Life history evolution in turquoise killifish

Inaugural-Dissertation
zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln
vorgelegt von

David Willemsen
aus Kleve

Köln 2020

Gutachter: Dr. Dario R. Valenzano
Prof. Dr. Thomas Wiehe

Tag der mündlichen Prüfung: 06.06.2019

für Johann

Kurzzusammenfassung

Die klassischen evolutionären Theorien des Alterns sagen voraus, dass sich das Altern aufgrund der Abnahme der Stärke der natürlichen Selektion mit dem Alter als Nebenprodukt von Mutationen mit schädlichen Auswirkungen im späten Leben entwickelt hat. Die Evolution der Unterschiede im Altern und in der Lebensspanne innerhalb natürlicher Populationen ist jedoch noch nicht ausreichend verstanden. Um zu untersuchen, wie Unterschiede in der späten Lebensgeschichte in natürlichen Populationen entstehen können, nutzte ich den natürlich kurzlebigen türkisen Prachtgrundkärpfling (*Nothobranchius furzeri*), ein aufstrebender Modellorganismus in der Altersforschung. Natürliche Populationen des türkisen Prachtgrundkärpflings kommen entlang eines extremen Trockenheitsgradienten in der afrikanischen Savanne vor. Populationen entlang dieses Trockenheitsgradienten zeigen Unterschiede in der späten Lebensgeschichte, wie Lebensspanne und Alterungsrate, aber nicht in der frühen Lebensgeschichte, wie z.B. in der Entwicklungsrate. Mithilfe eines populationsgenetischen Ansatzes untersuchte ich die Evolution der Unterschiede sowohl im Altern als auch in der Lebensspanne von Populationen des türkisen Prachtgrundkärpflings und studierte Anpassungen an den rauen Lebensraum. Dafür habe ich eine gepoolte Sequenzierung an insgesamt 118 Individuen, die aus drei verschiedenen Regionen entlang des Trockenheitsgradienten stammen, durchgeführt. Zusätzlich habe ich eine individuelle Re-Sequenzierung mit hoher Abdeckung angewendet, um die demographische Geschichte der Population zu untersuchen.

Ich habe entdeckt, dass Gene, die an mitochondrialen und ribosomalen Funktionen beteiligt sind, eine außergewöhnlich geringe genetische Divergenz zwischen Populationen des türkisen Prachtgrundkärpflings zeigten. Dies weist auf eine wichtige Rolle in der Biologie des türkisen Prachtgrundkärpflings hin. Darüber hinaus habe ich herausgefunden, dass die Population aus der trockensten Region, assoziiert mit einer kürzeren Lebensdauer und einem beschleunigten Altern, eine kleinere rezente und anzestrale effektive Populationsgröße hatte, eine entspanntere natürliche Selektion und ein höheres Maß an schädlichen Mutationen, insbesondere in Genen, die mit Signalwegen des Alterns und altersbedingten Krankheiten verbunden sind. Die Ergebnisse dieser Arbeit heben die Bedeutung der Demographie für die Evolution der Unterschiede in der späten Lebensgeschichte der natürlichen Populationen hervor. Sie zeigen, dass die Evolution dieser Unterschiede mit den klassischen evolutionären Theorien des Alterns übereinstimmt und verbinden diese Theorien mit der Nahezu Neutralen Theorie der molekularen Evolution. Diese Arbeit zeigt die Bedeutung einer entspannten Selektion im Kontext der molekularen und phänotypischen Evolution und kann somit helfen, den Alterungsprozess und altersbedingte Krankheiten bei anderen Wirbeltierarten mit begrenzter effektiver Populationsgröße, wie z.B. dem Menschen, zu verstehen.

Abstract

The classical evolutionary theories of aging predict that aging has evolved as the byproduct of mutations with deleterious effects late in life, due to the decline in force of natural selection with age. However, the evolution of differences in aging and lifespan in natural populations is still not well understood. To investigate how differences in late life history can evolve in natural populations, I made use of the naturally short-lived turquoise killifish (*Nothobranchius furzeri*), an emerging model organism in aging research. Natural populations of the turquoise killifish occur along an extreme aridity gradient in the African savannah. Populations along this aridity gradient show differences in late life history, such as lifespan and the rate of aging, but not in early life history, such as developmental rate. Using a population genetics approach, I investigated the evolution of differences in aging and lifespan of turquoise killifish populations and studied adaptations to the harsh environment. I therefore performed pooled sequencing on a total of 118 individuals from three different locations along the aridity gradient. Additionally, to investigate population demographic history, I performed high-coverage individual resequencing.

I discovered that genes involved in mitochondrial and ribosomal functions have exceptionally low genetic divergence between all turquoise killifish population, indicating an important role in turquoise killifish biology. Furthermore, I found that the population from the most arid region, associated with a shorter lifespan and more accelerated aging, had a smaller recent and ancestral effective population size, relaxed natural selection and a higher level of deleterious mutations, especially in genes associated with aging and age-related disease pathways. My findings highlight the importance of demography in driving the evolution of late life history differences in natural populations, demonstrate that the evolution of these differences is consistent with the predictions of the classical evolutionary theories of aging, and connect these theories to the nearly neutral theory of molecular evolution. This work highlights the importance of relaxed selection in the context of molecular and phenotypic evolution and can help to understand aging and age-related diseases in other vertebrate species with limited effective population size, such as humans.

Acknowledgements

This doctoral thesis would not have been possible without the constant help, be it scientific or not, of the people around me. First of all, my sincere thanks to Dario Valenzano, who gave me the opportunity to work on this great project. His everlasting fascination with science and his constant support as a supervisor and mentor have helped me become a better scientist. Our scientific discussions, in the lab or in the savannah, gave me a broader view about evolution and beyond.

Further, I would like to thank every former and current member of the Valenzano group. I will start chronologically. Patrick Smith and Yumi Kim introduced me to the turquoise killifish and supported me getting started in the laboratory. Rongfeng "Ray" Cui gave me tremendous and invaluable support in population genetics and bioinformatics. He has contributed a lot to my dissertation, not just by providing the new, improved genome assembly. He supervised me throughout the years. Special thanks to my fellow students Will Bradshaw, Franzi Metge, Jens Seidel, Miriam Popkes and Daniel Davila. You helped not only as scientists with ideas and expertise in my project, but also as friends. In the same sense I would like to thank Sam Kean, Davina Patel and Michael Poeschla. Especially, Michael and Jens for the numerous discussions of the neutral theory of molecular evolution. Not to forget, I would like to thank the rest of the Valenzano group, Ola, Alix, Mattias and Aysan, who keep this lab running!

Special thanks to Dr. Martin Denzel and Prof. Dr. Thomas Wiehe for supporting my work as members of my thesis committee and in the disputation. Many thanks also to Prof. Dr. Siegfried Roth for taking over the chairmanship of my disputation.

Many thanks to Dr. Martin Reichard for the supply of samples from Mozambique. Without these samples, the work would not have been possible. Thanks also to the Graduate School for Biological Sciences, especially to Dr. Isabell Witt; the bioinformatics core facility of the MPI and the sequencing core at CCG; our collaborators from Zimbabwe.
My deepest thanks to my family and friends for their endless support and love. Mama, Papa,

Julia and Rebecca, without you that would never have been possible. Thank you and I love you. Finally, Carmen, thank you for always being by my side. You always believed in me and supported me. I love you.

Table of contents

List of figures **xi**

List of tables **xiii**

List of abbreviations **xv**

1 Introduction **1**

1.1 Introduction to aging 2

1.2 Evolutionary theories of aging 6

1.2.1 Mutation accumulation 7

1.2.2 Antagonistic pleiotropy 8

1.2.3 Testable predictions from the evolutionary theories of aging 10

1.3 The turquoise killifish as an aging model organism 16

1.4 Aims of this thesis 19

2 Methods & Materials **21**

2.1 Methods 22

2.1.1 Molecular methods 22

2.1.2 Bioinformatic methods 24

2.2 Material 49

3 Results **51**

3.1 Sequencing of turquoise killifish populations 53

3.2 Effective population size decreases with higher aridity 57

3.3 Contrasting population history in wet and dry populations 63

3.4 Identifiying genomic regions associated with lifespan 69

3.4.1 Recent evolutionary events in the region associated with lifespan . . 69

3.4.2 Genetic differentiation in QTL regions 73

3.5 Genome-wide signatures of local selection 75

3.5.1 Regions of low genetic differentiation 75

3.5.2 Regions of high genetic differentiation 78

3.5.3 F_{ST} Gene Ontology analysis 85

3.6 Molecular evolution in populations of the turquoise killifish 89

3.6.1 Test for positive selection in turquoise killifish populations 90

3.6.2 Candidate regions of recent positive selection 93

3.6.3 Purifying selection is weaker in dry population 97

3.6.4 More slightly deleterious mutations in dry population 99

3.6.5 Higher proportion of SNPs with detrimental effects in dry population 101

3.6.6 Relaxed purifying selection in age-related pathways 104

4 Discussion 108

4.1 Evolution of aging in turquoise killifish 109

4.1.1 Smaller effective population size is associated with short lifespan . 109

4.1.2 Mutation accumulation theory of aging 110

4.1.3 Antagonistic pleiotropy theory of aging 114

4.2 Adaptations to harsh environments 117

4.2.1 Low differentiation in mitochondrial and ribosomal genes 118

4.2.2 Regions of exceptionally low genetic differentiation 119

5 Conclusions and future perspectives 122

References 126

Appendix A Comparison of `MSMC2` estimates 141

Appendix B Synteny medaka and platyfish 142

Appendix C F_{ST} overrepresentation analysis 144

Appendix D Rate of molecular evolution 151

Appendix E Distribution of fitness effects 152

Appendix F DoS GO slim families 155

Appendix G Pathway overrepresentation DoS 170

List of figures

1.1 Demographic change in age classes 2
1.2 Worldwide causes of death 4
1.3 Evolutionary theories of aging 9
1.4 Life cycle of the turquoise killifish 18

3.1 Geographical distribution and climate classification 53
3.2 Genetic differentiation of turquoise killifish populations 54
3.3 Genetic distance and geographical distance of turquoise killifish populations 55
3.4 Genetic differentiation of turquoise killifish populations 56
3.5 Estimates of current effective population size *Ne* 58
3.6 Watterson's estimator of Theta θ. 60
3.7 Nucleotide diversity π 61
3.8 Tajima's D. 62
3.9 Ancestral effective population size *Ne* 65
3.10 Neutral site frequency spectrum. 67
3.11 Synonymous site frequency spectrum. 67
3.12 Non-synonymous site frequency spectrum. 68
3.13 Synteny 71
3.14 Synteny of sex chromosome 72
3.15 Genetic differentiation in QTL regions 74
3.16 `Circos` plot of genome-wide $Z_{F_{ST}}$ 76
3.17 $Z_{F_{ST}}$ on chromosome 3 77
3.18 $Z_{F_{ST}}$ on chromosome 9 77
3.19 $Z_{F_{ST}}$ on chromosome 6 79
3.20 $Z_{F_{ST}}$ on chromosome 10 80
3.21 Gene ontology low F_{ST}: GNP vs. NF414. 86
3.22 Gene ontology low F_{ST}: GNP vs. NF303. 86

3.23 Gene ontology low F_{ST}: NF414 vs. NF303. 87
3.24 Gene ontology high F_{ST}: GNP vs. NF303. 88
3.25 Asymptotic McDonald-Kreitman α: *N. orthonotus* 91
3.26 Asymptotic McDonald-Kreitman α: *N. rachovii* 92
3.27 NF303 Tajima's D on chromosome 3. 94
3.28 NF303 Tajima's D on chromosome 4. 95
3.29 NF303 Tajima's D on chromosome 10. 96
3.30 Distribution of direction of selection: *N. orthonotus* 98
3.31 Distribution of direction of selection: *N. rachovii* 98
3.32 Negative distribution of fitness effects 100
3.33 Effect on protein function of SNPs in coding sequence 102
3.34 Frequency of detrimental mutations 103
3.35 Enrichment of Gene Ontology families based on DoS 105
3.36 Pathway overrepresentation: Direction of selection 107

A.1 Comparison of `MSMC2` estimates 141

B.1 Synteny medaka and platyfish 143

List of tables

2.1 *Nothobranchius furzeri* individual resequencing samples 23
2.2 *Nothobranchius furzeri* population samples 50

3.1 Summary of pooled sequencing of turquoise killifish populations 53
3.2 Genome-wide estimates of π, θ, Ne and Tajima's D 59
3.3 Mean coverage of individual resequencing 63
3.4 Summary of genes in lowly differentiated windows (0.5^{th} percentile of F_{ST}) 81
3.5 Summary of genes in lowly differentiated windows (1^{st} percentile of F_{ST}) . 82
3.6 Summary of highly differentiated genes between GNP and NF303 83
3.7 Summary of highly differentiated genes between GNP and NF414 83
3.8 Summary of highly differentiated genes between NF414 and NF303 84
3.9 List of genes in target region on chromosome 3 94
3.10 List of genes in target region on chromosome 4 95
3.11 List of genes in target region on chromosome 10 96
3.12 Proportion of fitness effects 100

C.1 Low F_{ST} overrepresentation: GNP and NF414 (q-value < 0.05) 144
C.2 Low F_{ST} overrepresentation: GNP and NF303 (q-value < 0.05) 148
C.3 Low F_{ST} overrepresentation: NF303 and NF414 (q-value < 0.05) 149
C.4 High F_{ST} overrepresentation: NF303 and GNP (q-value < 0.05) 150

D.1 Substitutions in synonymous and non-synonymous sites 151
D.2 Efficiency of natural selection in turquoise killifish populations 151

E.1 DFE NF303 model choice 152
E.2 DFE NF414 model choice 153
E.3 DFE GNP model choice 154

F.1 DoS Gene Ontology slim families: GNP *N. rachovii* 155

F.2 DoS Gene Ontology slim families: NF414 *N. rachovii* 160

F.3 DoS Gene Ontology slim families: NF303 *N. rachovii* 165

G.1 DoS <2.5^{th} percentile pathway overrepresentation with outgroup *N. orthonotus* 170

G.2 DoS <2.5^{th} percentile pathway overrepresentation with outgroup *N. rachovii* 173

G.3 DoS >97.5^{th} percentile pathway overrepresentation with outgroup *N. orthonotus* 174

G.4 DoS >97.5^{th} percentile pathway overrepresentation with outgroup *N. rachovii* 175

List of abbreviations

A	adenine
AIC	Akaike information criterion
AP	antagonistic pleiotropy
A. limnaeus	*Austrofundulus limnaeus*
BAC	bacterial artificial chromosome
bp	base pair(s)
BR	broad range
C	cytosine
°C	degrees Celsius
CCG	Cologne Center for Genomics
CDS	coding sequence
chr	chromosome
C. elegans	*Caenorhabditis elegans*
DAF	derived allele frequency
del	deleterious
df	degrees of freedom
DFE	distribution of fitness effects
D_n	number of non-synonymous substitutions per non-synonymous sites
DNA	deoxyribonucleic acid
D. melanogaster	*Drosophila melanogaster*
DoS	direction of selection
D_s	number of synonymous substitutions per synonymous sites
dsDNA	double-stranded deoxyribonucleic acid
EDTA	ethylenediaminetetraacetic acid
e.g.	*exempli gratia* (for example)

et al.	*et alia* (and others)
FDR	false discovery rate
F_{ST}	population differentiation index
G	guanine
Gb	gigabase(s)
GFF	gene file format
GNP	Gonarezhou National Park, Zimbabwe
GO	Gene Ontology
GRZ	*N. furzeri* strain derived from GNP
GRZ-AD	substrain of GRZ
GTF	gene transfer format
HCL	hydrochloric acid
i.e.	*id est* (that is)
INDEL	insertion or deletion
k	thousand
kb	kilobase(s)
km	kilometer(s)
LG	linkage group
lk	likelihood
lss	lanosterol synthase
M	molar
MA	mutation accumulation
MAF	minor allele frequency
Mb	megabase(s)
min	minute(s)
MK-α	McDonald-Kreitman's α
ml	milliliter(s)
mM	millimolar
MRCA	most recent common ancestor
mRNA	messenger ribonucleic acid
MSMC	multiple sequential markovian coalescence
N. furzeri	*Nothobranchius furzeri*

N. orthonotus	*Nothobranchius orthonotus*
N. rachovii	*Nothobranchius rachovii*
NaCl	sodium chloride
N	number of individuals
Ne	effective population size
ng	nanogram(s)
NGS	next generation sequencing
PSMC	pairwise sequential markovian coalescence
PBS	phosphate buffered saline
PCI	phenol-chloroform-isoamylalcohol
PCR	polymerase chain reaction
PE	paired-end
PEG	polyethylene glycol
P_n	number of non-synonymous polymorphisms per non-synonymous sites
P_s	number of synonymous polymorphisms per synonymous sites
Pool-Seq	sequencing pools of individuals
QTL	quantitative trait loci
RAD	restriction site associated DNA
RNA	ribonucleic acid
RNA-Seq	RNA-sequencing
rpm	rounds per minute
RT	room temperature
s	selection coefficient
sd	standard deviation
SDR	sex-determining region
SDS	sodiumdodecylsulfate
SFS	site frequency spectrum
SGA	string graph assembler
SNP	single-nucleotide polymorphism
SRPI	solid phase reversible immobilization
T	thymine
TE	Tris-EDTA buffer

TRIS	Tris-(hydroxymethyl)-methylamin
vs.	*versus* (against, compared to)
WGS	whole genome sequencing
π	nucleotide diversity
θ	Watterson's estimator of theta
0-fold	zero-fold degenerated/non-synonymous
4-fold	four-fold degenerated/synonymous
µ	mutation rate
µg	microgram(s)
µl	microliter(s)
µM	micromolar
%	percent

Chapter 1

Introduction

1.1 Introduction to aging

Human life expectancy in all countries in the world has increased substantially over the last 200 years [50]. In Europe, life expectancy has doubled since the 19th century and is still increasing with no limit in sight [128, 204]. This increase in duration of life is a result of improved health care and hygiene [128]. However, this is not the longevity scientists have been seeking for centuries [60]. The improvements of modern medicine are able to lower the mortality rate at all ages [204], but are not a solution to the functional decline of the human organism, a process that is called aging [109]. The percentage of people reaching higher ages in developed countries is rising (see Figure 1.1), thus we can find more people suffering from a decline of physiological and cognitive function with age.

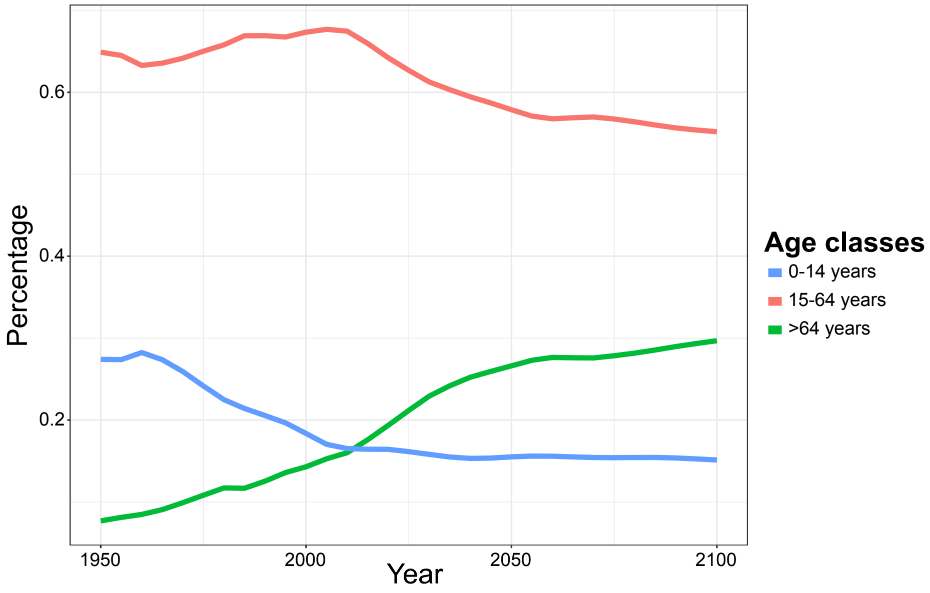

Figure 1.1: Demographic change. The figure shows the recent demographic development and the future demographic development for different age classes (Blue line shows the age class from 0-14 years; Red line shows the age class from 15-64 years; Green line shows the age class older than 64 years). Data from United Nations, Department of Economic and Social Affairs, Population Division (2017). World Population Prospects: The 2017 Revision. Custom data was acquired via website.

The recent demographic development brings new pressure to explore the aging process as it is one of the main risk factors for all prevalent lethal diseases in industrialized countries ([139, 109]; see Figure 1.2). The common definition of aging is a decline in function and vitality with age. In biological terms, aging, or senescence, as it is called, is defined in two different ways. A distinction is made between functional aging and demographic aging. Functional aging means the decline in performance and fitness with age [77, 47, 153, 157] and is defined at the level of a single person. Demographic aging, on the other hand, is defined at the population level and means an increase in mortality rate, including a reduction in fertility [23, 24, 136]. It has long been believed that aging is inevitable. Aging was thought to be a product of modern society, reflecting the favorable conditions in which societies live today. This would also reflect the situation of laboratory animals living under controlled and favorable conditions. Therefore, aging in nature was considered extremely rare due to the extreme extrinsic hazard [35].

But the evidence for the effects of aging, a decline in survival and reproduction with age, has been increasing for natural populations in recent years [127]. However, at the beginning of the 20th century, aging research was still in its infancy and it was doubted that there are genes that influence the aging process [85]. The breakthrough for aging research was the identification of single gene variants that could extend lifespan in a small roundworm, the nematode *Caenorhabditis elegans* [86, 94, 53]. In addition to the lifespan prolongation, these worms stayed healthy and youthful for longer [86]. The results of these studies showed that the ultimate goal of aging research is achievable. Namely, the ultimate goal of aging research is not just to extend lifespan, but to age healthily [109].

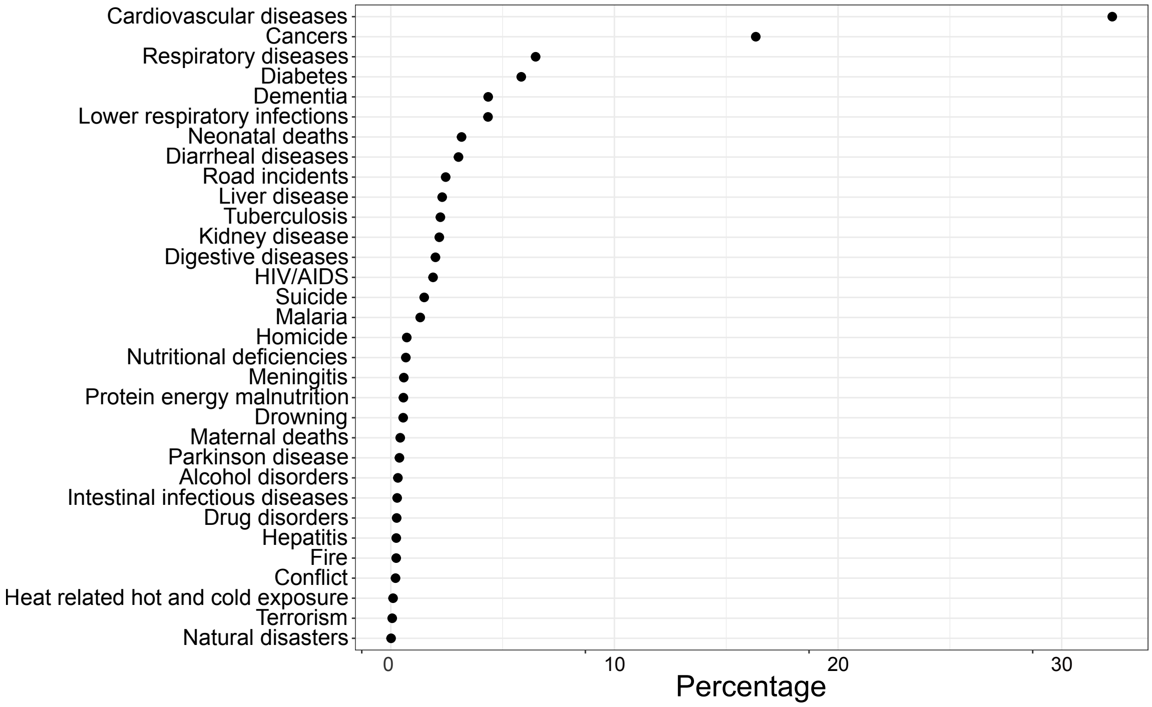

Figure 1.2: Worldwide causes of death. The figure shows the global percentage of most common causes of death in 2016. Data from www.ourworldindata.org (IHME, Global Burden of Disease).

The effects of single gene mutations on longevity and aging have been confirmed for several model organisms, from invertebrates such as *C. elegans* and the fruit fly *Drosophila melanogaster* to small vertebrate rodents such as the house mouse *Mus musculus* [163, 140, 176]. Interestingly, population genetic studies provide initial evidence that the identified modifiers of aging in model organisms also modulate the aging process in humans [123, 102, 202]. The abundance of evidence that mutations in the same pathway or even in the same gene are life-prolonging to various organisms, from yeast to humans, means that the mechanism is evolutionary conserved [51, 139, 50].

In the Section 1.2.2, I will discuss why such gene variants are not selected by natural selection and do not replace the wild-type alleles in natural population.

However, aging is not inevitable for every living organism. The small freshwater polyp *Hydra* shows no signs of decline in survival and reproduction with age in a period of four years under laboratory conditions [115]. Likewise, germline cells show no discernible signs of senescence [47, 92]. Therefore, aging seems neither necessary nor mandatory, but aging can be genetically altered. Thus, if it is partially genetically controlled, it must be explainable in terms of evolution. Theodosius Dobzhansky postulated in 1973 that nothing makes sense in biology except in the light of evolution [40]. Based on the quintessence of this argument, we must ask ourselves why aging has arisen. From the perspective of natural selection, only individuals who survive and reproduce in their specific environment can contribute to the next generation. Therefore, any optimization of survival or reproduction would benefit the individual harboring the causal genetic variant. However, aging is defined as an age-related decline in survival and reproduction [23, 24, 136]. This paradox led to the question, why has such an adverse process as aging evolved [5, 157]?

Although aging in the twentieth century was considered a "black box" for most molecular biologists, the theory of the evolution of aging addressing how such a harmful process could develop was formulated before the first long-lived mutant was identified. In the next section, I will discuss the history and the two main theories of the evolution of aging.

1.2 Evolutionary theories of aging

The first to explore the evolution of aging were Alfred Russel Wallace and August Weismann. Weismann, who was inspired by Wallace's ideas, stated that old individuals provide no value to the population, but rather impose costs [196, 197]. Weismann saw aging as an adaptive process that subordinated the good of the individual to the good of the population. However, he also pointed out that aging may be due to a limitation in the number of somatic cell divisions [196, 197, 203]. Nowadays, it is widely accepted that natural selection acts on the level of a single individual and not on a population level [117]. In addition, Weismann's argument that old people are a cost to the population because of *e.g.* less reproduction is a circular argument [92]. Aging can not be explained by its own definition, an age-related decline in reproduction and survival. But Weismann's distinction between somatic and germline cells is a crucial requirement for aging. There should be distinction between parents and offspring for aging to occur [50, 135, 203, 23, 24]. Weismann's approach was the starting point for the evolutionary theories of aging. In his seminal book, "The Genetical Theory of Natural Selection", R.A. Fisher noted that the force of selection should decrease with age [49]. This is consistent with the observation of J. B. S. Haldane in 1941 that a harmful disease such as Huntington's disease persists in the human population. Huntington's disease is a lethal dominantly inherited disease. If the strength of natural selection should remain constant at any age, no mutation with fatal consequences should occur in the population regardless of the age of the effect of the mutations. However, the example of Huntington's disease supports the assumption that the force of natural selection decreases with age. The onset of the phenotype of Huntington's disease is after reproduction and therefore not selected against by natural selection [65]. Thus, Haldane found an example for Fisher's observation that the force of natural selection declines in an age-dependent manner. Based on this insight, two separate evolutionary theories of aging were developed. These theories are the theory of "mutation accumulation" (Section 1.2.1) and the theory of "antagonistic pleiotropy" (Section 1.2.2).

The formal mathematical description of these theories were later developed by William D. Hamilton [69, 159] and Brian Charlesworth [23, 24, 28, 26].

1.2.1 Mutation accumulation

The mutation accumulation theory states that due to the decreasing force of natural selection with age, deleterious mutations with effects late in life are more prevalent than deleterious mutations with effects early in life (see Figure 1.3) [120, 121]. Any deleterious mutation that impacts early life and reduces the fitness of the carrier is filtered out by natural selection. But if the effect of the deleterious mutation acts after reproduction or at the peak of reproduction, thus after being transmitted to the next generation, this deleterious mutation is considered to be relatively neutral to natural selection. The later the effect of the mutation occurs, the lower the probability of the mutation being eliminated by natural selection. Thus, aging is the sum of many harmful mutations that act late in life. To illustrate this model, Gavrilov and Gavrilova [55] used two different diseases and their genetic consequences. Patients with progeria suffer from premature aging and usually live to only 12 years of age. Progeria is a genetic disease that has its effects early in life and therefore can not be inherited into the next generation. The premature death means that natural selection prevents accumulation of the causal genetic mutation in the population. On the other hand, Gavrilov and Gavrilova have used the case of Alzheimer's disease as an example of a late-acting deleterious mutation [55]. The phenotype of Alzheimer's disease occurs only after the patient has the chance to pass on their genetic material to the next generation. As a result, the frequency of mutations leading to Alzheimer's disease is higher than that of mutations leading to the progeria phenotype.

As a direct consequence of the mutational accumulation theory of aging, the additive genetic variance for age-related phenotypes should increase with age [23, 24]. The additive genetic variance is the genetic variability between individuals from different parents. For age-related traits, the force of natural selection is weak, leading to stochastic occurrence of late-acting mutations. The stochasticity leads to high genetic variation between individuals from different parents. In contrast, traits under strong selection, *e.g.* genes that regulate development, are shared by all individuals of the population. Hence, the additive genetic variance between individuals of different parents is low. This prediction of the mutation accumulation theory was confirmed by studying the genealogy of European families [55]. However, early empirical evidence for the validation of the mutational accumulation theory of aging was mixed, but the number of studies supporting the mutation accumulation theory is increasing [55, 138, 50, 77, 29].

1.2.2 Antagonistic pleiotropy

The antagonistic pleiotropy theory of aging was initially formulated by George C. Williams in 1957 [203]. The idea is based on the work of Sir Peter Brian Medawar, who linked aging (senescence) to the concept of pleiotropy [121]. Pleiotropy itself means that a gene can have different effects on unrelated phenotypes. Williams states that a gene that has beneficial effects early in life and detrimental effects later in life will be favored by natural selection if the benefit exceeds the detriment ([203], see Figure 1.3). Due to the fact that natural selection decreases with age, not only the magnitude but also the timing of the detrimental effect is crucial. He noted "that natural selection will maximize vigor in youth at the expense of vigor later on and thereby produce a declining vigor (senescence) during adult life" [203]. The advantage of any mutation thereby depends on the effect on reproduction. If the mutation confers an advantage when reproduction is the highest, natural selection will favor this mutation, if the detrimental effect has little impact on the total reproductive value. Under this concept, aging emerged as a byproduct of natural selection, but it is not actively selected [50]. A key assumption of this theory is the trade-off mechanism between early life traits and late life traits. One consequence is that selection for longevity should have reducing effects on vitality early in life. In fact, this assumption could be confirmed experimentally. Experiments with the fruit fly *D. melanogaster* were able to show that the selection for late reproduction prolonged life expectancy, but reduced early life fecundity [155, 158, 156, 110, 137]. The theory of antagonistic pleiotropy provides an explanation for why single gene mutations that prolong life are not selected by natural selection (see Section 1.1). Indeed, most of the single gene variants found to promote longevity had the effect of reduction in fitness early in life. Either the mutant individuals showed reduced body size or fertility, reduced responsiveness to stressful situations or could not compete with the wild-type individuals in mixed culture [33, 179, 86, 56, 79, 191, 187]. These clearly negative effects early in life explain why the genetic variants are not replacing the wild-type allele under natural conditions.

