutionary forces shaping life history trait divergence within species are largely unknown. Killifish (oviparous Cyprinodontiformes) evolved an annual life cycle as an exceptional adaptation to life in arid savannah environments characterized by seasonal water availability. The turquoise killifish (*Nothobranchius furzeri*) is the shortest-lived vertebrate known to science and displays differences in lifespan among wild populations, representing an ideal natural experiment in the evolution and diversification of life history. Here, by combining genome sequencing and population genetics, we investigate the evolutionary forces shaping lifespan among turquoise killifish populations. We generate an improved reference assembly for the turquoise killifish genome, trace the evolutionary origin of the sex chromosome, and identify genes under strong positive and purifying selection, as well as those evolving neutrally. We find that the shortest-lived turquoise killifish populations, which dwell in fragmented and isolated habitats at the outer margin of the geographical range of the species, are characterized by small effective population size and accumulate throughout the genome several small to large-effect deleterious mutations due to genetic drift. The genes most affected by drift in the shortest-lived turquoise killifish populations are involved in the WNT signalling pathway, neurodegenerative disorders, cancer and the mTOR pathway. As the populations under stronger genetic drift are the shortest-lived ones, we propose that limited population size due to habitat fragmentation and repeated population bottlenecks, by causing the genome-wide accumulation of deleterious mutations, cumulatively contribute to the short adult lifespan in turquoise killifish populations.

## Main

Killifish are a highly successful teleost group, dwelling in a range of diverse habitats and, uniquely among teleosts, evolved an annual life cycle to survive in arid climates characterized by cycles of rainfalls and drought[1]. While on the one hand intermittent precipitation and periodic drought pose strong selective pressures leading to the evolution of embryonic diapause, an adaptation that enables killifish to survive in absence of water[2,3], on the other hand they cause habitat and population fragmentation, promoting genetic drift. The co-occurrence of strong selective pressure for early-life and extensive drift characterizes life history evolution in African annual killifishes[4].

The turquoise killifish (*Nothobranchius furzeri*) is the shortest-lived vertebrate with a thoroughly documented[a] post-embryonic life, which, in the shortest-lived strains, amounts to four months[2,3,5,6]. Turquoise killifish has recently emerged as a powerful new laboratory model to study experimental biology of aging due to its short lifespan and to its wide range of aging-related changes, which include neoplasias[7], decreased regenerative capacity[8], cellular senescence[9,10], and loss of microbial diversity[11]. At the same time, while sharing physiological adaptations that enable embryonic diapause and rapid sexual maturation, different wild turquoise killifish populations display differences in lifespan, both in the wild and in captivity[12-14], making this species an ideal evolutionary model to study the genetic basis underlying life history trait divergence within species.

[a] *Eviota sigillata* has been reported to be the shortest-known living vertebrate[5], but no information is available regarding the duration of its larval stage and its complete post-hatching lifespan in laboratory conditions.

51 Characterisation of life history traits in wild-derived laboratory strains of turquoise killifish
52 revealed that while different populations have similar rates of sexual maturation[6], populations
53 from arid regions exhibit the shortest lifespans, while populations from more semi-arid regions
54 exhibit longer lifespans[6,12]. Hence, speed of sexual maturation and adult lifespan appear to be
55 independent in turquoise killifish populations. The evolutionary mechanisms responsible for the
56 lifespan differences among turquoise killifish populations are not yet clearly understood.
57 Mapping genetic loci associated with lifespan differences among turquoise killifish populations
58 showed that adult survival has a complex genetic architecture[13,15]. Here, combining genome
59 sequencing and population genetics, we investigate to what extent genomic divergence in natural
60 turquoise killifish populations that differ in lifespan is driven by adaptive or neutral evolution.

61 ***Genome assembly improvement and gene annotation***

62 To identify the genomic mechanism that led to the evolution of differences in lifespan between
63 natural populations of the turquoise killifish (*Nothobranchius furzeri*), we combined the currently
64 available reference genomes[13,16] into an improved reference turquoise killifish genome assembly.
65 Due to the high repeat content, assembly from short reads required a highly integrated and multi-
66 platform approach. We ran Allpaths-LG with all the available pair-end sequences, producing a
67 combined assembly with a contig N50 of 7.8kb, corresponding to a ~2kb improvement from the
68 previous versions. Two newly obtained 10X Genomics linked read libraries were used to correct
69 and link scaffolds, resulting in a scaffold N50 of 1.5Mb, i.e. a three-fold improvement from the
70 best previous assembly. With the improved continuity, we assigned 92.2% of assembled bases to
71 the 19 linkage groups using two RAD-tag maps[13]. Gene content assessment using the BUSCO
72 method improved "complete" BUSCOs from 91.43%[13] and 94.59%[16] to 95.20%. We mapped

73 Genbank *N. furzeri* RefSeq RNA to the new assembly to predict gene models. The predicted gene
74 model set is 96.1% for "complete" BUSCOs.

75 ***Population genetics of natural turquoise killifish populations***

76 Natural populations of turquoise killifish occur along an aridity gradient in Zimbabwe and
77 Mozambique and populations from more arid regions are associated with shorter captive
78 lifespan[6,12]. A QTL study performed between short-lived and long-lived turquoise killifish
79 populations showed a complex genetic architecture of lifespan (measured as age at death), with
80 several genome-wide loci associated with lifespan differences among long-lived and short-lived
81 populations[13]. To further investigate the evolutionary forces shaping genetic differentiation in the
82 loci associated with lifespan among wild turquoise killifish populations, we performed pooled
83 whole-genome-sequencing (WGS) of killifish collected from four sampling sites within the
84 natural turquoise killifish species distribution, which vary in altitude, annual precipitation and
85 aridity (**Figure S1**, **Table S1**). Population GNP is located within the Gonarezhou National Park
86 at high altitude and in an arid climate (Koeppen-Geiger classification "BWh", **Figure S1**), in a
87 region at the outer edge of the turquoise killifish distribution (**Figure S1**)[17-19], which corresponds
88 to the place of origin of the "GRZ" laboratory strain, which has the shortest lifespan of all
89 laboratory strains of turquoise killifish[12,13]. Population NF414 is located in an arid area in the
90 center of the Chefu river drainage in Mozambique ("BWh", **Figure S1**)[17-19], and population
91 NF303 is located in a semi-arid area in transition to more humid climate zones in the center of the
92 Limpopo river drainage system (Koeppen-Geiger classification "BSh", **Figure S1**)[17-19]. Altitude
93 among localities ranges from 344 m (GNP) to 68 m (NF303, **Figure S1a** and **Table S1**). The
94 temporary habitat of turquoise killifish populations differs in terms of altitude and aridity, as the
95 ephemeral pools at higher altitude are drained earlier and persist for shorter time, while water

96 bodies in habitats at lower altitude last longer[12]. Population GNP is therefore named "dry",
97 population NF414 is named "intermediate" and population NF303 "wet" throughout the
98 manuscript.

99

100 ***High genetic differentiation and contrasting population demography between dry and wet***
101 ***populations***

102 We asked whether populations from dry, intermediate and wet areas, corresponding to shorter
103 and progressively longer lifespan, differ in genetic variability. We calculated genome-wide
104 estimates of average pairwise difference ($\pi$) and genetic diversity ($\theta_{Watterson}$) based on 50kb-non-
105 overlapping sliding windows using PoPoolation[20]. We found that $\pi$ and $\theta_{Watterson}$ decrease from
106 wet to dry population ($\theta_{Watterson\ GNP}$: 0.0011, $\theta_{Watterson\ NF414}$: 0.0036, $\theta_{Watterson\ NF303}$: 0.0072; $\pi_{GNP}$:
107 0.0009, $\pi_{NF414}$: 0.0031, and $\pi_{NF303}$: 0.0054). To infer the genetic distance between the
108 populations, we computed the genome-wide pairwise genetic differentiation between populations
109 using $F_{ST}$[21]. Overall, the genetic differentiation between populations ranged between 0.14 and
110 0.26 and was the highest between population GNP (dry) and population NF303 (wet) (**Figure**
111 **1a**).

112 Next, we inferred the demographic history of the populations using pairwise sequentially
113 Markovian coalescent (PSMC) by resequencing at high-coverage single individuals for each
114 population[22]. The population GNP (dry) experienced a strong population decline starting
115 approximately 150k generations ago, a result consistent in both sequenced individuals from the
116 two sampling sites (GNP-G1-3 and GNP-G4, **Figure 1a**). In contrast to the demographic history
117 in GNP, we found indications for recent population expansions in populations from the center of
118 the Chefu and Limpopo basins clades. Analysis of population NF414 (intermediate) (**Figure 1a**,

119 NF414-Y and NF414-R) and NF303 (wet) (**Figure 1a**, blue line) shows population expansion
120 until recent time (~50k generations ago). To infer the effective population size ($N_e$) of the
121 populations, we used the published mutational rate of 2.6321e−9 per base pair per generation for
122 *Nothobranchius* computed via dated phylogeny[4] and $\theta_{\text{Watterson}}$. In line with the decrease in genetic
123 diversity from wet to dry population, we found a decrease in $N_e$ estimates (107221.8, 338849.48
124 and 683693.25 for GNP, NF414 and NF303, respectively; **Figure 1b**). Hence, our findings show
125 that dry populations from the outer edge of the species distribution show lower genetic diversity
126 and smaller effective population size compared to population from intermediate and more wet
127 regions.

128

129 ***Genetic differentiation among turquoise killifish populations***

130 To test whether regions underlying longevity QTL in turquoise killifish[13,15] display a genetic
131 signature for positive or purifying selection, we took advantage of the improved turquoise
132 killifish genome assembly and the newly sequenced wild turquoise killifish populations (**Figure**
133 **2**). The strongest QTL for lifespan differences among long-lived and short-lived populations
134 mapped on the sex chromosome[13,15], in proximity to the sex determining locus[13]. To identify a
135 genomic signature of strong selection, we performed an outlier approach based on the pairwise
136 genetic differentiation index ($F_{ST}$). To find highly differentiated regions that may underlie
137 positive selection in natural turquoise killifish populations, we scanned for regions with elevated
138 genetic differentiation between pairs of populations, i.e. exceeding the 0.995 quantile of Z-
139 transformed non-overlapping 50kb sliding windows of $F_{ST}$. To find regions under purifying
140 selection, we scanned for regions with lowered genetic differentiation among populations, i.e.
141 below the 0.005 quantile of Z-transformed non-overlapping 50kb sliding windows of $F_{ST}$

142 (**TableS7**). The outlier approach did not reveal clear signatures of positive or purifying selection
143 based on genetic differentiation in the four main clusters associated with lifespan in experimental
144 strains of turquoise killifish (**Figure 2**). We then analysed genomic regions carrying signatures of
145 positive and purifying selection in the natural turquoise killifish populations irrespective of the
146 QTL regions (**Figure 2**). The $F_{ST}$ outlier approach led to the identification of several potential
147 regions under divergent selection between populations, in particular between the intermediate and
148 wet populations (**Table S4**) and only two between the dry and wet populations (**Table S5)**. Genes
149 significantly different and within regions of larger genetic differentiation based on Z-transformed
150 non-overlapping sliding windows of $F_{ST}$ were located on chromosomes 6 and 10. The region on
151 chromosome 6 includes the gene *slc8a1*, which contains mutations with significant difference in
152 allele frequencies between the wet and intermediate population (Fisher's exact test implemented
153 in PoPoolation; adjusted p value < 0.001). The region on chromosome 10 contains four genes:
154 XM_015941868, XM_015941869, *lss* and *hibch.* All genes under the major $F_{ST}$ peak on
155 chromosome 10 showed significant difference in allele frequencies between the intermediate and
156 wet population (Fisher's exact test; adjusted p value < 0.001) and additionally, *hibch* had
157 significantly different allele frequencies between the dry and wet population (Fisher's exact test;
158 adjusted p value < 0.001). Genes under $F_{ST}$ peaks between populations that differ in lifespan, are
159 not necessarily causally involved in lifespan differences between populations, as sequence
160 differences could segregate in populations due to population structure and drift. However, to test
161 whether the genes located in genomic regions that are significantly divergent between
162 populations could be functionally involved in age-related phenotypes, we investigated whether
163 gene expression in these genes varied as a function of age. Analysing available turquoise killifish
164 longitudinal RNA-datasets generated in liver, brain and skin[23], we found that *hibch*, *lss* and
165 *slc8a1* are differentially expressed between adult and old killifish (**Table S10**, adjusted p value <

166 0.01). *hibch*, *lss and slc8a1* are involved in amino acid metabolism[24], biosynthesis of
167 cholesterol[25], and proton-mediated accelerated aging[26], respectively. Gene XM_015956265
168 (ZBTB14) is the only gene that is an $F_{ST}$ outlier and that is differentially expressed in adult vs.
169 old individuals between at least two populations in all tissues (liver, brain and skin).
170 XM_015956265 encodes a transcriptional modulator with ubiquitous functions, ranging from
171 activation of dopamine transporter to repression of MYC, FMR1 and thymidine kinase
172 promoters[27]. However, although genomic regions that have sequence divergence between
173 turquoise killifish populations contain genes that are differentially expressed during ageing in
174 different tissues, whether any of these genes are causally involved in modulating ageing-related
175 changes between turquoise killifish wild populations still remains to be assessed. Based on the
176 outlier approach, we found two genomic regions with low genetic differentiation between all
177 pairs of populations, indicating strong purifying selection. The first region is located on the sex
178 chromosome and contains the putative sex determining gene *gdf6*[16], which is hence conserved
179 among these populations. This same region also contains *sybu*, a maternal-effect gene associated
180 with the establishment of embryo polarity[28]. The second region under low genetic differentiation
181 is located on chromosome 9 and harbours the genes XM_015965812 (*abi2-like*), *cnot11* and *lcp1*,
182 which are involved in phagocytosis[29], mRNA degradation[30] and cell motility[31], respectively.