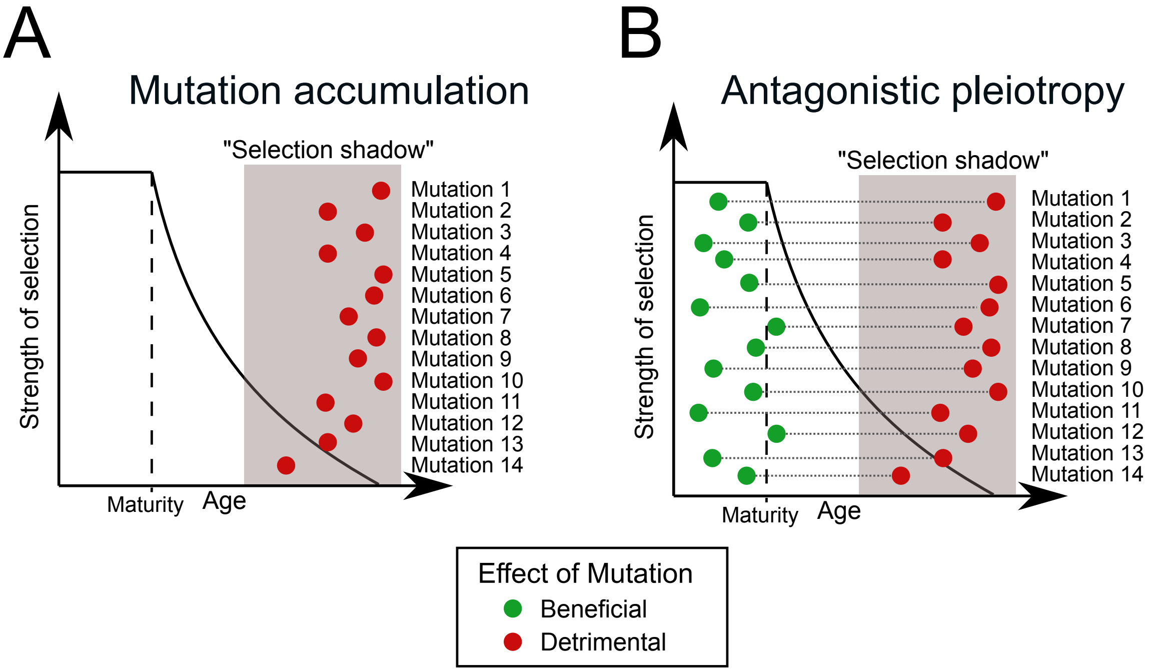

Figure 1.3: Evolutionary theories of aging. A) The mutation accumulation theory of aging. Mutations that act late in life and have deleterious effects can not be removed from the population. Therefore, the deleterious mutations are in the "selection shadow". The build-up of deleterious mutation is called mutation accumulation and is causal for the onset of aging. B) The antagonistic pleiotropy theory of aging. Mutations whose beneficial effect in early life exceeds the detrimental effect later in life are favored by natural selection and are responsible for the onset of aging. The pleiotropic effect is illustrated with dotted lines (Figure according to Fabian and Flatt [45]).

1.2.3 Testable predictions from the evolutionary theories of aging

The two major theories of aging, mutation accumulation and antagonistic pleiotropy, suggest that aging is not selected by natural selection, but rather is a byproduct of the decline in force of selection with age [50]. In the case of the antagonistic pleiotropy theory of aging, this is a result of the selection for "vigor in youth" [203], despite the negative effects later in life. In both theories, aging has evolved due to the accumulation of mutations in the genome with detrimental effects late in life. In this section I will discuss predictions of the two evolutionary theories of aging.

Extrinsic hazard and aging rate

A direct implication of the evolutionary theories of aging is that the level of extrinsic hazard influences the rate of aging [4, 23, 24, 42, 203]. Therefore, when starvation, predation or climatic stress are high, the high extrinsic risk leads to a stronger decline of the force of natural selection with age, because fewer individuals will survive until old age. The stronger decline in the force of natural selection under high extrinsic risk therefore leads to accelerated aging [4]. Conversely, if external risk factors such as starvation, predation or climatic stress are less pronounced, the decline in force of natural selection and thus aging slows down [4]. Stearns and colleagues showed that this prediction holds true for populations of *D. melanogaster* [174]. The genetically identical populations were exposed to different extrinsic mortality rates. The population with a higher extrinsic mortality rate developed a higher intrinsic mortality rate compared to the population with a lower extrinsic mortality rate [174]. Thus, the increase in intrinsic mortality with age is defined as aging, and as the evolutionary theories of aging predicted, the rate was higher in the population with higher extrinsic mortality. Additionally, the population subjected to high extrinsic mortality rate developed faster and had higher fecundity early in life, confirming life history theory [174, 173]. A natural experiment is provided by an island opossum species *Didelphis* *virginiana* [3]. This species lives longer than a comparable mainland species, as the predator risk - the extrinsic mortality risk - is lower [3]. Although the correlation between rate of aging and extrinsic mortality is supported by many studies [3, 4, 113, 167, 174], studies in guppy *Poecilia reticulata* showed mixed results [152, 151]. In natural populations of guppies, it was shown that populations with different risk for adult or juvenile predatory differ in life history traits. As predicted by the evolutionary theories, the population with a higher risk for adult predatory mature earlier and invest more in early reproduction [151]. However, in another experiment with natural populations of guppies that differ in predatory risk, the population with higher predation did not show an earlier onset of senescence [152]. This finding is in contrast with predictions from the evolutionary theories of aging. As one explanation for this result, Reznick and colleagues argued that the prediction of earlier onset of senescence with higher mortality risk is dependent on whether or not the population is subjected to density regulation, as shown by Charlesworth [23, 24] and Abrams [1]. Therefore, a higher mortality rate, which should lead to an earlier onset of senescence, had an indirect effect on mortality rates for old age due to density dependence. The lower density at old ages increased the resource availability which led to an decrease in mortality rate for old ages. This benefit could have resulted in delayed senescence, which counteracts the predicted earlier onset of senescence due to higher mortality at younger ages.

Effective population size and aging

The evolutionary theories of aging indicate that aging is a byproduct of the accumulation of detrimental mutations that act late in life. For the mutation accumulation theory to be correct, two conditions are required. First, the mutation needs to be neutral early in life, or it would be removed by natural selection. Second, the same mutation must be relatively neutral late in life in order to accumulate in the population and not be removed by natural selection. This means that the mutation is in the "selection shadow", therefore, not seen by natural selection, as shown in Figure 1.3. For the antagonistic pleiotropy hypothesis to be true, the demands on the effects early and late in life are linked. The beneficial effect of the mutation in early life must exceed the detrimental effect later in life. Thus, the mutation is still in the "selection shadow" (see Figure 1.3) at later life stages, due to the comparisons of effects early and late in life.

To understand when a mutation is considered neutral, we need to look at the neutral theory of molecular evolution [88]. The neutral theory of molecular evolution from Motoo Kimura is based on the modern evolutionary synthesis developed by Haldane, Fisher, and Wright [64, 49, 207]. The neutral theory of molecular evolution states that most genetic variance is due to mutations, genetic drift, and purifying selection [88]. A new mutation at a given locus, without any selective advantage or disadvantage, entering a population has a probability of fixation of one divided by the total number of copies of that genetic locus (referred to as genetic drift). The total number of copies of a genetic locus is defined as the total number of chromosomes that harbor this locus. In a diploid system the number of chromosomes is two times the number of individuals in the population (2N).
Therefore, the fixation probability of a new mutation equals its frequency:

$$\text{Fixation probability: } \frac{1}{2N} \tag{1.1}$$

The occurence of new mutations is determined by the mutation rate of the species, denoted as μ. In a population with 2N chromosomes the number of mutations arising every new generation is:

$$\text{New mutations per generation: } \mu * 2N \tag{1.2}$$

Combining Equations 1.1 and 1.2, we get the rate of neutral mutations that reach fixation (Equation 1.3). This rate is called the rate of neutral substitutions. Therefore, under neutrality the rate of neutral substitutions equals the mutation rate.

$$\text{Rate of neutral substitutions: } \mu * 2N * \frac{1}{2N} = \mu \tag{1.3}$$

Under neutral theory, most mutations are selectively neutral, and therefore - as stated above - the rate of substitutions for most mutations is the rate of neutral substitutions μ, independent of population size [88, 89, 91]. But does population size play a role in molecular evolution?

Mutations can not only be neutral, the can also be deleterious or beneficial. For example, if a mutation results in a change in the amino acid sequence of a protein, this change is termed non-synonymous substitution. These changes are usually detrimental to protein function and are therefore deleterious. Conversely, if a mutation results in a change in the genetic code, but not at the amino acid level, this mutation is called a synonymous substitution. This is an essential feature of the genetic code of life [37]. How deleterious a mutation is depends on the selection coefficient s. The selection coefficient s is a measure of relative fitness of the new variant compared to the existing variant. The selection coefficient can be positive if the variant provides an advantage or negative if the variant provides a disadvantage. Haldane showed that the fixation probability of a single copy of an allele with a constant selection coefficient s in a diploid system is approximately $2s$ [63]. This represents the chance of fixation of a new mutation under the force of natural selection [68]. If natural selection and genetic drift are in equilibrium, then the fate of an allele to reach fixation is equally influenced by both forces ([68], Equation 1.4).

$$\text{Equilibrium: } 2s = \frac{1}{2N} \tag{1.4}$$

As a result of the Equation 1.4, it is now possible to define when a mutation is considered neutral or non-neutral with respect to the selection coefficient. A mutation is considered neutral if the absolute value of its selection coefficient multiplied by four times the effective population size (similar to population size, explained below) is much less than one (see Equation 1.5 , [68]).

$$\text{Neutral mutation: } |s * 4Ne| << 1 \tag{1.5}$$

A mutation is considered non-neutral, either advantageous or deleterious, when the product of Equation 1.5 is much greater than one (see Equation 1.6 , [68]):

$$\text{Non-neutral mutation: } |s * 4Ne| >> 1 \tag{1.6}$$

Hence, the effective population size (*Ne*) is equal to the number of individuals in an ideal population that produces the same genetic variance seen in the real population [54]. *Ne* is used as an improved estimate of the population size seen in previous equations (Equations 1.1, 1.2, 1.3). The usage of *Ne* is based on the idea that in ideal populations - those with random mating, equal sex ratio and non-overlapping generations - genetic drift is defined as the reciprocal of population size 2N (diploid system). In nature, most populations violate the assumptions of an ideal population and therefore, *Ne* was introduced by Wright [207].

An extension to the neutral theory is that slightly deleterious mutations can be considered "nearly neutral" if they fulfill the following condition:

$$\text{Nearly neutral mutations: } |s * 4Ne| \approx 1 \tag{1.7}$$

The nearly neutral theory of molecular evolution states that a substantial proportion of substitutions at the molecular level are caused by random fixation of very slightly deleterious mutations [129–133]. The fixation of very slightly deleterious mutations implies that the rate of molecular evolution is not independent of *Ne*. Ohta proposed that molecular evolution in small populations is faster than in large populations [129].

In conclusion, whether a mutation is considered neutral, nearly neutral or non-neutral depends on *Ne* and s. This has a clear application for the requirements of the evolutionary theories of aging. For the mutation accumulation theory of aging, a mutation must be effectively neutral late in life and therefore fulfill Equation 1.7. To achieve this, $|s|$ must be close to the reciprocal value of four times *Ne*. Therefore, $|s|$ can be large yet effectively neutral if *Ne* is correspondingly small. The antagonistic pleiotropy theory of aging requires that the advantage of a mutation early in life exceeds the disadvantage late in life. Mathematically, this would lead to following equation with s_{young}, s_{old} being the age-dependent selection coefficients and Ne_{young}, Ne_{old} being the age-dependent effective population sizes.

$$|s_{young} * 4Ne_{young}| > |s_{old} * 4Ne_{old}| \tag{1.8}$$

Equation 1.8 holds true either when $|s_{young|}$ is sufficiently larger than $|s_{old}|$, with Ne_{young} equals Ne_{old} or when Ne_{young} is sufficiently larger than Ne_{old}, with $|s_{young}|$ being equal to $|s_{old}|$. In line with that, Hughes recently hypothesized that mutation accumulation should be a stronger factor in the evolution of aging in a species with small *Ne* [76].

In this thesis, I will explore the evolution of aging in light of the nearly neutral theory of molecular evolution. I will investigate the connection between extrinsic hazard, effective population size (*Ne*) and the rate of aging. For this, I will use natural populations of the turquoise killifish, *Nothobranchius furzeri*. In the next section, I will explain why the turquoise killifish is a good model to study aging in terms of population genetics.

1.3 The turquoise killifish as an aging model organism

The turquoise killifish (*Nothobranchius furzeri*) inhabits temporal ponds in Zimbabwe and Mozambique, in southeastern Africa ([193]; see Figure 1.4). It is a teleost fish that belongs to the order of Cyprinodontiformes [201]. Yearly desiccation of its habitat during dry season in the savannah led to the evolution of the unique life cycle of the turquoise killifish (see Figure 1.4). During the short rainy season, turquoise killifish hatch, develop, mature and reproduce [15, 189]. In nature, populations of turquoise killifish occur along an aridity gradient [185, 181]. To successfully reproduce in more arid regions, the life cycle of turquoise killifish can be completed in as short as two weeks [189]. Although the lifespan of the adult turquoise killifish is bound to the seasonality of the environment, the eggs of turquoise killifish can sustain the dry season in a developmental arrested state, called diapause [205].

The median lifespan of turquoise killifish ranges between 9 and 30 weeks [183, 57, 141, 150, 185, 93]. Thus far, the highly inbred strain GRZ (derived from the Gonarezhou National Park, Zimbabwe) is the shortest-lived vertebrate that can be kept and raised in captivity [185, 183]. Despite its short lifespan, turquoise killifish display molecular, cellular and physiological aging phenotypes [87, 74, 20, 184]. Old turquoise killifish have an increased risk of cancer [10], show a decline in motor and cognitive abilities [57, 186] and a decline in fertility [39]. Overall, the aging phenotypes of turquoise killifish are comparable to the aging phenotypes of mice and humans [87]. Thus, the short lifespan and normal aging phenotypes make turquoise killifish a suitable model for aging research [87, 74, 20, 184].

In addition, captive strains of turquoise killifish substantially vary in lifespan [185, 181]. Captive strains derived from more arid areas in nature live shorter than strains derived from more wet areas [185, 181]. This difference in lifespan is not an artifact of captivity, as the difference in lifespan has recently been observed for natural populations of turquoise killifish

[14]. Natural populations from drier regions have a shorter lifespan and a steeper increase in aging rate [14]. Therefore, populations living in harsher environments - with a higher extrinsic mortality risk (see Section 1.2.3) - show faster aging. In addition, quantitative trait loci (QTL) mapping identified genomic regions associated to the difference in lifespan in captive strains [185, 93]. The identified regions provide targets for studies in natural populations that vary in lifespan.

In summary, the unique combination of a short-lived vertebrate model for aging research and the natural difference in lifespan and aging make the turquoise killifish an ideal model to study the evolution of aging.

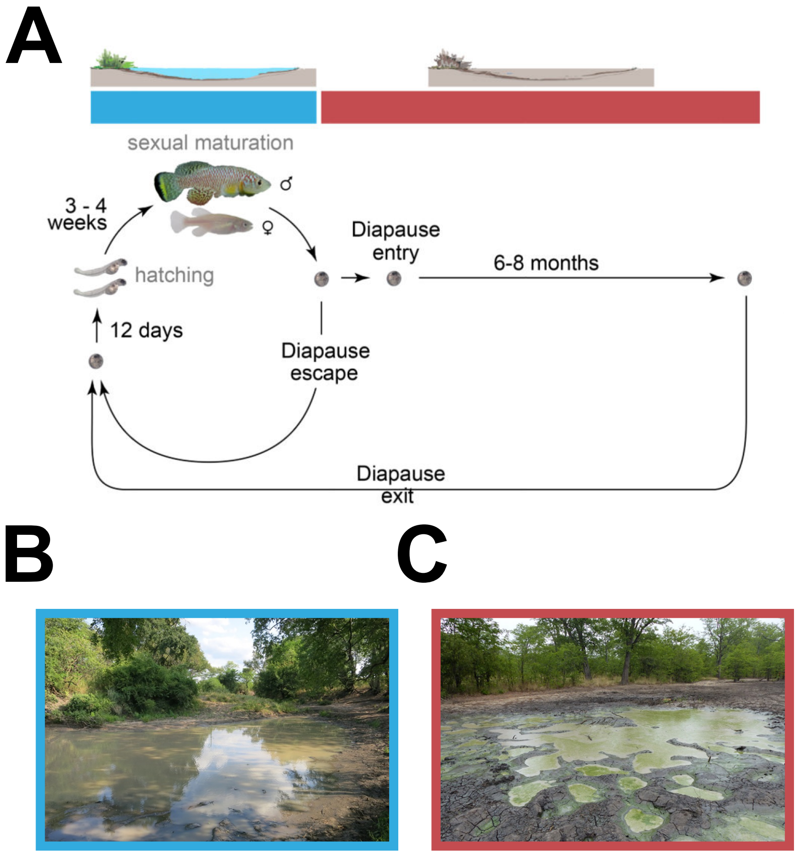

Figure 1.4: Life cycle of the turquoise killifish. A) At the beginning of the rainy season, the turquoise killifish hatches, develops and becomes sexually mature within 5-6 weeks. After successful reproduction, the eggs of the turquoise killifish can survive the dry season in a developmental break state (diapause) until the next rainy season (adapted from [185]). B) An example of a pond during rainy season. C) An example of a pond at the start of the dry season (Pictures from D.Willemsen).

1.4 Aims of this thesis

Exploring the evolution of life history

The theoretical background for the evolution of aging and lifespan has been extensively studied and developed in the last decades [23, 24, 28, 26]. The classical theories of the evolution of aging, such as the mutation accumulation (MA) theory of aging [120, 121] and the antagonistic pleiotropy (AP) theory of aging [203], represent the theoretical background of this thesis. Further, the advent of next generation sequencing (NGS) makes it possible to investigate the theoretical framework of the evolution of aging with whole genome sequencing (WGS) data. The high resolution of WGS makes it possible to study the classical theories of the evolution of aging in light of modern evolutionary genomics using population genetics. The neutral theory of molecular evolution provides a strong foundation for modern evolutionary genomics [80]. The general aim for this thesis is to explore the evolution of life history traits using a population genetics approach in a short-lived vertebrate, the turquoise killifish *Nothobranchius furzeri*.

Exploring differences in aging and lifespan

The turquoise killifish is an emerging model for aging research [87, 74, 20, 184]. It has recently been shown that natural populations of the turquoise killifish occurring along an aridity gradient display differences in late life history traits, such as aging and lifespan [14, 181]. The first aim of my thesis is therefore to use population genetic methods to identify the evolutionary forces that explain the differences in late life history.

Exploring selection pressures in harsh environments

Natural populations of the turquoise killifish live in harsh environments that exert a high selection pressure on early life history traits such as rapid development, maturation and sexual reproduction. Despite high genetic structuring between populations in close proximity [8, 9], genetic regions that regulate these core processes should show comparatively low genetic differentiation between populations of the turquoise killifish. The expectation of identifying these core processes based on low genetic differentiation is in line with the observation that turquoise killifish populations along an aridity gradient differ in late life history traits, but not in early life history traits [14]. Therefore, the second aim of this thesis is to identify genes potentially involved in adaptations to the harsh environment.

Chapter 2

Methods & Materials

2.1 Methods

2.1.1 Molecular methods

DNA isolation

To extract high-molecular-weight genomic DNA, I modified a published protocol [199]. The ethanol preserved fin tissue was washed with 1X PBS and placed into a new 1.5 ml tube. Fin tissue was digested with 10 µg/ml proteinase K (Thermo Fisher) in 10 mM TRIS pH 8; 10 mM EDTA; 0.5% SDS at 50°C overnight. The DNA was extracted with phenol-chloroform-isoamylalcohol (25:24:1, PCI, Sigma) followed by a washing step with chloroform (Sigma). Next, DNA was precipitated by adding 2.5 volume of chilled 100% ethanol and 0.26 volume of 7.5 M ammoniumacetat (Sigma) at -20°C overnight. DNA was collected via centrifugation at 4°C at 12000rpm for 20 min. After a final washing step with 70% ice-cold ethanol and air drying, DNA was eluted in 30µl of nuclease-free water. DNA quality was checked on 1% agarose gels stained with RotiSafe (Roth) and a UV-VIS spectrometer (Nanodrop2000c, Thermo Scientific). DNA concentration was measured with a Qubit fluorometer (BR dsDNA Assay Kit, Invitrogen). For each population, the DNA of the individuals were pooled at equimolar contribution (GNP-G1-3 N=29, GNP-G4 N=29; NF414, NF303 N=30). The total amount of DNA provided to the sequencing facility was 3.2 µg per pooled population sample.

Library production and pooled whole genome sequencing

Pooled whole genome sequencing (WGS) was conducted commercially at the Cologne Center of Genomic (CCG, Cologne, Germany). The DNA pools were given to the CCG for library preparation. Libraries were sequenced with 150 bp x 2 paired-end (PE) on the Illumina HiSeq4000.

Individual resequencing (in collaboration with Dr. Rongfeng Cui)

The resequencing of individuals was performed on the extracted DNA from individuals of the pooled sequencing run. To generate the Illumina short-insert library, I adapted a published protocol [160]. The extracted DNA was digested (500 ng) with fragmentase (New England Biolabs) for 20 min at 37°C, followed by end-repair and A-tailing (1 µl NEB End-repair buffer, 0.5 µl Klenow fragment, 0.5 µl Taq. Polymerase, 0.2 µl T4 polynucleotide kinase, 10µl reaction volume. 30 min at 25°C, 30 min at 75°C) and adapter ligation (NEB Quick ligase buffer 12.5 µl, Quick ligase 0.5 µl, 1 µl adapter P1 (D50X, see Table 2.1), 1µl adapter P2 (universal), 5 µM each; 20 min at 20°C, 25 µl reaction volume). Next, the ligation mix was diluted to 50 µl and used 0.583:1 volume of home-brewed SPRI beads (SPRI binding buffer: 2.5 M NaCl; 20 mM PEG 8000; 10 mM TRIS-HCL; 1 mM EDTA,pH 8, 1 ml TE-washed SpeedMag beads GE Healthcare, product number: 65152105050250, per 100 ml buffer) for purification. Next, ligation products were amplified with 9 PCR cycles using the KAPA Hifi kit (Roche, P5 universal primer and P7 indexed primer D7XX, see Table 2.1). The samples were pooled and sequenced on Hiseq X. Due to low coverage of samples NF414-Y1 and NF414-R3, these samples were sequenced again and the sequencing files were merged together for downstream analyses (see below).

Table 2.1: *Nothobranchius furzeri* individual resequencing samples

Population	Species	ID	i5 barcode	i7 barcode	DNA concentration [ng/µl]
GNP-G1-3	*N. furzeri*	M1	D503	D706	282.00
GNP-G4	*N. furzeri*	M9	D507	D705	256.00
NF414	*N. furzeri*	R3	D504/D502	D701/D711	280.00
NF414	*N. furzeri*	Y1	D505/D501	D703/D710	322.00
NF303	*N. furzeri*	M7	D506	D704	476.00

2.1.2 Bioinformatic methods

Assembly of the updated *N. furzeri* genome (performed by Dr. Rongfeng Cui)

10x Genomics read clouds

A single GRZ male individual was sacrificed with MS-222 (Sigma-Aldrich, Steinheim, Germany). Blood was drawn from the heart and high molecular weight DNA was isolated with Qiagen MagAttract kit following manufacturer's instructions. Gemcode v2 DNA library generation was performed by Novogene (Beijing, China). Briefly, a proportion of the sample was run on a pulse field agarose gel to confirm high molecular weight > 100 kb. Based on a genome size estimate of 1.54 Gb, 0.6 ng of DNA was used to construct 2 Gemcode libraries, sequenced on two HiSeq X lanes to obtain a raw coverage of approximately 60X each. The reported input molecular length by `SuperNova` [195] was 118 kb for library 1 and 60.73 kb for library 2. Both libraries were used to correct and scaffold the `Allpath-LG` assembly ([58], see below), and library 1 was also de novo assembled with the `SuperNova` assembler v.2 with default parameters. The `SuperNova` assembly totaled 802.6 Mb, with a contig N50 of 19.65 kb, scaffolded into 6.78 thousand scaffolds with an N50 of 3.83 Mb. Despite high continuity, however, the `BUSCO` [168] metrics are much lower than the `Allpath-LG` assemblies.

Nanopore long reads

DNA was extracted from a single GRZ male individual's muscle tissue by grinding in liquid nitrogen followed by phenol-chlorofom extraction (Sigma). Two kits, the rapid sequencing kit (SQK-RAD004) and the ligation kit (SQK-LSK108), were used to prepare 6 libraries, followed by sequencing on 6 MinION flow cells (R9.4.1). These runs yielded a total of 3.3 Gb of sequences after trimming and correction by `HALC` [7]. For correction, we used the `Allpath-LG` [58] contigs (see below) and short reads from the 10X genomic run.

Allpath-LG assembly

Two independent short read datasets were previous collected for the GRZ strain of *Nothobranchius furzeri*. Allpath-LG [58] was used on the pooled datasets. Together, 4 Illumina short read PE libraries with a fragment size distribution from 158 bp to 179 bp were used to construct the contigs (sequence coverage 191.9X, physical coverage 153.5X), and 22 pair-end and mate pair libraries distributed at 92 bp, 135 bp, 141 bp, 176 bp, 267 bp, 2 kb, 3 kb, 5 kb and 10 kb were used for the scaffolding step (sequence coverage 135.7X, physical coverage 453.8X). The published BAC library ends [150] with an insert size of 112 kb were also included in the ALLPaths-LG run (physical coverage 0.6X). The resulting assembly has a total contig length of 823,583,106 bp distributed in 151,307 contigs > 1 kb, with an N50 of 7.8 kb. The total scaffold length is 943,793,727 bp distributed in 7830 scaffolds with an N50 of 421 kb (with gaps). The resulting assembly was further scaffolded by ARKS ([36], version 1.0) and LINKS ([192], version 1.8.5) with the following parameters:

```
# ARKS
arks        −e 50000
            −c 3
            −r 0.05
            −s 98
# LINKS
links       −m
            −d 4000
            −k 20
            −e 0.1
            −l 3
            −a 0.3
            −t 2
            −o 0
            −z 500
            −r
            −p 0.001
            −x 0
```

This increased the scaffold N50 to 1.527 Mb. Scaffolds were assigned to the Restriction site associated DNA tag (RADtag) linkage map collected from a previous study [185] with Allmaps [178], using equal weight for the two independent mapping crosses. This procedure assigned 90.6% of the assembled bases in 1131 scaffolds to 19 linkage groups, in which 76.6% can be oriented. Next, the 10X genomic read clouds were used to correct misassemblies. Read clouds were mapped to the preliminary assembly with Long Ranger (version 2.1.6) using default parameters, and custom script was used to scan for sudden drops in barcode shares along the assembled linkage groups. The scaffolds were broken at the nearest gap of the drop in 10x barcodes. The same ARKS [36] and LINKS [192] pipeline was again run on the broken scaffolds, increasing the scaffold N50 to 1.823 Mb. Next, BESST_RNA (https://github.com/ksahlin/BESST_RNA) was used to further scaffold the assembly with RNA-Seq libraries, Allmaps [178] was again used to assign the fixed scaffolds back to linkage groups, increasing the assignable bases to 92.2% (879 Mb) with 80.3% (765 Mb) with determined orientation. The assembly was broken again with Long Ranger and reassigned to linkage groups (LGs) with Allmaps [178], and further partitioned the scaffolds to LG due to linkage of some left-over scaffolds with an assigned scaffold. Each partitioned scaffold group was subjected to the ARKS [36] and LINKS [192] pipeline again, to constraint the previously unassigned scaffolds onto the same linkage group. Allmaps [178] was run again on the improved scaffolds, and now assigned 94.5% (903.4 Mb) of bases and oriented 89.1% (852 Mb) of bases. Long Ranger was ran again, visually checked and compared with the RADtag markers and this way 11 misoriented positions were identified and corrected. Gaps were further patched with 2X of HALC corrected nanopore long reads and the BGI500 short PE reads with GMCloser [99] with the following parameters:

```
# GMCloser
gmcloser     --blast
             --long_read
             --lr_cov 2
              -l 100
              -i 466
              -d 13
             --min_subcon 1
             --min_gap_size 10
             --iterate 2
             --mq 1
              -c
```

The corrected long reads not mapped by `GMCloser` [99] were assembled by `CANU` [97] into 7.9 Mb of sequences, which are likely unassigned repeats.

Meta assembly

Five assemblies were integrated by `MetAssembler` [198] in the following order (ranked by `BUSCO` [168] scores) using a 20 kb mate pair library: 1) The improved `Allpaths-LG` assembly assigned to linkage groups produced in this study, 2) A previously published assembly with `Allpaths-LG` and optical map [150] 3) A previously published assembly using `SGA` (string graph assembler, used in Valenzano *et al.* [185]), 4) The `SuperNova` assembly with only 10x Genomic reads and 5) Unassigned nanopore contigs from `CANU`. The final assembly NFZ V2.0 has 911.5 Mb of scaffolds assigned to linkage groups. Unassigned scaffolds summed up to 142.2 Mb, yielding a total assembly length of 1053.7 Mb, approximately 2/3 of the total genome size of 1.53 Gb. The final assembly has 95.2% complete and 2.24% missing BUSCOs [168].

Mapping of NCBI GenBank gene annotations

RefSeq mRNAs for the GRZ strain (PRJNA314891, PRJEB5837) were downloaded from GenBank [34], and aligned them to the assembly with `exonerate` [170]. The RefSeq mRNAs have a `BUSCO` score of 98.0% complete, 0.9% missing. The mapped gene models resulted in a `BUSCO` score of 96.1% complete, 2.1% missing.

Human ortholog

Human ENSEMBL gene identifier were identified using the Compara database from ENSEMBL v.87 [209]. The killifish orthologs were first mapped to fish species present in the ENSEMBLE v.87 database. Next, the fish species identifier were mapped to their human orthologs. If multiple human orthologs were available for a fish gene, one identifier was arbitrarily picked.

Classification of synonymous and non-synonymous sites

Four-fold degenerated (4-fold) were treated as synonymous and zerofold-degenerated sites (0-fold) were treated as non-synonymous. Synonymous and non-synonymous sites were categorized using the NFZ V2.0 gene annotations. During this process, sites overlapping unreliable parts of the gene model based on the extended GFF annotations were excluded. Specifically, regions annotated as unreliable in the 3' and 5' ends, as well as regions with poor alignment to NCBI reference proteins were excluded from further analyses.

Pseudogenome assembly generation (performed by Dr. Rongfeng Cui)

The *pseudogenomes* for *Nothobranchius orthonotus* and *Nothobranchius rachovii* were generated from sequencing data from Cui *et al.* [38] using the same method used in Cui *et al.* [38]. Briefly, the sequencing data were mapped to the NFZ V2.0 reference genome by BWA-mem (version 0.7.12, [104, 105]) in PE mode. PCR duplicates were marked with MarkDuplicates tool in the Picard (version 1.119, http://broadinstitute.github.io/picard/) package. Reads were realigned around INDELs with the IndelRealigner tool in GATK (version 3.4.46, [119]). Variants were called with SAMTOOLS (version 1.2, [106]) mpileup command, requiring a minimal mapping quality of 20 and a minimal base quality of 25. A *pseudogenome* assembly was generated by substituting reference bases with the alternative base in the reads. Uncovered regions, INDELs and sites with >2 alleles were masked as unknown "N". The allele with more supporting reads was chosen at biallelic sites.

Mapping of longevity and sex quantitative trait loci

The quantitative trait loci (QTL) markers published in Valenzano *et al.* [185] were obtained directly from Dario Valenzano. In order to map the markers associated to longevity and sex, I created a reference database using BLAST [2]. The nucleotide database was created with the new reference genome of *N. furzeri* (NFZ V2.0). Subsequently, I mapped the QTL marker sequences to the database. Only markers with full support for the total length of 95 bp were considered as QTL markers. Parameters were used as described below.

```
# Building the database
makeblastdb -in NFZv2.0.fasta -dbtype nucl

# Use BLAST to find the QTL sequences in the genome
blastn -db NFZv2.0.fasta -outfmt 6 -query QTL.fa -evalue 0.0000001
```

Synteny analysis

Synteny analysis was performed using orthologous information from Cui *et al.* [38] determined by the UPhO pipeline [6]. For this, I compared the 1-to-1 orthologous gene positions of the new turquoise killifish reference genome (NFZ V2.0) and two closely related teleost species, *Xiphophorus maculatus* and *Oryzias latipes*. I visualized the result using Circos [101] for the genome-wide comparison and the *genoPlotR* package [62] in R

[146] for the sex chromosome synteny analysis. Synteny plots for orthologous chromosomes of *Xiphophorus maculatus* and *Oryzias latipes* were generated with `Synteny DB` (http://syntenydb.uoregon.edu, [18]).

General mapping pipeline

Raw sequencing reads were trimmed using `Trimmomatic-0.32` [16]. Data files were inspected with `FastQC` (version 0.11.22, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Trimmed reads were subsequently mapped to the reference genome with `BWA-MEM` (version 0.7.12, [104, 105]). The `SAM` output was converted into `BAM` format, sorted, and indexed via `SAMTOOLS` (version 1.3.1, [106]). Filtering and realignment was conducted with `PICARD` (version 1.119, http://broadinstitute.github.io/picard/) and `GATK`[119]. Briefly, the reads were relabeled, sorted, and indexed with `AddOrReplaceReadGroups`. Duplicated reads were marked with the `PICARD` feature `MarkDuplicates` and reads were realigned with first creating a target list with `RealignerTargetCreator`, second by `IndelRealigner` from the `GATK` suite [119]. Resulting reads were again sorted and indexed with `SAMTOOLS` [106]. Parameters were used as described below.

```
# Trimming
java -Xmx3000m -jar trimmomatic-0.32.jar PE -threads 10
                    {input.R1} {input.R2}
                    {output.paired1} {output.unpaired1}
                    {output.paired2} {output.unpaired2}
                    ILLUMINACLIP:illumina-adaptors.fa:3:7:7:1:true
                    LEADING:20
                    TRAILING:20
                    SLIDINGWINDOW:4:20
                    MINLEN:50
# Mapping
bwa mem -t 10 -k 17 -w 500 -E 0 {input}
# Conversion into BAM format, sorting and indexing
samtools view -b -S {input} > {output}
samtools sort -m 4G -T {sample} {input.file} -o {output}
samtools index {input}
* continued on next page
```

```
# AddOrReplaceReadGroups
java -Dsnappy.disable=true -Xmx12g -jar AddOrReplaceReadGroups.jar
                          I={input.bam}
                          O={output}
                          SORT_ORDER=coordinate
                          RGID=sample
                          RGLB=sample
                          RGPL=illumina
                          RGSM=sample
                          RGPU=sample
                          CREATE_INDEX=True
                          VALIDATION_STRINGENCY=SILENT

# MarkDuplicates
java  -Dsnappy.disable=true -Xmx12g -jar  MarkDuplicates.jar
                          INPUT={input.bam}
                          OUTPUT={output}
                          METRICS_FILE={sample}.dupmetrics
                          VALIDATION_STRINGENCY=SILENT
                          MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1000
                          TMP_DIR=./data/realigned_BAM/
# Indexing
samtools index {output.MarkDuplicates}

# RealignerTargetCreator
java -Djava.io.tmpdir=~/tmp -Xmx12g -jar GenomeAnalysisTK.jar
                          -T RealignerTargetCreator
                          -R {reference.genome}
                          -I {input.bam}
                          -o {output}

# IndelRealigner
java -Xmx12g   -jar GenomeAnalysisTK.jar
                          -T IndelRealigner
                          -R {reference.genome}
                          -I {input.bam}
                          -targetIntervals {input.intervals}
                          -o {output}
                          --maxReadsForRealignment 120000

# Sorting & indexing
samtools sort  -m 4G -T {.sample} {input.file} -o {output} -@ 10
samtools index {input}
```

Inference of population history with individual resequencing data

I inferred the demographic history from single individual resequencing data using Pairwise Sequential Markovian Coalescence (`PSMC'` mode from `MSMC2`, [165]). Raw sequencing reads were mapped according to the mapping pipeline (Section 2.1.2). Next, I followed the guidance for `PSMC'` (https://github.com/stschiff/msmc/blob/master/guide.md) and generated VCF-files and masked files with the *bamCaller.py* script (MSMC-tools package). This step requires the chromosome coverage information to mask regions with too low or too high coverage. As recommended in the guidelines, I computed the coverage per chromosome using `SAMTOOLS` (1.3.1, [106]). In addition, I performed this step using a coverage threshold of 18 as recommended by Nadachowska-Brzyska *et al.* [124]. I generated the final input data using the *generate_multihetsep.py* script (MSMC-tools package). Subsequently, for each sample `PSMC'` was run independently. Bootstrapping was performed for 30 samples per individual and input files were generated with the *multihetsep_bootstrap.py* script (MSMC-tools package).

Generating input files for PoPoolation, PoPoolation2 and NPSTAT

For population genetic bioinformatics analyses the BAM files of the populations were converted into the required MPILEUP format via the SAMTOOLS mpileup command [106]. Low quality reads were excluded by setting a minimum mapping quality of 20 and a minimum base quality of 20. Further, possible insertion and deletions (INDELs) were identified with *identify-genomic-indel-regions.pl* script from the PoPoolation package [95] and were subsequently removed via the *filter-pileup-by-gtf.pl* script [95]. Coding sequence positions that were identified to be putative ambiguous were removed by providing the *filter-pileup-by-gtf.pl* script a custom modified GTF file with the corresponding coordinates.

Parameters were used as described below.

```
# Generating population specific pileup
samtools mpileup  -A -B -t DP,DV,DPR,INFO/DPR,DP4,SP
                  --adjust-MQ 0
                  --min-MQ 20
                  --min-BQ 20
                  -o {output}
                  -f {reference.genome} {input.file}

# Identify INDEL regions
perl identify-genomic-indel-regions.pl
                  --input {input.file}
                  --output {output}
# Exclude identified INDEL regions
perl filter-pileup-by-gtf.pl
                  --input {input.file}
                  --gtf {input.gtf}
                  --output {output}
# Exclude ambigious coding sequence
perl filter-pileup-by-gtf.pl
                  --input {input.file}
                  --gtf mask_ambigous_cds.gtf
                  --output {output}
```

Merging sequencing reads of populations from the Gonarezhou National Park

To combine the sequencing reads of the two populations from the Gonarezhou National Park (GNP), I used the `SAMTOOLS merge` command [106]. The populations GNP-G1-3 and GNP-G4 were merged together and this population was subsequently denoted as GNP.

Estimating genetic diversity

Genetic diversity in the populations was estimated by calculating the nucleotide diversity π [125] and Wattersons's estimator θ [194]. The nucleotide diversity π is defined as the average pairwise differences between any two pair of sequences in the sample. Therefore, π is the average number of nucleotide difference per site and is referred to as the observed genetic diversity. Wattersons's estimator θ is based on the total number of segregating sites scaled to the sample size. In population genetics, θ is referred to as the expected genetic diversity. Calculation of π and θ was done with a sliding window approach by using the *Variance-sliding.pl* script from the `PoPoolation` program [95]. Non-overlapping windows were calculated with a length of 50 kb with a minimum count of two per single-nucleotide polymorphism (SNP), minimum quality of 20 and the population specific haploid pool size. Low covered regions that fall below half the mean coverage of each population were excluded, as well as regions that exceed a two times higher coverage than the mean coverage. The upper threshold is set to avoid repetitive regions with possible wrong assemblies. Mean coverage was estimated on filtered `MPILEUP` files. Each window had to be at least covered to 30% to be included in the estimation.

Parameters were used as described below (see next page).

```
# Nucleotide diversity calculation
perl Variance-sliding.pl
            --input {input.pileup}
            --output {output}
            --measure {pi;theta,D}
            --window-size 50000
            --step-size 50000
            --min-count 2
            --min-coverage {GNP= 23; NF303=18; NF414=19}
            --max-coverage {GNP= 94; NF303=70; NF414=77}
            --min-qual 20
            --pool-size {GNP=116; NF303=60; NF414=60}
            --min-covered-fraction 0.3
            --fastq-type sanger
```

Estimating Tajima's D

Deviations from neutrality were detected via calculation of Tajima's D [177]. The assumption underlying Tajima's D is that under neutrality the two estimates of genetic diversity, π and θ, should be the same. The observed genetic diversity should equal the expected genetic diversity. Deviations from the neutral expectation of the neutral theory of molecular evolution lead to the observed genetic diversity (π) being greater or smaller than the expected genetic diversity (θ). Tajima's D is calculated by taken the difference between these two estimates divided by the standard deviation (sd) of the difference ([177], see Equation 2.1).

Significant deviation from zero was tested with a t-test on the observed distribution of Tajima's D against a random normal distribution (18421 values based on number of investigated windows, functions *t.test()* and *rnorm()* in R [146]) with an average of zero and the population specific standard deviation (according to [48]).

$$D = \frac{\pi - \theta}{sd(\pi - \theta)} \tag{2.1}$$

For the reason that estimates of Tajima's D can be strongly biased by sequencing errors, I computed Tajima's D with NPSTAT [46]. NPSTAT takes sequencing error and ascertainment error into account and is specifically designed for pooled sequencing [46]. Tajima's D was calculated for 50 kb non-overlapping sliding windows with a minimum mapping quality of 20 and a minimum allele count of 4. The minimum allele count of 4 was set to further reduce the effect of sequencing errors. Low covered regions that fall below half the mean coverage of each population were excluded, as well as regions that exceed a two times higher coverage than the mean coverage.

To detect recent signatures of positive selection, I performed outlier detection based on windows with a value below a threshold of 0.5% of all genome-wide values. To reduce false-positives results, I only considered regions with at least 3 consecutive windows below the 0.5% threshold as potentially under positive selection. The results were visualized with the *ggplot2* package in R [200, 146]. For easier visualization, I averaged Tajima's D values over four windows with the *rollmean()* function of the *zoo* package [208].

Parameters were used as described below.

```
# Calculation of Tajima's D with NPSTAT
npstat      -n 60
            -l 50000
            -minqual 20
            -nolowfreq 4
            -mincov {GNP= 23; NF303=18; NF414=19}
            -maxcov {GNP= 94; NF303=70; NF414=77}
            {input.pileup}
```

Estimation of effective population size *Ne*

Wattersons's estimator of θ [194] is referred to as the population mutation rate. The estimate is a compound parameter that is calculated as the product of the effective population size (*Ne*), the ploidy (2p, with p is ploidy) and the mutational rate μ (Equation 2.2).

$$\theta = 2p * Ne * \mu \tag{2.2}$$

Therefore, *Ne* can be obtained when θ, the ploidy and the mutational rate μ are known. The turquoise killifish is a diploid organism with a mutational rate of $2.6321e^{-9}$ per base pair per generation (assuming one generation per year in Killifish [38]) and θ estimates were obtained with `PoPoolation` ([95], see Section 2.1.2). According to the following equation I calculated *Ne* for all populations:

$$Ne = \frac{\theta}{4 * \mu} \tag{2.3}$$

Genetic differentation index F_{ST}

The genetic differentiation index F_{ST} is a measure of population subdivision. Population subdivision can cause a reduction in heterozygosity due to limited exchange of genetic information between the subpopulations. F_{ST} measures the decrease in heterozygosity of the subpopulation (H_S) in relation to the total population (H_T, Equation 2.4, [206]).

According to Wright:

$$F_{ST} = \frac{H_T - H_S}{H_T} \tag{2.4}$$

In general, F_{ST} in the range from 0 to 0.05 indicates little genetic differentiation, from 0.05 to 0.15 moderate genetic differentiation, from 0.15 to 0.25 strong genetic differentiation and a F_{ST} above 0.25 very strong genetic differentiation [71].

In order to calculate F_{ST} from pooled sequencing data, I used the program `PoPoolation2` [96]. The filtered and realigned `BAM` files of each population were merged into a single pileup file with `SAMTOOLS` `mpileup` [106], with a minimum mapping quality and a minimum base quality of 20. The pileup was synchronized using the *mpileup2sync.jar* program from the `PoPoolation2` program, requiring a minimum base quality of 20 [96]. Insertions and deletions were identified and removed with the *identify-indel-regions.pl* and *filter-sync-by-gtf.pl* scripts of `PoPoolation2` [96]. Again, coding sequence positions that were identified to be putative ambiguous were removed by providing a custom modified `GTF` file with the corresponding coordinates to the *filter-pileup-by-gtf.pl* script. Further, I generated a synchronized pileup file for only genes by providing a `GTF` file with genes coordinates to the *create-genewise-sync.pl* from `PoPoolation2` [96]. I calculated F_{ST} for each pairwise comparison (GNP vs NF303, GNP vs NF414, NF414 vs NF303) in a genome-wide approach using non-overlapping sliding windows of 50 kb with a minimum count of four per SNP, a minimum coverage of 20, a maximum coverage of 94 for GNP, 77 for NF414, and 70 for NF303 and the corresponding pool size of each population (N= 116 ; 60; 60). Each sliding window had to be at least covered to 30% to be included in the estimation. The same thresholds, except the minimum covered fraction, with different sliding window sizes were used to calculate the gene-wise F_{ST} for the complete gene body (window-size of 2000000, step-size of 2000000) and single SNPs within genes (window-size of 1, step-size of 1). The non-informative positions were excluded from the output. Significance of allele differences per base pair within the gene-coordinates were calculated with the Fisher's exact test implemented in the *fisher-test.pl* script of PoPoolation2 [96].

Calculation of unrooted neighbor joining tree based on the genome-wide pairwise F_{ST} averages was performed with the *ape* package in R [134].

Parameters were used as described below.

```
# Generating merged pileup
samtools mpileup   -A -B -t DP,DV,DPR,INFO/DPR,DP4,SP
                   --adjust-MQ 0
                   --min-MQ 20
                   --min-BQ 20
                   -o {output}
                   -f {reference.genome} {input.files}
# Synchronize merged pileup
java -jar mpileup2sync.jar
                   --input {input.file}
                   --output {output}
                   --fastq-type sanger
                   --min-qual 20
                   --threads 10
# Synchronize merged pileup for genes
perl create-genewise-sync.pl
                   --input {input.file}
                   --gtf exons.gtf
                   --output {output}
* continued on next page
```

```
# Calculate population differentiation FST
# Genome-wide sliding windows
perl fst-sliding.pl
                  --input {input.file}
                  --output {output}
                  --min-count 4
                  --min-coverage 20
                  --max-coverage 94,77,70
                  --min-covered-fraction 0.3
                  --window-size 50000
                  --step-size 50000
                  --pool-size 116:60:60
# Genes: complete gene body
perl fst-sliding.pl
                  --input {input.file}
                  --output {output}
                  --suppress-noninformative
                  --min-count 4
                  --min-coverage 20
                  --max-coverage 94,77,70
                  --window-size 2000000
                  --step-size 2000000
                  --pool-size 116:60:60
# Genes: SNPs
perl fst-sliding.pl
                  --input {input.file}
                  --output {output}
                  --suppress-noninformative
                  --min-count 4
                  --min-coverage 20
                  --max-coverage 94,77,70
                  --window-size 1
                  --step-size 1
                  --pool-size 116:60:60
# Fisher's exact test on SNPs
perl  fisher-test.pl
                  --input {input.file}
                  --output {output}
                  --suppress-noninformative
                  --min-count 4
                  --min-coverage 20
                  --max-coverage 94,77,70
                  --min-covered-fraction 0.0
                  --window-size 1
                  --step-size 1
```

F_{ST} Outlier detection

To detect regions of high and low genetic differentiation, I performed genome-wide outlier detection. The pairwise F_{ST} values calculated with `PoPoolation2` were standardized by Z-transformation (see equation 2.5).

$$Z_{F_{ST}} = \frac{F_{ST} - \overline{F_{ST}}}{\sigma F_{ST}} \tag{2.5}$$

The Z-transformation allowed to put all pairwise comparison into the same framework because $Z_{F_{ST}}$ values indicate the number of standard deviations σ by which F_{ST} deviates from the mean $\overline{F_{ST}}$ (method comparable to Rubin *et al.* [162]). Outliers were identified as non-overlapping windows of 50 kb within the 0.5% of lowest and highest genetic differentiation per comparison. To reduce the number of false-positive results, the outlier threshold was chosen at 0.5% highest and lowest percentile of each pairwise genetic differentiation [61, 142]. To find candidate genes within windows of highest differentiation, I used a total of three selection criteria. First, the window-based $Z_{F_{ST}}$ value had to be above the 99.5^{th} percentile of pairwise genetic differentiation. Second, the gene F_{ST} value had to be above the 99.5^{th} percentile of pairwise genetic differentiation and last, the gene needed to include at least one SNP with significant differentiation based on Fisher's exact test (calculated with `PoPoolation2`; P<0.001, Benjamini-Hochberg corrected P-values [12]).

Identifying polymorphic sites from pooled sequencing

I performed SNP calling with `Snape` [147]. This program was specifically designed to infer polymorphic positions from pooled sequencing data [147]. The program requires information of the prior probability of genetic diversity (θ) for the Bayesian model. Hence, the initial values of genetic diversity obtained with `PoPoolation` [95] were used. The inference with `SNAPE` was done on the folded spectrum and the prior type was set to informative. As `Snape` requires the `MPILEUP` format, the previously obtained `MPILEUP` files generated with `PoPoolation` [95] and `SAMTOOLS` [106] were used as input files. SNP calling was separately performed on coding and non-coding parts of the genome. Therefore, each population `MPILEUP` file was filtered by coding sequence position with the *filter-pileup-by-gtf.pl* script of `PoPoolation` [95]. For coding sequences the *–keep-mode* was set to retain all coding sequences. The non-coding sequences were obtained by using the default option and discarding the coding sequences from the `MPILEUP` file. `Snape` produces a posterior probability of segregation for each position. The posterior probability of segregation was used to filter low-confidence SNPs and indicated in the specific section.

Site frequency spectrum

I generated the site frequency spectrum (SFS) of the minor allele frequency (MAF) for each population with SNPs called with `SNAPE`. For this, the SNPs were filtered based on population coverage. Therefore, regions with too low coverage (below half the mean population coverage) and too high coverage (above two times the mean coverage) were excluded. In addition, positions with only one supporting read for an allele were treated as monomorphic sites. Further, only SNPs with a posterior probability of segregation above 0.75 were considered as true SNPs. The SFS was generated for SNPs with a frequency between 0.05 and 0.95. The SFS was visualized with *ggplot2* [200] using bins from 0.05 to 0.95 in steps of 0.05.

Divergence and polymorphisms in synonymous and non-synonymous sites

Polarization of SNPs can be done based on major/minor frequency, reference/non-reference frequency or ancestral/derived frequency. For subsequent analyses, SNPs were polarized based on ancestral/derived frequency. Therefore, polarization of synonymous sites (four-fold degenerated sites) and non-synonymous sites (zero-fold degenerated sites) was done using the *pseudogenomes* of outgroups *Nothobranchius orthonotus* and *Nothobranchius rachovii*, respectively. For each population the genomic information of the respective *pseudogenomes* was extracted with `bedtools getfasta` [144, 145]. Next, I inferred the derived allele frequency (DAF) for every position with a custom `R` script. Briefly, I only included sites with bases *A*, *G*, *T* or *C* present in the outgroup *pseudogenomes* and I checked if the position has an alternate allele in each of the investigated populations. Occasions with an alternate allele present in the population data were treated as possible divergent or polymorphic sites. Divergent sites are positions in the genome were the outgroup allele is different from the allele present in the population. Polymorphic sites are sites in the genome that have more than one allele segregating in the population. Only biallelic polymorphic sites were used in this analysis. The DAF was determined as frequency of the allele not shared with the respective outgroup. In general, postions with only one supporting read for an allele were treated as monomorphic sites. SNPs with a DAF $< 5\%$ or $\geq 95\%$ were treated as fixed mutations. Further filtering was done based on the threshold of the posterior probability of >0.75 calculated with `Snape` (see previous subsection), combined with the minimum and maximum coverage threshold per population. Low covered regions that fall below half the mean coverage of each population were excluded, as well as regions that exceed a two times higher coverage than the mean coverage of each population.

Asymptotic McDonald-Kreitman α

The rate of substitutions that were driven to fixation by positive selection was evaluated with a method based on the McDonald-Kreitman test [118]. The test is based on the assumption that the proportion of non-synonymous mutations that are neutral has the same fixation rate as synonymous mutations. Therefore, under neutrality the ratio between non-synonymous to synonymous substitutions ($\frac{D_n}{D_s}$) between species is equal to the ratio of non-synonymous to synonymous polymorphisms within species ($\frac{P_n}{P_s}$). If positive selection takes place, the ratio between non-synonymous to synonymous substitutions between species is larger than the ratio of non-synonymous to synonymous polymorphism within species [118]. The concept behind this is that the selected variant reaches fixation in a shorter time than by random drift. Therefore, the selected variant increases D_n, not P_n. The proportion of non-synonymous substitutions that were fixed by positive selection (α) was estimated with an extension of the McDonald-Kreitman test ([172], see Equation 2.6).

$$\alpha = 1 - \frac{P_n * D_s}{P_s * D_n} \tag{2.6}$$

Due to the presence of slightly deleterious mutations the estimate of α can be underestimated. For this reason, I implemented the method used by Messer and Petrov to calculate α as a function of the derived allele frequency x ([122, 66], see Equation 2.7).

$$\alpha(x) = 1 - \frac{P_{n(x)} * D_s}{P_{s(x)} * D_n} \tag{2.7}$$

With this method the true value of α can be inferred as the asymptote of the function of α. Additionally, the value of $\alpha(x)$ for low derived frequencies should give an estimate of the number of slightly deleterious mutations that segregate in the population. From now on I refer to the McDonald-Kreitman's α as MK-α.

Direction of selection

To further investigate the signature of selection, I computed the direction of selection (DoS) index for every gene [175]. DoS standardizes α values to a value between -1 and 1 (see Equation 2.8). A positive value of DoS indicates adaptive evolution (positive selection) and a negative value indicates the segregation of slightly deleterious alleles, therefore weaker purifying selection [175]. This ratio is undefined for genes without any information about polymorphic or substituted sites. Therefore, I only included genes with at least one polymorphic and one substituted site.

$$DoS = \frac{D_n}{D_n + D_s} - \frac{P_n}{P_n + P_s} \tag{2.8}$$

Direction of selection: Gene Ontology slim families

To test for significant differences in the distribution of direction of selection (DoS) within genes belonging to certain Gene Ontology (GO) families, I downloaded the human GO slim annotation (GRCh38.p7) from ENSEMBL BioMart v87 [171]. GO slims are a subset of all GO ontologies and give a broad overview of the ontology content. Only GO slim terms with at least 10 genes were included in the analysis. Calculation of DoS values of genes belonging to each GO slim term per population was done with a custom R script. Significant deviation of the DoS distribution of a single GO slim term from the total DoS distribution was assessed with the *wilcox.test()* function in R [146]. GO slim terms that deviated significantly (*P*-value <0.05) and were below the median DoS of the total distribution were classified in the significantly relaxed selection category. GO slim terms that deviated significantly (*P*-value <0.05) and were above the median DoS of the total distribution were classified in the significantly intensified selection category. Only GO slim terms that were significant in at least one population were used for visualization with the `R` packages *ggplot2* [200], *ggpubr* [82] and *extrafont* [22].

Inference of distribution of fitness effects

To infer the distribution of fitness effects (DFE), I used the program `polyDFE2.0` [180]. For this analysis the unfolded site frequency spectra (SFS) of non-synonymous (0-fold) and synonymous sites (4-fold) were projected into 10 chromosomes for each population. Information about the fixed derived sites was included in this analysis (using *Nothobranchius orthonotus*). `PolyDFE2.0` estimates either the full DFE, containing deleterious, neutral and beneficial mutations, or only the deleterious DFE. To obtain the best model for each population, I performed a model testing approach with three different models implemented in `PolyDFE2.0`:

- model A: using a reflected displaced gamma distribution
- model B: using a mixture of gamma distribution and a discrete distribution
- model C: using a mixture of gamma distribution and an exponential distribution

Due to possible biases from erroneous polarization or unknown demography, I included runs accounting for polarization errors and demography (+eps, +r). Initial parameters were automatically estimated with the *–e* option, as recommended. To ensure that the parameter space is explored thoroughly, I applied the basin hopping option with maximum 10 iterations (*-b*). The best model for each population was chosen based on the Akaike Information Criterion (AIC). Since the best model for GNP was the deleterious DFE (Table E.3), I tested only the deleterious DFE for NF414 and NF303 (see Tables E.1, E.2). In the case of deleterious DFE inference, model B and C should give similar results (see Tables E.3, E.1, E.2). Confidence intervals were generated by running 500 bootstrap datasets with the same parameters used to infer the best model.

Variant annotation of coding-sequence single-nucleotide polymorphisms

To classify changes in the coding-sequence (CDS), I performed variant annotation with the variant annotator SnpEFF [32]. Due to the limitation that there is no available variant annotation database for *N. furzeri*, the database was generated with a feature within SnpEFF and included the new genome of *N. furzeri* (NFZ V2.0) to the SnpEFF pipeline [32]. The database for CDS variant annotation was done with the genome NFZ V2.0 FASTA file and the annotation gene transfer file (GTF). For variant annotation the population specific synonymous and non-synonymous sites with a change in respect to the reference genome NFZ V2.0 were used to infer the impact of these sites. The possible annotation impact classes were low, moderate, and high. SNPs with a frequency below 5% were excluded for this analysis. Information about the derived frequency is not required for the analysis with SNPEFF. However, to be consistent with the analysis of the distribution of fitness effects, I included only positions also found to be present in the *N. orthonotus* pseudogenome. The command used to build the database for variant calling with the genome NFZ V2.0 FASTA and GTF file was following:

```
# Building SnpEFF database
java -jar snpEff.jar build -gtf22 -v NFZ_v2.0
```

Overrepresentation analysis

I performed Gene ontology (GO) and pathway overrepresentation analysis with the online tool `ConsensusPathDB` (http://cpdb.molgen.mpg.de;version33)[73]. Briefly, each gene present in the outlier list was provided with an ENSEMBL human gene identifier [209], if available, and entered as the target list into the user interface. All genes included in the analysis and with available human ENSEMBL identifier were used as the background gene list. `ConsensusPathDB` maps the entries to the databases and calculates the enrichment score for each entity by comparing the proportion of target genes in the entity over the proportion of background genes in the entity. For each of the enrichment a P-value is calculated based on a hyper geometric model and is corrected for multiple testing using the false discovery rate (FDR). Only GO terms and pathways with more than two genes were counted as enriched. Overrepresentation analysis was performed on genes falling below the 2.5^{th} percentile or above the 97.5^{th} percentile thresholds. The percentiles for either F_{ST} or DoS values were calculated with the *quantile()* function in `R`.

Statistical analysis and data processing

Statistical analyses were performed using `R studio` version 1.0.136 (R version 3.3.2, [146, 161]) on a local computer and `R studio` version 1.1.456 (R version 3.5.1) in a cluster environment at the Max-Planck-Institute for Biology of Ageing (Cologne). Unless otherwise stated, the functions *t.test()* and *wilcox.test()* in `R` [146] have been used to evaluate statistical significance. Snakemake was used to generate a pipeline for data processing [98]. Data visualization in `R` was done using *ggplot2* [200] or base `R` visualization functions. Figure style was modified using `Inkscape` version 0.92.4. Circular visualization of genomic data was performed with `Circos` [101].

2.2 Material

Equipment

- Balance (Sartorius, Göttingen, Germany)
- Centrifuge 5424 R (Eppendorf , Hamburg, Germany)
- Chemidoc MP (Bio Rad, München, Germany)
- Freezer -80 °C (Thermo Scientific, Waltham, Massachusetts, USA)
- Freezer -20 °C (Liebheer, Bulle, Germany)
- Fridge 4 °C (Liebheer, Bulle, Germany)
- Gel electrophoresis system (Bio Rad, München, Germany)
- Ice machine (Scotsman, Mailand, Italy)
- Lab dancer (VWR, Darmstadt, Germany)
- Microcentrifuge GmCLab (Gilson, Limburg, Germany)
- Microwave (Severin, Sundern, Germany)
- Nanodrop 2000c (Thermo Scientific, Waltham, Massachusetts, USA)
- Qubit (Thermo Scientific, Waltham, Massachusetts, USA)
- Pipetboy (Integra, Konstanz, Germany)
- Pipets (Eppendorf , Hamburg, Germany)
- Pipet tips + filter tips (Tip One, Hamburg, Germany)
- Plastic pipettes 2, 5, 10, 25, 50ml (Greiner Bio-One, Österreich)
- Power supply (Bio Rad, München, Germany)
- PCR tubes 0.2ml (Biozym, Oldendorf, Germany)
- Thermal cycler C1000 touch (Bio Rad, München, Germany)
- Thermal cycler 7900HT Fast Real Time PCR (Applied Biosystems, Darmstadt, Germany)
- Thermomixer comfort (Eppendorf, Hamburg, Germany)
- Tubes 1.5 & 2.0 ml (Sarstedt, Nümbrecht, Germany)

Chemicals

The chemicals used were purchased from the companies Sigma-Aldrich (Steinheim, Germany), New England Biolabs (Ipswich, Massachusetts, USA), Thermo Scientific (Waltham, Massachusetts, USA), Roche (Mannheim, Germany), Roth (Karlsruhe, Germany) and Life Technologies (Darmstadt, Germany). Plastic- and glassware were ordered from Sarstedt (Nümbrecht, Germany), Schott (Mainz, Germany) and VWR International (Darmstadt, Germany). All the non-sterile chemicals and materials were autoclaved at 121°C under 1.2 bar.

Samples

The samples for the population genetic study of the turquoise killifish were collected in the Gonarezhou National Park (GNP) in Zimbabwe in 2015 (Permit NO.:23(1)(C) (II)30/2015, see Table 2.2). I collected the samples together with Dr. Dario Riccardo Valenzano. Additional samples from Mozambique (NF414+NF303) were provided by Dr. Martin Reichard (Institute of Vertebrate Biology, Brno, Czech Republic). For population GNP-G1-3 it was not possible to obtain equal numbers of male and females. Population NF414 is polymorphic with regard to tail color of male individuals. As indicated in Table 2.2, I used seven red-tailed male individuals (R) and eight yellow-tailed male individuals (Y).

Table 2.2: *Nothobranchius furzeri* population samples

ID	Species	Latitude	Longitude	Country	Female	Male	Library-prep	Platform
GNP-G1-3	*N. furzeri*	-21.80	31.92	Zimbabwe	21	8	pooledWGS	Illumina HiSeq4000
GNP-G4	*N. furzeri*	-21.77	31.88	Zimbabwe	14	15	pooledWGS	Illumina HiSeq4000
NF414	*N. furzeri*	-22.55	32.73	Mozambique	15	15 (7R,8Y)	pooledWGS	Illumina HiSeq4000
NF303	*N. furzeri*	-24.11	32.77	Mozambique	15	15	pooledWGS	Illumina HiSeq4000

Chapter 3

Results

To find out what evolutionary mechanisms led to differences in life history traits between the turquoise killifish populations, I sequenced four populations of the turquoise killifish along an aridity gradient (Figure 3.1). The populations NF303 (N=30) and NF414 (N=30) were collected in Mozambique and correspond to a semi-arid climate classification, with NF303 being the population derived from the least arid location in this study (Figure 3.1). In addition to the NF303 and NF414 populations, I also collected two populations in the Gonarezhou National Park (GNP-G1-3 and GNP-G4; N=29 each), the driest region within the species distribution of the turquoise killifish [185]. The populations GNP-G1-3 and GNP-G4 are located near the collection site of the type specimen of *Nothobranchius furzeri* [81]. Moreover, the most commonly used laboratory strain, GRZ-AD, is derived from near GNP locations and is the shortest-lived turquoise killifish strain in captivity [185], whereas the longer-lived laboratory strains are derived from the vicinity of population NF303 [185]. When the terms wet or dry population are used in this study, I refer to the population NF303 or GNP. The population NF414 is classified as intermediate between wet and dry (Figure 3.1).