183 ***Evolutionary origin of the sex chromosome***

184 Since we found reduced genetic differentiation among populations in the chromosomal region
185 containing the putative sex-determining gene in the sex chromosome, we used synteny analysis
186 and the new genome assembly to investigate the genomic events that led to evolution of this
187 chromosomal region (**Figure 3**). We found that the structure of the turquoise killifish sex
188 chromosome is compatible with a chromosomal translocation within an ancestral chromosome

189 and a fusion event between two chromosomes. The translocation event within an ancestral
190 chromosome corresponding to medaka´s chromosome 16 and platyfish´s linkage group 3 led to a
191 repositioning of a chromosomal region containing the putative sex-determining gene *gdf6*
192 (**Figure 3b**). The fusion of the translocated chromosome with a chromosome corresponding to
193 medaka chromosome 8 and platyfish linkage group 16, possibly led to the origin of turquoise
194 killifish sex chromosome. We could hence reconstruct a model for the origin of the turquoise
195 killifish sex chromosome (**Figure 3c**), which parsimoniously places a translocation event before a
196 fusion event. The occurrence of two major chromosomal rearrangements, namely a translocation
197 and a fusion, could have then contributed to suppressing recombination around the sex-
198 determining region[13,32].

199 ***Relaxed selection in turquoise killifish populations***

200 Since we could not identify specific signatures of genetic differentiation in the genomic regions
201 associated to longevity from previous QTL mapping, we asked whether other evolutionary forces
202 than directional selection may underlie differences in survival among wild turquoise killifish
203 populations. The difference in the recent and past demography between populations (**Figure 1**)
204 led us to ask whether demography could have led to evolutionary changes on genome-wide scale
205 between natural populations. For each population, we calculated the fraction of substitutions
206 driven to fixation by positive selection since divergence from the outgroup species
207 *Nothobranchius orthonotus* (NOR) using the asymptotic McDonald-Kreitman $\boldsymbol{\alpha}$[33]. Using NOR as
208 an outgroup, we infer the fraction of positive selection by pooling all coding sites (**Figure 4a**).
209 SNPs were called with the program SNAPE[34], which specifically deals with pooled sequencing.
210 We only included SNPs with a derived frequency between 0.05-0.95 and performed stringent
211 filtering. The asymptotic McDonald-Kreitman $\boldsymbol{\alpha}$ ranged from -0.21 to -0.01 in comparison to the

212 very closely related sister species *N. orthonotus*, confirming limited genome-wide positive
213 selection since divergence from *N. orthonotus* (**Figure 4a**). The population GNP, located in an
214 arid region at higher altitude and associated with the shortest recorded lifespan, shows the lowest
215 asymptotic McDonald-Kreitman $\boldsymbol{\alpha}$, as well as lower McDonald-Kreitman $\boldsymbol{\alpha}$ values throughout all
216 derived frequency bins, potentially suggesting a higher load of slightly deleterious mutations
217 segregating in this population (**Figure 4a**). Using as an outgroup species another annual killifish
218 species, *Nothobranchius rachovii* (NRC), we confirmed the lowest asymptotic McDonald-
219 Kreitman $\boldsymbol{\alpha}$ value in the dry population GNP (**Figure 4b**). Additionally, using *Nothobranchius*
220 *rachovii* (NRC) as outgroup species, the asymptotic McDonald-Kreitman $\boldsymbol{\alpha}$ ranged from -0.06 to
221 0.23 among populations, indicating that more alleles were driven to fixation by positive selection
222 in the ancestral lineage leading to *Nothobranchius furzeri* and *Nothobranchius orthonotus*. In
223 particular, the wet population NF303 had the highest asymptotic McDonald-Kreitman $\boldsymbol{\alpha}$ value
224 (**Figure 4b**). Using both *N. orthonotus* and *N. rachovii* as outgroups, we found that the dry GNP
225 population had the lowest McDonald-Kreitman $\boldsymbol{\alpha}$ values at the low derived frequency bins,
226 potentially consistent with a genome-wide accumulation of slightly deleterious mutations in these
227 isolated populations.

228 To directly estimate the fitness effect of gene variants associated with each population, we
229 analysed population-specific genetic polymorphisms to assign mutations as beneficial, neutral or
230 detrimental, and determine the distribution of fitness effect (DFE)[35]. Consistently with the overall
231 lower McDonald-Kreitman $\boldsymbol{\alpha}$ values throughout all derived frequency bins, we found more
232 mutations assigned as the slightly deleterious category in the dry GNP population, compared to
233 the other two populations (indicated by the higher number of deleterious SNPs in proximity to
234 $4N_eS \sim 0$ in the GNP population, **Figure 4d, TableS8**). To further infer the effect of the putative
235 deleterious mutations on protein function, we used the new turquoise killifish genome assembly

236 as a reference and adopted an approach that, by analysing sequence polymorphism among
237 populations, predicts functional consequences at the protein level[36]. We found that the proportion
238 of mutations causing a change in protein function is significantly larger in the GNP population
239 compared to populations NF414 and NF303 (Chi-square test: $P_{GNP\text{-}NF303}$<1.87e-119, $P_{GNP\text{-}NF414}$<
240 4.96e-57, $P_{NF303\text{-}NF414}$< 3.51e-35, **Figure 4e**). Additionally, the mutations with predicted
241 deleterious effects on protein function reached also higher frequencies in the dry population GNP
242 (**Figure 4c**). To further investigate the impact of mutations on protein function, we calculated the
243 Consurf [37-40] score, which determines the evolutionary constraint on an amino acid, based on
244 sequence conservation. Mutations at amino acid positions with high Consurf score (i.e. otherwise
245 highly conserved) are considered to be more deleterious. We found that the dry population GNP
246 had a significantly higher mean Consurf score for mutations at non-synonymous sites in
247 frequency bins from 5%-20% up to 40%-60%, compared to populations NF414 (intermediate)
248 and NF303 (wet) (**Figure S2**). The mutations in the dry GNP population had significantly higher
249 Consurf scores than the other populations using both outgroup species *N. orthonotus* and *N.*
250 *rachovii* (**Figure S2**). Upon exclusion of potential mutations at neighbouring sites (CMD: codons
251 with multiple differences), CpG hypermutation and genes containing mutations with highly
252 detrimental effect on protein function based on SnpEFF analysis, the dry population GNP had
253 higher mean Consurf score at the low frequency bin (**Figure S2**, **Table S11-12**). To note, we also
254 found a significantly higher average Consurf score at synonymous sites in GNP at low derived
255 frequencies (Figure S2, Supplementary Table S11 and S12), possibly suggestive of an overall
256 higher mutational rate in GNP.

257 ***Relaxation of selection in age-related disease pathways***

258 Computing the gene-wise direction of selection (DoS)[41] index, which enables to score the
259 strength of selection based on the count of mutations in non-synonymous and synonymous sites,
260 we found support to the hypothesis that the dry, short-lived population GNP has significantly
261 more slightly deleterious mutations segregating in the population, compared to the populations
262 NF414 and NF303 (**Figure 5a**, Median NOR: GNP: -0.17, NF414: -0.02, NF303: -0.01; Median
263 NRC: GNP: -0.14, NF414: 0.00, NF303: 0.00; Wilcoxon rank sum test: NOR: $P_{GNPNF303}$<2.21e-
264 105, $P_{GNP\text{-}NF414}$< 1.19e-76, $P_{NF303\text{-}NF414}$< 1.39e-06; NRC: $P_{GNP\text{-}NF303}$<4.61e-179, $P_{GNP\text{-}NF414}$< 1.42e-
265 100, $P_{NF303\text{-}NF414}$< 5.96e-22), indicating that purifying selection is relaxed in GNP. We calculated
266 DoS in all populations using independently as outgroup species *N. orthonotus* and *N. rachovii*
267 (**Figure 5a**).

268 To assess whether specific biological pathways were significantly more impacted by the
269 accumulation of slightly deleterious mutations, we performed pathway overrepresentation
270 analysis. We found a significant overrepresentation in the lower 2.5$^{th}$ DoS (i.e. genes under
271 relaxation of selection) in the GNP population for pathways associated with age-related diseases,
272 including *gastric cancer, breast cancer, neurodegenerative disease, mTOR signalling* and WNT
273 signalling (q-value <0.05, **Figure5b**, **Table S9**). Overall, relaxed selection in the dry GNP
274 population affected accumulation of deleterious mutations in age-related and in the WNT
275 pathway. Analysing the pathways affected by genes within the upper 2.5$^{th}$ DoS values –
276 corresponding to genes undergoing adaptive evolution – we found a significant enrichment for
277 mitochondrial pathways – potentially compensatory[42] – in population NF303 (**Figure5b**, **Table**
278 **S9**). Overall, our results show that differences in effective population size among wild turquoise
279 killifish are associated with an extensive relaxation of purifying selection, significantly affecting
280 genes involved in age-related diseases, and which could have cumulatively contributed to
281 reducing survival.

282 ## Discussion

283 The turquoise killifish (*Nothobranchius furzeri*) is the shortest-lived known vertebrate and while
284 its natural populations show similar timing for sexual maturation, exhibit differences in lifespan
285 along a cline of altitude and aridity in south-eastern Africa[6,12]. Here we generate an improved
286 genome assembly (NFZ v2.0) in turquoise killifish (*Nothobranchius furzeri*) and study the
287 evolutionary forces shaping genome evolution among natural populations.

288 Using the new turquoise killifish genome assembly and synteny analysis with medaka and
289 platyfish, we reconstructed the origin of the turquoise killifish sex chromosome, which appears to
290 have evolved through two independent chromosomal events, i.e. a fusion and a translocation
291 event.

292 Using the new genome assembly and pooled sequencing of natural turquoise killifish populations,
293 we found that genetic differentiation among populations of the short-lived turquoise killifish is
294 consistent with differences in demographic constraints. While we found that strong purifying
295 selection maintains low genetic diversity among populations at genomic regions underlying key
296 species-specific traits, such as in proximity to the sex-determining region, demography and
297 genetic drift largely shape genome evolution, leading to relaxation of selection and the
298 accumulation of deleterious mutations. We showed that isolated populations from an arid region,
299 dwelling at higher altitude and characterised by shorter lifespan, experienced extensive
300 population bottlenecking and a sharp decline in effective population size. We found that
301 relaxation of selection in highly drifted populations significantly affected the accumulation of
302 deleterious gene variants in pathways associated with neurodegenerative diseases and WNT-
303 signalling (**Figure 5**). While simple traits, such as male tail colour and sex have a simple genetic
304 architecture among turquoise killifish populations[13,32], we find that the complex genetic

305 architecture of lifespan differences among killifish populations[13] is entirely compatible with
306 genome-wide relaxation of selection. Additionally, the absence of genomic signature of positive
307 selection in genomic regions underlying survival QTL in killifish suggest that, rather than
308 directional selection, the neutral accumulation of deleterious mutations in short-lived populations
309 may be the evolutionary mechanism underlying survival differences among turquoise killifish
310 populations. The "antagonistic pleiotropy" evolutionary theory of ageing states that positive
311 selection could lead to the fixation of gene variants that, while overall beneficial for fitness, could
312 reduce survival and reproductive capacity in late life[43]. The lack of genomic signature of positive
313 selection at the genomic regions underlying survival QTL in turquoise killifish rather suggests
314 that accumulation of deleterious mutations may have played a key role in shaping genome and
315 phenotype differences among natural turquoise killifish populations. Historical fluctuations in the
316 size of natural turquoise killifish populations, especially in isolated and populations living in
317 more arid and elevated habitats, cause decreased efficiency of the strength of natural selection,
318 ultimately contributing to increased load of deleterious gene variants, preferentially in genes
319 associated with ageing-related diseases and in the WNT pathway.

320 Our findings highlight the role of demographic constraints in shaping life history within species.

321

322

323

324 ## Materials and Methods

325 Merging and Improvement of the Turquoise killifish genome assembly

326 ***10x Genomics read clouds***

327 A single GRZ male individual was sacrificed with MS222 (Sigma-Aldrich, Steinheim, Germany).
328 Blood was drawn from the heart and high molecular weight DNA was isolated with Qiagen
329 MagAttract kit following manufacturer's instructions. Gemcode v2 DNA library generation was
330 performed by Novogene (Beijing, China). Briefly, a proportion of the sample was run on a pulse
331 field agarose gel to confirm high molecularity > 100kb. Based on a genome size estimate of
332 1.54Gb (half of human genome), 0.6ng of DNA was used to construct 2 Gemcode libraries,
333 sequenced on two HiSeq X lanes to obtain a raw coverage of approximately 60X each. The
334 reported input molecular length by SuperNova[44] was 118kb for library 1 and 60.73kb for library
335 2. Both libraries were used to correct and scaffold the Allpath-LG assembly (see below), and
336 library 1 was also de novo assembled with the SuperNova assembler v.2 with default parameters.
337 The SuperNova assembly totaled 802.6Mb, with a contig N50 of 19.65kb, scaffolded into 6.78
338 thousand scaffolds with an N50 of 3.83Mb. Despite high continuity, however, the BUSCO[45]
339 metrics are much lower than the Allpath-LG assemblies.

340 ***Nanopore long reads***

341 DNA was extracted from a single GRZ male individual's muscle tissue by griding in liquid
342 nitrogen followed by phenol-chlorofom extraction (Sigma). The rapid sequencing kit (SQK-
343 RAD004) and the ligation kit (SQK-LSK108) were sued to prepare 6 libraries and were
344 sequenced on 6 MinION flow cells (R9.4.1). These runs yielded a total of 3.3 Gb of sequences

345 after trimming and correction by HALC[46]. For correction, Allpath-LG contigs (see below) and
346 short reads from the 10X genomic run were used.