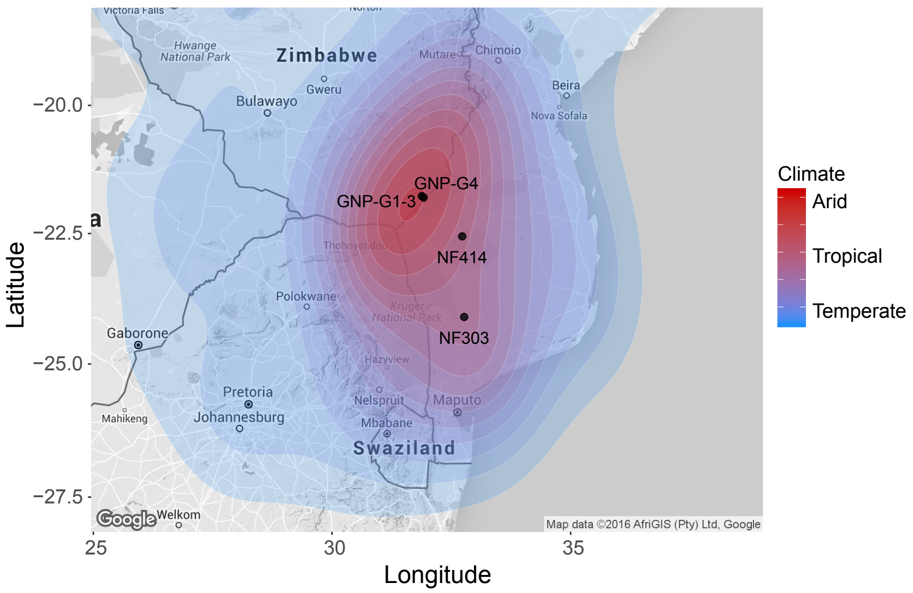

Figure 3.1: Geographical distribution and climate classification. Geographical distribution of sampled turquoise killifish population in the wild (black dots). Shown in concentric contours is the climate classification of the area measured by the Koeppen-Geiger index. Modified from Valenzano *et al.* [185].

3.1 Sequencing of turquoise killifish populations

Pooled sequencing of the natural turquoise killifish populations resulted in a range of 418-517 million paired-end reads per population (see Table 3.1). Reference mapping was conducted by aligning sequence reads from each population to the updated reference genome (NFZ V2.0, unpublished). After quality filtering, the average coverage for all four populations ranged between 22-39x (see Table 3.1).

Table 3.1: Summary of pooled sequencing of turquoise killifish populations

Population	**Read pairs [in million]**	**Mean coverage BAM**	**Mean coverage MPILEUP**	**Mean coverage filtered MPILEUP**
GNP-G1-3	517.2	55.8	32.1	22.1
GNP-G4	498.3	58.6	35.4	26.5
NF414	418.7	60.9	41.9	38.7
NF303	420.7	56.9	37.5	35.1

In nature, turquoise killifish populations are highly structured and exhibit high genetic differentiation between populations [8, 9]. Even in populations that are in close proximity, high genetic differentiation can be observed [8, 9]. To find out whether the GNP populations are genetically distinct or genetically similar enough to be treated as one population, I first calculated the genome-wide pairwise genetic differentiation index F_{ST} (see Section 2.1.2). The pairwise F_{ST} values ranged between 0.04 and 0.27 (see Figure 3.2), with low genetic differentiation between both GNP populations (0.04), moderate genetic differentiation between NF414 and NF303 (0.14), strong genetic differentiation between GNP populations and the NF414 population (both 0.21) and very strong genetic differentiation between GNP populations and NF303 (0.27, according to [71]). In general, genetic distance increased with geographic distance (Figure 3.3).

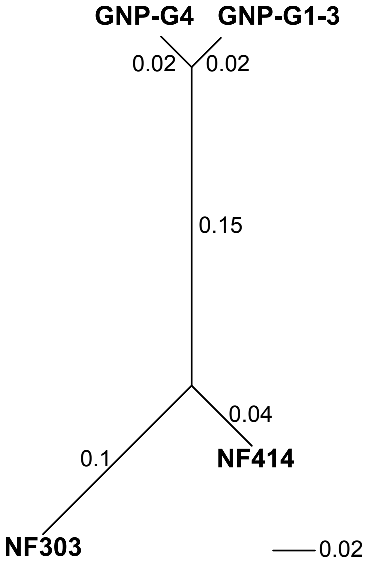

Figure 3.2: Genetic differentiation of all sampled turquoise killifish populations. Unrooted neighbor joining tree based on pairwise F_{ST} values of all sampled populations of the turquoise killifish.

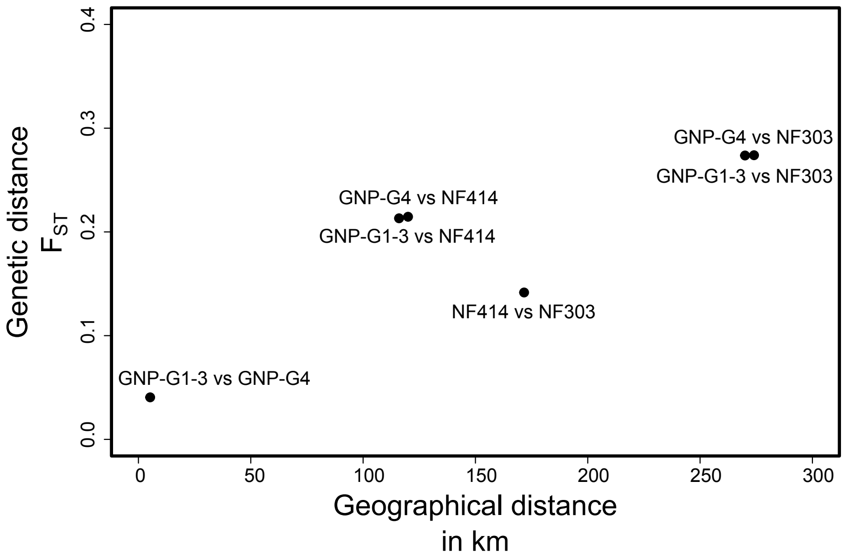

Figure 3.3: Genetic distance and geographical distance of turquoise killifish populations. Shown are the pairwise genetic differentiation values of F_{ST} on the y-axis and the geographical distance on the x-axis.

Based on the low genetic differentiation between the GNP populations, I decided to treat them as one population throughout this study. Therefore, population GNP-G1-3 and population GNP-G4 will be denoted as population GNP. Subsequently, the sequencing files of populations GNP-G1-3 and GNP-G4 were combined via SAMTOOLS (see Chapter 2.1.2). The merged sequencing file had a mean coverage of 47x. To test if the merged population shows the same genetic differentiation to populations NF414 and NF303, I calculated the pairwise F_{ST} again. As shown in Figure 3.4, the genetic differentiation ranged between 0.14 and 0.26 and is highly similar to the previously obtained genetic differentiation between populations (Figure 3.2).

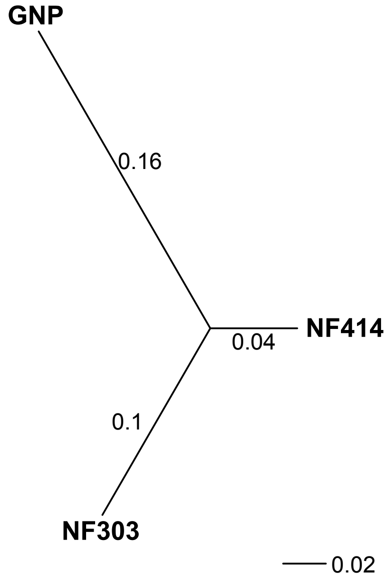

Figure 3.4: Genetic differentiation of turquoise killifish populations. Unrooted neighbor joining tree based on pairwise F_{ST} values for GNP, NF414 and NF303.

In summary, I used three distinct populations (GNP, NF414 and NF303) of the turquoise killifish in this study, which in nature occur along an aridity gradient (see Figure 3.1). F_{ST} analysis revealed that the populations along the gradient show a moderate (NF414 to NF303) to very strong genetic differentiation (GNP to NF303). Further, F_{ST} increased with geographical distance between the populations.

As introduced in Section 1.2.3, extrinsic hazard, such as climate stress, can lead to accelerated aging. In theory, high extrinsic hazard leads to fewer individuals surviving until late age, reducing the force of natural selection for late life traits [23, 24]. In both evolutionary theories of aging, the antagonistic pleiotropy (AP) theory and the mutation accumulation (MA) theory, a difference in the severity of extrinsic hazard between populations can lead to accelerated aging in one of these populations. In the case of the turquoise killifish, it was previously shown that higher aridity correlates with accelerated aging and shorter lifespan [181, 14]. I further hypothesized that the aridity of the environment exerts a limit on population size.

Therefore, I set out to explore whether the differences in aridity between the populations of turquoise killifish led to differences in genetic diversity and effective population size *Ne*.

3.2 Effective population size decreases with higher aridity

The effective population size (*Ne*) determines whether the fate of a mutation is more influenced by natural selection or genetic drift [27]. To test whether differences in *Ne* could play a role in the differences in life history traits of the turquoise killifish populations, I estimated the *Ne* for all populations used in this study. In population genetics, *Ne* is directly correlated to Watterson's estimator of θ (see Section 2.1.2). In order to obtain the genome-wide average of θ, I used the program `PoPoolation` [95]. Watterson's estimator of θ was calculated based on sliding windows with a length of 50 kb. This resulted in θ estimates of 0.0011 (GNP), 0.0036 (NF414), and 0.0072 (NF303, Table 3.2). Next, based on a mutational rate for the turquoise killifish of $2.6321e^{-9}$ per base pair and generation [38], I calculated *Ne* (see Section 2.1.2). Finally, I obtained estimates of *Ne* ranging between approximately 107,000 and 684,000 (Table 3.2). Population GNP had the smallest *Ne*, followed by NF414 with intermediate *Ne* and NF303 with the largest *Ne* (see Figure 3.5). Therefore, the analysis showed that the effective population size correlates negatively with the aridity of the environment.

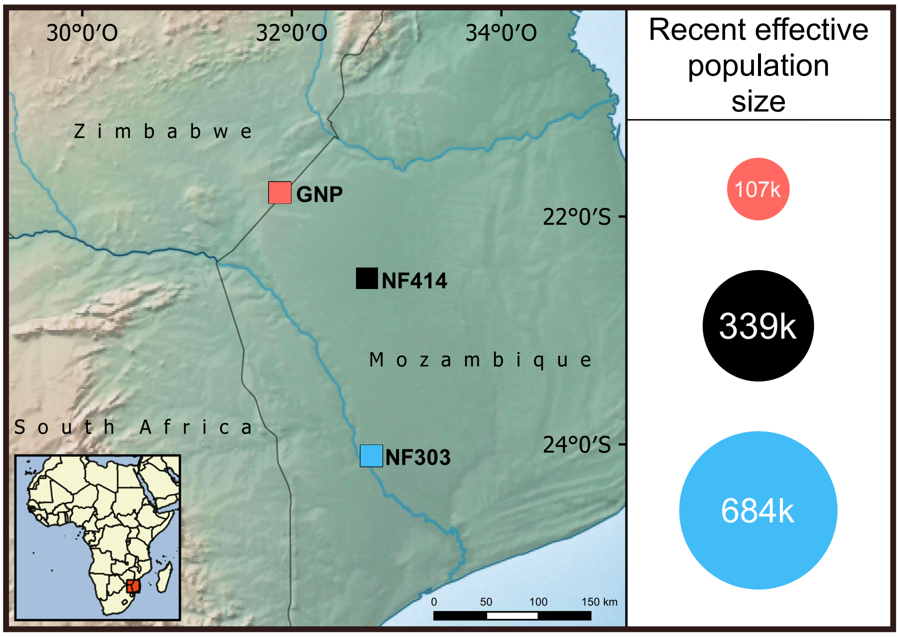

Figure 3.5: Estimates of current effective population size *Ne*. Geographical location of the turquoise killifish populations GNP, NF414 and NF303 (left panel) and estimated effective population size *Ne* (right panel). *Ne* was estimated with Wattersons's estimator of theta (θ). Area of each circle correlates to *Ne* of each population. Map was made with QGIS version 2.18.20 [143] combined with GRASS version 7.4 [126] and data from Natural Earth.

Changes in population size can be a result of demographic events. Therefore, the observed differences in *Ne* between the populations could be caused by contrasting demography. To get a closer look at population demography, I compared two measures of genetic diversity: the expected genetic diversity and the observed genetic diversity. Watterson's estimator of θ is considered as the expected genetic diversity and the nucleotide diversity π is considered as the observed genetic diversity (see Section 2.1.2).

According to the calculation of θ, I calculated π with `PoPoolation` [95] and obtained genome-wide averages of 0.0009, 0.0031 and 0.0054 for GNP, NF414 and NF303, respectively (Table 3.2). In agreement with the low *Ne* in GNP, the distribution of θ and π were L-shaped, due to the low genetic diversity found in this population (Figures 3.6, 3.7). In contrast, NF414 and NF303 showed normally distributed θ and π values (Figures 3.6, 3.7).

Interestingly, I found that in all three populations the observed genetic diversity π was lower than the expected genetic diversity θ (Table 3.2). Under neutrality, the difference between π and θ is expected to be zero, but demography or selection can cause a departure from neutrality. To test if the populations deviate from neutrality, I calculated Tajima's D (see Section 2.1.2). The genome-wide values for Tajima's D were -0.47, -0.47 and -0.97 for GNP, NF414 and NF303, respectively (Table 3.2). All populations showed a significant deviation from zero (t-test, P-value<0.001, Figure 3.8). Moreover, GNP showed the highest standard deviation in Tajima's D with 0.99 compared to NF414 with 0.37 and NF303 with 0.40 (Table 3.2).

A negative genome-wide Tajima's D can indicate a recent expansion after a bottleneck [13, 177]. Therefore, in the next section, I investigated the demographic history of the populations in more detail using individual resequencing.

Table 3.2: Genome-wide estimates of π, θ and Tajima's D

Population	π	θ	***Ne***	**Tajima's D**	**sd Tajima's D**
GNP	0.0009	0.0011	107221.8	-0.47	0.99
NF414	0.0031	0.0036	338849.48	-0.47	0.37
NF303	0.0054	0.0072	683693.25	-0.97	0.40

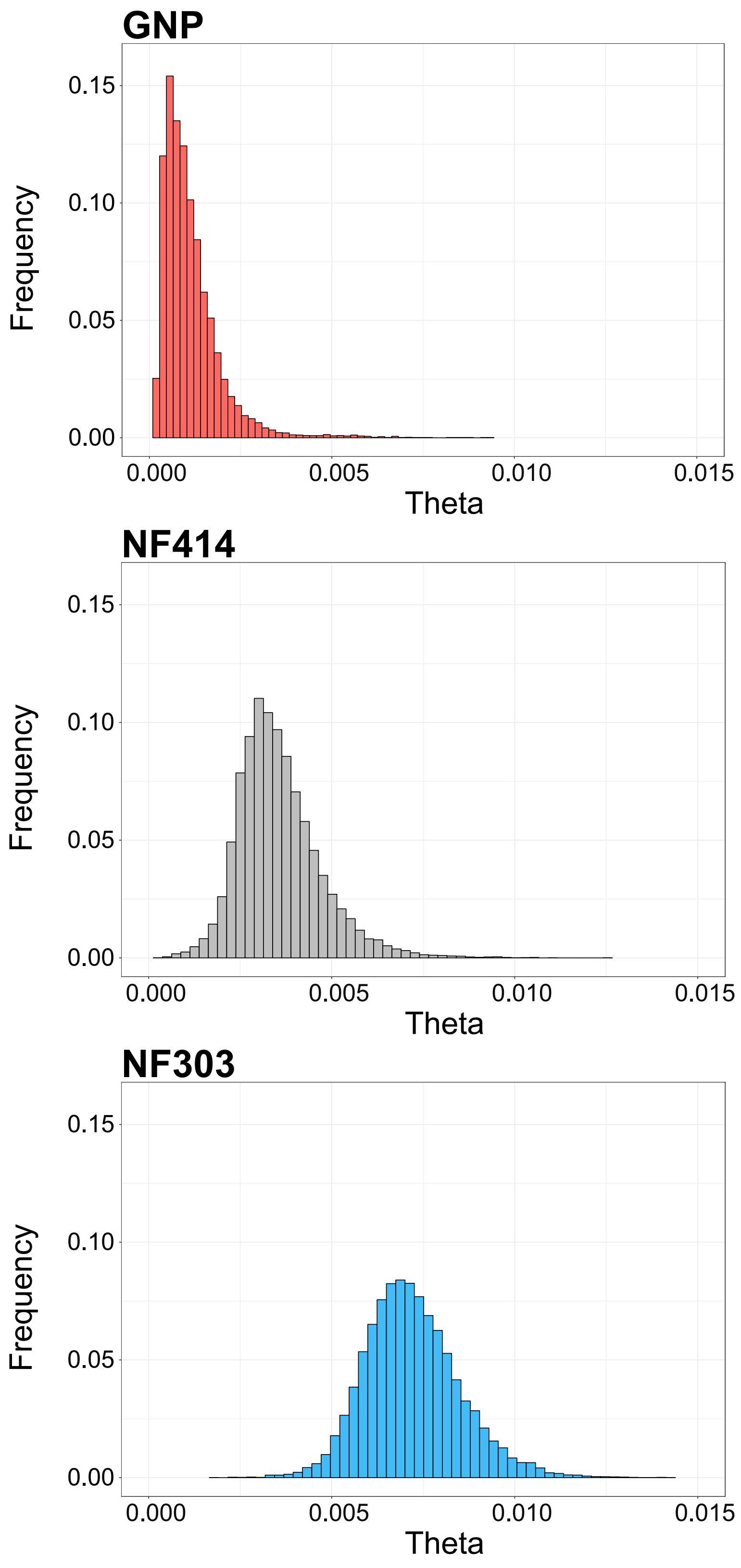

Figure 3.6: Histogram of Watterson's estimator of theta θ. Shown are the distributions of Watterson's estimator of theta per population (based on 50 kb sliding windows). Shown are the populations GNP, NF414 and NF303 in the top, middle and bottom panel. Frequency is shown on the y-axis and the value of θ on the x-axis.

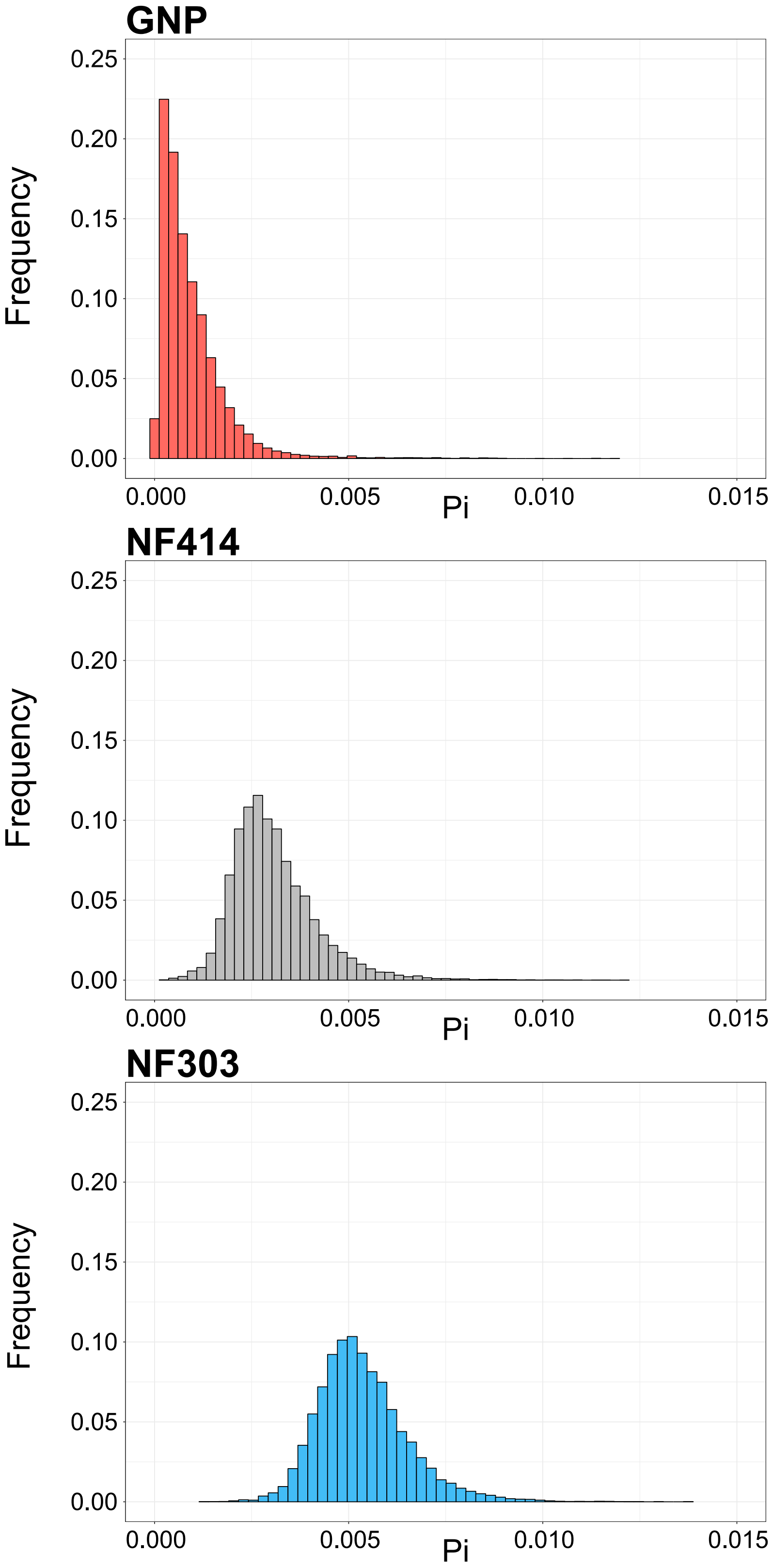

Figure 3.7: Nucleotide diversity π. Histogram of nucleotide diversity π per population (based on 50 kb sliding windows). Shown are the populations GNP, NF414 and NF303 in the top, middle and bottom panel. Frequency is shown on the y-axis and the value of π on the x-axis.

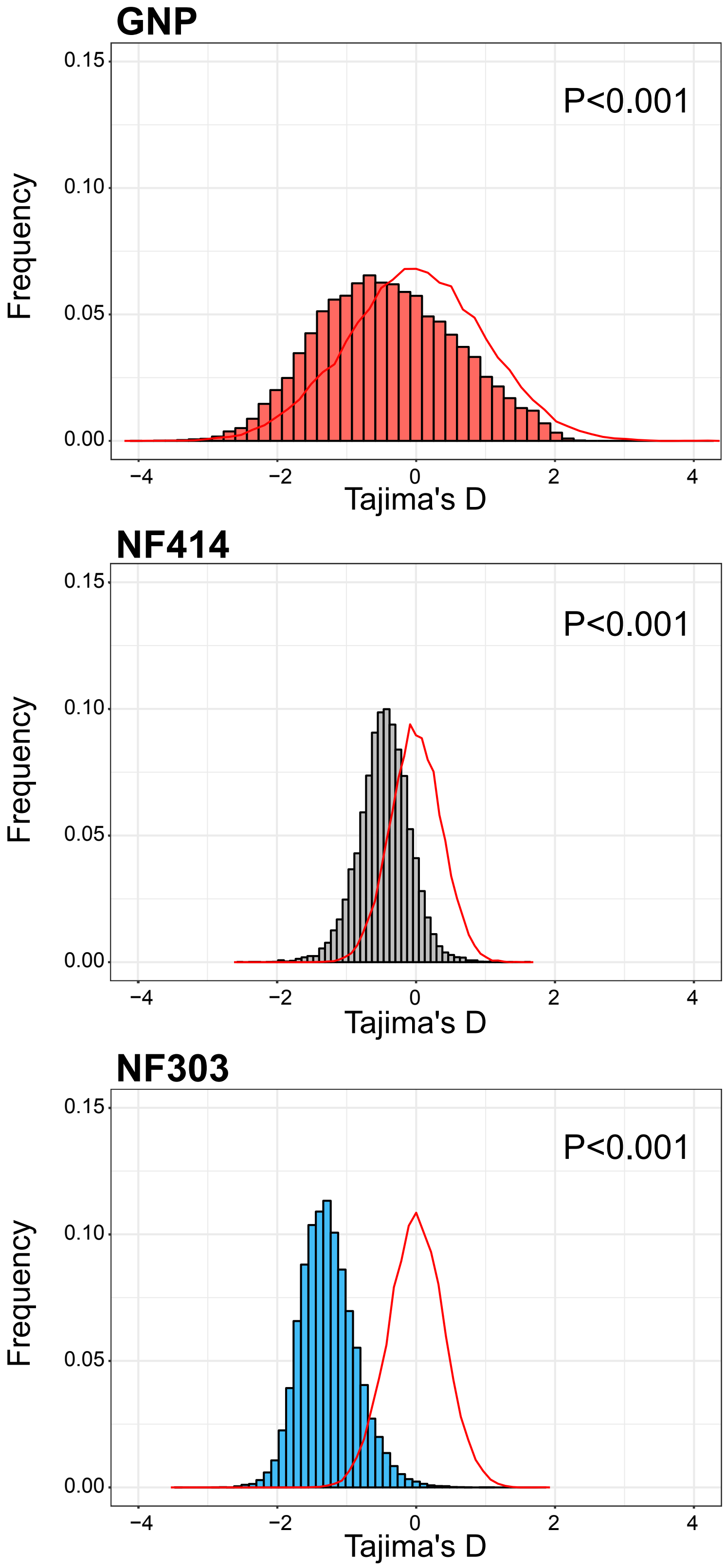

Figure 3.8: Tajima's D. Histogram of observed Tajima's D values per population (based on 50 kb sliding windows). Shown are the populations GNP, NF414 and NF303 in the top, middle and bottom panel. Frequency is shown on the y-axis and the value of Tajima's D on the x-axis. The expected normal distribution consists of 18,241 values with an average of zero and the same standard deviation as the observed dataset (red solid line). *P*-values indicate whether there was a significant deviation of the average from zero using a t-test.

3.3 Contrasting population history in wet and dry populations

To explore the demographic history of the populations, I performed single individual resequencing and population history inference with `PSMC'` [165]. For this analysis I used one individual from NF303 and two individuals each from the NF414 and GNP populations. Two individuals were taken for population NF414 because this population is polymorphic with respect to male tail color. Therefore, to avoid overestimation due to possible admixture, I sequenced one individual per tail color morph, red (NF414-R) and yellow (NF414-Y). In the case of the population GNP, two individuals were taken to verify that the two sampling locations have the same population history. The resequencing of single individuals resulted in a mean coverage between 13.2-21.5 (Table 3.3).

Table 3.3: Mean coverage of individual resequencing

Population	**Species**	**ID**	**Mean coverage**
GNP-G1-3	*N. furzeri*	M1	20.8
GNP-G4	*N. furzeri*	M9	13.2
NF414	*N. furzeri*	R3	16.6
NF414	*N. furzeri*	Y1	21.5
NF303	*N. furzeri*	M7	18.0

Population history inference is based on heterozygosity and can be biased by miscalled heterozygosity. Therefore, the inference was done only on sites that passed a coverage filter (see Section 2.1.2). In general, sites that are below half or above double the mean coverage per chromosome were discared. To ensure optimal inference, I performed this analysis in two different ways. First, with the specific mean coverage per chromosome (see Figure A.1) and second, with a set threshold for coverage of 18x (as recommended in [124]).

Inference of population history revealed that the population with the largest current *Ne*, NF303, underwent a population expansion. The expansion started between 200k and 500k generations ago (lightblue line, Figure 3.9). This expansion could have been in the shared ancestral population of all populations used in this study, because the shape of the inferred ancestral effective population size is equal in all populations. Further, the expansion continued in NF303 from 150k generations ago until approximately 75k generations ago (lightblue line, Figure 3.9). This was followed by a minor reduction in ancestral *Ne* in NF303. The same trend is observable for NF414 (grey and black line). The estimates for more recent time points (<100k) are in general more uncertain and show a high variance in bootstrap estimates (dotted lines).

In contrast, population GNP showed the same increase of the putative shared ancestral population, but then continuously decreased in *Ne* until recent time (red and orange line, Figure 3.9). This is consistent with the low current genetic diversity of this population (Table 3.2). The inference of the two individuals from population GNP were highly similar (individuals from sampling site GNP-G1-3 and GNP-G4).

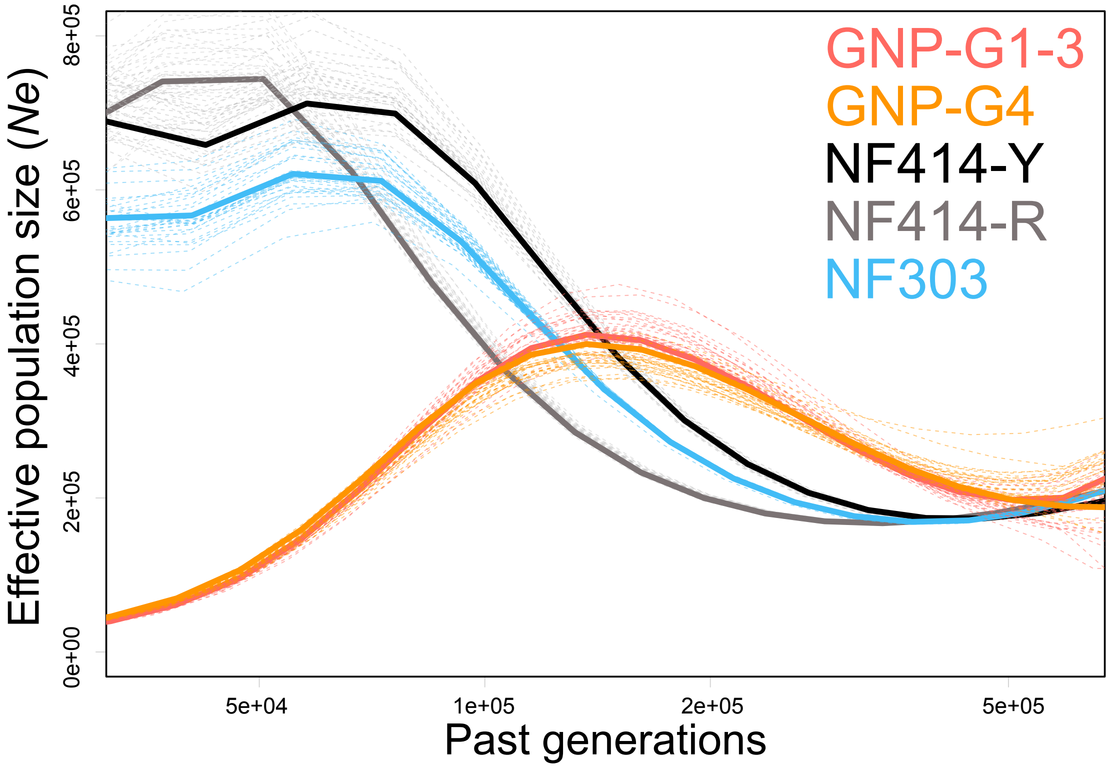

Figure 3.9: Ancestral effective population size *Ne*. Estimated ancestral effective population size *Ne* based on `PSMC'` analysis. Shown is time in past generations on the x-axis and the ancestral *Ne* on the y-axis. Inference of population NF303 in lightblue and population GNP in red and orange. For population NF414 the ancestral *Ne* was inferred based on a red-tailed male (grey) and a yellow-tailed male (black).

Demographic events, such as population expansion and population bottlenecks should have a genome-wide impact on the distribution of allele frequencies. A recent expansion causes an increase in rare variants in a population, whereas population that underwent a bottleneck should be more susceptible to genetic drift [27]. To explore the distribution of allele frequencies, I calculated the site frequency spectra (SFS) for minor allele frequencies (MAF) of neutral, synonymous and non-synonymous sites in the genome. For all populations and all SFS, the proportion of single-nucleotide polymorphisms (SNPs) decreased at higher MAF bins (Figures 3.10, 3.11, 3.12). Further, NF303 showed an increased proportion of rare variants up to a frequency of 0.2 in all SFS compared to population NF414 and GNP. This is in line with the population expansion seen in Figure 3.9. In contrast, I observed that more SNPs reached higher frequencies (0.2-0.5 MAF) in GNP compared to NF414 and NF303, potentially indicating genetic drift. However, in all populations the proportion of SNPs at higher frequencies was higher in synonymous sites compared to non-synonymous sites (Figures 3.11, 3.12). Examination of the recent and ancestral *Ne* in populations associated with different lifespans and aging rate, revealed that the populations were subject to contrasting demographic histories. The analysis showed that the population from the driest region, GNP, constantly decreased in *Ne*, whereas populations NF414 and NF303 from less arid regions recently expanded. In summary, genetic diversity was inversely correlated to aridity of the environment.

As mentioned previously, differences in *Ne* can impact the efficacy of natural selection or genetic drift [27]. Therefore, I wanted to check whether contrasting population histories led to signatures of genetic differentiation that could potentially be causal for the differences in life history traits. Under the antagonistic pleiotropy theory (AP) of aging, a mutation with beneficial effects early in life that exceed the detrimental effects later in life should be selected by natural selection through positive selection. In addition, a region that has been positively selected in one population but not in another should be detectable by genetic differentiation analysis. I hypothesized that under the AP theory of aging these regions should have a clear signal of high F_{ST} above the genome-average of each pairwise comparison between populations of the turquoise killifish. In particular, I set out to examine whether the populations are genetically different in regions known to be associated with lifespan in this species [185]. Further, I hypothesized that if AP is responsible for the differences in life history traits, the local signal of genetic differentiation should be strongest between the dry and wet population and hence between GNP and NF303.

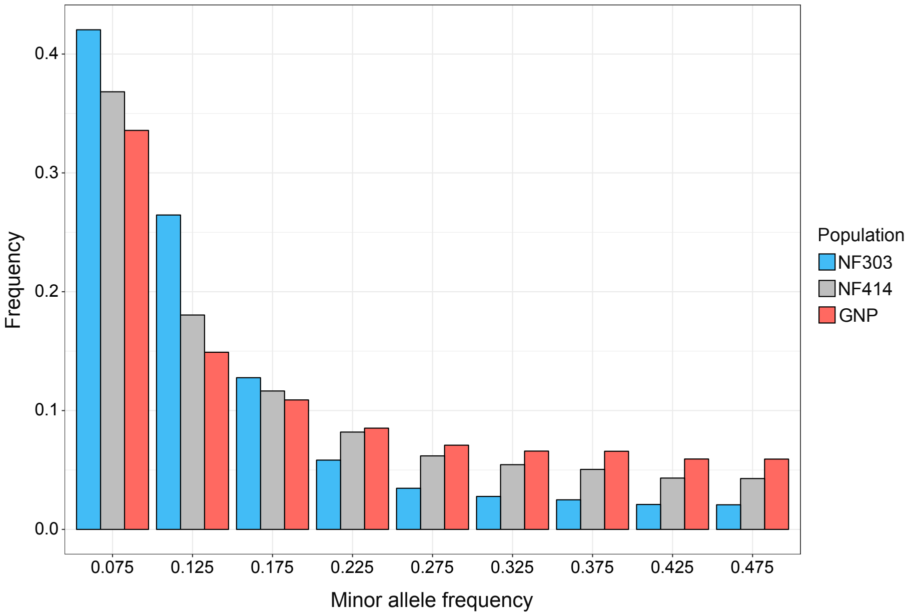

Figure 3.10: Neutral site frequency spectrum. Barplot of the site frequency spectra of non-coding sites. Shown are the populations GNP, NF414 and NF303 in red, grey and blue color. Frequency is shown on the y-axis and binned minor allele frequency on the x-axis.

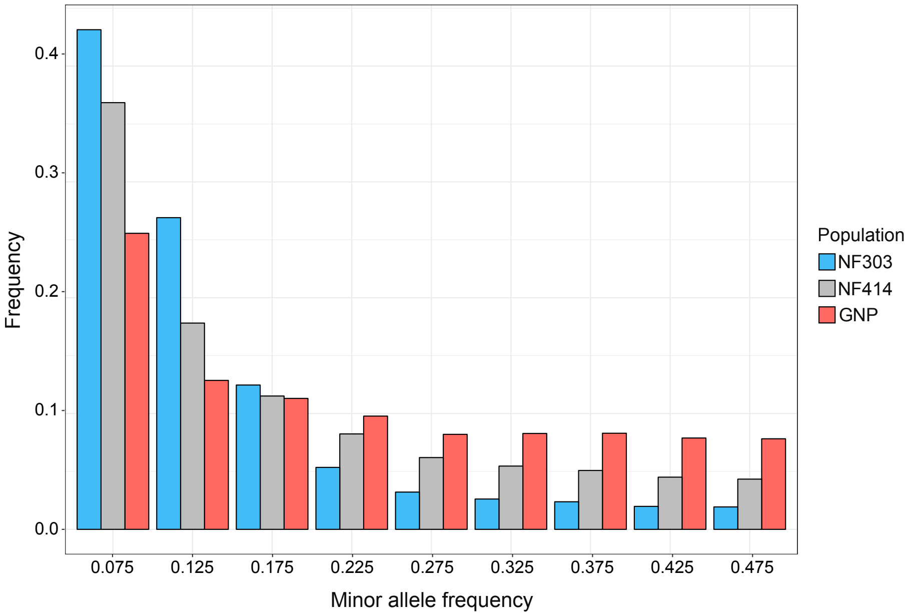

Figure 3.11: Synonymous site frequency spectrum. Barplot of the site frequency spectra of synonymous coding sites. Shown are the populations GNP, NF414 and NF303 in red, grey and blue color. Frequency is shown on the y-axis and binned minor allele frequency on the x-axis.

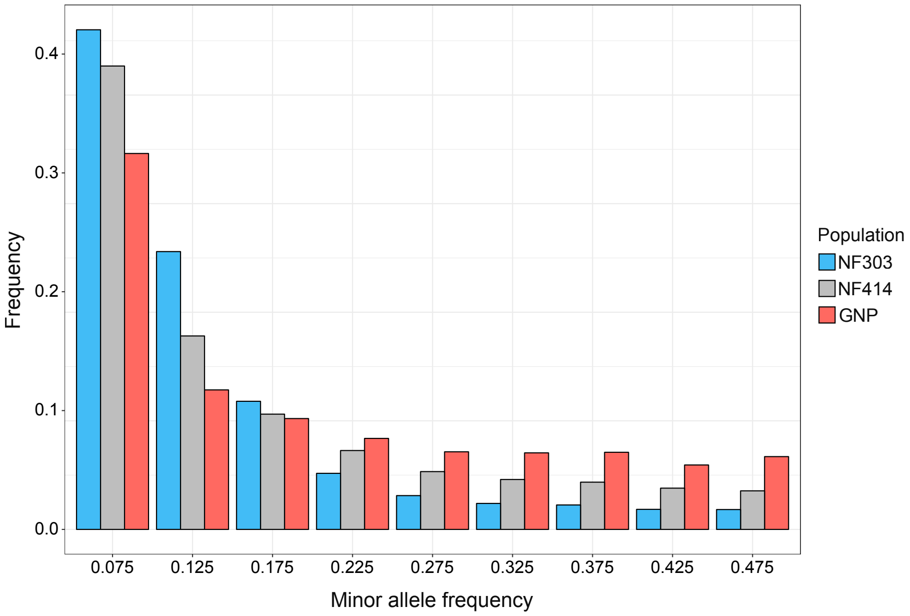

Figure 3.12: Non-synonymous site frequency spectrum. Barplot of the site frequency spectra of non-synonymous coding sites. Shown are the populations GNP, NF414 and NF303 in red, grey and blue color. Frequency is shown on the y-axis and binned minor allele frequency on the x-axis.

3.4 Identifiying genomic regions associated with lifespan

3.4.1 Recent evolutionary events in the region associated with lifespan

To test whether genetic differentiation in loci that are known to be associated with lifespan could be responsible for the lifespan difference between populations along the aridity gradient, I identified the genomic locations of the quantitative trait loci (QTL) markers published in Valenzano *et al.* [185]. For this, I used BLAST [2] to map the lifespan and sex QTL markers to the new genome version of the turquoise killifish (NFZ V2.0). The strongest association with lifespan was identified on the sex chromosome (Chromosome 3, Figure 3.13), as reported in previous findings [185]. In total, based on a significance threshold of -log(q-value)>1.5, I identified four main clusters of lifespan located on chromosomes 3, 5, 6 and 14 (Figure 3.13).

Since the strongest association with lifespan is with the sex chromosome and sex chromosomes in teleost fishes are highly dynamic [114], I performed synteny analysis between the new genome assembly of the turquoise killifish (NFZ V2.0) and two closely related teleost species, platyfish (*Xiphophorus maculatus*) and medaka (*Oryzias latipes*). I used 1-to-1 orthologous gene positions to identify chromosomes in synteny between the turquoise killifish and each of the other species. In both comparisons, the sex chromosome of the new genome NFZ V2.0 assembly showed signatures of two distinct evolutionary events (Figure 3.14). The synteny analysis showed a fusion event of two ancestral autosomal chromosomes, combined with a translocation event in one of these ancestral autosomes. Figure 3.14 shows that the sex chromosome of the turquoise killifish is in synteny with two autosomal chromosome of each related species. In platyfish, LG 16 corresponds to the first part of the sex chromosome of the turquoise killifish and LG 3 corresponds to the second part of it (Figure 3.14A). Further, synteny to LG 3 splits in two linear blocks, suggestive of a translocation event (Figure 3.14A). Moreover, the strongest association with sex is located near the translocation point of LG

3, in proximity to the gene *gdf6* on the sex chromosome of the turquoise killifish, which is suggested to be the sex-determining gene in the turquoise killifish [150]. The strongest association with lifespan is located on the first part of the turquoise killifish chromosomes, which corresponds to LG 16 of platyfish.

In medaka, chromosome 8 corresponds to the first part of the sex chromosome, whereas chromosome 16 corresponds to the second part of the sex chromosome. Analogously, a translocation of the two blocks of chromosome 16 occurred in proximity to the putative sex-determining region. Importantly, the two regions on medaka and platyfish are syntenic with each another (see Figure B.1).

The synteny alignment revealed that the sex chromosome of the turquoise killifish evolved through two different evolutionary events. Figure 3.14B shows the putative chronological order of the two evolutionary events based on the lowest level of complexity. I postulate that previous to the fusion event of the ancestral autosomes, the translocation within one of the ancestral autosomes occurred. Valenzano and colleagues showed that the lifespan QTL is distinct from the sex QTL, but within the region of suppressed recombination [185]. This is in line with the clear distinction between lifespan QTL and sex QTL in the conducted synteny analysis. The lifespan QTL is located on the part of the sex chromosome that is derived from a different ancestral chromosome than the sex QTL (Figure 3.14A). Next, I looked at genetic differentiation in regions associated with lifespan to see if I can detect signatures of selection.

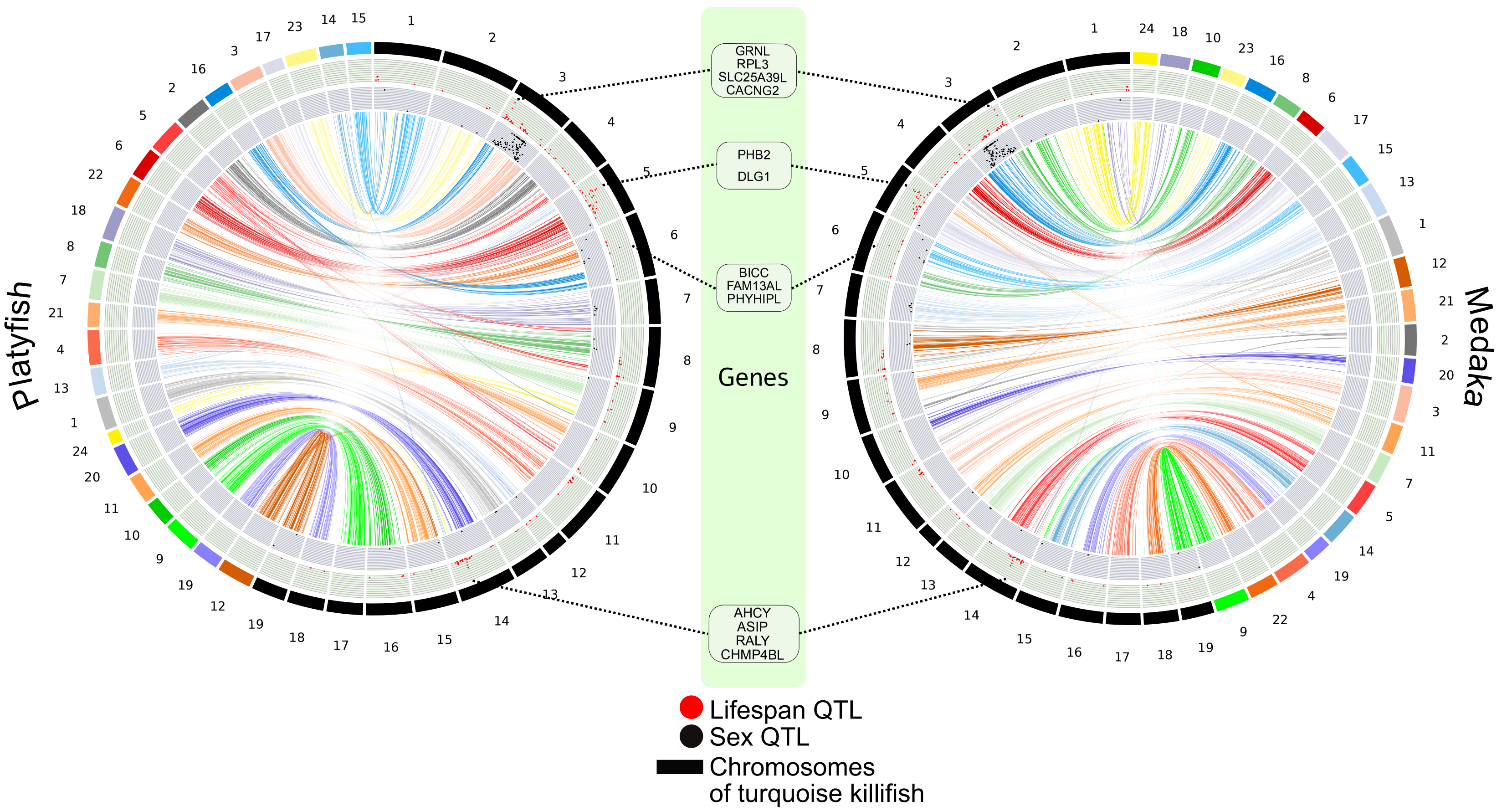

Figure 3.13: Synteny. `Circos` plot of syntenic regions between the turquoise killifish and platyfish (left `Circos` plot) and turquoise killifish and medaka (right `Circos` plot). The ribbons show syntenic regions between the chromosomes of each comparisons. Ribbon colors correspond to chromosomes coming from either platyfish or medaka. Chromosomes of the turquoise killifish are colored in black. Shown are -log(q-values) of the quantitative trait loci markers for sex (middle layer) and lifespan (outer layer). Genes within regions of highest association with lifespan QTL are shown in the green box.

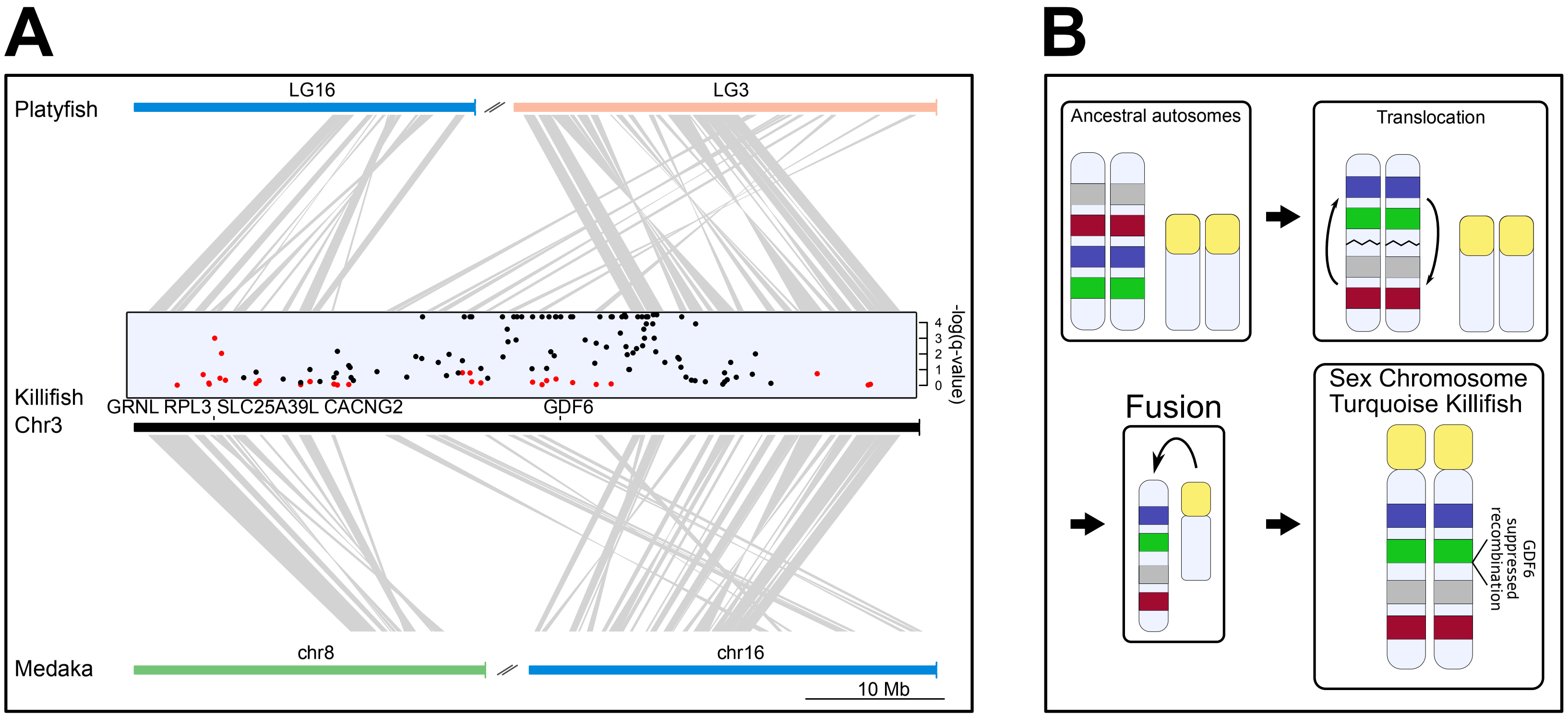

Figure 3.14: Synteny of sex chromosome. A) Synteny between turquoise killifish sex chromosome (Chromosome 3, black) and the corresponding chromosomes of platyfish (top, LG16 & LG3) or the corresponding chromosomes of medaka (bottom, chr8 & chr16). The dot plot shows the -log(q-values) of the quantitative trait loci markers for sex (black dots) and lifespan (red dots). B) Model of sex chromosome evolution in the turquoise killifish. Previous to the fusion of two ancestral chromosomes, one autosomal chromosome underwent a translocation event. The putative sex-determining gene *gdf6* [150] is located in proximity to the fusion site of the translocated parts of one ancestral chromosome. The sex chromosome of the turquoise killifish has a region of suppressed recombination [185].

3.4.2 Genetic differentiation in QTL regions

Next, I set out to explore whether the regions associated with lifespan in turquoise killifish are responsible for the lifespan differences between natural populations of the turquoise killifish. Therefore, I searched for signatures of local selection based on genetic differentiation. If the lifespan-associated regions in the populations are subject to different selection regimens, they should show a higher pairwise genetic differentiation locally than the genome-wide background. I computed the pairwise genetic differentiation index F_{ST} for non-overlapping sliding windows with a length of 50 kb. Significance was determined via an outlier approach based on the genome-wide highest differentiation above 99.5^{th} percentile in each pairwise comparison. To visualize the different pairwise comparisons in the same scale, F_{ST} windows were Z-transformed ($Z_{F_{ST}}$). As shown in Figure 3.15, the $Z_{F_{ST}}$ values of each pairwise comparison in regions harboring QTL markers with association to lifespan did not reach the genome-wide significance threshold. Thus, I could not find signatures for local selection in regions containing lifespan QTLs.

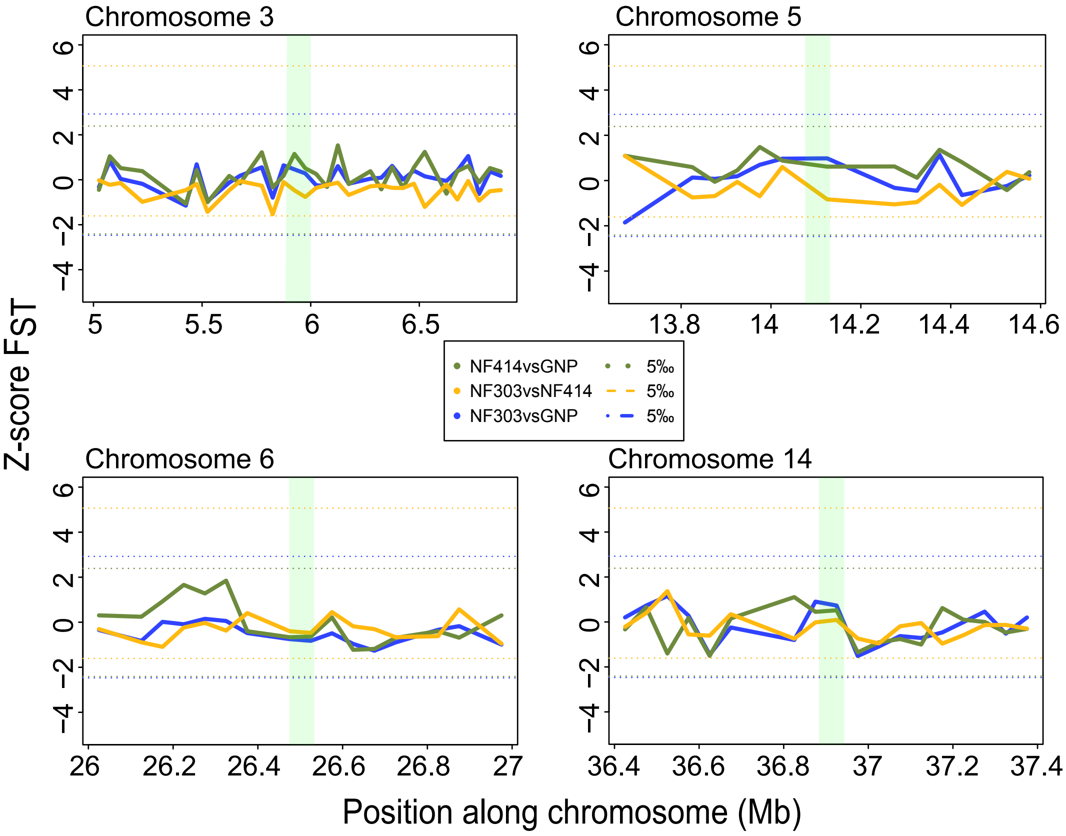

Figure 3.15: Genetic differentiation in QTL regions. Z-transformed pairwise genetic difference F_{ST} of all pairwise comparisons between the populations of the turquoise killifish in QTL regions associated with lifespan (QTL region highlighted with lightgreen background). The Z-transformed F_{ST} values are shown in solid lines and the 0.5^{th} or 99.5^{th} percentile thresholds of F_{ST} in dotted lines with NF303 vs. GNP in blue, NF414 vs. GNP in green, and NF303 vs. NF414 in yellow.

3.5 Genome-wide signatures of local selection

3.5.1 Regions of low genetic differentiation

The genetic differentiation in regions associated with lifespan showed no significant differences between the populations. To find regions under local selection, I searched for regions across the genome that had either low genetic differentiation (<0.5^{th} percentile) or high genetic differentiation (>99.5^{th} percentile, Figure 3.16) per pairwise comparison (Figure 3.16). Regions important for the species, and thus under strong purifying selection, should exhibit low genetic divergence between all investigated populations. Hence, I identified these regions by overlapping all windows below the 0.5^{th} percentile of $Z_{F_{ST}}$ of each pairwise comparison.

Based on this method I identified two regions that showed low genetic differentiation between all populations (see Table 3.4). The first region is located on chromosome 3, the sex chromosome of this species (see Chapter 3.4.1) and contains two genes, *gdf6* and *sybu* (Figure 3.17). As mentioned in the previous section, *gdf6* is reported to be the sex-determining gene in the turquoise killifish [150]. The second region is located on chromosome 9 and contains three genes, *LOC107389593*, *cnot11* and *lcp1* (Figure 3.18). In addition, I identified more regions of low genetic differentiation shared between all pairwise comparisons by using a $Z_{F_{ST}}$ threshold of 1% . Table 3.5 shows the list of genes within these regions.

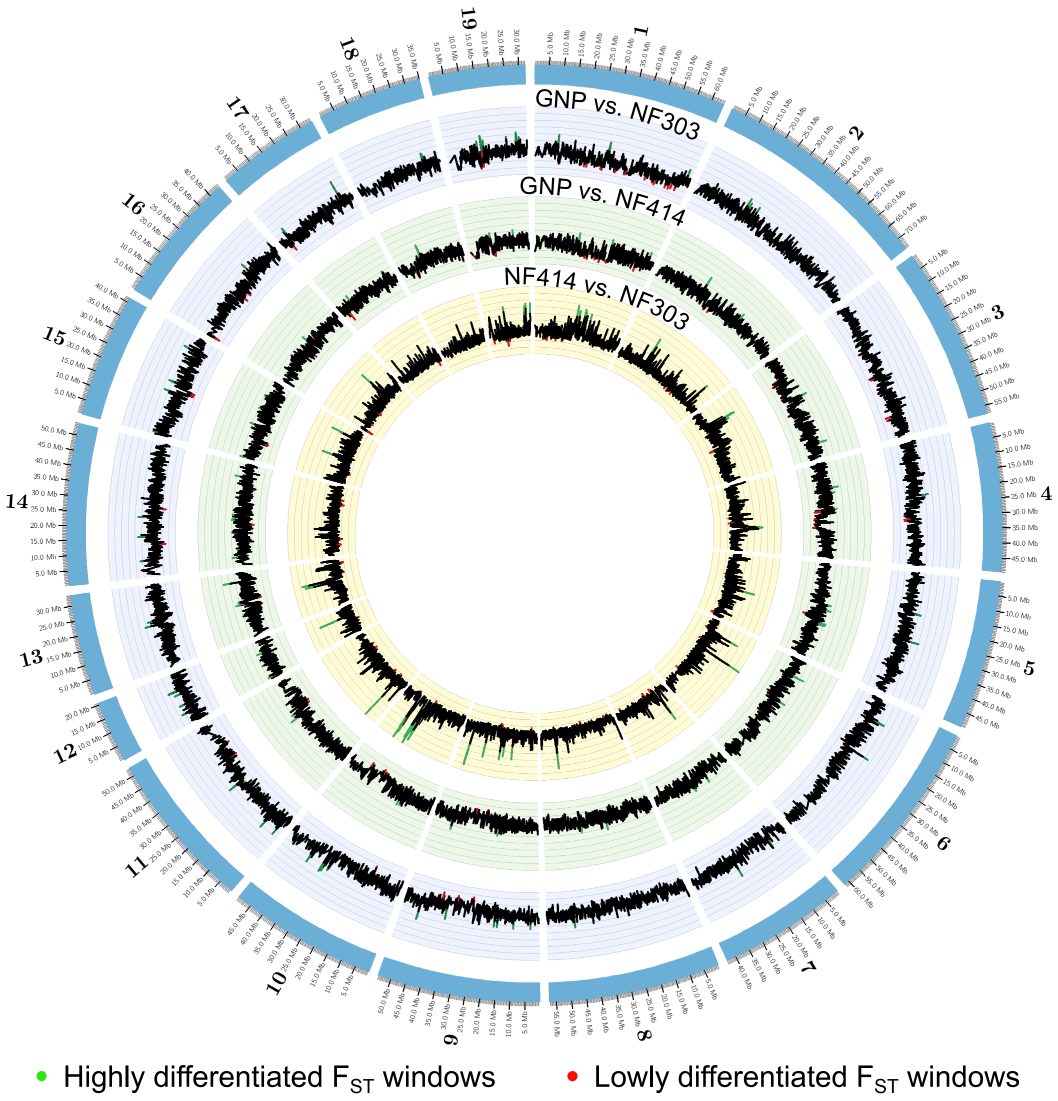

Figure 3.16: `Circos` **plot of genome-wide** $Z_{F_{ST}}$**.** Shown are the genome-wide Z-transformed pairwise genetic differentiation F_{ST} ($Z_{F_{ST}}$) values of all pairwise comparisons between the populations of the turquoise killifish. The 19 chromosomes of the turquoise killifish genome are shown clock-wise with the pairwise comparisons of GNP vs. NF303 (outer most circle, blue background), GNP vs. NF414 (middle circle, green background) and NF414 vs. NF303 (inner most circle, yellow background). $Z_{F_{ST}}$ values are based on 50 kb windows. Highly differentiated windows (>99.5th percentile) are labelled in green and lowly differentiated windows (< 0.5th percentile) are labelled in red.

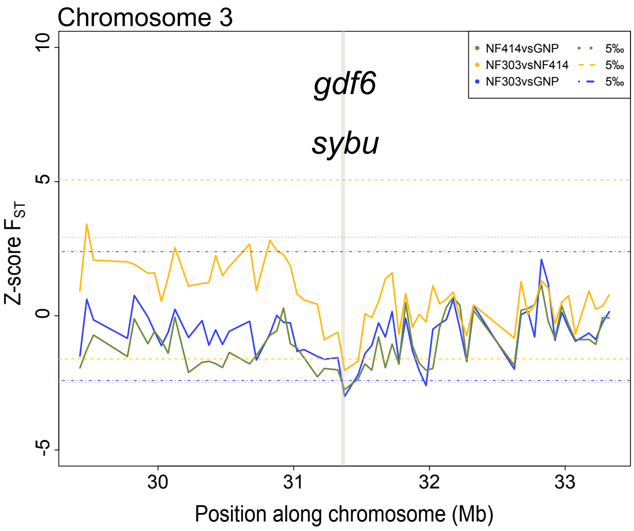

Figure 3.17: $Z_{F_{ST}}$ **on chromosome 3.** Z-transformed pairwise genetic differentiation F_{ST} ($Z_{F_{ST}}$) of all pairwise comparisons between the populations of the turquoise killifish along chromosome 3. Shown are the $Z_{F_{ST}}$ values in solid lines and the 0.5^{th} or 99.5^{th} percentile thresholds of $Z_{F_{ST}}$ in dotted lines with NF303 vs. GNP in blue, NF414 vs. GNP in green, and NF303 vs. NF414 in yellow.