347 ***Allpath-LG assembly***

348 Two independent short read datasets were previous collected for the GRZ strain of
349 *Nothobranchius furzeri*. Allpath-LG[47] was used on the pooled datasets. Together, 4 illumina short
350 read pair-end libraries with a fragment size distribution from 158bp to 179bp were used to
351 construct the contigs (sequence coverage 191.9X, physical coverage 153.5X), and 22 pair-end
352 and mate pair libraries distributed at 92bp, 135bp, 141bp, 176bp, 267bp, 2kb, 3kb 5kb and 10kb
353 were used for the scaffolding step (sequence coverage 135.7X, physical coverage 453.8X). The
354 published BAC library ends[16] with an insert size of 112kb were also included in the ALLPaths-
355 LG run (physical coverage 0.6X). The resulting assembly has a total contig length of
356 823,583,106bp distributed in 151,307 contigs > 1kb, with an N50 of 7.8kb. The total scaffold
357 length is 943,793,727bp distributed in 7830 scaffolds with an N50 of 421kb (with gaps). The
358 resulting assembly was further scaffolded by ARCS v1.0[48] + LINKS v1.8.5[49] with the following
359 parameters: arcs -e 50000 -c 3 -r 0.05 -s 98 and LINKS -m -d 4000 -k 20 -e 0.1 -l 3 -a 0.3 -t 2 -o
360 0 -z 500 -r -p 0.001 -x 0. This increased the scaffold N50 to 1.527 Mb. Next, scaffolds were
361 assigned to the RAD-tag linkage map[13] collected from a previous study with Allmaps [50], using
362 equal weight for the two independent mapping crosses. This procedure assigned 90.6% of the
363 assembled bases in 1131 scaffolds to 19 linkage groups, in which 76.6% can be oriented.
364 Misassemblies were corrected with the 10X genomic read cloud. Read clouds were mapped to the
365 preliminary assembly with longranger v2.1.6 using default parameters, and a custom script was
366 used to scan for sudden drops in barcode shares along the assembled linkage groups. The
367 scaffolds were broken at the nearest gap of the drop in 10x barcodes. The same ARCS + LINKS

368 pipeline was again run on the broken scaffolds, increasing the scaffold N50 to 1.823 Mb. Next,
369 BESST_RNA (https://github.com/ksahlin/BESST_RNA) was used to further scaffold the
370 assembly with RNASeq libraries, Allmaps was again used to assign the fixed scaffolds back to
371 linkage groups, increasing the assignable bases to 92.2% (879Mb) with 80.3% (765Mb) with
372 determined orientation. The assembly was again broken with longranger and reassigned to LG
373 with Allmaps, and the scaffolds were further partitioned to linkage groups due to linkage of some
374 left-over scaffolds with an assigned scaffold. Each partitioned scaffold groups were subjected to
375 the ARCS + LINKS pipeline again, to constraint the previously unassigned scaffolds onto the
376 same linkage group. Allmaps was run again on the improved scaffolds, resulting in 94.5%
377 (903.4Mb) of bases assigned and 89.1% (852Mb) of bases oriented. Longranger was run again,
378 visually checked and compared with the RADtag markers. Eleven mis-oriented positions were
379 identified and corrected. Gaps were further patched by GMCloser[51] with ~2X of nanopore long
380 reads corrected by HALC using BGI500 short PE reads with the following parameters: gmcloser
381 --blast --long_read --lr_cov 2 -l 100 -i 466 -d 13 --min_subcon 1 --min_gap_size 10 --iterate 2 --
382 mq 1 -c. The corrected long reads not mapped by GMCloser were assembled by CANU [52] into
383 7.9Mb of sequences, which are likely unassigned repeats.

384 ***Meta Assembly***

385 Five assemblies were integrated by MetAssembler[53] in the following order (ranked by BUSCO
386 scores) using a 20kb mate pair library: 1) The improved Allpaths-LG assembly assigned to
387 linkage groups produced in this study, 2) A previously published assembly with Allpaths-LG and
388 optical map[16] 3) A previously published assembly using SGA[13], 4) The SuperNova assembly
389 with only 10x Genomic reads and 5) Unassigned nanopore contigs from CANU. The final
390 assembly NFZ v2.0 has 911.5Mb of scaffolds assigned to linkage groups. Unassigned scaffolds

391 summed up to 142.2Mb, yielding a total assembly length of 1053.7Mb, approximately 2/3 of the
392 total genome size of 1.53Gb. The final assembly has 95.2% complete and 2.24% missing
393 BUSCOs.

394 ***Mapping of NCBI genbank gene annotations***

395 RefSeq mRNAs for the GRZ strain (PRJNA314891, PRJEB5837) were downloaded from
396 GenBank[54], and aligned to the assembly with Exonerate[55]. The RefSeq mRNAs have a BUSCO
397 score of 98.0% complete, 0.9% missing. The mapped gene models resulted in a BUSCO score of
398 96.1% complete, 2.1% missing.

399 ***Pseudogenome assembly generation***

400 The pseudogenomes for *Nothobranchius orthonotus* and *Nothobranchius rachovii* were generated
401 from sequencing data and the same method used in Cui *et al.* [4]. Briefly, the sequencing data were
402 mapped to the NFZ v2.0 reference genome by BWA-mem v0.7.12 in PE mode[56,57]. PCR
403 duplicates were marked with MarkDuplicates tool in the Picard (version 1.119,
404 http://broadinstitute.github.io/picard/) package. Reads were realigned around INDELs with the
405 IndelRealigner tool in GATK v3.4.46[58]. Variants were called with SAMTOOLS v1.2[59] mpileup
406 command, requiring a minimal mapping quality of 20 and a minimal base quality of 25. A
407 pseudogenome assembly was generated by substituting reference bases with the alternative base
408 in the reads. Uncovered regions, INDELs and sites with >2 alleles were masked as unknown "N".
409 The allele with more supporting reads was chosen at biallelic sites.

410 ***Mapping of longevity and sex quantitative trait loci***

411 The quantitative trait loci (QTL) markers published in Valenzano *et al.*[13] were directly provided
412 by Dario Riccardo Valenzano. In order to map the markers associated to longevity and sex, a

413 reference database was created using BLAST[60]. The nucleotide database was created with the
414 new reference genome of *N. furzeri* (NFZ v2.0). Subsequently, the QTL marker sequences were
415 mapped to the database. Only markers with full support for the total length of 95 bp were
416 considered as QTL markers.

417 ***Synteny analysis***

418 Synteny analysis was performed using orthologous information from Cui *et al.* [4] determined by
419 the UPhO pipeline[61]. For this, the 1-to-1 orthologous gene positions of the new turquoise killifish
420 reference genome (NFZ v2.0) were compared to two closely related teleost species, *Xiphophorus*
421 *maculatus* and *Oryzias latipes*. Result were visualized using Circos[62] for the genome-wide
422 comparison and the *genoPlotR* package[63] in R for the sex chromosome synteny analysis. Synteny
423 plots for orthologous chromosomes of *Xiphophorus maculatus* and *Oryzias latipes* were
424 generated with Synteny DB (http://syntenydb.uoregon.edu) [64].

425 ***Koeppen-Geiger index and bioclimatic variables***

426 The Koeppen-Geiger classification data was taken from Peel *et al.*[65] and the altitude, precipitation
427 per month, and the bioclimatic variables were obtained from the Worldclim database (v2.0[66]).
428 The monthly evapotranspiration was obtained from Trabucco and Zomer[67]. Aridity index was
429 calculated based on the sum of monthly precipitation divided by sum of monthly
430 evapotranspiration. Maps in FigureS1 were generated with QGIS version 2.18.20 combined with
431 GRASS version 7.4[68], the Koeppen-Geiger raster file, data from Natural Earth, and the river
432 systems database from Lehner *et al.*[69].

433 ***DNA isolation and pooled population sequencing***

434 The ethanol preserved fin tissue was washed with 1X PBS before extraction. Fin tissue was
435 digested with 10 µg/mL Proteinase K (Thermo Fisher) in 10 mM TRIS pH 8; 10mM EDTA; 0.5
436 SDS at 50°C overnight. DNA was extracted with phenol-chloroform-isoamylalcohol (Sigma)
437 followed by a washing step with chloroform (Sigma). Next, DNA was precipitated by adding 2.5
438 volume of chilled 100% ethanol and 0.26 volume of 7.5M Ammonium Acetate (Sigma) at -20°C
439 overnight. DNA was collected via centrifugation at 4°C at 12000rpm for 20 minutes. After a final
440 washing step with 70% ice-cold ethanol and air drying, DNA was eluted in 30µl of nuclease-free
441 water. DNA quality was checked on 1 agarose gels stained with RotiSafe (Roth) and a UV-VIS
442 spectrometer (Nanodrop2000c, Thermo Scientific). DNA concentration was measured with Qubit
443 fluorometer (BR dsDNA Assay Kit, Invitrogen). For each population, the DNA of the individuals
444 were pooled at equimolar contribution (GNP_G1_3, GNP_G4 N=29; NF414, NF303 N=30).
445 DNA pools were given to the Cologne Center of Genomic (CCG, Cologne, Germany) for library
446 preparation. The total amount of DNA provided to the sequencing facility was 3.2 µg per pooled
447 population sample. Libraries were sequenced with 150bp x 2 paired-ends on the HiSeq4000.
448 Sequencing of pooled samples resulted in a range of 419 - 517 million paired-end reads for each
449 population (**Table S2**).

450 ***Mapping of pooled sequencing reads***

451 Raw sequencing reads were trimmed using Trimmomatic-0.32 (ILLUMINACLIP:illumina-
452 adaptors.fa:3:7:7:1:true, LEADING:20, TRAILING:20, SLIDINGWINDOW:4:20,
453 MINLEN:50[70]. Data files were inspected with FastQC (version 0.11.22,
454 https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Trimmed reads were subsequently
455 mapped to the reference genome with BWA-MEM v0.7.12[56,57]. The SAM output was converted
456 into BAM format, sorted, and indexed via SAMTOOLS v1.3.1[59]. Filtering and realignment was

457 conducted with PICARD v1.119 (http://broadinstitute.github.io/picard/) and and GATK[58].
458 Briefly, the reads were relabelled, sorted, and indexed with AddOrReplaceReadGroups.
459 Duplicated reads were marked with the PICARD feature MarkDuplicates and reads were
460 realigned with first creating a target list with RealignerTargetCreator, second by IndelRealigner
461 from the GATK suite. Resulting reads were again sorted and indexed with SAMTOOLS. For
462 population genetic bioinformatics analyses the BAM files of the pooled populations were
463 converted into the required MPILEUP format via the SAMTOOLS mpileup command. Low
464 quality reads were excluded by setting a minimum mapping quality of 20 and a minimum base
465 quality of 20. Further, possible insertion and deletions (INDELs) were identified with
466 *identifygenomic-indel-regions.pl* script from the PoPoolation package[20] and were subsequently
467 removed via the *filter-pileup-by-gtf.pl* script[20]. Coding sequence positions that were identified to
468 be putative ambiguous were removed by providing the *filter-pileup-by-gtf.pl* script a custom
469 modified GTF file with the corresponding coordinates. After adapter and quality filtering,
470 mapping to the newly assembled reference genome resulted in mean genome coverage of 35x,
471 39x, and 47x for the population NF303, NF414, and GNP, respectively (**Table S2**).

472 ***Merging sequencing reads of populations from the Gonarezhou National Park***

473 Population GNP consists of two sampling sites (GNP-G1_3, GNP-G4) with very low genetic
474 differentiation (**Figure S1c**, **Table S3**). Sequencing reads of the two populations from the
475 Gonarezhou National Park (GNP) were combined used the SAMTOOLS 'merge' command. The
476 populations GNP-G1-3 and GNP-G4 were merged together and this population was subsequently
477 denoted as GNP.

478

479 ***Estimating genetic diversity***

480 Genetic diversity in the populations was estimated by calculating the nucleotide diversity $\pi$[71] and
481 Wattersons's estimator $\theta$[72]. Calculation of $\pi$ and $\theta$ was done with a sliding window approach by
482 using the *Variance-sliding.pl* script from the PoPoolation program[20]. Non-overlapping windows
483 with a length of 50 kb with a minimum count of two per SNP, minimum quality of 20 and the
484 population specific haploid pool size were used (GNP=116; NF414=60; NF303=60). Low
485 covered regions that fall below half the mean coverage of each population were excluded
486 (GNP=23; NF414=19; NF303=18), as well as regions that exceed a two times higher coverage
487 than the mean coverage (GNP=94; NF414=77; NF303=70). The upper threshold is set to avoid
488 regions with possible wrong assemblies. Mean coverage was estimated on filtered MPILEUP
489 files. Each window had to be at least covered to 30% to be included in the estimation.

490 ***Estimation of effective population size***

491 Wattersons's estimator of $\theta$[72] is referred to as the population mutation rate. The estimate is a
492 compound parameter that is calculated as the product of the effective population size ($N_e$), the
493 ploidy (2p, with p is ploidy) and the mutational rate $\mu$ ($\theta = 2pN_e\mu$). Therefore, $N_e$ can be obtained
494 when $\theta$, the ploidy and the mutational rate $\mu$ are known. The turquoise killifish is a diploid
495 organism with a mutational rate of 2.6321e−9 per base pair per generation (assuming one
496 generation per year in killifish[4] and $\theta$ estimates were obtained with PoPoolation (see Section
497 2.1.2)[20].

498 ***Estimating population differentiation index $F_{ST}$***

499 The filtered and realigned BAM files of each population were merged into a single pileup file
500 with SAMTOOLS mpileup, with a minimum mapping quality and a minimum base quality of 20.
501 The pileup was synchronized using the *mpileup2sync.jar* script from the PoPoolation2 program
502 [21]. Insertions and deletions were identified and removed with the *identify-indel-regions.pl* and

503 *filter-sync-by-gtf.pl* scripts of PoPoolation2 [21]. Again, coding sequence positions that were
504 identified to be putative ambiguous were removed by providing the *filter-pileup-by-gtf.pl* script a
505 custom modified GTF file with the corresponding coordinates. Further a synchronized pileup file
506 for genes only were generated by providing a GTF file with genes coordinates to the *create-*
507 *genewise-sync.pl* from PoPoolation2[21]. $F_{ST}$ was calculated for each pairwise comparison (GNP vs
508 NF303, GNP vs NF414, NF414 vs NF303) in a genome-wide approach using non-overlapping
509 sliding windows of 50kb with a minimum count of four per SNP, a minimum coverage of 20, a
510 maximum coverage of 94 for GNP, 77 for NF414, and 70 for NF303 and the corresponding pool
511 size of each population (N= 116; 60; 60). Each sliding window had to be at least covered to 30%
512 to be included in the estimation. The same thresholds, except the minimum covered fraction, with
513 different sliding window sizes were used to calculate the gene-wise $F_{ST}$ for the complete gene
514 body (window-size of 2000000, step-size of 2000000) and single SNPs within genes (window-
515 size of 1, step-size of 1). The non-informative positions were excluded from the output.
516 Significance of allele differences per base-pair within the gene-coordinates were calculated with
517 the fisher´s exact test implemented in *the fisher-test.pl* script of PoPoolation2 [21]. Calculation of
518 unrooted neighbor joining tree based on the genome-wide pairwise $F_{ST}$ averages was performed
519 with the ape package in R[73].

520 ***Detecting signatures of selection based on $F_{ST}$ outliers***

521 For $F_{ST}$-outlier detection, the pairwise 50kb-window $F_{ST}$-values for each comparison were Z-
522 transformed ($Z_{FST}$). Next, regions potentially under strong selection were identified by applying
523 an outlier approach. Outliers were identified as non-overlapping windows of 50 kb within the
524 0.5% of lowest and highest genetic differentiation per comparison. To reduce the number of
525 false-positive results, the outlier threshold was chosen at 0.5% highest and lowest percentile of

526 each pairwise genetic differentiation[74,75]. To find candidate genes within windows of highest
527 differentiation, a total of three selection criteria were used. First, the window-based $Z_{FST}$ value
528 had to be above the 99.5$^{th}$ percentile of pairwise genetic differentiation. Second, the gene $F_{ST}$
529 value had to be above the 99.5$^{th}$ percentile of pairwise genetic differentiation and last, the gene
530 needed to include at least one SNP with significant differentiation based on Fisher's exact test
531 (calculated with PoPoolation2[21]; P<0.001, Benjamini-Hochberg corrected P-values [76]).

532 ***Identifying polymorphic sites***

533 SNP calling was performed with Snape[34]. The program requires information of the prior
534 nucleotide diversity $\theta$. Hence, the initial values of nucleotide diversity obtained with PoPoolation
535 were used. Snape was run with folded spectrum and prior type informative. As Snape requires the
536 MPILEUP format, the previously generated MPILEUP files were used. SNP calling was
537 separately performed on coding and non-coding parts of the genome. Therefore, each population
538 MPILEUP file was filtered by coding sequence position with the *filter-pileup-by-gtf.pl* script of
539 PoPoolation. For coding sequences the --keep-mode was set to retain all coding sequences. The
540 non-coding sequences were obtained by using the default option and thus discarding the coding
541 sequences from the MPILEUP file. Snape produces a posterior probability of segregation for
542 each position. The posterior probability of segregation was used to filter low-confidence SNPs
543 and indicated in the specific section.

544 ***Divergence and polymorphisms in 0-fold and 4-fold sites***

545 Polarization of synonymous sites (four-fold degenerated sites) and non-synonymous sites (zero-
546 fold degenerated sites) was done using the pseudogenomes of outgroups *Nothobranchius*
547 *orthonotus* and *Nothobranchius rachovii*. For each population the genomic information of the
548 respective pseudogenome was extracted with bedtools getfasta command [77,78] and the derived

549 allele frequency of every position was inferred with a custom R script. Briefly, only sites with the
550 bases A, G, T or C in the outgroup pseudogenome were included and checked whether the
551 position has an alternative allele in each of the investigated populations. Positions with an
552 alternative allele present in the population data were treated as possible divergent or polymorphic
553 sites. The derived frequency was determined as frequency of the allele not present in the
554 outgroup. Occasions with an alternate allele present in the population data were treated as
555 possible divergent or polymorphic sites. Divergent sites are positions in the genome were the
556 outgroup allele is different from the allele present in the population. Polymorphic sites are sites in
557 the genome that have more than one allele segregating in the population. Only biallelic
558 polymorphic sites were used in this analysis. The DAF was determined as frequency of the allele
559 not shared with the respective outgroups. In general, positions with only one supporting read for
560 an allele were treated as monomorphic sites. SNPs with a DAF < 5% or > 95% were treated as
561 fixed mutations. Further filtering was done based on the threshold of the posterior probability of
562 >0.9 calculated with Snape (see previous subsection), combined with a minimum and maximum
563 coverage threshold per population (GNP: 24, 94; NF414:19, 77; NF303: 18, 70).

564

565 ***Asymptotic McDonald-Kreitman α***

566 The rate of substitutions that were driven to fixation by positive selection was evaluated with an
567 improved method based on the McDonald-Kreitman test[79]. The test assumes that the proportion
568 of non-synonymous mutations that are neutral has the same fixation rate as synonymous
569 mutations. Therefore, under neutrality the ratio between non-synonymous to synonymous
570 substitutions (Dn/Ds) between species is equal to the ratio of non-synonymous to synonymous
571 polymorphisms within species (Pn/Ps). If positive selection takes place, the ratio between non-

572 synonymous to synonymous substitutions between species is larger than the ratio of non-
573 synonymous to synonymous polymorphism within species[79]. The concept behind this is that the
574 selected variant reaches fixation in a shorter time than by random drift. Therefore, the selected
575 variant increases Dn, not Pn. The proportion of non-synonymous substitutions that were fixed by
576 positive selection (α) was estimated with an extension of the McDonald-Kreitman test[80]. Due to
577 the presence of slightly deleterious mutations the estimate of α can be underestimated. For this
578 reason, the method used by Messer and Petrov was implemented to calculate α as a function of
579 the derived allele frequency x[33,81]. With this method the true value of α can be inferred as the
580 asymptote of the function of α. Additionally, the value of α(x) for low derived frequencies should
581 give an estimate of the number of slightly deleterious mutations that segregate in the population.

582 ***Direction of selection (DoS)***

583 To further investigate the signature of selection, the direction of selection (DoS) index for every
584 gene was calculated[41]. DoS standardizes α to a value between -1 and 1. A positive value of DoS
585 indicates adaptive evolution (positive selection) and a negative value indicates the segregation of
586 slightly deleterious alleles, therefore weaker purifying selection[41]. This ratio is undefined for
587 genes without any information about polymorphic or substituted sites. Therefore, only genes with
588 at least one polymorphic and one substituted site were included.

589 ***Inference of distribution of fitness effects***

590 The distribution of fitness effects (DFE) was inferred using the program polyDFE2.0[35]. For this
591 analysis the unfolded site frequency spectra (SFS) of non-synonymous (0-fold) and synonymous
592 sites (4-fold) were projected into 10 chromosomes for each population. Information about the
593 fixed derived sites was included in this analysis (using *Nothobranchius orthonotus*). PolyDFE2.0
594 estimates either the full DFE, containing deleterious, neutral and beneficial mutations, or only the

595 deleterious DFE. The best model for each population was obtained using a model testing
596 approach with three different models implemented in PolyDFE2.0 (Model A, B, C). Due to
597 possible biases from erroneous polarization or unknown demography, runs accounting for
598 polarization errors and demography (+eps, +r) were included. Initial parameters were
599 automatically estimated with the –e option, as recommended. To ensure that the parameter space
600 is explored thoroughly, the basin hopping option was applied with a maximum of 500 iterations
601 (-b). The best model for each population was chosen based on the Akaike Information Criterion
602 (AIC). Confidence intervals were generated by running 200 bootstrap datasets with the same
603 parameters used to infer the best model.
604

605 ***Variant annotation***

606 Classification of changes in the coding-sequence (CDS) was done with the variant annotator
607 SnpEFF [36]. The new genome of *Nothobranchius furzeri* (NFZ v2.0) was implemented to the
608 SnpEFF pipeline. Subsequently, a database for variant annotation with the genome NFZ v2.0
609 FASTA file and the annotation GTF file was generated. For variant annotation the population
610 specific synonymous and non-synonymous sites with a change in respect to the reference genome
611 NFZ v2.0 were used to infer the impact of these sites. The possible annotation impact classes
612 were low, moderate, and high. SNPs with a frequency below 5% or above 95% were excluded for
613 this analysis. To be consistent with the analysis of the distribution of fitness effects, only
614 positions also found to be present in the *N. orthonotus* pseudogenome were considered. Positions
615 with warnings in the variant annotation were removed.

616 **Consurf analysis**

617 The Consurf score was calculated accordingly to the method used in Cui *et al.*[4]. We used the
618 Consurf [37-40] package to assign each AA a conservation score based on the evolutionary rate in
619 homologs of other vertebrates. Consurf scores were estimated for 12575 genes of *N. furzeri* and
620 synonymous and non-synonymous genomic positions were matched with the derived allele
621 frequency of *N. orthonotus* and *N. rachovii*, respectively. The derived frequencies were binned in
622 five bins and we used pairwise Wilcoxon rank sum test to assess significance after correcting for
623 multiple testing (Benjamini & Hochberg adjustment) between each subsequent bin per population
624 and matching bins between populations.

625 ***Over-representation analysis***

626 Gene ontology (GO) and pathway overrepresentation analysis was performed with the online tool
627 ConsensusPathDB (http://cpdb.molgen.mpg.de;version34)[82] using "KEGG" and "REACTOME"
628 databases. Briefly, each gene present in the outlier list was provided with an ENSEMBL human
629 gene identifier[83], if available, and entered as the target list into the user interface. All genes
630 included in the analysis and with available human ENSEMBL identifier were used as the
631 background gene list. ConsensusPathDB maps the entries to the databases and calculates the
632 enrichment score for each entity by comparing the proportion of target genes in the entity over
633 the proportion of background genes in the entity. For each of the enrichment a P-value is
634 calculated based on a hyper geometric model and is corrected for multiple testing using the false
635 discovery rate (FDR). Only GO terms and pathways with more than two genes were included.
636 Overrepresentation analysis was performed on genes falling below the 2.5th percentile or above
637 the 97.5th percentile thresholds. The percentiles for either $F_{ST}$ or DoS values were calculated with
638 the quantile() function in R.

639 **Statistical analysis and data processing**

640 Statistical analyses were performed using R studio version 1.0.136 (R version 3.3.2[84,85]) on a
641 local computer and R studio version 1.1.456 (R version 3.5.1) in a cluster environment at the
642 Max-Planck-Institute for Biology of Ageing (Cologne). Unless otherwise stated, the functions
643 t.test() and wilcox.test() in R have been used to evaluate statistical significance. To generate a
644 pipeline for data processing we used Snakemake[86]. Figure style was modified using Inkscape
645 version 0.92.4. For circular visualization of genomic data we used Circos[62].

646 **Inference of demographic population history with individual resequencing data**

647 To infer the demographic history, we performed whole genome re-sequencing of single
648 individuals from all populations resulting in mean genome coverage between 13-21x (**Table S2**).
649 Demographic history was inferred from single individual sequencing data using Pairwise
650 Sequential Markovian Coalescence (PSMC' mode from MSMC2[22]). Re-sequencing of single
651 individuals was performed with the DNA of single individuals extracted for the pooled
652 sequencing for each examined population. The Illumina short-insert library was constructed
653 based on a published protocol[87]. Extracted DNA (500ng) was digested with fragmentase (New
654 England Biolabs) for 20min at 37°C, followed by end-repair and A-tailing (1.0µl NEB End-repair
655 buffer, 0.5µl Klenow fragment, 0.5µl Taq.Polymerase, 0.2µl T4 polynucleotide kinase, 10µl
656 reaction volume, 30min at 25°C, 30min at 75°C) and adapter ligation (NEB Quick ligase buffer
657 12.5µl, Quick ligase 0.5µl, 1µl adapter P1 (D50X), 1µl adapter P2 (universal), 5µM each; 20min
658 at 20°C, 25µl reaction volume). Next, ligation mix was diluted to 50µl and used 0.583:1 volume
659 of home-brewed SPRI beads (SPRI binding buffer: 2.5M NaCL, 20mM PEG 8000, 10mM Tric-
660 HCL, 1mM EDTA,oh=8, 1mL TE-washed SpeedMag beads GE Healthcare, 65152105050250
661 per 100mL buffer) for purification. The ligation products were amplified with 9 PCR cycles using
662 KAPA Hifi kit (Roche, P5 universal primer and P7 indexed primer D7XX). The samples were

663 pooled and sequenced on Hiseq X. Raw sequencing reads were trimmed using Trimmomatic-
664 0.32 (ILLUMINACLIP:illumina-adaptors.fa:3:7:7:1:true, LEADING:20, TRAILING:20,
665 SLIDINGWINDOW:4:20, MINLEN:50)[70]. Data files were inspected with FastQC v0.11.22.
666 Trimmed reads were subsequently mapped to the reference genome with BWA-MEM (version
667 0.7.12). The SAM output was converted into BAM format, sorted, and indexed via SAMTOOLS
668 v1.3.1[59]. Filtering and realignment was conducted with PICARD v1.119 and GATK[58]. Briefly,
669 the reads were relabeled, sorted, and indexed with AddOrReplaceReadGroups. Duplicated reads
670 were marked with the PICARD feature MarkDuplicates and reads were realigned with first
671 creating a target list with RealignerTargetCreator, second by IndelRealigner from the GATK
672 suite. Resulting reads were again sorted and indexed with SAMTOOLS. Next, the guidance for
673 PSMC' (https://github.com/stschiff/msmc/blob/master/guide.md) was followed; VCF-files and
674 masked files were generated with the *bamCaller.py* script (MSMC-tools package). This step
675 requires the chromosome coverage information to mask regions with too low or too high
676 coverage. As recommended in the guidelines, the average coverage per chromosome was
677 calculated using SAMTOOLS. In addition, this step was performed using a coverage threshold of
678 18 as recommended by Nadachowska-Brzyska *et al*.[88]. Final input data was generated using the
679 *generate_multihetsep.py* script (MSMC-tools package). Subsequently, for each sample PSMC'
680 was run independently. Bootstrapping was performed for 30 samples per individual and input
681 files were generated with the *multihetsep_bootstrap.py* script (MSMCtools package).

682 ***Analysis of differential expressed genes with age***

683 We downloaded the previously published RNaseq data from a longitudinal study of
684 *Nothobranchius furzeri*[16]. The data set contains five time points (5w, 12w, 20w, 27w, 39w) in
685 three different tissues (liver, brain, skin). The raw reads were mapped to the NFZ v2.0 reference

genome and subsequently counted using STAR (version 2.6.0.c)[89] and FeatureCounts (version 1.6.2)[90]. We performed statistical analysis of differential expression with age using DESeq2[91] and age as factor. Genes are classified as upregulated in young (log(FoldChange) < 0, adjusted p < 0.01), upregulated in old (log(FoldChange) > 0,adjusted p < 0.01).