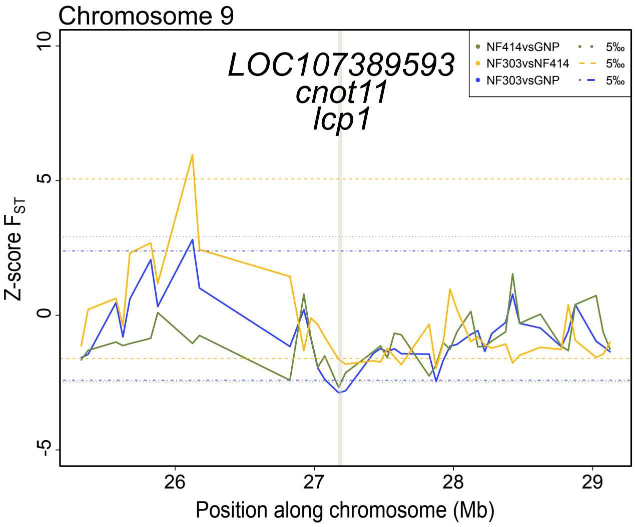

Figure 3.18: $Z_{F_{ST}}$ **on chromosome 9.** Z-transformed pairwise genetic differentiation F_{ST} ($Z_{F_{ST}}$) of all pairwise comparisons between the populations of the turquoise killifish along chromosome 9. Shown are the $Z_{F_{ST}}$ values in solid lines and the 0.5^{th} or 99.5^{th} percentile thresholds of $Z_{F_{ST}}$ in dotted lines with NF303 vs. GNP in blue, NF414 vs. GNP in green, and NF303 vs. NF414 in yellow.

3.5.2 Regions of high genetic differentiation

Next, to identify candidate genes that are under different selection regimes in the turquoise killifish populations, I identified 50 kb windows of high genetic differentiation (>99.5th percentile of each comparison). In total, 63 windows per pairwise comparison were found to fall into this category. To further narrow down the list of candidate genes, I used a total of three selection criteria. First, the window-based $Z_{F_{ST}}$ value had to be above the 99.5th percentile of genetic differentiation. Second, the gene F_{ST} value had to be above the 99.5th percentile of genetic differentiation and last, the gene needed to include at least one SNP with significant differentiation based on Fisher's exact test (based on `PoPoolation2`).

This approach enabled me to only find genes with strong genetic differentiation between the populations and potentially reduce the number of false positive hits. Although GNP showed high genome-wide genetic differentiation to NF303 and NF414 (Figure 3.4), I only identified two genes fulfilling the selection criteria in comparison to NF303 (Table 3.6) and three genes fulfilling the selection criteria in comparison to NF414 (Table 3.7). Unexpectedly, the comparison between NF414 and NF303 resulted in more genes with high genetic differentiation. This was inverse to genome-wide genetic differentiation seen between all populations (Figure 3.4). In total, I identified 18 genes that are located on 9 different chromosomes (Table 3.8). Chromosome 10 contains eight genes that showed high genetic differentiation between NF414 and NF303. The two regions with highest genetic differentiation between population NF414 and NF303 are located on chromosome 6 and chromosome 10. Moreover, both regions showed high genetic differentiation between GNP and NF303 based on the 50 kb window $Z_{F_{ST}}$ value. The region on chromosome 6 contains the gene *slc8a1* (Figure 3.19).

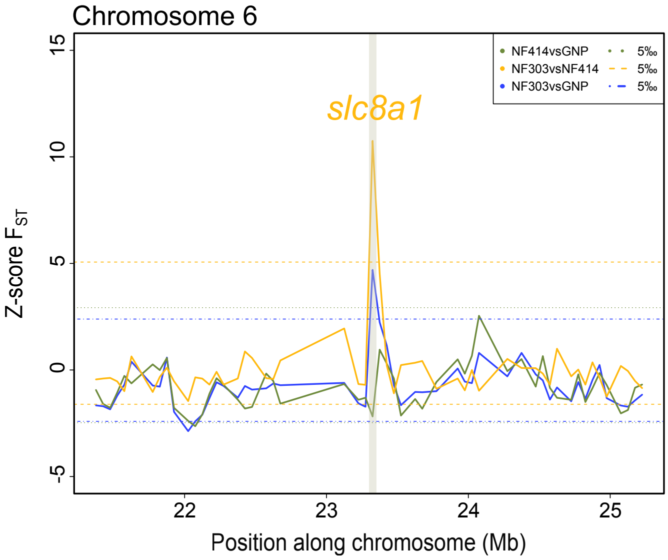

Figure 3.19: $Z_{F_{ST}}$ **on chromosome 6.** Z-transformed pairwise genetic differentiation F_{ST} ($Z_{F_{ST}}$) of all pairwise comparisons between the populations of the turquoise killifish along chromosome 6. Shown are the $Z_{F_{ST}}$ values in solid lines and the 0.5^{th} or 99.5^{th} percentile thresholds of $Z_{F_{ST}}$ in dotted lines with NF303 vs. GNP in blue, NF414 vs. GNP in green, and NF303 vs. NF414 in yellow. Gene names are colored in the color of the comparison they were found to be significant.

The region on chromosome 10 contains four genes that were highly differentiated between NF303 and NF414, including the gene *hibch* that was also highly differentiated between NF303 and GNP (Figure 3.20). Together, the F_{ST}-based outlier approach enabled me to find regions under purifying selection in all populations of the turquoise killifish and under local intensified selection between the populations of the turquoise killifish. To gain insight into which biological processes show either genetic convergence or genetic differentiation between the populations of the turquoise killifish, I performed Gene Ontology (GO) overrepresentation analysis based on gene-wise F_{ST} values.

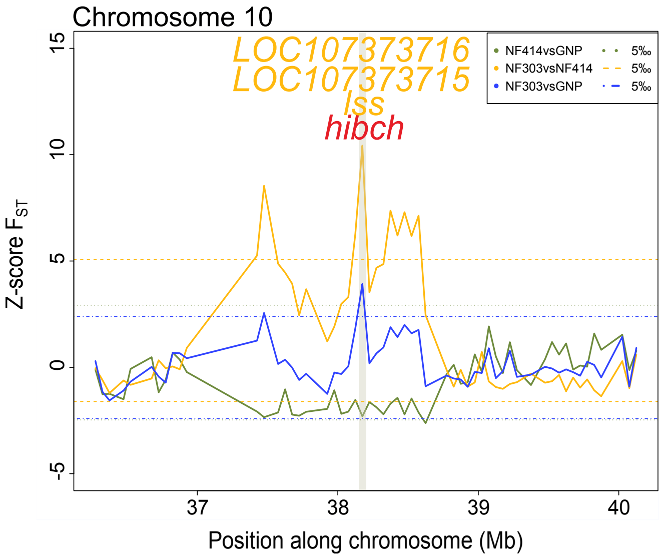

Figure 3.20: $Z_{F_{ST}}$ on chromosome 10. Z-transformed pairwise genetic differentiation F_{ST} ($Z_{F_{ST}}$) of all pairwise comparisons between the populations of the turquoise killifish along chromosome 10. Shown are the $Z_{F_{ST}}$ values in solid lines and the 0.5^{th} or 99.5^{th} percentile thresholds of $Z_{F_{ST}}$ in dotted lines with NF303 vs. GNP in blue, NF414 vs. GNP in green, and NF303 vs. NF414 in yellow .Gene names are colored in the color of the comparison they were found to be significant. Gene names colored in red reached significance in comparisons of NF303 vs. GNP and NF303 vs. NF414.

Table 3.4: Summary of genes in lowly differentiated windows (0.5th percentile of F_{ST})

Chromosome	GeneID	Name	Description	Human Ensembl ID
chr3	XM_015949386	*gdf6*	growth differentiation factor 6	
chr3	XM_015949383	*sybu*	syntabulin	
chr9	XM_015965805	*lcp1*	lymphocyte cytosolic protein 1 (L-plastin)	ENSG00000136167
chr9	XM_015965807	*cnot11*	CCR4-NOT transcription complex subunit 11	ENSG00000158435
chr9	XM_015965812	*LOC107389593*	abl interactor 2-like	

Table 3.5: Summary of genes in lowly differentiated windows (1^{st} percentile of F_{ST})

Chromosome	GeneID	Name	Description	Human Ensembl ID
chr1	XM_015940864	*LOC107372657*	butyrophilin subfamily 2 member A2-like	
chr1	XM_015940862	*LOC107372655*	uncharacterized LOC107372655	
chr14	XM_015957072	*LOC107384047*	sodium channel protein type 8 subunit alpha-like	
chr14	XM_015957075	*LOC107384048*	fidgetin	
chr19	XM_015972257	*hectd4*	HECT domain E3 ubiquitin protein ligase 4	ENSG00000173064
chr6	XM_015944666	*LOC107375860*	arylsulfatase I-like	
chr6	XM_015944667	*LOC107375861*	synaptopodin 2-like protein	ENSG00000166317
chr7	XM_015950935	*LOC107379975*	protein FAM222B-like	ENSG00000173065
chr7	XM_015950936	*LOC107379976*	fructose-bisphosphate aldolase C-like	ENSG00000109107
chr7	XM_015950937	*LOC107379977*	serine/arginine-rich splicing factor 1	ENSG00000136450
chr7	XM_015950939	*LOC107379978*	dynein light chain 2	

Table 3.6: Summary of highly differentiated genes between GNP and NF303

Chromosome	GeneID	Name	Description	Human Ensembl ID
chr10	XM_015941863	*hibch*	3-hydroxyisobutyryl-CoA hydrolase	ENSG00000198130
chr6	XM_015944547	*inpp5a*	inositol polyphosphate-5-phosphatase A	

Table 3.7: Summary of highly differentiated genes between GNP and NF414

Chromosome	GeneID	Name	Description	Human Ensembl ID
chr5	XM_015956265	*LOC107383550*	zinc finger and BTB domain-containing protein 14-like	ENSG00000198081
chr4	XM_015964158	*LOC107388594*	potassium voltage-gated channel subfamily D member 2-like	ENSG00000184408
chr13	XM_015967982	*LOC107390959*	patatin-like phospholipase domain-containing protein 2	ENSG00000180316

Table 3.8: Summary of highly differentiated genes between NF414 and NF303

Chromosome	GeneID	Name	Description	Human Ensembl ID
chr3	XM_015941178	*LOC107372989*	protein-lysine N-methyltransferase EEF2KMT	
chr4	XM_015941768	*insc*	inscuteable homolog (Drosophila)	
chr10	XM_015941863	*hibch*	3-hydroxyisobutyryl-CoA hydrolase	ENSG00000198130
chr10	XM_015941866	*lss*	lanosterol synthase	
chr10	XM_015941868	*LOC107373715*	chromosome unknown open reading frame 2C human C2orf88	
chr10	XM_015941869	*LOC107373716*	uncharacterized LOC107373716	
chr10	XM_015941982	*cops8*	COP9 signalosome subunit 8	ENSG00000198612
chr4	XM_015942501	*LOC107374351*	protein inscuteable homolog	
chr6	XM_015944700	*slc8a1*	solute carrier family 8 member A 1	
chr1	XM_015946918	*phf23*	PHD finger protein 23	ENSG00000040633
chr8	XM_015966934	*casc3*	cancer susceptibility candidate 3	ENSG00000108352
chr13	XM_015967972	*LOC107390951*	transcriptional enhancer factor TEF-5-like	
chr19	XM_015972986	*LOC107394182*	solute carrier family 23 member 2-like	ENSG00000089057
chr12	XM_015973765	*jarid2*	jumonji and AT-rich interaction domain containing 2	ENSG00000008083
chr10	XM_015974725	*LOC107395351*	dipeptidyl aminopeptidase-like protein 6	
chr3	XM_015976081	*LOC107396381*	serine/arginine repetitive matrix protein 2-like	ENSG00000167978
chr10	XM_015977084	*LOC107397167*	growth/differentiation factor 8-like	
chr10	XM_015977085	*LOC107397168*	glutathione S-transferase theta-1-like	ENSG00000099984

3.5.3 F_{ST} Gene Ontology analysis

To get insights into the biological processes that exhibit either low or high genetic differentiation between the populations, I performed gene ontology (GO) overrepresentation analysis. For the analysis of biological processes that are shared between the populations, I used only genes below the 2.5^{th} percentile of each pairwise comparison. For the analysis of differentiated biological process, I used only genes above the 97.5^{th} percentile of each comparison. Only genes with a corresponding human ENSEMBL identifier were used for GO overrepresentation analysis (see Section 2.1.2).

Interestingly and consistently in all pairwise comparisons, I found that GO terms associated with both mitochondrial and ribosomal functions were significantly overrepresented in low-differentiated genes (q-value < 0.05). Figure 3.21 shows the significantly overrepresented GO terms of the comparison between GNP and NF414, which have reached significance in at least one other comparison.

In both comparisons of GNP and either NF414 or NF303, a total of 19 GO terms were shared and significantly overrepresented. Among the GO terms with highest significance in both comparisons were *Ribosomal subunit* (q-value GNP vs. NF414: $6.23e^{-12}$, q-value GNP vs. NF303: 0.00057), *Ribosome* (q-value GNP vs. NF414: $1.12e^{-10}$, q-value GNP vs. NF303: $1.91e^{-04}$) and *Stuctural constitutent of ribosome* (q-value GNP vs. NF414: $4.8e^{-10}$, q-value GNP vs. NF303: $7.9e^{-05}$; see Figures 3.21, 3.22, Tables C.1, C.2).

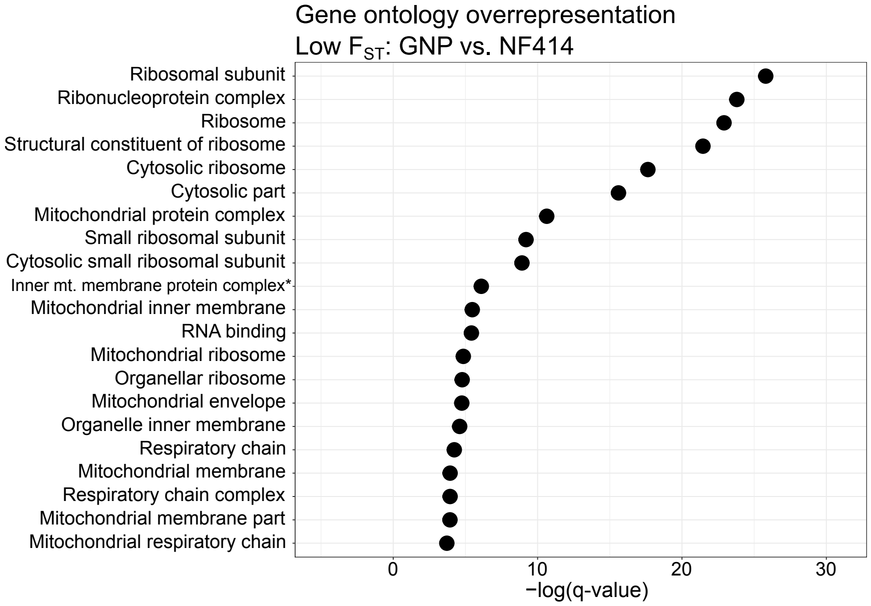

Figure 3.21: Gene ontology low F_{ST}: GNP vs. NF414. Shown are the overrepresented GO terms of genes with a F_{ST} value below the 2.5^{th} percentile of all F_{ST} values. The overrepresented GO terms have a q-value < 0.05 and are significant in at least one other comparison (* mitochondrial abbreviated as mt.).

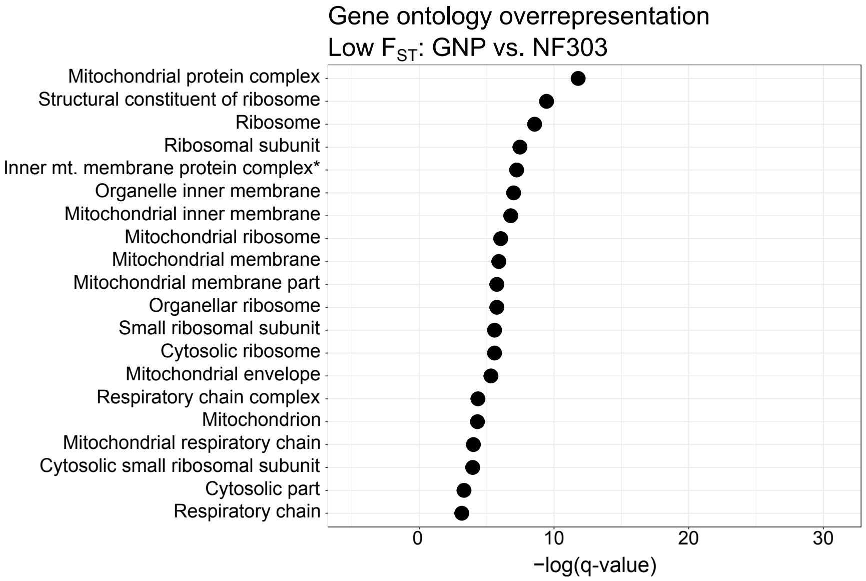

Figure 3.22: Gene ontology low F_{ST}: GNP vs. NF303. Shown are the overrepresented GO terms of genes with a F_{ST} value below the 2.5^{th} percentile of all F_{ST} values. The overrepresented GO terms have a q-value < 0.05 and are significant in at least one other comparison (* mitochondrial abbreviated as mt.).

The GO term *Structural constituent of ribosome* was significantly overrepresented in all three comparisons (q-value NF414 vs. NF303: 0.024; see Figure 3.23, Table C.3).

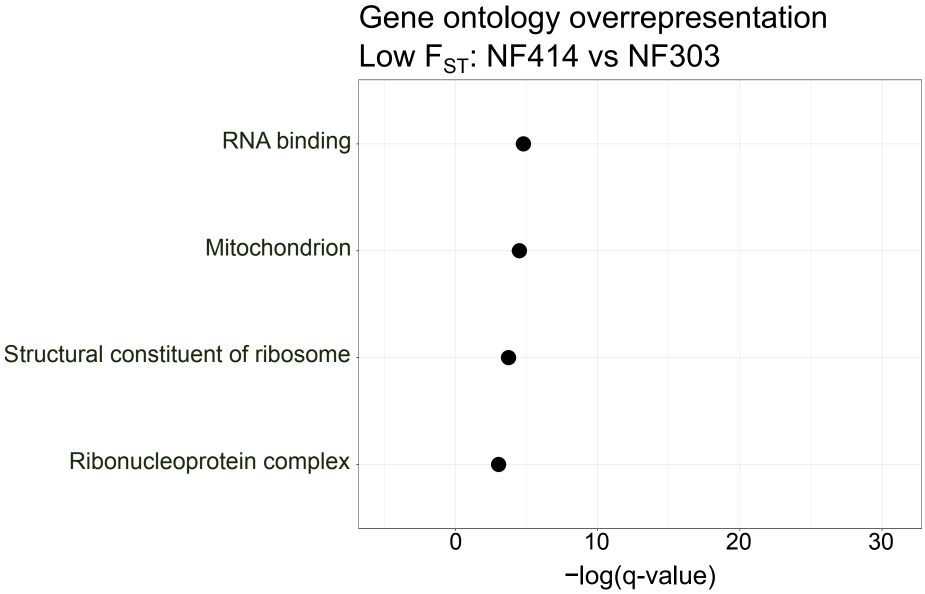

Figure 3.23: Gene ontology low F_{ST}: NF414 vs. NF303. Shown are the overrepresented GO terms of genes with a F_{ST} value below the 2.5^{th} percentile of all F_{ST} values. The overrepresented GO terms have a q-value < 0.05 and are significant in at least one other comparison.

In contrast to the GO overrepresentation analysis of genes with low genetic differentiation (below the 2.5^{th} percentile of gene-wise F_{ST} values), in which GO terms were overrepresented in each comparison, only one comparison reached significance (q-value < 0.05) in genes with high genetic differentiation (>97.5^{th} percentile of gene-wise F_{ST} values). Only the comparison between GNP and NF303 showed significant overrepresentation of GO terms in highly differentiated genes. The overrepresented GO terms were associated with lipid metabolism, such as *Long-chain fatty acid binding* (q-value: 0.028) and *Fatty acid derivative binding* (q-value: 0.036, Figure 3.24, Table C.4). In addition, the GO term *Structural constituent of ribosome* (q-value: 0.043) showed significant overrepresentation in genes with low and high genetic differentiation between GNP and NF303 (Figure 3.24, Table C.4).

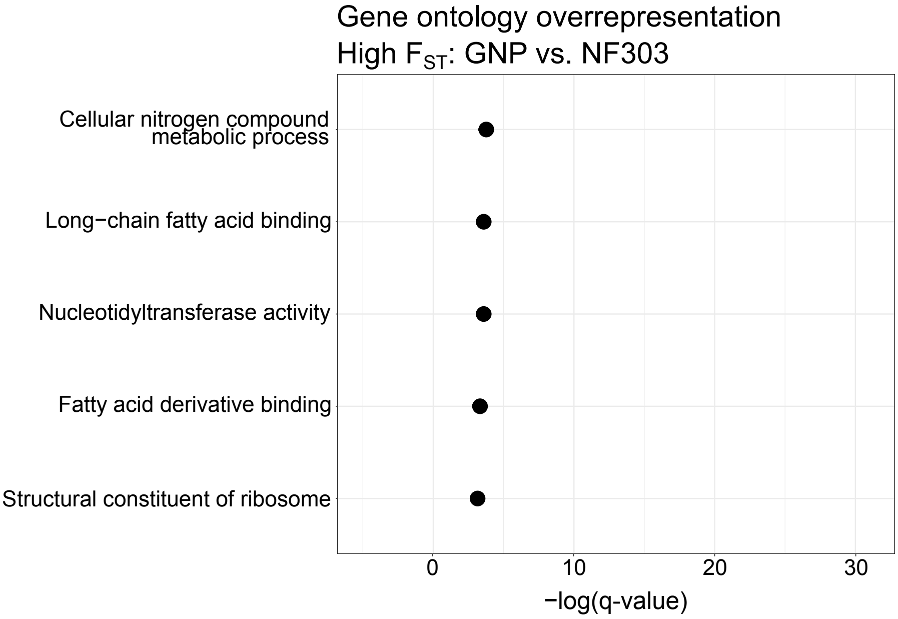

Figure 3.24: Gene ontology high F_{ST}: GNP vs. NF303. Shown are the overrepresented GO terms of genes with a F_{ST} value above the 97.5^{th} percentile of all F_{ST} values. The overrepresented GO terms have a q-value < 0.05.

Using the genetic differentiation index F_{ST} enabled me to find genes, genomic regions or biological processes that were either genetically similar or genetically different between the populations used in this study. Genes involved in mitochondrial and ribosomal function showed low genetic differentiation between the populations. However, I could not detect a clear signal of local selection in line with the differences in life history traits. Although the genome-wide genetic differentiation was the highest between GNP and NF303 (Figure 3.4), the strongest local genetic differentiation was observed between NF414 and NF303. Therefore, the findings did not support the hypothesis for the AP theory of aging, because one would expect the strongest signals of local genetic differentiation between the shortest-lived and longest-lived population. In the case of the studied populations, this refers to population GNP and population NF303. Thus, I could not find clear evidence that support the hypothesis of AP.

If strong local selection cannot explain the differences in late life history characteristics of these populations, I hypothesized that genome-wide patterns could explain the differences. Therefore, I set out to explore whether the contrasting population histories (Section 3.3) led to differences in the rate of molecular evolution as proposed by Ohta [129].

3.6 Molecular evolution in populations of the turquoise killifish

Mutations that occur within the coding sequence (CDS) of a gene can either be neutral, beneficial or detrimental. Neutral mutations can reach fixation via a random process called genetic drift. A beneficial mutation will rapidly reach fixation driven by positive selection. On the contrary, a detrimental mutation is likely to be purged out of the population by purifying selection. Most non-neutral mutations in the CDS have a detrimental effect [44] and are selected against by purifying selection. Thus, it should be expected that the ratio of fixed non-neutral mutations (non-neutral substitution) per non-neutral sites compared to fixed neutral mutations (neutral substitutions) per neutral sites is less than one. In general, this ratio is used as a measure of the evolutionary rate of genes [78, 21]. In this section, I investigated whether the expectation of purifying selection at non-neutral sites applies to the populations of the turquoise killifish and whether I can find differences in the rate of molecular evolution between the populations.

To calculate the ratio, I computed for the entire genome the number of non-synonymous substitutions per non-synonymous sites (non-neutral, D_n) and the number of synonymous substitutions per synonymous sites (neutral, D_s). I computed $\frac{D_n}{D_s}$ for all populations based on the derived frequency of two different outgroups, *Nothobranchius orthonotus* and *Nothobranchius rachovii* (see Section 2.1.2). Both species belong to the southern clade of the *Notho-*

branchius species, within which *N. orthonotus* is phylogenetically closer to the turquoise killifish than *N. rachovii* [41]. The genome-wide values for $\frac{D_n}{D_s}$ with *N. orthonotus* as outgroup were 0.222 (GNP), 0.221 (NF414) and 0.224 (NF303, see Table D.1). Using *N. rachovii* as outgroup, the genome-wide values for $\frac{D_n}{D_s}$ were 0.258 (GNP), 0.262 (NF414) and 0.277 (NF303, see Table D.1). The ratio increased slightly in the comparison between the turquoise killifish and *N. rachovii*.

In summary, the $\frac{D_n}{D_s}$ values observed for populations of turquoise killifish are in line with the expectation that purifying selection removes deleterious mutations in non-synonymous sites of the CDS. However, the $\frac{D_n}{D_s}$ ratio is preferentially used when investigating long-term evolutionary scales [100]. For exploring short-term evolutionary scales, like intra-species comparisons, information about polymorphisms should be included [100]. Therefore, I included information about polymorphisms in the subsequent analyses. Next, to assess what proportion of fixed mutations in non-synonymous sites are caused by positive selection and whether this proportion differs between the population of the turquoise killifish, I used the McDonald-Kreitman's α (MK-α, see Section 2.1.2).

3.6.1 Test for positive selection in turquoise killifish populations

To test for positive selection, I computed asymptotic MK-α for all populations [122, 66]. The estimate of MK-α gives the proportion of sites that reached fixation due to positive selection since divergence from the outgroup. Again, I used the closely related species of *N. orthonotus* as an outgroup. Among the three populations used in this study, the asymptotic MK-α values were -0.211, -0.027 and 0.042 for GNP, NF414 and NF303, respectively (Table D.2). Further, GNP showed a more negative MK-α compared to NF414 and NF303 at all calculated intervals. The most negative value was reached at lowest derived frequency bin (0.1-0.15, Figure 3.25).

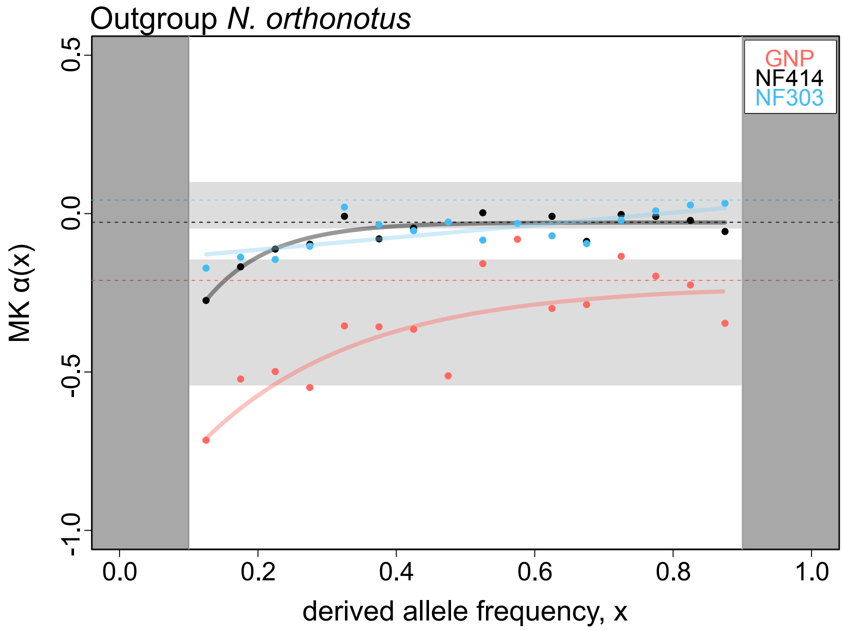

Figure 3.25: Asymptotic McDonald-Kreitman α: *N. orthonotus*. Shown is the asymptotic McDonald-Kreitman α (MK-α) as a function of binned derived frequencies on the x-axis and the MK-α on the y-axis. Population GNP is shown in red, NF414 in black and NF303 in blue.

Next, I calculated the asymptotic MK- α using *N. rachovii* as the outgroup. Consistent with the other comparison, GNP showed a more negative MK-α compared to population NF414 and NF303 at all calculated intervals (Figure 3.26). However, the asymptotic MK-α values were higher than the asymptotic MK-α values computed with *N. orthonotus* as outgroup. Populations GNP, NF414 and NF303 had asymptotic MK-α values of -0.057, 0.149 and 0.227, respectively (Table D.2). Hence, NF414 and NF303 showed evidence for recent molecular adaptation via positive selection. In contrast, the negative values of MK-α at lower derived frequency bins can be caused by slightly deleterious mutations segregating in the populations [122].

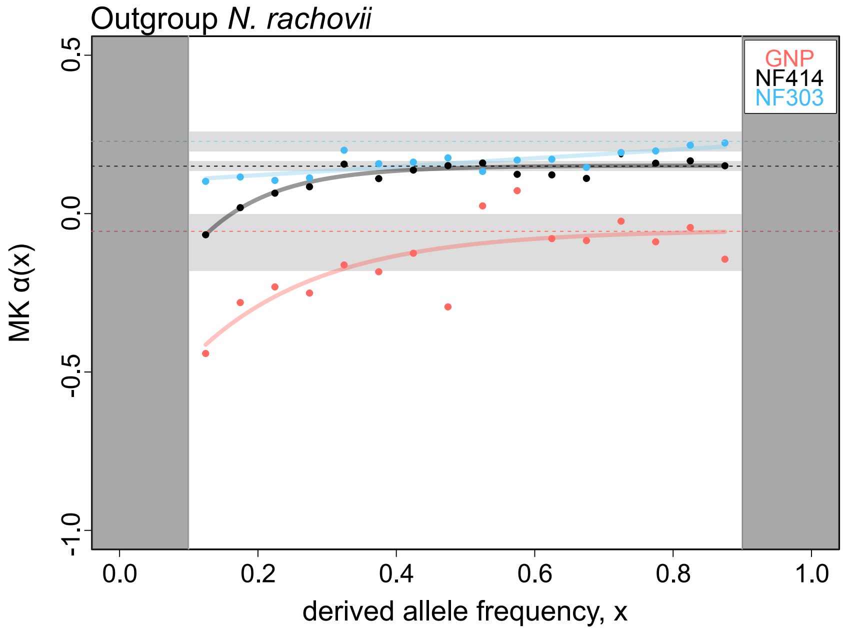

Figure 3.26: Asymptotic McDonald-Kreitman α: *N. rachovii*. Shown is the asymptotic McDonald-Kreitman α (MK-α) as a function of binned derived frequencies on the x-axis and the MK-α on the y-axis. Population GNP is shown in red, NF414 in black and NF303 in blue.

The MK-α analysis in turquoise killifish therefore revealed two interesting findings. First, I found evidence for recent positive selection in the populations NF414 and NF303, with NF303 showing a higher percentage of mutations that were fixed by positive selection. Second, I found that the populations studied may have different numbers of slightly deleterious mutations that segregate within each population. To further understand each aspect, I looked at signatures for recent positive selection based on Tajima's D and explored the efficacy of natural selection to remove deleterious mutations.

3.6.2 Candidate regions of recent positive selection

The estimate of Tajima's D can infer deviation from neutrality (see Chapter 2.1.2). This deviation can be caused either by selection or demography. Both a positive selection event and recent population growth could lead to an increase in low frequency variants and thus a negative value of Tajima's D [30]. To differentiate between demography and selection, I performed outlier detection in all populations. Demography should affect the entire genome, while recent positive selection should have a only a local effect.