691 ## Figures

692 **Figure 1. Demography and natural occurrence of turquoise killifish populations.** a) Inferred
693 ancestral effective population size ($N_e$) (using PSMC') on y-axis and past generations on x-axis in
694 GNP (red, orange), NF414 (black, grey) and NF303 (blue). Inset: unrooted neighbour joining tree
695 based on pairwise genetic differentiation ($F_{ST}$) values. b) Geographical locations of sampled
696 natural population of turquoise killifish (*Nothobranchius furzeri)*. The area of the coloured circles
697 represents the estimated effective population size ($N_e$) based on $\theta_{\text{Watterson}}$. c) Natural environment
698 of turquoise killifish and schematic of the annual life cycle. Figure 1 was partly made with
699 Biorender®.

700 **Figure 2. Genomic regions of high and low genetic divergence between pairs turquoise**
701 **killifish populations.** Left) Genomic regions with high or low genetic differentiation between
702 turquoise killifish populations identified with an $F_{ST}$ outlier approach. Z-transformed $F_{ST}$ values
703 of all pairwise comparison in solid lines, with "NF303vsNF414" in yellow, "NF303vsGNP" in
704 blue, and "NF414vsGNP" in green. The significance thresholds of upper and lower 5‰ are
705 shown in dotted lines with same colour coding. Center) Circos plot of Z-transformed $F_{ST}$ values
706 between all pairwise comparisons with "NF303vsNF414" in the inner circle (yellow),
707 "NF414vsGNP" in the middle circle (green), and "NF303vsGNP" in the outer circle (blue).
708 Right) Pairwise genetic differentiation based on $F_{ST}$ in the four main clusters associated with
709 lifespan (QTL from Valenzano *et al.*[13]).

710 **Figure 3. Synteny and sex chromosome evolution in turquoise killifish.** a) Synteny circos
711 plots based on 1-to-1 orthologous gene location between the new turquoise killifish assembly
712 (black chromosomes) and platyfish (*Xiphophorus maculatus*, coloured chromosomes, left circos
713 plot) and between the new turquoise killifish assembly (black chromosomes) and medaka

714 (*Oryzias latipes*, coloured chromosomes, right circos plot). Orthologous genes in concordant
715 order are visualized as one syntenic block. Synteny regions are connected via colour-coded
716 ribbons, based on their chromosomal location in platyfish or medaka. If the direction of the
717 syntenic sequence is inverted compared to the compared species, the ribbon is twisted. Outer data
718 plot shows –log(q-value) of survival quantitative trait loci (QTL, ordinate value between 0 and
719 3.5, every value above 3.5 is visualized at 3.5[13]) and the inner data plot shows –log(q-value) of
720 the sex QTL (ordinate value between 0 and 3.5, every value above 3.5 is visualized at 3.5). Boxes
721 between the two circos plots show genes within the peak regions of the four highest –log(q-value)
722 of survival QTL on independent chromosomes (red box) and the highest association to sex (black
723 box). b) High resolution synteny map between the sex-chromosome of the turquoise killifish
724 (Chr3) with platyfish chromosome 16 and 3 in the upper plot, and between the turquoise killifish
725 and medaka chromosome 8 and 16 (lower plot). The middle plot shows the QTLs for survival and
726 sex along the turquoise killifish sex chromosome. c) Model of sex chromosome evolution in the
727 turquoise killifish. A translocation event within one ancestral autosome led to the emergence of a
728 chromosomal region harbouring a new sex-determining-gene (SDG). The fusion of a second
729 autosome led to the formation of the current structure of the turquoise killifish sex chromosome.

730 **Figure 4. Genome-wide signatures of natural and relaxed selection in turquoise killifish**
731 **populations.** Asymptotic McDonald-Kreitman alpha (MK $\boldsymbol{\alpha}$) analysis based on derived
732 frequency bins using as outgroups a) *Nothobranchius orthonotus* and b) *Nothonbranchius*
733 *rachovii*. Population GNP is shown in red, NF414 in black, and NF303 in blue. c) Proportion of
734 non-synonymous SNPs binned in allele frequencies of non-reference (alternative) alleles for GNP
735 (red), NF414 (black) and NF303 (blue). d) Negative distribution of fitness effects of populations
736 GNP (red), NF414 (black) and NF303 (blue) with cumulative proportion of deleterious SNPs on
737 y-axis and the compound measure of $4N_es$ on x-axis. e) Proportion of different effect types of

738 SNPs in coding sequences of all populations. The effect on amino acid sequence for each genetic
739 variant is represented by colours (legend). Significance is based on ratio between synonymous
740 effects to non-synonymous effects (significance based on Chi-square test).

741 **Figure 5. Pathway enrichment in genes under adaptive and neutral evolution in turquoise**
742 **killifish populations.** a) Distribution of direction of selection (DoS) represented with median of
743 distribution for population GNP (red), NF414 (grey) and NF303 (blue). Left panel shows DoS
744 distribution computed using *Nothobranchius orthonotus* as outgroup and right panel shows DoS
745 distribution computed using *Nothobranchius rachovii* as outgroup. Significance based on
746 Wilcoxon-Rank-Sum test. b) Pathway over-representation analysis of genes below the 2.5% level
747 of gene-wise DoS values are shown with red background and above the 97.5% level of gene-wise
748 DoS values are shown with green background. Only pathway terms with significance level of
749 FDR corrected q-value < 0.05 are shown (in -log(q-value)). Terms enriched in population GNP
750 have red dots, enriched in population NF414 have black dots, and enriched in population NF303
751 have blue dots, respectively.

752 ## Acknowledgments

753 We would like to thank Patience and Edson Gandiwa for their administrative support, Tamuka
754 Nhiwatiwa for helping with logistics and samples handling; Evious Mpofu, Hugo and Elsabe van
755 der Westhuizen and all the rangers of the Gonarezhou National Park for their support in the field.
756 We are thankful to Zimbabwe National Parks for allowing our team to conduct research in the
757 Gonarezhou National Park; Itamar Harel, Matej Polacik and Radim Blazek for hands-on
758 contribution with the field work. We further thank all members of the Valenzano lab for their
759 continuous scientific input and support. The Czech Science Foundation provided financial
760 support to MR for sampling Mozambican populations (19-01789S). This project was funded by

761 the Max Planck Institute for Biology of Ageing, the Max Planck Society and the CECAD at the
762 University of Cologne.

763 ## References


764 1 Furness, A. I. The evolution of an annual life cycle in killifish: adaptation to ephemeral aquatic
765 environments through embryonic diapause. *Biol. Rev. Camb. Philos. Soc.* **91**, 796-812,
766 doi:10.1111/brv.12194 (2016).
767 2 Cellerino, A., Valenzano, D. R. & Reichard, M. From the bush to the bench: the annual
768 *Nothobranchius* fishes as a new model system in biology. *Biol Rev Camb Philos Soc* **91**, 511-533,
769 doi:10.1111/brv.12183 (2016).
770 3 Hu, C. K. & Brunet, A. The African turquoise killifish: A research organism to study vertebrate aging
771 and diapause. *Aging Cell* **17**, e12757, doi:10.1111/acel.12757 (2018).
772 4 Cui, R. *et al.* Relaxed Selection Limits Lifespan by Increasing Mutation Load. *Cell*,
773 doi:10.1016/j.cell.2019.06.004 (2019).
774 5 Kim, Y., Nam, H. G. & Valenzano, D. R. The short-lived African turquoise killifish: an emerging
775 experimental model for ageing. *Dis Model Mech* **9**, 115-129, doi:10.1242/dmm.023226 (2016).
776 6 Blazek, R. *et al.* Repeated intraspecific divergence in life span and aging of African annual fishes
777 along an aridity gradient. *Evolution* **71**, 386-402, doi:10.1111/evo.13127 (2017).
778 7 Di Cicco, E., Tozzini, E. T., Rossi, G. & Cellerino, A. The short-lived annual fish *Nothobranchius furzeri*
779 shows a typical teleost aging process reinforced by high incidence of age-dependent neoplasias.
780 *Exp Gerontol* **46**, 249-256, doi:10.1016/j.exger.2010.10.011 (2011).
781 8 Wendler, S., Hartmann, N., Hoppe, B. & Englert, C. Age-dependent decline in fin regenerative
782 capacity in the short-lived fish *Nothobranchius furzeri*. *Aging Cell* **14**, 857-866,
783 doi:10.1111/acel.12367 (2015).
784 9 Ahuja, G. *et al.* Loss of genomic integrity induced by lysosphingolipid imbalance drives ageing in
785 the heart. *EMBO Rep* **20**, doi:10.15252/embr.201847407 (2019).
786 10 Valenzano, D. R., Terzibasi, E., Cattaneo, A., Domenici, L. & Cellerino, A. Temperature affects
787 longevity and age-related locomotor and cognitive decay in the short-lived fish *Nothobranchius*
788 *furzeri*. *Aging Cell* **5**, 275-278, doi:10.1111/j.1474-9726.2006.00212.x (2006).
789 11 Smith, P. *et al.* Regulation of life span by the gut microbiota in the short-lived African turquoise
790 killifish. *Elife* **6**, doi:10.7554/eLife.27014 (2017).
791 12 Terzibasi, E. *et al.* Large differences in aging phenotype between strains of the short-lived annual
792 fish *Nothobranchius furzeri*. *PLoS One* **3**, e3866, doi:10.1371/journal.pone.0003866 (2008).
793 13 Valenzano, D. R. *et al.* The African turquoise killifish genome provides insights into evolution and
794 genetic architecture of lifespan. *Cell* **163**, 1539-1554, doi:10.1016/j.cell.2015.11.008 (2015).
795 14 Vrtilek, M., Zak, J., Polacik, M., Blazek, R. & Reichard, M. Longitudinal demographic study of wild
796 populations of African annual killifish. *Sci Rep* **8**, 4774, doi:10.1038/s41598-018-22878-6 (2018).
797 15 Kirschner, J. *et al.* Mapping of quantitative trait loci controlling lifespan in the short-lived fish
798 *Nothobranchius furzeri*--a new vertebrate model for age research. *Aging Cell* **11**, 252-261,
799 doi:10.1111/j.1474-9726.2011.00780.x (2012).
800 16 Reichwald, K. *et al.* Insights into sex chromosome evolution and aging from the genome of a short-
801 lived fish. *Cell* **163**, 1527-1538, doi:10.1016/j.cell.2015.10.071 (2015).
802 17 Dorn, A. *et al.* Phylogeny, genetic variability and colour polymorphism of an emerging animal
803 model: the short-lived annual *Nothobranchius* fishes from southern Mozambique. *Mol Phylogenet*
804 *Evol* **61**, 739-749, doi:10.1016/j.ympev.2011.06.010 (2011).
805 18 Bartakova, V. *et al.* Strong population genetic structuring in an annual fish, *Nothobranchius furzeri*,
806 suggests multiple savannah refugia in southern Mozambique. *BMC Evol Biol* **13**, 196,
807 doi:10.1186/1471-2148-13-196 (2013).
808 19 Bartáková, V., Reichard, M., Blažek, R., Polačik, M. & Bryja, J. Terrestrial fishes: rivers are barriers
809 to gene flow in annual fishes from the African savanna. *Journal of Biogeography* **42**, 1832-1844,
810 doi:10.1111/jbi.12567 (2015).

811 20 Kofler, R. *et al.* PoPoolation: a toolbox for population genetic analysis of next generation
812 sequencing data from pooled individuals. *PLoS One* **6**, e15925, doi:10.1371/journal.pone.0015925
813 (2011).
814 21 Kofler, R., Pandey, R. V. & Schlotterer, C. PoPoolation2: identifying differentiation between
815 populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* **27**, 3435-3436,
816 doi:10.1093/bioinformatics/btr589 (2011).
817 22 Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple
818 genome sequences. *Nat Genet* **46**, 919-925, doi:10.1038/ng.3015 (2014).
819 23 Baumgart, M. *et al.* A miRNA catalogue and ncRNA annotation of the short-living fish
820 Nothobranchius furzeri. *BMC Genomics* **18**, 693, doi:10.1186/s12864-017-3951-8 (2017).
821 24 Ferdinandusse, S. *et al.* HIBCH mutations can cause Leigh-like disease with combined deficiency of
822 multiple mitochondrial respiratory chain enzymes and pyruvate dehydrogenase. *Orphanet J Rare*
823 *Dis* **8**, 188, doi:10.1186/1750-1172-8-188 (2013).
824 25 Huff, M. W. & Telford, D. E. Lord of the rings--the mechanism for oxidosqualene:lanosterol cyclase
825 becomes crystal clear. *Trends Pharmacol Sci* **26**, 335-340, doi:10.1016/j.tips.2005.05.004 (2005).
826 26 Osanai, T. *et al.* Novel anti-aging gene NM_026333 contributes to proton-induced aging via NCX1-
827 pathway. *J Mol Cell Cardiol* **125**, 174-184, doi:10.1016/j.yjmcc.2018.10.021 (2018).
828 27 Orlov, S. V. *et al.* Novel repressor of the human FMR1 gene - identification of p56 human (GCC)(n)-
829 binding protein as a Kruppel-like transcription factor ZF5. *FEBS J* **274**, 4848-4862,
830 doi:10.1111/j.1742-4658.2007.06006.x (2007).
831 28 Nojima, H. *et al.* Syntabulin, a motor protein linker, controls dorsal determination. *Development*
832 **137**, 923-933, doi:10.1242/dev.046425 (2010).
833 29 Ulvila, J. *et al.* Cofilin regulator 14-3-3zeta is an evolutionarily conserved protein required for
834 phagocytosis and microbial resistance. *J Leukoc Biol* **89**, 649-659, doi:10.1189/jlb.0410195 (2011).
835 30 Mauxion, F., Preve, B. & Seraphin, B. C2ORF29/CNOT11 and CNOT10 form a new module of the
836 CCR4-NOT complex. *RNA Biol* **10**, 267-276, doi:10.4161/rna.23065 (2013).
837 31 Kell, M. J. *et al.* Targeted deletion of the zebrafish actin-bundling protein L-plastin (lcp1). *PLoS One*
838 **13**, e0190353, doi:10.1371/journal.pone.0190353 (2018).
839 32 Valenzano, D. R. *et al.* Mapping loci associated with tail color and sex determination in the short-
840 lived fish *Nothobranchius furzeri*. *Genetics* **183**, 1385-1395, doi:10.1534/genetics.109.108670
841 (2009).
842 33 Messer, P. W. & Petrov, D. A. Frequent adaptation and the McDonald-Kreitman test. *Proc Natl*
843 *Acad Sci U S A* **110**, 8615-8620, doi:10.1073/pnas.1220835110 (2013).
844 34 Raineri, E. *et al.* SNP calling by sequencing pooled samples. *BMC Bioinformatics* **13**, 239,
845 doi:10.1186/1471-2105-13-239 (2012).
846 35 Tataru, P., Mollion, M., Glemin, S. & Bataillon, T. Inference of Distribution of Fitness Effects and
847 Proportion of Adaptive Substitutions from Polymorphism Data. *Genetics* **207**, 1103-1119,
848 doi:10.1534/genetics.117.300323 (2017).
849 36 Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide
850 polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-
851 3. *Fly (Austin)* **6**, 80-92, doi:10.4161/fly.19695 (2012).
852 37 Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. Rate4Site: an algorithmic tool for the
853 identification of functional regions in proteins by surface mapping of evolutionary determinants
854 within their homologues. *Bioinformatics* **18 Suppl 1**, S71-77,
855 doi:10.1093/bioinformatics/18.suppl_1.s71 (2002).
856 38 Mayrose, I., Graur, D., Ben-Tal, N. & Pupko, T. Comparison of site-specific rate-inference methods
857 for protein sequences: Empirical Bayesian methods are superior. *Molecular Biology and Evolution*
858 **21**, 1781-1791, doi:10.1093/molbev/msh194 (2004).

859 39 Glaser, F. *et al.* ConSurf: identification of functional regions in proteins by surface-mapping of
860 phylogenetic information. *Bioinformatics* **19**, 163-164, doi:10.1093/bioinformatics/19.1.163
861 (2003).
862 40 Ashkenazy, H. *et al.* ConSurf 2016: an improved methodology to estimate and visualize
863 evolutionary conservation in macromolecules. *Nucleic Acids Res.* **44**, W344-350,
864 doi:10.1093/nar/gkw408 (2016).
865 41 Stoletzki, N. & Eyre-Walker, A. Estimation of the neutrality index. *Mol Biol Evol* **28**, 63-70,
866 doi:10.1093/molbev/msq249 (2011).
867 42 Cui, R. *et al.* Relaxed Selection Limits Lifespan by Increasing Mutation Load. *Cell* **178**, 385-399 e320,
868 doi:10.1016/j.cell.2019.06.004 (2019).
869 43 Williams, G. C. Pleiotropy, natural selection, and the evolution of senescence. *Evolution* **11**, 398-
870 411 (1957).
871 44 Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid
872 genome sequences. *Genome Res* **27**, 757-767, doi:10.1101/gr.214874.116 (2017).
873 45 Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing
874 genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**,
875 3210-3212, doi:10.1093/bioinformatics/btv351 (2015).
876 46 Bao, E. & Lan, L. HALC: High throughput algorithm for long read error correction. *BMC*
877 *Bioinformatics* **18**, 204, doi:10.1186/s12859-017-1610-3 (2017).
878 47 Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel
879 sequence data. *Proc Natl Acad Sci U S A* **108**, 1513-1518, doi:10.1073/pnas.1017351108 (2011).
880 48 Coombe, L. *et al.* ARKS: chromosome-scale scaffolding of human genome drafts with linked read
881 kmers. *BMC Bioinformatics* **19**, 234, doi:10.1186/s12859-018-2243-x (2018).
882 49 Warren, R. L. *et al.* LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads.
883 *Gigascience* **4**, 35, doi:10.1186/s13742-015-0076-3 (2015).
884 50 Tang, H. *et al.* ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol* **16**, 3,
885 doi:10.1186/s13059-014-0573-1 (2015).
886 51 Kosugi, S., Hirakawa, H. & Tabata, S. GMcloser: closing gaps in assemblies accurately with a
887 likelihood-based selection of contig or long-read alignments. *Bioinformatics* **31**, 3733-3741,
888 doi:10.1093/bioinformatics/btv465 (2015).
889 52 Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and
890 repeat separation. *Genome Res* **27**, 722-736, doi:10.1101/gr.215087.116 (2017).
891 53 Wences, A. H. & Schatz, M. C. Metassembler: merging and optimizing de novo genome assemblies.
892 *Genome Biol.* **16** (2015).
893 54 Sayers, E. W. *et al.* GenBank. *Nucleic Acids Res* **47**, D94-D99, doi:10.1093/nar/gky989 (2019).
894 55 Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison.
895 *BMC Bioinformatics* **6** (2005).
896 56 Li, H. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. *arXiv*
897 *preprint*, doi:eprint arXiv:1303.3997 (2013).
898 57 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform.
899 *Bioinformatics* **26**, 589-595 (2010).
900 58 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-
901 generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).
902 59 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079,
903 doi:10.1093/bioinformatics/btp352 (2009).
904 60 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool.
905 *J Mol Biol* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).
906 61 Ballesteros, J. A. & Hormiga, G. A New Orthology Assessment Method for Phylogenomic Data:
907 Unrooted Phylogenetic Orthology. *Mol Biol Evol* **33**, 2481, doi:10.1093/molbev/msw153 (2016).

62 Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-1645, doi:10.1101/gr.092759.109 (2009).
63 Guy, L., Kultima, J. R. & Andersson, S. G. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**, 2334-2335, doi:10.1093/bioinformatics/btq413 (2010).
64 Catchen, J. M., Conery, J. S. & Postlethwait, J. H. Automated identification of conserved synteny after whole-genome duplication. *Genome Res* **19**, 1497-1505, doi:10.1101/gr.090480.108 (2009).
65 Peel, M. C., Finlayson, B. L. & McMahon, T. A. Updated world map of the Koppen-Geiger climate classification. *Hydrol Earth Syst Sc* **11**, 1633-1644, doi:DOI 10.5194/hess-11-1633-2007 (2007).
66 Fick, S. E. & Hijmans, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302-4315, doi:10.1002/joc.5086 (2017).
67 Trabucco, A. & Zomer, R. Global Aridity Index and Potential Evapotranspiration Climate Database v2. *CGIAR Consortium for Spatial Information* **available at: https://cgiarcsi.community/2019/01/24/global-aridity-index-andpotential-evapotranspiration-climate-database-v2/** (2019).
68 Neteler, M., Bowman, M. H., Landa, M. & Metz, M. GRASS GIS: A multi-purpose open source GIS. *Environ Modell Softw* **31**, 124-130, doi:10.1016/j.envsoft.2011.11.014 (2012).
69 Lehner, B., Verdin, K. & A., J. HydroSHEDS Technical Documentation. *World Wildlife Fund US, Washington* **available at http://hydrosheds.cr.usgs.gov** (2006).
70 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
71 Nei, M. & Li, W. H. Mathematical-Model for Studying Genetic-Variation in Terms of Restriction Endonucleases. *P Natl Acad Sci USA* **76**, 5269-5273, doi:DOI 10.1073/pnas.76.10.5269 (1979).
72 Watterson, G. A. Number of Segregating Sites in Genetic Models without Recombination. *Theoretical Population Biology* **7**, 256-276, doi:Doi 10.1016/0040-5809(75)90020-9 (1975).
73 Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290, doi:10.1093/bioinformatics/btg412 (2004).
74 Pruisscher, P., Nylin, S., Gotthard, K. & Wheat, C. W. Genetic variation underlying local adaptation of diapause induction along a cline in a butterfly. *Mol Ecol*, doi:10.1111/mec.14829 (2018).
75 Guo, B., Li, Z. & Merila, J. Population genomic evidence for adaptive differentiation in the Baltic Sea herring. *Mol Ecol* **25**, 2833-2852, doi:10.1111/mec.13657 (2016).
76 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J R Stat Soc B* **57**, 289-300 (1995).
77 Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **47**, 11 12 11-34, doi:10.1002/0471250953.bi1112s47 (2014).
78 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).
79 McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351**, 652-654, doi:10.1038/351652a0 (1991).
80 Smith, N. G. & Eyre-Walker, A. Adaptive protein evolution in *Drosophila*. *Nature* **415**, 1022-1024, doi:10.1038/4151022a (2002).
81 Haller, B. C. & Messer, P. W. asymptoticMK: A Web-Based Tool for the Asymptotic McDonald-Kreitman Test. *G3 (Bethesda)* **7**, 1569-1575, doi:10.1534/g3.117.039693 (2017).
82 Herwig, R., Hardt, C., Lienhard, M. & Kamburov, A. Analyzing and interpreting genome data at the network level with ConsensusPathDB. *Nat Protoc* **11**, 1889-1907, doi:10.1038/nprot.2016.117 (2016).
83 Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res* **46**, D754-D761, doi:10.1093/nar/gkx1098 (2018).
84 RStudio: Integrated Development for R (RStudio, Inc. , Boston, MA, 2015).
85 R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing

957 , Vienna, Austra, 2014).
958 86 Koster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics*
959 **28**, 2520-2522, doi:10.1093/bioinformatics/bts480 (2012).
960 87 Rowan, B. A., Patel, V., Weigel, D. & Schneeberger, K. Rapid and inexpensive whole-genome
961 genotyping-by-sequencing for crossover localization and fine-scale genetic mapping. *G3*
962 *(Bethesda)* **5**, 385-398, doi:10.1534/g3.114.016501 (2015).
963 88 Nadachowska-Brzyska, K., Burri, R., Smeds, L. & Ellegren, H. PSMC analysis of effective population
964 sizes in molecular ecology and its application to black-and-white *Ficedula* flycatchers. *Mol Ecol* **25**,
965 1058-1072, doi:10.1111/mec.13540 (2016).
966 89 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
967 90 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning
968 sequence reads to genomic features. *Bioinformatics* **30**, 923-930,
969 doi:10.1093/bioinformatics/btt656 (2014).
970 91 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-
971 seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

972

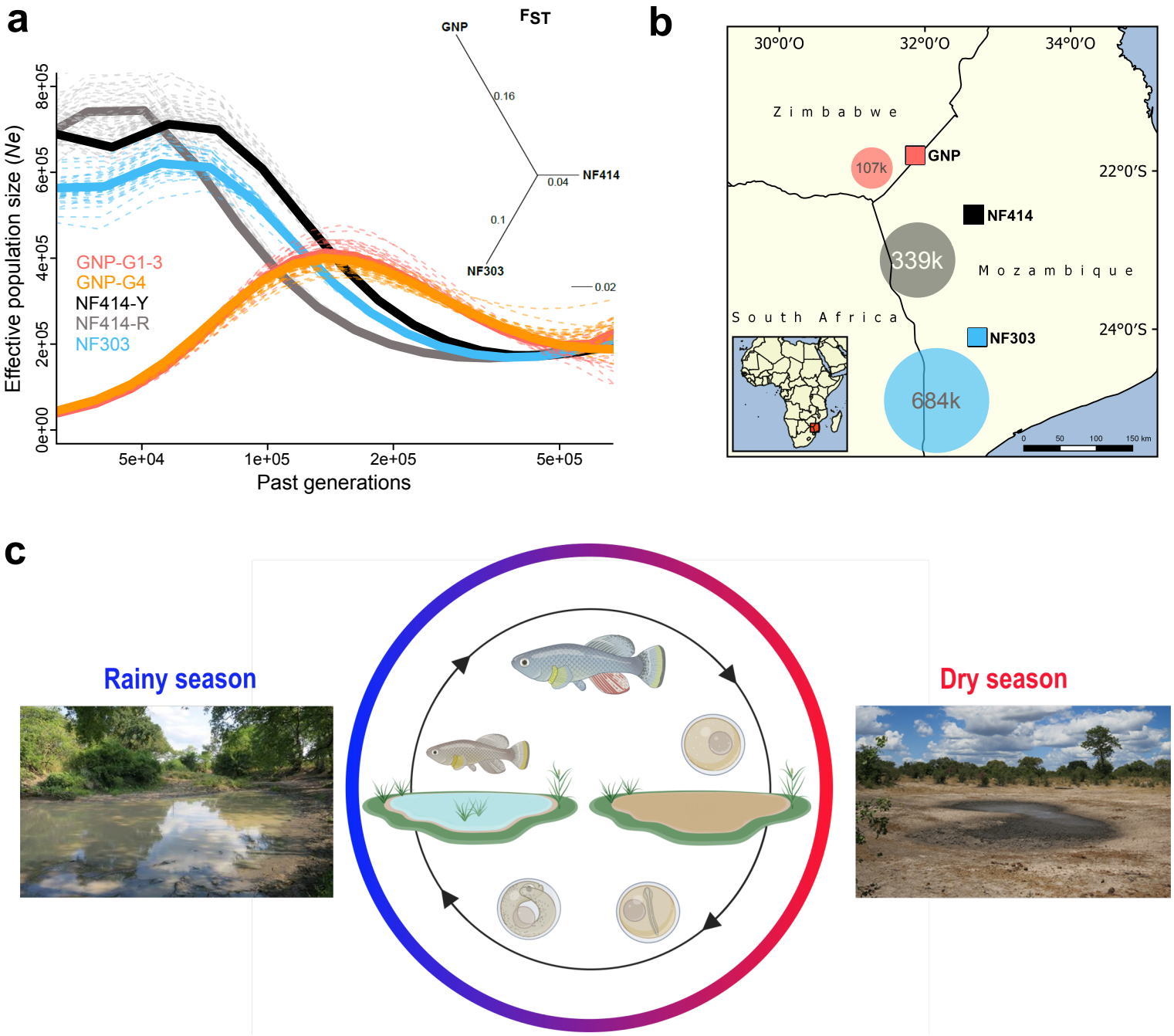


**Figure 1. Demography and natural occurence of turquoise killifish populations.** a) Inferred ancestral effective population size ($N_e$) (using PSMC') on y-axis and past generations on x-axis in GNP (red, orange), NF414 (black, grey) and NF303 (blue). Inset: unrooted neighbour joining tree based on pairwise genetic differentiation ($F_{ST}$) values. b) Geographical locations of sampled natural population of turquoise killifish (*Nothobranchius furzeri*). The area of coloured circles represent the estimated effective population size ($N_e$) based on $\theta_{Watterson}$. c) Natural environment of turquoise killifish and schematic of the annual life cycle. Figure 1 partly made with Biorender®.