To identify candidate regions of recent positive selection, I used an outlier approach to identify genomic regions that contained at least three consecutive 50 kb windows that fall below the 0.5^{th} percentile of Tajima's D values. This approach enabled me to characterize three independent genomic regions in NF303 that potentially experienced recent positive selection. The genomic regions were located on chromosomes 3, 4 and 10 (Figures 3.27, 3.28, 3.29). A detailed list of genes within each region of interest is provided in Table 3.9 for chromosome 3, Table 3.10 for chromosome 4 and Table 3.11 for chromosome 10. However, with this outlier approach I was not able to identify regions with signatures of recent positive selection in populations NF414 and GNP.

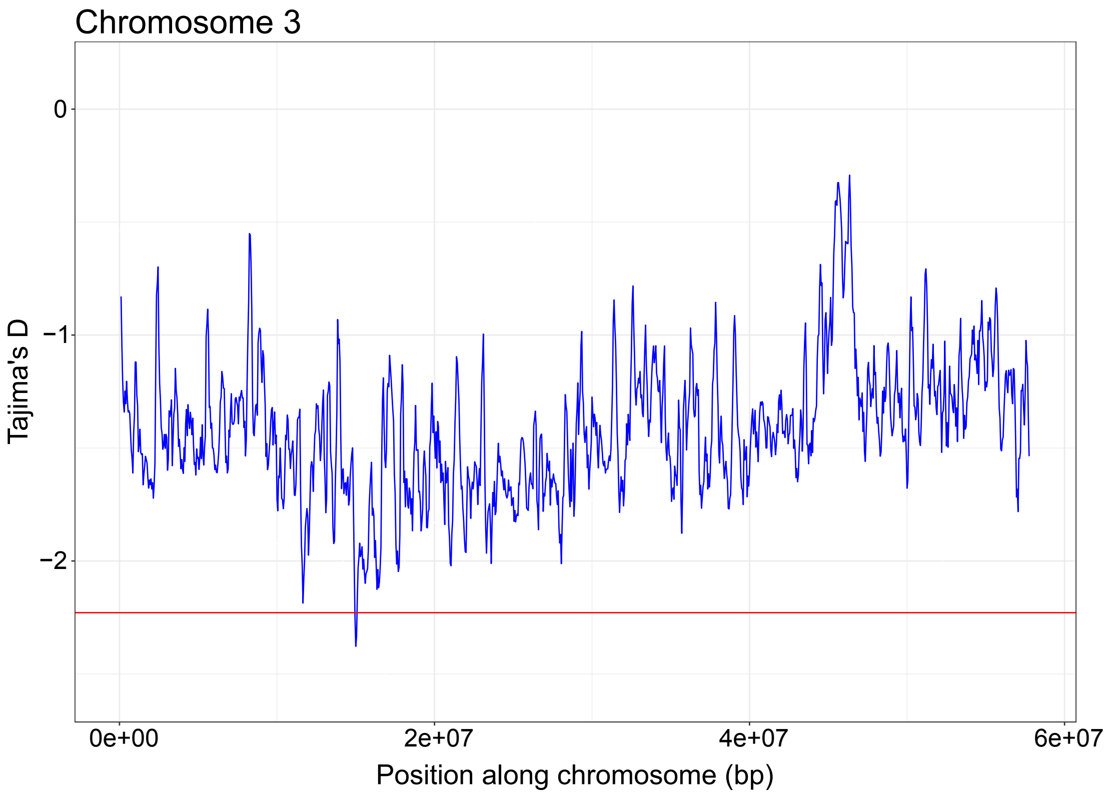

Figure 3.27: NF303 Tajima's D on chromosome 3. Shown are smoothed Tajima's D values of population NF303 on the y-axis and the position along chromosome 3 on the x-axis (in bp). Tajima's D values are calculated based on 50 kb non-overlapping windows. The red line indicates the 0.5^{th} percentile genome-wide threshold of all Tajima's D values.

Table 3.9: List of genes in target region on chromosome 3

Chromosome	GeneID	Name	Description	Human Ensembl ID
chr3	XM_015948909	LOC107378608	glutamate receptor ionotropic, NMDA 2C-like	NA
chr3	XM_015948913	rbfox3	RNA binding protein 2C fox-1 homolog (C. elegans) 3	NA

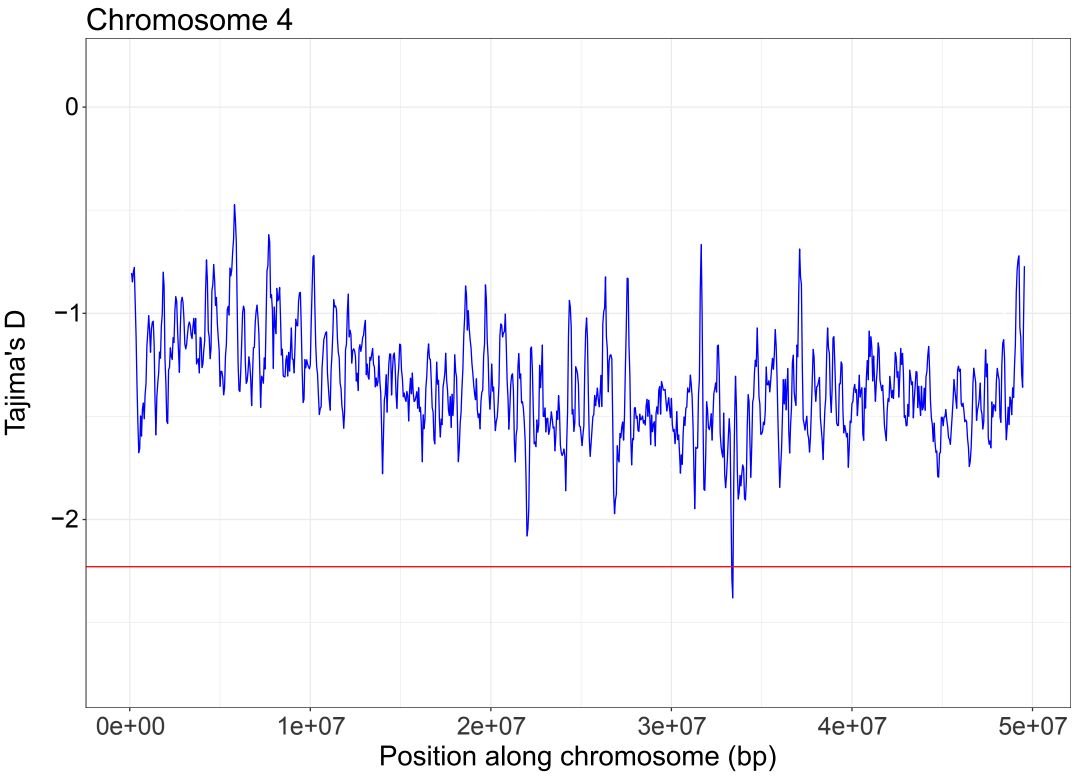

Figure 3.28: NF303 Tajima's D on chromosome 4. Shown are smoothed Tajima's D values of population NF303 on the y-axis and the position along chromosome 4 on the x-axis (in bp). Tajima's D values are calculated based on 50 kb non-overlapping windows. The red line indicates the 0.5^{th} percentile genome-wide threshold of all Tajima's D values.

Table 3.10: List of genes in target region on chromosome 4

Chromosome	GeneID	Name	Description	Human Ensembl ID
chr4	XM_015963928	LOC107388409	switch-associated protein 70-like	ENSG00000133789
chr4	XM_015963930	wee1	WEE1 G2 checkpoint kinase	ENSG00000166483
chr4	XM_015963932	znf143	zinc finger protein 143	NA
chr4	XM_015963927	akip1	A-kinase interacting protein 1	ENSG00000166452
chr4	XM_015963934	rpl27a	ribosomal protein L27a	ENSG00000166441
chr4	XM_015941825	LOC107373687	switch-associated protein 70-like	NA

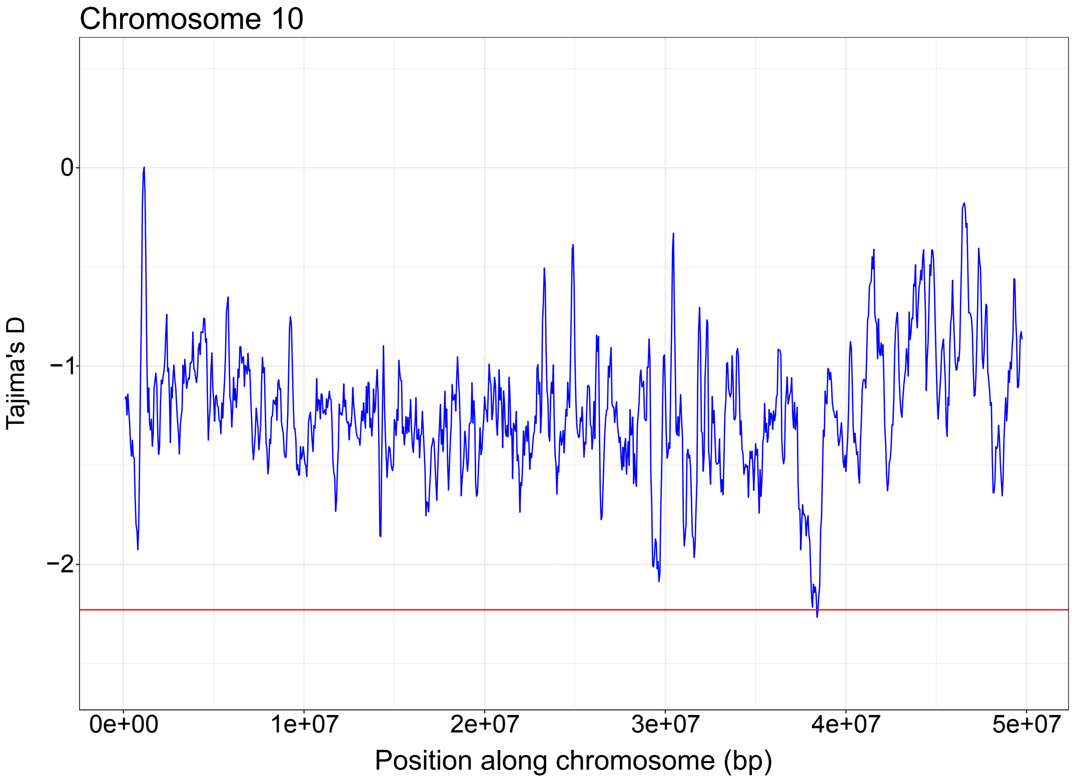

Figure 3.29: NF303 Tajima's D on chromosome 10. Shown are smoothed Tajima's D values of population NF303 on the y-axis and the position along chromosome 10 on the x-axis (in bp). Tajima's D values are calculated based on 50 kb non-overlapping windows. The red line indicates the 0.5^{th} percentile genome-wide threshold of all Tajima's D values.

Table 3.11: List of genes in target region on chromosome 10

Chromosome	GeneID	Name	Description	Human Ensembl ID
chr10	XM_015976205	LOC107396454	mitochondrial chaperone BCS1	NA
chr10	XM_015976202	LOC107396451	sterol 26-hydroxylase, mitochondrial	ENSG00000135929
chr10	XM_015976193	tns1	tensin 1	NA

3.6.3 Purifying selection is weaker in dry population

In Section 3.6.1, I showed that GNP has a more negative MK-α compared to NF414 and NF303. To assess whether the negative MK-α is a result of more slightly deleterious mutation segregating in GNP, I computed the direction of selection (DoS) estimate (see Section 2.1.2). DoS ranges between minus one and one; a negative DoS is a result of slightly deleterious mutations not being purged from the population due to weaker purifying selection and a positive DoS is caused by intensified selection [175].

I used the same comparative framework as in Section 3.6.1 and calculated DoS for each gene separately with either *N. orthonotus* or *N. rachovii* as outgroup. Using *N. orthonotus* as outgroup resulted in a median DoS of -0.167 for GNP (Number of genes included N=8438), -0.021 for NF414 (N=17284) and -0.014 for NF303 (N=16183, see Table D.2). The DoS values were significantly different between each pairwise comparison of the populations, with a clear shift to the negative range of DoS for GNP (two-sided Wilcoxon rank-sum test, GNP vs. NF303: P<$4.35e^{-107}$, GNP vs. NF414: P<$2.9e^{-78}$, NF414 vs. NF303:$1.72e^{-06}$, Figure 3.30). The results remained unchanged when using *N. rachovii* as outgroup. The median DoS of GNP, NF414 and NF303 were -0.143 (N=8580), 0.000 (N=18828) and 0.000 (N=17790), respectively (see Table D.2). As shown in Figure 3.31, the populations were significantly different to each other, with GNPs DoS distribution shifted to the negative range of DoS (two-sided Wilcoxon rank-sum test, GNP vs. NF303: P<$5.75e^{-182}$, GNP vs. NF414: P<$1.39e^{-102}$, NF414 vs. NF303:$6.15e^{-22}$).

In summary, the finding of a significantly more negative DoS estimate for the GNP population indicates that the force of purifying selection is weaker in GNP, in line with the negative MK-α values observed in Section 3.6.1.

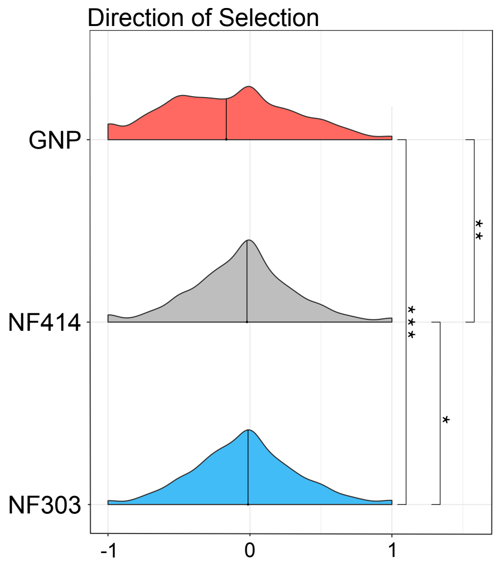

Figure 3.30: Distribution of direction of selection: *N. orthonotus*. Direction of selection (DoS) distribution of the populations GNP (in red), NF414 (in grey) and NF303 (in blue). DoS ranges between -1 and 1. Results of Wilcoxon rank-sum tests are indicated with asterisks (*P<1.72e^{-06}, **P<2.9e^{-78}, ***P<4.35e^{-107}).

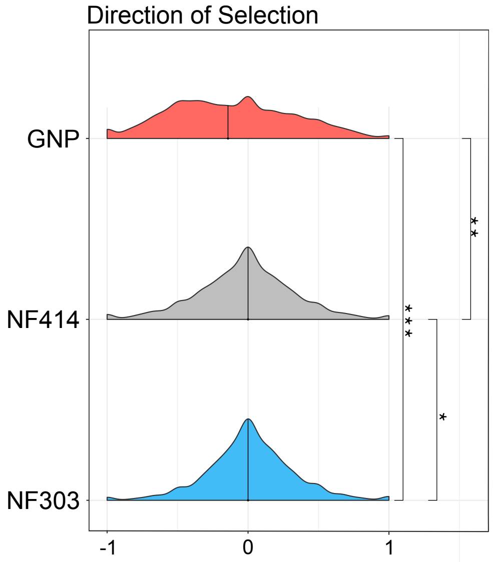

Figure 3.31: Distribution of direction of selection: *N. rachovii*. Direction of selection (DoS) distribution of the populations GNP (in red), NF414 (in grey) and NF303 (in blue). DoS ranges between -1 and 1. Results of Wilcoxon rank-sum tests are indicated with asterisks (*P<6.15e^{-22}, **P<1.39e^{-102}, ***P<5.75e^{-182}).

3.6.4 More slightly deleterious mutations in dry population

The DoS analysis revealed that purifying selection is weaker in GNP. Weaker purifying selection means that mutations with deleterious effects (negative selection coefficient s) are not recognized as such and therefore not removed from the population. To quantify the genome-wide proportions of deleterious and neutral mutations, I calculated the distribution of fitness effect (DFE) using `polyDFE2.0` [180]. This method uses the SFS of synonymous and non-synonymous sites to estimate the DFE. The fitness effect is represented as product of $4Ne$ multiplied by the selection coefficient s (see Section 1.2.3). Based on this product, the mutations are categorized as effectively neutral ($0 > 4Ne*s > -1$), slightly deleterious ($-1 > 4Ne*s > -10$), deleterious ($-10 > 4Ne*s > -100$) or strongly deleterious ($-100 > 4Ne*s$, based on [90, 111]). Importantly, the DFE does not account for the effect of non-synonymous mutations present in the real data, it predicts the effect of new non-synonymous mutations entering the population. Therefore, the larger the proportion in the strongly deleterious category, the more likely that the mutations with deleterious effects are removed by natural selection (see Section 1.2.3).

As shown in Figure 3.32, the DFE for NF414 and NF303 had similar shapes, with approximately 70% of the mutations in the deleterious to strongly deleterious category and 21% in the effectively neutral category (Table 3.12). The proportion of slightly deleterious mutations in NF414 and NF303 was 9.1% and 7.7%, respectively. This is in contrast to GNP, with approximately 43% of mutations in the deleterious to strongly deleterious category and 37.6% of the mutations in slightly deleterious category. The proportion of neutral mutations of GNP was 19%, comparable to NF414 and NF303 (21.1% and 21.2%, Table 3.12).

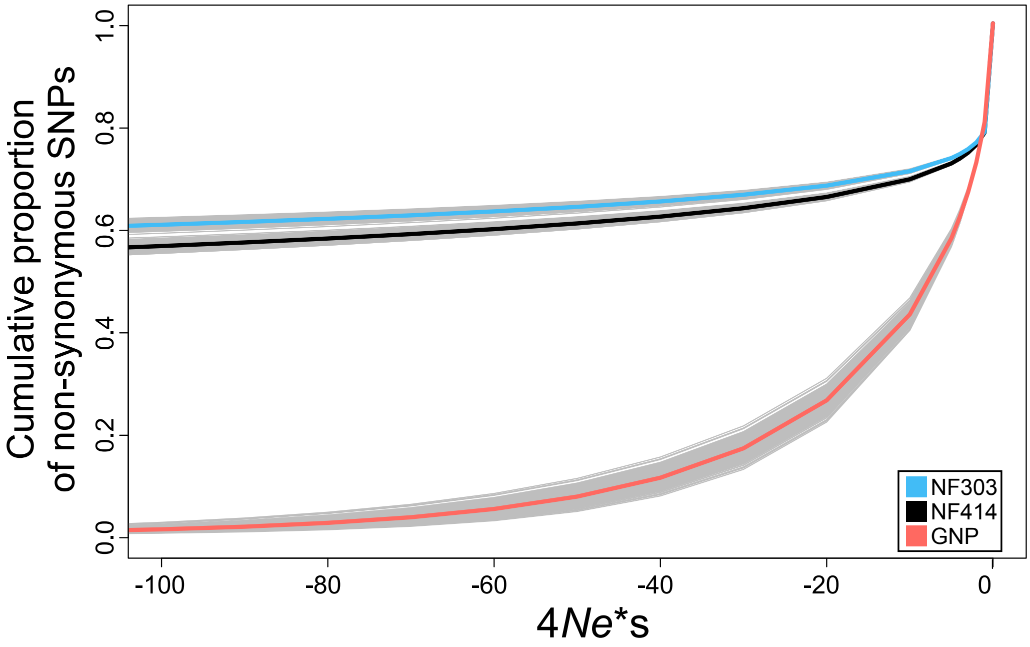

Figure 3.32: Deleterious distribution of fitness effects. The deleterious distribution of fitness effects (DFE) is shown for each population. The cumulative proportion of non-synonymous SNPs is shown on the y-axis and the fitness effect of the mutations in 4*Ne**s on the x-axis.

Table 3.12: Proportion of fitness effects

Category	**GNP [95%CI]**	**NF414 [95%CI]**	**NF303 [95%CI]**
Strongly deleterious	0.012 [0.006-0.019]	0.565 [0.554-0.577]	0.607 [0.597-0.616]
Deleterious	0.420 [0.404-0.430]	0.130 [0.122-0.138]	0.105 [0.098-0.112]
Slightly deleterious	0.376 [0.355-0.401]	0.091 [0.087-0.095]	0.077 [0.073-0.081]
Effectively neutral	0.193 [0.189-0.196]	0.213 [0.212-0.216]	0.212 [0.210-0.214]

3.6.5 Higher proportion of SNPs with detrimental effects in dry population

The frequency of slightly deleterious mutations is more determined by genetic drift and less by natural selection, the closer the mutation is to neutrality (Section 1.2.3, Equation 1.7). In previous sections, I showed that in GNP the strength of natural selection is decreased and that GNP contained more mutations categorized as slightly deleterious. These results confirmed that natural selection is less efficient in populations with smaller *Ne*. However, the effect of the mutations on protein function remains unknown. In order to examine the effect of mutations on protein function, I used the program `SnpEFF` [32]. `SnpEFF` categorizes SNP effects in three different classes, with low, moderate, and high impact on protein function. These classes refer to the change that the SNP would impose on the reference amino acid sequence. A mutation that causes a change in the amino acid sequence, *i.e.*, a non-synonymous mutation, is categorized in a higher impact class than a mutation that does not change the amino acid sequence, a synonymous mutation.

As shown in Figure 3.33 the proportion of SNPs with detrimental effect on protein function (moderate or high impact) was significantly higher in GNP compared to NF414 (Chi-square test, $P < 1.2e^{-57}$) and NF303 (Chi-square test, $P < 5.9e^{-121}$). Interestingly, the proportion of SNPs with detrimental effects on protein function decreased with an increased *Ne*. Accordingly, NF303, the population with the largest *Ne*, showed the lowest proportion of SNPs with detrimental effects on protein function (NF414 vs. NF303, Chi-square test, $P < 2e^{-35}$). In total, approximately 1.4% of the mutation in the CDS in GNP belong to the high impact class, in contrast to 0.44% in NF414 and 0.37% in NF303.

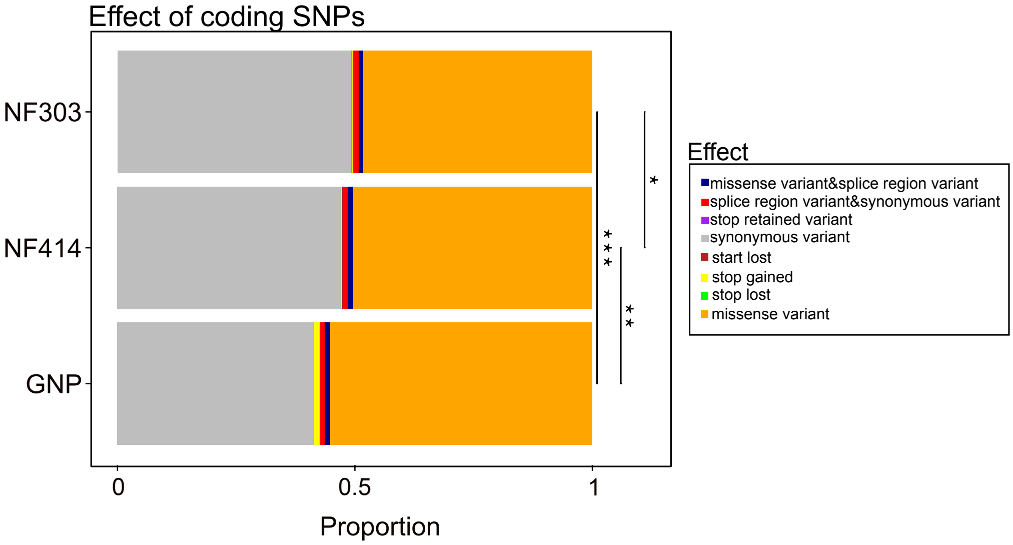

Figure 3.33: Effect on protein function of SNPs in coding sequence. Variant effects were categorized using SnpEff [32] based on their position in the annotated turquoise killifish genome. Results of Chi-square tests are indicated next to the bars (*P<$2e^{-35}$, **P<$1.2e^{-57}$, ***P<$5.9e^{-121}$).

Moreover, SNPs with moderate or high impact on protein function (i.e., detrimental SNPs) segregate at higher allele frequencies in GNP. More than 45% of detrimental SNPs reach an allele frequency above 20% (Figure 3.34). Again, the trend decreases with respect to a larger *Ne*. In both populations, NF414 and NF303, the percentage of detrimental SNPs reaching 20% allele frequency is lower compared to GNP (34% in NF414, 19% in NF303 and 45% in GNP, Figure 3.34).

In total, I found that weaker purifying selection in the GNP population resulted in an excess of SNPs with detrimental effects on protein function and these SNPs reached high frequencies. Next, I wondered whether the detrimental SNPs accumulate randomly or if they are enriched in specific biological processes. I hypothesized that under the MA theory of aging, population GNP should show enrichment of detrimental SNPs acting in late life processes.

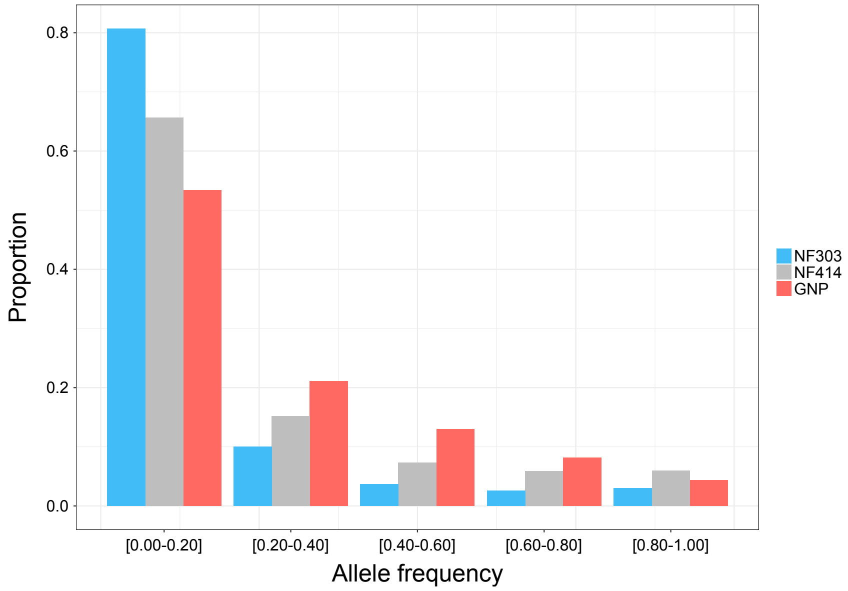

Figure 3.34: Frequency of detrimental mutations. Shown are the frequencies of the mutations with moderate or high impact. The alternative allele frequency compared to the reference genome is shown on the x-axis and the proportion of the effect causing mutations is shown on the y-axis.

3.6.6 Relaxed purifying selection in age-related pathways

To assess which biological processes are enriched for SNPs with detrimental effects on protein function, I performed enrichment analysis on GO families based on the DoS distribution. I used the GO slim annotation from ENSEMBL BioMart v87 [171, 209] and computed the DoS distribution for all GO slim families with a gene count of at least 10 (see Section 2.1.2). As shown in Figure 3.25 and Figure 3.26, the comparison with *N. rachovii* has more power to detect differences, as the time to most recent common ancestor (MRCA) between the two species is longer. Therefore, I computed the GO slim family DoS distribution based on *N. rachovii* as outgroup. If the DoS distribution is significantly different from the total DoS distribution of the population (Wilcoxon rank-sum test), it is considered either as being under relaxed purifying selection (negative DoS, DoS GO slim family < DoS total) or intensified selection (positive DoS, DoS GO slim family > DoS total). Figure 3.35 shows that in GNP the GO slim families *aging*, *membrane organization*, *growth*, *cell death*, *peroxisome* and *transferase activity, transferring glycosyl groups* are under relaxed purifying selection. The GO terms had a negative DoS median and a significantly more negative DoS distribution compared to the overall DoS distribution (Wilcoxon rank-sum test, *P*-value <0.05, not significant after FDR correction, see Table F.1). Conversely, the GO term *histone binding* showed a signature for intensified selection, with a positive DoS and a significantly more positive DoS distribution (Wilcoxon rank-sum test, *P*-value <0.05, not significant after FDR correction, see Table F.1). In population NF414, the GO terms *ribosome*, *mRNA binding* and *structural constituent of ribosome* were identified as being under intensified selected (Wilcoxon rank-sum test, *P*-value <0.05, not significant after FDR correction, see Table F.2). Finally, population NF303 showed intensified selection in *nitrogen cycle metabolic process*, *rRNA binding*, *mitotic nuclear division*, *chromosome segregation*, *ribosome* and *structural constituent of ribosome*, with significance after FDR correction (FDR<0.1, Benjamini-Hochberg correction [12], see Table F.3). Further, in NF303 the GO terms *carbohydrate metabolic process* and *lyase activity* were identified as being under relaxed purifying selection (FDR<0.1, Benjamini-Hochberg correction [12], see Table F.3).

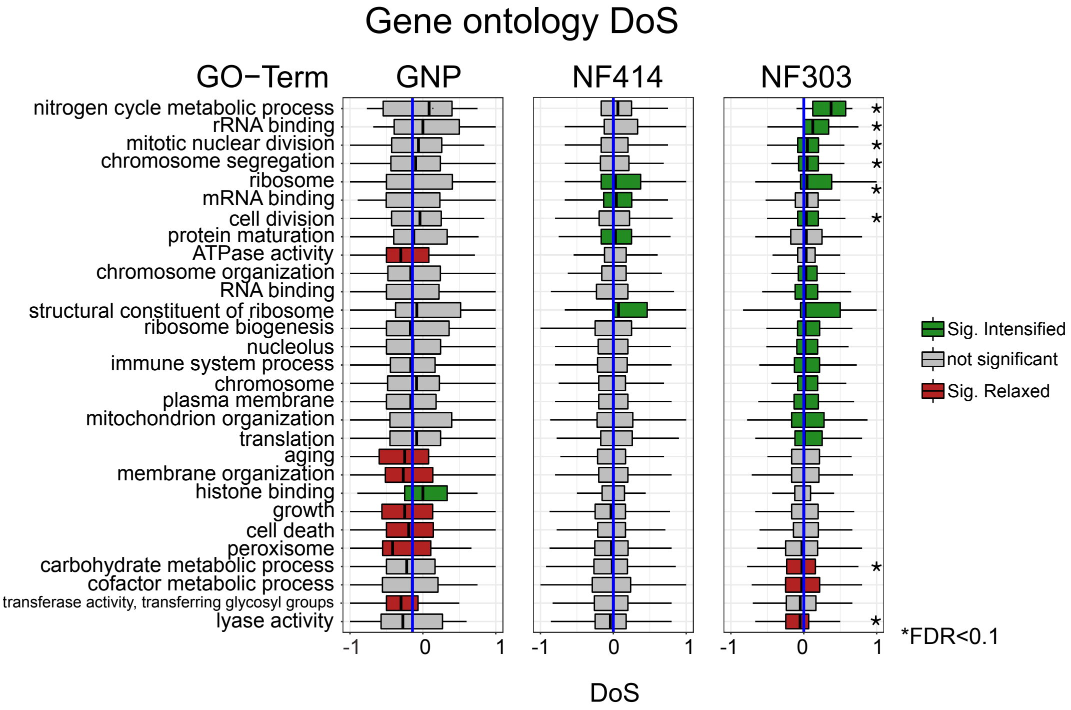

Figure 3.35: Enrichment of Gene Ontology families based on DoS. Boxplots for DoS distribution based on GO families per population. Significant GO families based on Wilcoxon rank-sum tests (*P*<0.05) are colored in red and green, with red colored GO families being under relaxed selection (DoS GO familiy < DoS population) and green colored GO families being under intensified selection (DoS GO familiy > DoS population). Significant values after FDR corrections are labelled with an asterisk.