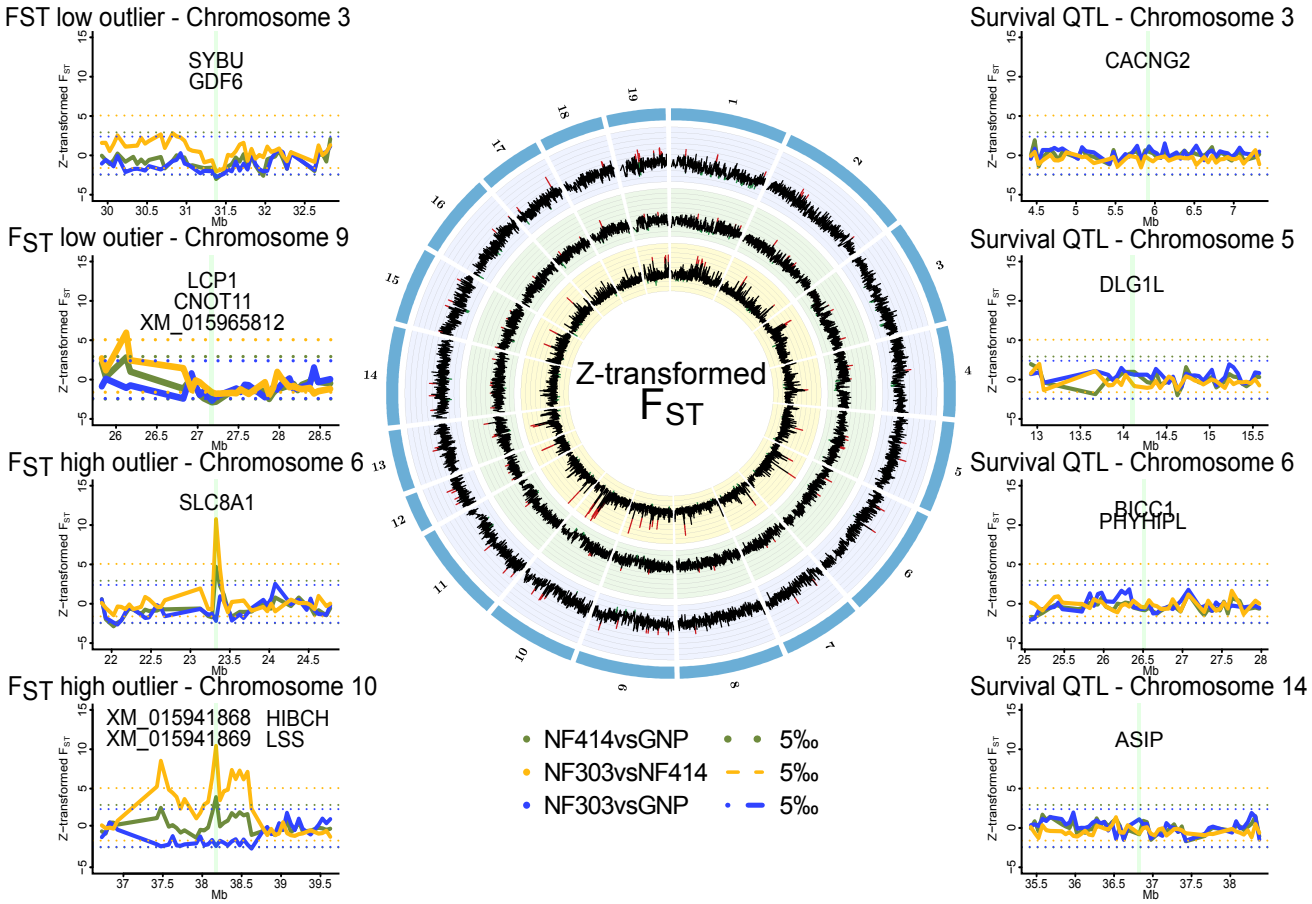


**Figure 2. Genomic regions of high and low genetic divergence between pairs of turquoise killifish populations.** Left) Genomic regions with high or low genetic differentiation between turquoise killifish populations identified with an $F_{ST}$ outlier approach. Z-transformed $F_{ST}$ values of all pairwise comparison in solid lines, with "NF303vsNF414" in yellow, "NF303vsGNP" in blue, and "NF414vsGNP" in green. The significance thresholds of upper and lower 5‰ are shown in spotted lines with same colour coding. Center) Circos plot of Z-transformed $F_{ST}$ values between all pairwise comparisons with "NF303vsNF414" in the inner circle (yellow), "NF414vsGNP" in the middle circle (green), and "NF303vsGNP" in the outer circle (blue). Right) Pairwise genetic differentiation based on FST in the four main clusters associated with lifespan (QTL from Valenzano et al.[13]).

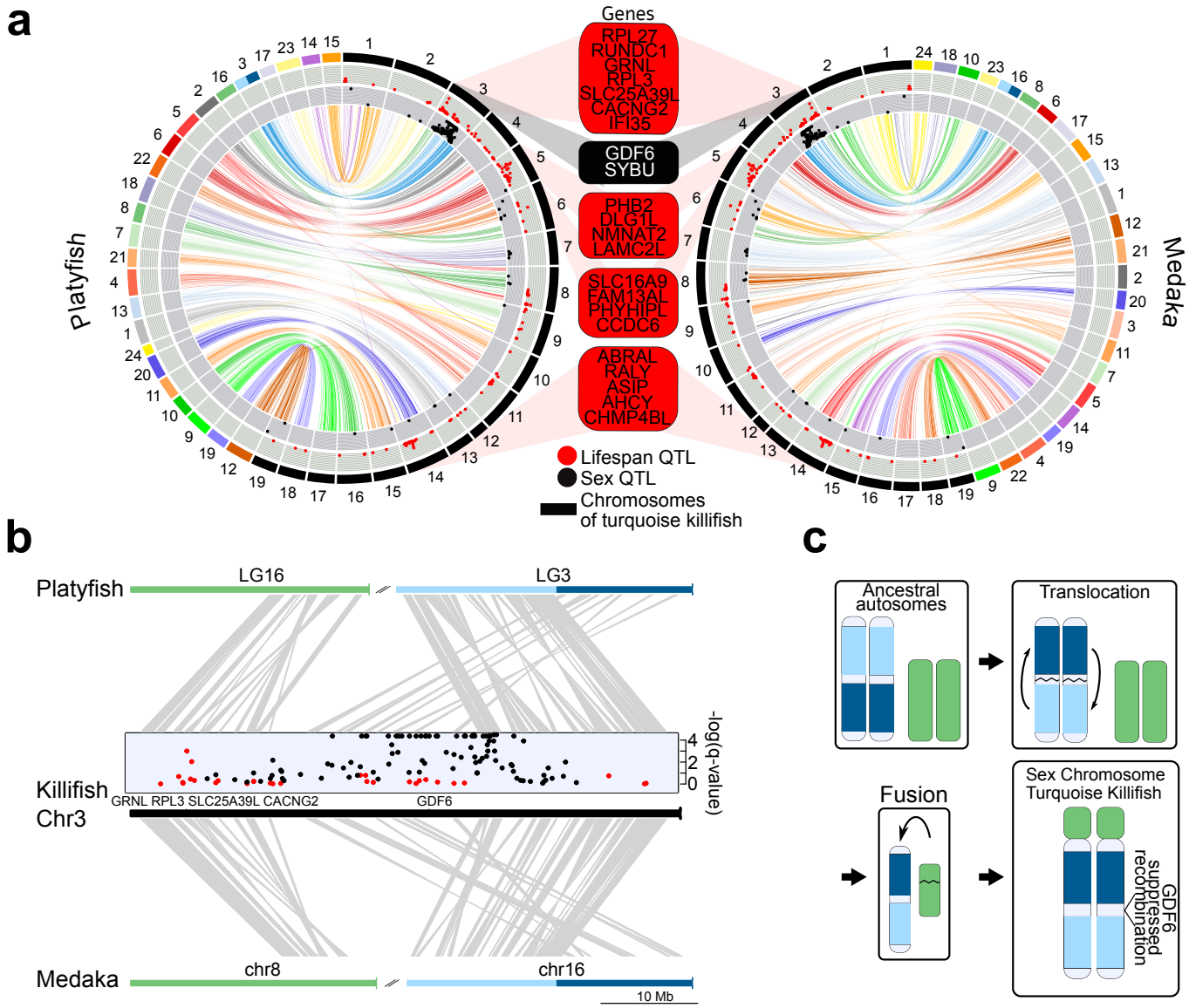


**Figure 3. Synteny and sex chromosome evolution in turquoise killifish.** a) Synteny circos plots based on 1-to-1 orthologous gene location between the new turquoise killifish assembly (black chromosomes) and platyfish (*Xiphophorus maculatus*, coloured chromosomes, left circos plot) and between the new turquoise killifish assembly (black chromosomes) and medaka (*Oryzias latipes*, coloured chromosomes, right circos plot). Orthologous genes in concordant order are visualized as one syntenic block. Synteny regions are connected via colour-coded ribbons, based on their chromosomal location in platyfish or medaka. If the direction of the syntenic sequence is inverted compared to the compared species, the ribbon is twisted. Outer data plot shows –log(q-value) of survival quantitative trait loci (QTL, ordinate value between 0 and 3.5, every value above 3.5 is visualized at 3.5[13]) and the inner data plot shows –log(q-value) of the sex QTL (ordinate value between 0 and 3.5, every value above 3.5 is visualized at 3.5). Boxes between the two circos plots show genes within the peak regions of the four highest –log(q-value) of survival QTL on independent chromosomes (red box) and the highest association to sex (black box) . b) High resolution synteny map between the sex-chromosome of the turquoise killifish (Chr3) with platyfish chromosome 16 and 3 in the upper plot, and between the turquoise killifish and medaka chromosome 8 and 16 (lower plot). The middle plot shows the QTLs for survival and sex along the turquoise killifish sex chromosome. c) Model of sex chromosome evolution in the turquoise killifish. A translocation event within one ancestral autosome led to the emergence of a chromosomal region harbouring a new sex-determining-gene (SDG). The fusion of a second autosome led to the formation of the current structure of the turquoise killifish sex chromosome.

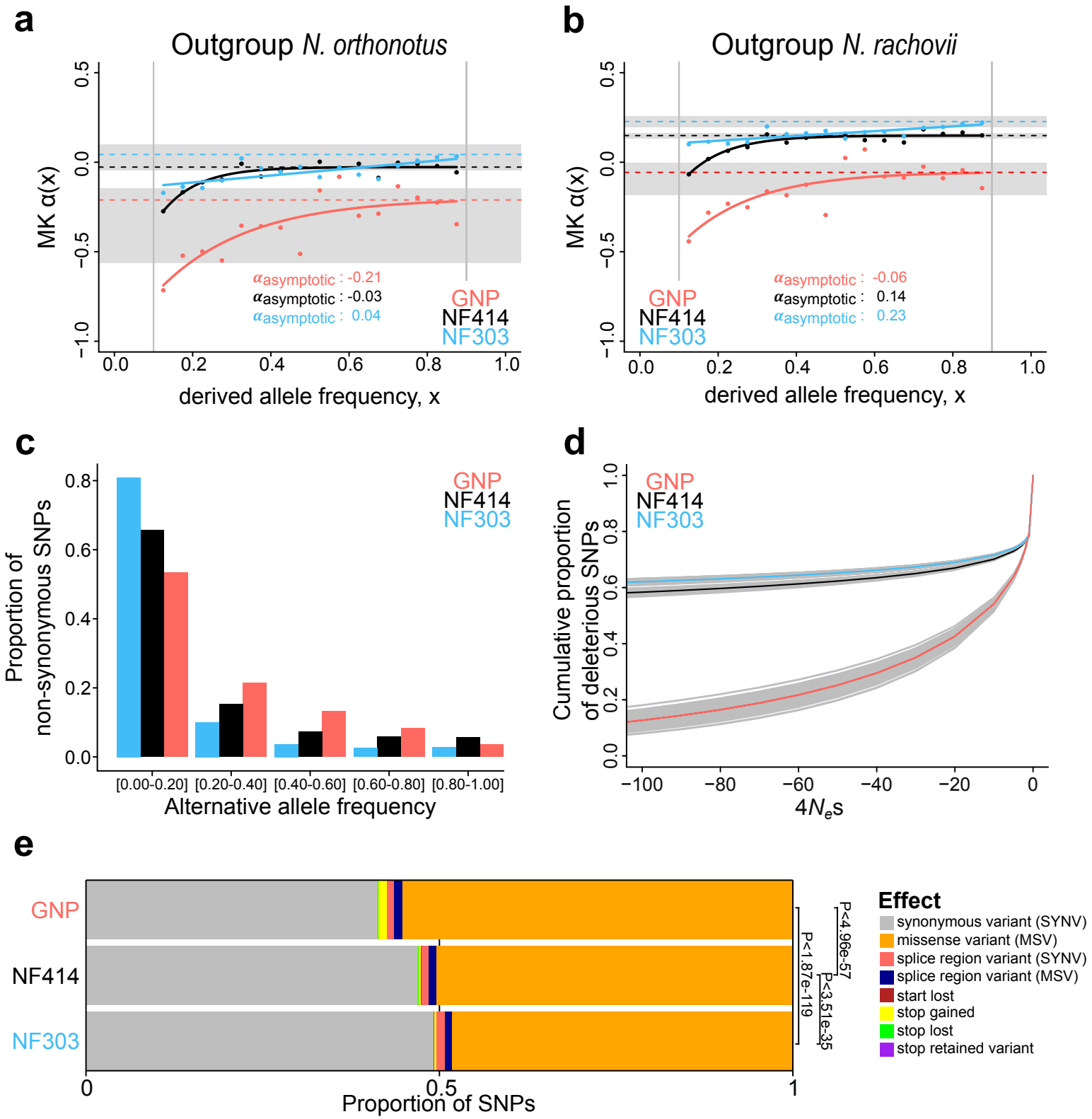


**Figure 4. Genome-wide signatures of natural and relaxed selection in turquoise killifish.** Asymptotic McDonald-Kreitman alpha (MK $\boldsymbol{\alpha}$) analysis based on derived frequency bins using as outgroups a) *Nothobranchius orthonotus* and b) *Nothonbranchius rachovii*. Population GNP is shown in red, NF414 in black, and NF303 in blue. c) Proportion of non-synonymous SNPs binned in allele frequencies of non-reference (alternative) alleles for GNP (red), NF414 (black) and NF303 (blue). d) Negative distribution of fitness effects of populations GNP (red), NF414 (black) and NF303 (blue) with cumulative proportion of deleterious SNPs on y-axis and the compound measure of $4N_es$ on x-axis. e) Proportion of different effect types of SNPs in coding sequences of all populations. The effect on amino acid sequence for each genetic variant is represented by colours (legend). Significance is based on ratio between synonymous effects to non-synonymous effects (significance based on Chi-square test).

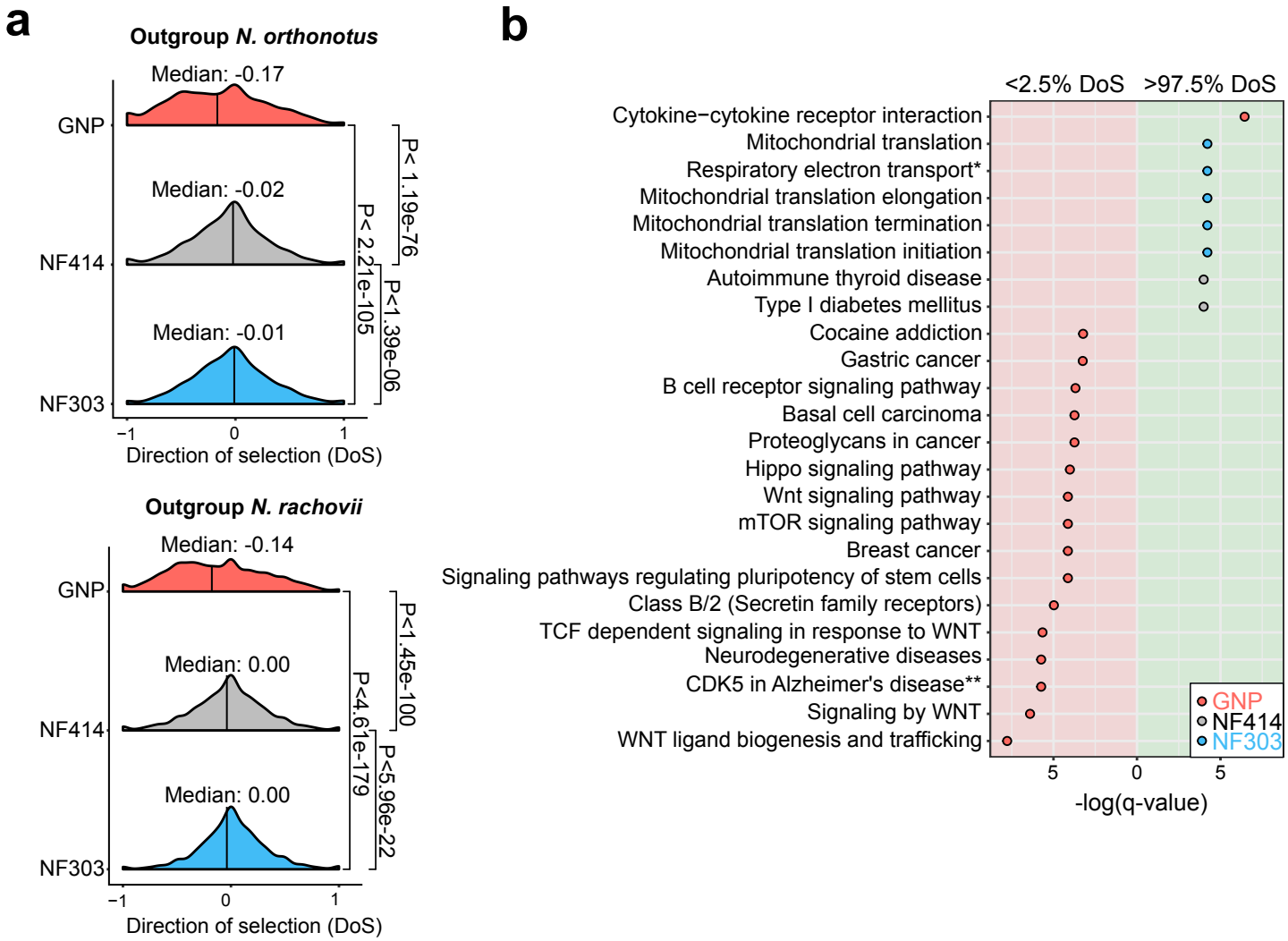


**Figure 5. Pathway enrichment in genes under adaptive and neutral evolution in turquoise killifish populations.** a) Distribution of direction of selection (DoS) represented with median of distribution for population GNP (red), NF414 (grey) and NF303 (blue). Left panel shows DoS distribution computed using *Nothobranchius orthonotus* as outgroup and right panel shows DoS distribution computed using *Nothobranchius rachovii* as outgroup. Significance based on Wilcoxon-Rank-Sum test. b) Pathway over-representation analysis of genes below the 2.5% level of gene-wise DoS values are shown with red background and above the 97.5% level of gene-wise DoS values are shown with green background. Only pathway terms with significance level of FDR corrected q-value < 0.05 are shown (in -log(q-value)). Terms enriched in population GNP have red dots, enriched in population NF414 have black dots, and enriched in population NF303 have blue dots, respectively. *ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins. ** Deregulated CDK5 triggers multiple neurodegenerative pathways in Alzheimer's disease models.

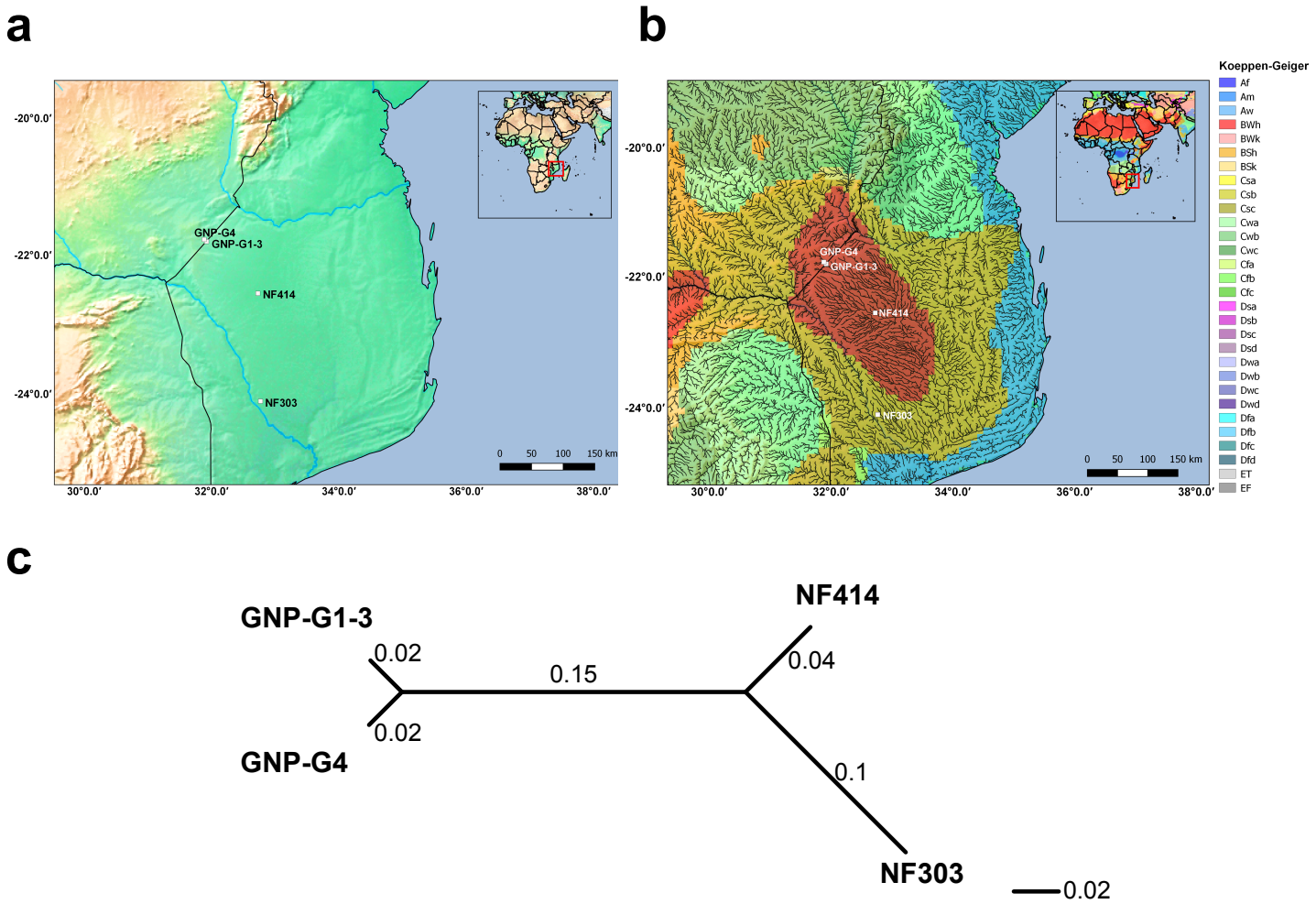


**Figure S1. Altitude, climate classification and genetic differentiation of studied samples.** a) Altitude elevation map with studied samples. b) Climate classification based on Koeppen-Geiger index combined with a high resolution river map. c) Unrooted neighbor joining tree based on pairwise genetic differentiation ($F_{ST}$) between all sample sites.

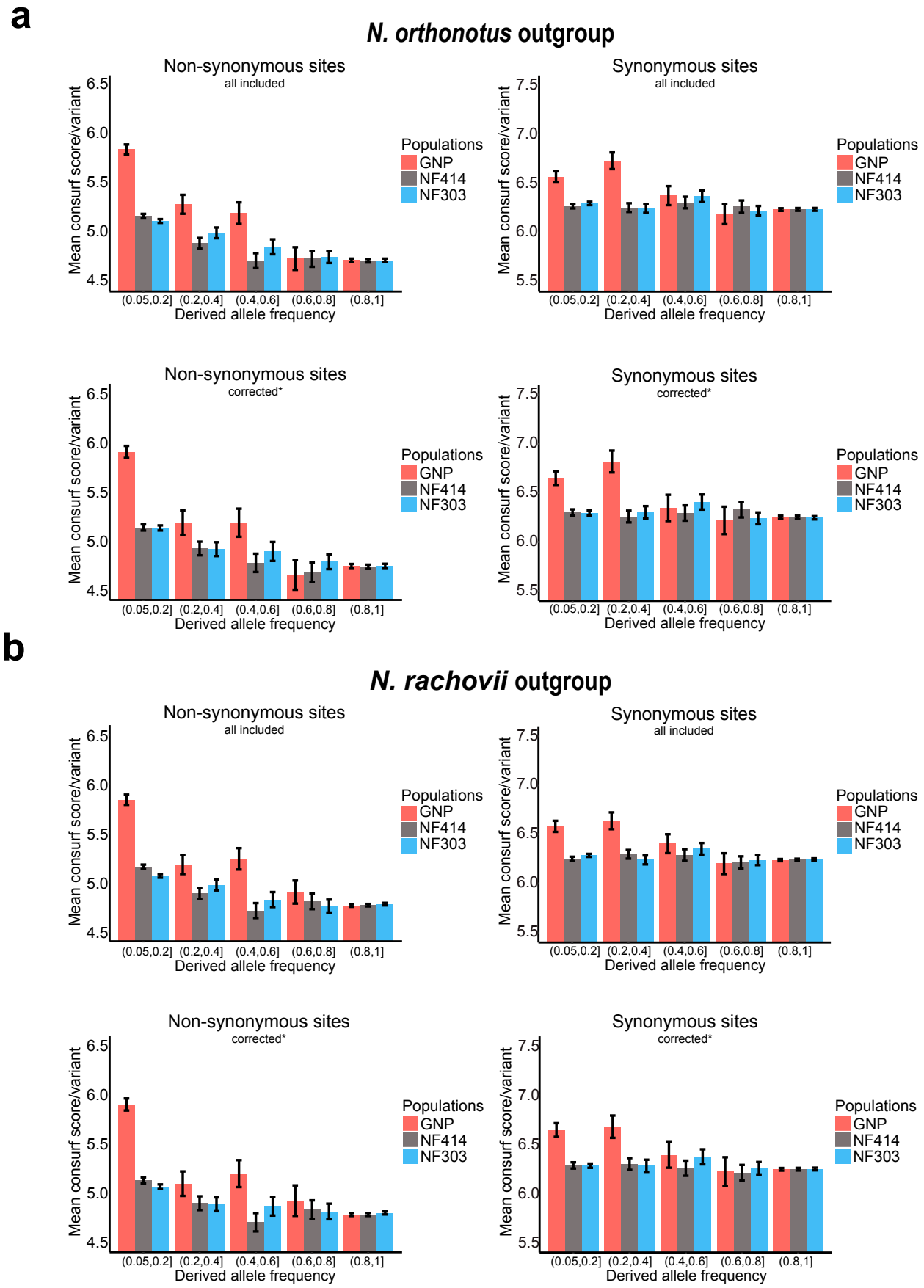


**Figure S2. Mean Consurf score per variant based on derived frequency bins. a)** Mean Consurf score based on derived frequency bins in non-synonymous (left) and synonymous (right) sites using *Nothobranchius orthonotus* as outgroup, including all available sites (upper panel) or only sites corrected for CMD, CpG hypermutation and highly detrimental effect based on SnpEFF analysis (lower panel). **b)** Mean Consurf score based on derived frequency bins in non-synonymous (left) and synonymous (right) sites using *Nothobranchius rachovii* as outgroup, including all available sites (upper panel) or only sites corrected for CMD, CpG hypermutation and highly detrimental effect based on SnpEFF analysis (lower panel). Mean consurf scores per variant are shown with SEM.