Genes involved in the aging process are potentially under relaxed purifying selection in GNP, whereas genes involved in ribosomal function are under intensified selection in NF414 and NF303. However, a limitation of this analysis is that for substitutions it cannot distinguish on which branch the mutation occurred. Therefore, intensified selection (DoS > 0, more non-synonymous substitutions) could have occured in either species used in the analysis. To address this limitation, I computed the intersection of genes with high or low DoS compared to both outgroups, *N. rachovii* and *N. orthonotus*. Therefore, for each population I used only genes that are present in both comparisons below the 2.5^{th} percentile or above the 97.5^{th} percentile. The intersection of both outgroups should enable me to identify pathways that are either enriched for genes under relaxed purifying selection or intensified selection in the turquoise killifish populations. Next, I performed pathway overrepresentation analysis (see Section 2.1.2). Interestingly, in the GNP population pathways associated with aging and age-related diseases were significantly overrepresented in the relaxed purifying selection category (<2.5% DoS , q-value <0.05, Figure 3.36). This included pathways such as *DNA damage response*, *mTOR signaling pathway* and several pathways related to cancer. The significant overrepresentation of aging and age-related diseases pathways in population GNP did not change when considering each outgroup separately (see Tables G.1, G.2). Further, population NF414 and NF303 showed no significantly enriched pathways in the relaxed purifying selection category (q-value < 0.05). NF303 showed significant overrepresentation of genes in mitochondrial pathways under intensified selection (Figure 3.36). Thereby, pathways of *mitochondrial translation*, *mitochondrial translation initiation*, *mitochondrial translation elongation* and *mitochondrial translation termination* were significantly overrepresented. In addition, I found *mitochondrial translation* and *mitochondrial translation termination* in GNP and *electron transport chain* pathway in NF414 as overrepresented in the intensified selection category when only using *N. rachovii* as outgroup and a q-value threshold of 0.1 (Table G.4; not with outgroup *N. orthonotus*, see Table G.3).

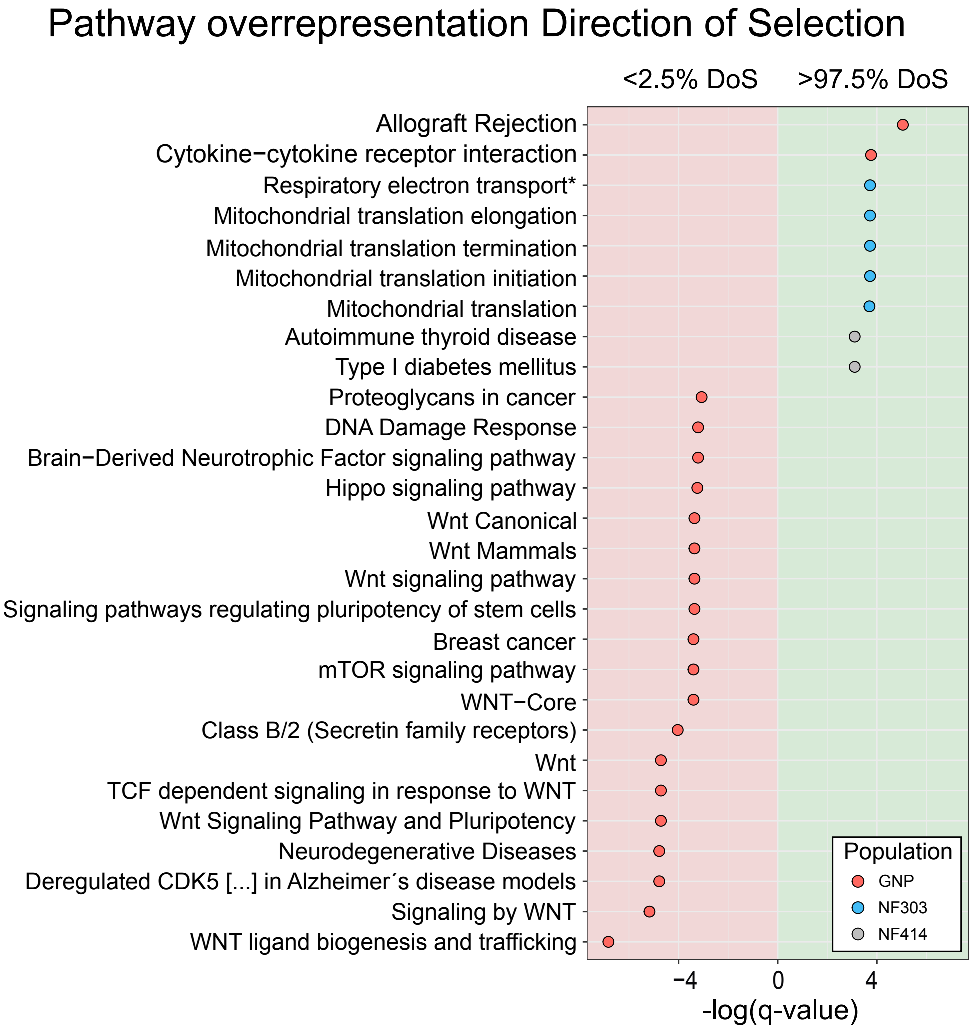

Figure 3.36: Pathway overrepresentation: Direction of selection. Results of pathway overrepresentation analysis based on genes below the 2.5^{th} percentile or above the 97.5^{th} percentile of DoS values of each population in both comparisons (with *N. orthonotus* and *N. rachovii*). The significantly overrepresented pathways were required to have a q-value below 0.05. Results for population GNP are shown in red dots, for NF414 in grey dots and for NF303 in blue dots. Overrepresented pathways below the 2.5^{th} percentile of DoS values are shown in red background and overrepresented pathways above the 97.5^{th} percentile of DoS values are shown in green background (* abbreviated : Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins).

Chapter 4

Discussion

In this thesis, I explored life history evolution in natural populations of the short-lived vertebrate *Nothobranchius furzeri* (turquoise killifish) using a population genetics approach. I studied the evolution of life history characteristics in two different ways. First, I asked which evolutionary forces - *i.e* adaptive or neutral evolution - can explain the naturally occurring differences in late life history traits (aging rate and lifespan) in wild populations of the turquoise killifish. Second, I studied the biological processes that are under intense selection in natural populations of turquoise killifish, which may reflect adaptations to the harsh environment in which the turquoise killifish lives.

4.1 Evolution of aging in turquoise killifish

4.1.1 Smaller effective population size is associated with short lifespan

Natural populations of the turquoise killifish show differences in the rate of aging and lifespan along an aridity gradient [181, 14]. Populations from more arid regions have a shorter lifespan and display a more rapid demographic aging, *e.g.* a higher increase in mortality rate. This is in line with the prediction from the classical evolutionary theories of aging, that extrinsic mortality is a main driver of the evolution of aging [25]. Using pooled whole genome sequencing of natural populations of the turquoise killifish, I found that populations along the aridity gradient (Figure 3.1) differ in genetic diversity and effective population size (*Ne*) (Figure 3.5). Populations associated with shorter lifespan and faster aging showed reduced genetic diversity and smaller *Ne*. Moreover, I found that populations from dry and wet regions have contrasting population demographic history (Figure 3.9). The dry population continuously declined in ancestral *Ne*, whereas the populations from more wet regions underwent a population expansion. Overall, my findings corroborate and extend previous results from Bartakova *et al.* [9, 8], who investigated the nucleotide diversity and demography among turquoise killifish populations based on 13 microsatellites and one mitochondrial DNA sequence. Bartakova *et al.* found evidence for population expansion in the main clades of the turquoise killifish and showed that populations at the periphery of the species distribution have reduced genetic diversity. My results confirm the previous results from Bartakova *et al.* and give novel insights into the decline in *Ne* along the aridity gradient. In addition, my study provides novel insights regarding genome-wide genetic diversity in turquoise killifish populations.

4.1.2 Mutation accumulation theory of aging

Relaxed purifying selection in populations from dry regions

The analysis of the rate of molecular evolution based on $\frac{D_n}{D_s}$ revealed that natural populations of turquoise killifish did not differ in the ratio of non-synonymous to synonymous substitutions (Section 3.6), indicating that turquoise killifish populations do not differ on the long evolutionary timescale. However, the analysis of asymptotic MK-α (Section 3.6.1) showed a clear difference in the proportion of sites driven to fixation by positive selection between the populations (Figures 3.25, 3.26), with populations from more dry regions showing a more negative MK-α value at all derived frequency bins in comparison to populations from more wet regions. More sites driven to fixation by positive selection in populations from more wet regions and an overall more negative MK-α value in populations from more dry regions were consistent in comparisons with both outgroup species *N. orthonotus* and *N.rachovii* (Figures 3.25, 3.26), indicating that more slightly deleterious mutations segregate in the dry population [122]. Populations with small *Ne* are expected to be greatly affected by the accumulation of deleterious mutations [76]. Therefore, accelerated aging or shorter lifespan in populations with small *Ne* should be largely driven by mutation accumulation (MA) than by antagonistic pleiotropy (AP) [76]. In other words, the overall efficiency of selection is generally reduced in populations with small *Ne*, which amplifies the decrease in the force of natural selection with age (see Section 1.2). An increase in the number of slightly deleterious mutations that segregate in populations from more dry regions is in line with selection being less efficient in populations with small *Ne* [27]. The decline in efficiency of selection is in line with the expectation that the rate of molecular evolution is expected to be higher in populations with small *Ne* [131].

Further, my analyses of direction of selection (DoS) revealed that natural selection in populations from more dry regions was significantly more relaxed than in populations from more wet regions (Figures 3.30, 3.31). In addition, the analysis of the distribution of fitness effects (DFE) supported the finding of more relaxation of selection in populations from more dry regions, with more mutations assigned to the slightly deleterious category (Figure 3.32, Table 3.12), indicating that new alleles are more influenced by genetic drift than by selection.

Overall, my findings that turquoise killifish populations from more dry regions had a smaller *Ne* and accumulated more slightly deleterious mutations by genetic drift is in line with the nearly neutral theory of molecular evolution proposed by Ohta ([130, 131], see Section 1.2.3).

Genetic drift and aging

Given that I found that selection was more relaxed and genetic drift was more pronounced in populations from more dry regions, I asked whether genetic drift is a major driver of the differences in late life history traits in turquoise killifish populations. To infer the impact of genetic drift, I asked whether the dry population has accumulated more mutations with detrimenal effects on protein function (Section 3.6.5). The approach utilized suffers from the limitation that the inference of the effect of a mutation is based on the turquoise killifish reference genome (NFZ V2.0). The genome sequence is from an individual of the short-lived

GRZ strain, a highly inbred laboratory strain [149] derived from the Gonarezhou National Park and might be genetically closer to populations from more dry regions used in this study (see Section 3), which could potentially bias the analysis towards detecting less detrimental effects in populations from more dry regions.

However, I found that populations from more dry regions had proportionally more mutations with detrimental effect on protein function (Figure 3.33), strengthening the results of the analysis. Furthermore, I found that the mutations with detrimental effect on protein function were present at higher frequencies in populations from more dry regions (Figure 3.34). Proportionally more detrimental mutations at higher frequencies in populations from more dry regions suggest that the mutational load in populations from more arid regions is higher and that the detrimental mutations shift to higher frequencies due to genetic drift. Overall, these findings are in accordance with predictions about the impact of a small population size on the allele frequency of deleterious mutations [112, 103]. In human out-of-Africa populations, which have a smaller *Ne* in comparison to African populations, it has been observed that genetic drift has shifted the allele frequency spectrum of deleterious mutations to higher frequencies [72]. The shift of deleterious mutations to higher frequencies in human out-of-Africa populations is in agreement with the shift of deleterious mutations to high frequencies in the dry population of turquoise killifish presented here. Whether the individual mutational load is higher in turquoise killifish from dry regions compared to wet regions is a question that cannot be addressed as the study design of pooled WGS (also referred to as "Pool-seq" [166]) prevents inference of the individual burden of mutation. Further evaluation of the individual mutational load, as described for the human populations in Simons *et al.* [169], would require more individual resequencing per population.

In general, the finding that genetic drift results in accumulation of deleterious mutations in populations with small *Ne* is known from experimental studies in various species, *e.g.* in *C. elegans* [84, 83] and *D. melanogaster* [67]. However, in line with the classical evolutionary theories of aging, pathway overrepresentation analysis showed that in populations from more dry regions pathways related to aging and age-related diseases were significantly enriched in the relaxed purifying selection category (Figure 3.36). The accumulation of deleterious mutations in aging-related genes was further supported by the DoS enrichment analysis, where the Gene Ontology (GO) slim term *aging* was under relaxed selection in populations from more dry regions (Figure 3.35). Hence, as predicted by the classical evolutionary theories of aging (Section 1.2), genes with effect late in life experienced less selective constraint [25]. In summary, the accumulation of deleterious mutations in genes involved in aging and age-related diseases are in line with the hypothesis that detrimental mutations with effects late in life should accumulate under the MA theory of aging [120, 121]. Therefore, the novelty of my study is the finding that accumulation of deleterious mutations due to genetic drift is a main driver of the late life history differences in turquoise killifish populations, providing a significant extension to the classical evolutionary theories of aging.

In summary, my work shows that in populations with small *Ne* the contribution of MA to the evolution of shorter lifespans and accelerated aging is more pronounced, in line with predictions from Hughes [76]. A similar pattern has been suggested to shape late life history traits in *Daphnia magna* [108]. The study by Lohr *et al.* experimentally showed that genetic drift can lead to a shorter life expectancy and increased aging in populations with smaller *Ne* [108]. The findings presented here confirm the connection between genetic drift and its effects on late life history traits, but are able to address questions about the patterns of molecular evolution in more detail. I showed that, in the shorter-lived population, genetic drift outweighs positive selection (Figures 3.25, 3.26, 3.30, 3.31). The prevalence of relaxed purifying selection in the shorter-lived population is in accordance with another prediction proposed by Hughes *et al.* [77], who stated that in light of the evolutionary theories of aging, patterns consistent with neutral evolution rather than positive selection would strongly support MA [77]. Therefore, my work combines the nearly neutral theory of molecular evolution with the MA theory of aging.

4.1.3 Antagonistic pleiotropy theory of aging

No signs of local selection in lifespan QTL regions

The genetic differentiation analysis based on F_{ST} in genomic regions associated with lifespan in experimental crosses [185] showed no significant genetic difference between the studied populations (Figure 3.15), suggesting that there is no antagonistic pleotropism underlying the lifespan QTL regions in the studied populations. However, a lack of significant genetic differentiation in lifespan QTL regions in natural populations does not preclude the potential importance of QTL regions in regard to lifespan in turquoise killifish and highlights that lifespan is a very complex trait. There are several possible explanations for the unexpected result of a lack of significant genetic differentiation in lifespan QTLs. It is possible that different genes were selected in different populations of the turquoise killifish, especially because population structuring is very strong in this species [8, 9]. Therefore, the QTL regions associated with lifespan in experimental crosses do not necessarily overlap with natural populations. Further, the lifespan-associated QTL regions could also be carrying deleterious alleles. Under this scenario, the effect would then be more likely to be assigned to the MA theory of aging than to the AP theory of aging. An extension of the classical MA theory of aging states that in distinct populations different deleterious alleles can reach fixation by chance [43]. The strength of the detrimental effect of the independently fixed deleterious alleles could vary between populations depending on *Ne* (see Section 1.2.3). In populations with small *Ne* the detrimental effect could be more severe than in populations with larger *Ne*. Therefore, the QTL region associated with lifespan could be a result of two possible scenarios. Either a fixation of deleterious alleles in only the short-lived laboratory strain or fixation of deleterious alleles with different effects in both short-/ and long-lived laboratory strains. However, because the fixation of the deleterious allele is due to chance, one would expect that the genetic differentiation of the QTL regions does not exceed the genome-wide value, since an intensified selection pressure - contrary to the assumptions of AP - is not active. The assumption of independent fixation of deleterious alleles in short-/ and long-lived experimental strains used in the QTL studies may be possible because the general transition of the genus *Nothobranchius* into the annual environment was largely driven by a relaxation of purifying selection [38].

Next, I used an unbiased genome-wide outlier approach to identify candidate regions that contribute to the differences in late life history traits between natural turquoise killifish populations. I detected these candidate regions based on highest genetic differentiation F_{ST} (Section 3.5). Unexpectedly, most of the regions identified were not in the contrast between dry and wet populations but between the intermediate (NF414) and wet populations of this study, not matching the general genetic differentiation, which was highest between dry and wet populations (Figure 3.4). The finding of the strongest outlier signal between the wet and intermediate populations does not agree with the hypothesis that in the case of AP, the strongest signal of the outlier approach should be between dry and wet populations. A possible explanation for the strong local genetic differentiation between the intermediate (NF414) and wet (NF303) population could be recent admixture in population NF414. In contrast to the populations GNP (dry) and NF303, population NF414 is polymorphic with respect to the tail color of male turquoise killifish. Therefore, possible admixture could result in a bias in the F_{ST} outlier approach [116]. In addition, the outlier regions identified here could regulate other biological processes that are independent of lifespan and aging.

For example, the identified region on chromosome 10 contains the gene *hibch* that was found to be highly differentiated between GNP vs. NF303 and NF414 vs. NF303 (Figure 3.20). This gene is involved in valine metabolism and levels of valine seems to play a role in response to environmental changes in mangrove killifish *Rivulus marmoratus* [52]. It is tempting to speculate that *hibch* plays a role in response to different environmental conditions in populations of the turquoise killifish. However, the biological basis of the genetic differentiation in *hibch* between the population needs further investigation. Interestingly, the gene *lanosterol synthase* (*lss*) was highly differentiated between population NF414 and NF303 (Figure 3.20). This gene encodes a catalytic protein that limits the biosynthesis of cholesterol, steroid hormones, and vitamin D [75]. It has been recently shown that vitamin D regulates diapause in *Austrofundulus limnaeus* [154], another annual killifish species from South America. It is possible that vitamin D plays an important role in diapause in turquoise killifish, but any influence of *lss* needs further investigations. However, diapause is a feature of early life history in turquoise killifish. If *lss* is involved in the regulation of diapause, this could be a suitable candidate gene for the AP theory of aging.

In conclusion, I found no clear evidence for AP to explain the differences in the late life history traits of the turquoise killifish. This does not rule out that AP is involved in the evolution of late life history trait differences in turquoise killifish populations, but suggests that MA may contribute more to the differences in late life history traits in natural turquoise killifish populations.

4.2 Adaptations to harsh environments

The second aim of this thesis was to explore which biological processes show low genetic differentiation between turquoise killifish populations and to identify genes involved in adaptations to dry environments. In general, I found strong genetic differentiation between the populations of the turquoise killifish (Figure 3.4) which increased with geographical distance (Figure 3.3), indicating low levels of gene flow between distant populations. High genetic differentiation and low levels of gene flow are consistent with previous results from Bartakova *et al.* [8, 9]. However, I argue that despite the high overall genetic differentiation between populations, genes responsible for adapting to the harsh environment should be under strong purifying selection and should show low inter-populational differentiation. Therefore, identifying the biological processes which are enriched in genes showing less inter-populational genetic differentiation can give important insights into species specific adaptations. The assumption of low inter-population genetic differentiation in genes involved in adaptations to the harsh environment is in agreement with the previous finding that turquoise killifish populations along the aridity gradient do not differ in in early life history characteristics, such as age at maturity or hatching size [14]. Interestingly, the lack of divergence in early life history traits contradicts the predictions of the evolutionary theory of life history, where it is expected that higher extrinsic mortality, for example, leads to faster development [173]. In order to explain the similarity in early life history traits despite difference in aridity, Blazek and colleagues suggested that the extremely rapid maturation of the turquoise killifish - possible in only two weeks [189] - is already at the limit of feasibility for a vertebrate species [14]. Alternatively, the hypothesis of a trade off between early and late life history traits might not be true for populations of turquoise killifish.

4.2.1 Low differentiation in mitochondrial and ribosomal genes

To identify biological processes involved in adaption to the harsh environment, I conducted Gene Ontology (GO) overrepresentation analysis of genes with low genetic differentiation among turquoise killifish populations. I found significant enrichment of genes involved in mitochondrial and ribosomal function (Figures 3.21, 3.22, 3.23, Tables C.1, C.2, C.3). In particular, the analyses between either of the two less dry populations, NF414 and NF303, and the dry population showed a strong enrichment for mitochondrial or ribosomal GO terms (Figures 3.21, 3.22). Interestingly, genes involved in translation elongation and ribosomal function are up-regulated in diapause and during aging in the turquoise killifish [150]. The high expression of ribosomal genes during diapause is suggested to be linked with rapid development after hatching [19], indicating a specific adaptation to their ephemeral habitat. Reduced ribosomal biogenesis is correlated with prolonged lifespan [182, 59]. It is plausible that there is antagonistic pleiotropism between ribosomal biogenesis early in life (*e.g.* diapause) and aging in turquoise killifish.

In line with the finding of low genetic differentiation in genes involved in mitochondrial function between turquoise killifish populations, I found evidence for positive selection in mitochondrial genes, especially in population NF303 (Figures 3.29, 3.36, Table G.4). Detection of positive selection in NF303, but not in the other populations, could be due to the larger *Ne* in NF303, as the frequency of adaptive substitution is expected to be higher in populations with larger *Ne* ([17], as shown in MK-α analysis, Figure 3.26). In general, positive selection in mitochondrial genes is consistent with other findings of positive selection in the genus *Nothobranchius* [38, 164]. However, Cui *et al.* [38] showed that mitochondrial genes additionally experienced relaxed purifying selection and suggests that positive selection might be a compensatory mechanism following relaxation of selection [38]. Furthermore, mitochondrial genes have recently been shown to be positively selected in another annual killifish, *A. limnaeus* [190]. Mitochondrial genes therefore may play an important role in the transition to annual environments. In addition, it has even been shown that the mitochondrial complex I is a modifier of lifespan in turquoise killifish [11], highlighting the connection between mitochondria and lifespan in turquoise killifish.

In summary, genes involved in mitochondrial and ribosomal function appear to play an important role in adaptations to the harsh environments. The potential role of mitochondrial and ribosomal genes in adaptations to the harsh environment of the turquoise killifish must be further supported by experimental studies in the future. The mitochondrial genes that showed signatures of positive selection in NF303 provide a good starting point for further studies.

4.2.2 Regions of exceptionally low genetic differentiation

Finally, F_{ST} outlier detection revealed two genomic regions with extremely low genetic differentiation among all populations studied. One region was found to be located on the sex chromosome and contains the putative sex-determining gene *gdf6* ([150], Figure 3.17). Low inter-populational genetic differentiation supports the importance of *gdf6* in a fundamental process such as sex determination. The second region, on chromosome 9, contains three genes *LOC107389593*, *cnot11* and *lcp1* (Figure 3.18). The exceptionally low genetic differentiation in this region between all investigated populations of turquoise killifish suggests that either the mentioned genes or potential non-coding elements of this region play an important role in the turquoise killifish. In the future, the potential role of the three candidate genes can be further explored using availabe expression dataset or gene editing techniques, which are already established in the turquoise killifish [70].

Limitations

A biological caveat of my study is that in populations from more wet regions, individual fitness could be more greatly affected by non-random (condition-dependent) mortality, such as sex-biased predation [20, 148, 188]. Non-random mortality complicates predictions made by the classical evolutionary theories of aging [31], because classical evolutionary theories are based on random (condition-independent) extrinsic mortality, such as habitat desiccation. For example, Chen *et al.* selected for more robust individuals by exposing the thermotolerant nematode *C. elegans* to high temperatures [31]. The non-random mortality source caused selection for more robust individuals, leading to the evolution of longer lifespan. In contrast, an experiment with increased random mortality led to the evolution of shorter lifespan when compared to the control group without increased random/non-random mortality [31]. This example illustrates that the source of mortality (random vs. non-random) must be taken into account when exploring the evolution of differences in lifespan and aging.

If non-random mortality played an important role in the evolution of differences in lifespan and aging between populations from dry and wet regions, one would expect to find a clear signature of positive selection in the wet population and strong genetic differentiation between the dry and wet populations in the positively selected genomic region. However, I could not find any such clear signatures between populations from wet and dry regions. Further, my findings indicate that relaxation of selection driven by demography rather than positive selection led to the evolution of shorter lifespan and accelerated aging in populations from more dry regions. Hence, despite possible limitations, I could show that the differences in lifespan are a result of demography in populations from more dry regions and dry regions are expected to be more influenced by random mortality [20] and therefore not conflicting with the basic predictions made by the classical evolutionary theories of aging.

Another technical limitation of this study is the "Pool-seq" approach in which samples from different individuals are pooled for sequencing. While "Pool-seq" offers a cost-effective approach to perform population genetics [166], the inference of rare SNPs can be biased by unequal contribution of single individuals, sequencing errors or low coverage [166]. To address these sources of error, I sequenced a higher number of individuals per population as recommended in Ferreti *et al.* [46] and performed additional recommended filtering steps [166]. Additionally, software designed for "Pool-seq" analysis was used to call SNPs [147] and SNPs with only one supporting read were treated as monomorphic. For the analysis of derived allele frequency, SNPs with a frequency below 5% or $\geq$ 95% were treated as fixed sites. Hence, despite potential limitations, I could show that "Pool-seq" is a powerful approach to perform population genetic analyses.

Chapter 5

Conclusions and future perspectives

The classical evolutionary theories of aging provide a plausible explanation for why organisms age. The mutation accumulation (MA, Section 1.2.1) and the antagonistic pleiotropy (AP, Section 1.2.2) theories of aging describe how aging can arise as a non-adaptive process, based on the fact that the force of natural selection decreases with age (Section 1.2). However, the contribution of either MA or AP to aging in natural vertebrate populations is still largely unexplored. Understanding the evolution of aging in nature can have a major impact on how we handle aging and age-related diseases in an aging society, such as ours (Section 1.1).

In this thesis, I used natural populations of turquoise killifish (*Nothobranchius furzeri*) to study the evolution of life history traits in a vertebrate species using a population genetics approach. The objectives of this thesis were twofold. The first objective was to find evolutionary forces that explain the differences in lifespan and aging in natural populations of turquoise killifish. I showed that populations from more arid regions - characterized by a shorter lifespan and accelerated aging - had a smaller effective population size (*Ne*) and a contrasting demographic history compared to populations from more wet regions (Sections 3.2, 3.3). In accordance with the nearly neutral theory of molecular evolution outlined in the Introduction (Section 1.2.3), I showed that natural selection in the population with smaller *Ne* was more relaxed (Sections 3.6.1, 3.6.3, 3.6.4) and resulted in an accumulation of deleterious mutations present at higher frequencies (Section 3.6.5). In Section 3.6.6, I showed that the deleterious mutations were overrepresented in aging or age-related disease pathways. Further, I could not find regions under positive selection explaining the differences in lifespan and aging (Sections 3.4.2, 3.5.2). The findings of my thesis showed evidence for MA but not AP contributing to the evolution of aging in natural populations of turquoise killifish. Further, the results showed that reduction in effective population size, *i.e* an increase in genetic drift, in more arid regions is the main driver of the evolution of differences in lifespan and aging. Therefore, my findings combine the nearly neutral theory of molecular evolution with the classical evolutionary MA theory of aging, highlighting the importance of relaxation of selection in phenotypic evolution. To my knowledge, this study merges for the first time the nearly neutral theory of molecular evolution with the classical evolutionary theories of aging to study aging in natural populations. Far from claiming that positive selection is not important, my work provides evidence that relaxation of selection importantly contributes to molecular and phenotypic evolution. In the case of the turquoise killifish populations, it would be interesting to investigate further whether the populations from drier regions would continue to reduce their *Ne* and whether this decrease would result in even shorter lifespans and accelerated aging. As mentioned in Section 4.1.2, a reduction in effective population size combined with a higher mutational load also occurs in human populations. Therefore, it will be interesting to explore potential differences in susceptibility to age-related diseases in human populations in light of relaxed selection.

The second objective of my thesis was to find biological pathways that were important for adaptations to the harsh environment in which natural populations of turquoise killifish live. I showed that, although genetic differentiation was generally high between populations (Section 3.1), genes involved in mitochondrial and ribosomal functions had low genetic differentiation between populations (Section 3.5.3), suggesting that mitochondrial and ribosomal pathways may play an important role in adaptations to the harsh environment. In Section 3.6.2, I showed that mitochondrial genes showed signatures of positive selection in the population with the largest *Ne*, stressing the importance of mitochondrial pathways in turquoise killifish and confirming that positive selection is stronger in populations with larger *Ne* (Sections 3.6.1, 3.6.3). Finding how mitochondrial and ribosomal genes are involved in adaptations to harsh environments is beyond this thesis aims. My goal was to identify important biological processes that might be involved in adaptations to the harsh environment. In addition, this study is the first population genetic study based on whole genome sequencing conducted in natural populations of turquoise killifish and together with the new genome assembly (from Dr. Rongfeng Cui, Section 2.1.2) used in this study provides valuable new resources for the research community. I hope that future studies will combine population genetics with molecular biology to functionally validate candidate genes involved in adaptations leading to the unique life cycle of the turquoise killifish. The results of my thesis showed that the life history evolution of turquoise killifish is a balance between early life and late life traits, between intensified selection and relaxed selection, and is amplified by demography.

References

[1] Abrams, P. A. (1993). Does increased mortality favor the evolution of more rapid senescence? *Evolution*, 47(3):877–887.

[2] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

[3] Austad, S. N. (1993). Retarded senescence in an insular population of virginia opossums (*Didelphis virginiana)*. *Journal of Zoology*, 229:695–708.

[4] Austad, S. N. (1997). Comparative aging and life histories in mammals. *Experimental Gerontology*, 32(1-2):23–38.

[5] Austad, S. N. and Hoffman, J. M. (2018). Is antagonistic pleiotropy ubiquitous in aging biology? *Evol Med Public Health*, 2018(1):287–294.

[6] Ballesteros, J. A. and Hormiga, G. (2016). A new orthology assessment method for phylogenomic data: unrooted phylogenetic orthology. *Molecular Biology and Evolution*, 33(8):2117–2134.

[7] Bao, E. and Lan, L. (2017). Halc: High throughput algorithm for long read error correction. *BMC Bioinformatics*, 18(1):204.

[8] Bartakova, V., Reichard, M., Blazek, R., Polacik, M., and Bryja, J. (2015). Terrestrial fishes: rivers are barriers to gene flow in annual fishes from the African savanna. *Journal of Biogeography*, 42(10):1832–1844.

[9] Bartakova, V., Reichard, M., Janko, K., Polacik, M., Blazek, R., Reichwald, K., Cellerino, A., and Bryja, J. (2013). Strong population genetic structuring in an annual fish, *Nothobranchius furzeri*, suggests multiple savannah refugia in southern Mozambique. *BMC Evolutionary Biology*, 13:196.

[10] Baumgart, M., Di Cicco, E., Rossi, G., Cellerino, A., and Tozzini, E. T. (2015). Comparison of captive lifespan, age-associated liver neoplasias and age-dependent gene expression between two annual fish species: *Nothobranchius furzeri* and *Nothobranchius korthause*. *Biogerontology*, 16(1):63–9.

[11] Baumgart, M., Priebe, S., Groth, M., Hartmann, N., Menzel, U., Pandolfini, L., Koch, P., Felder, M., Ristow, M., Englert, C., Guthke, R., Platzer, M., and Cellerino, A. (2016). Longitudinal RNA-Seq analysis of vertebrate aging identifies mitochondrial complex I as a small-molecule-sensitive modifier of lifespan. *Cell Systems*, 2(2):122–32.

[12] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 57(1):289–300.

[13] Biswas, S. and Akey, J. M. (2006). Genomic insights into positive selection. *Trends in Genetics*, 22(8):437–446.

[14] Blazek, R., Polacik, M., Kacer, P., Cellerino, A., Rezucha, R., Methling, C., Tomasek, O., Syslova, K., Terzibasi Tozzini, E., Albrecht, T., Vrtilek, M., and Reichard, M. (2017). Repeated intraspecific divergence in life span and aging of African annual fishes along an aridity gradient. *Evolution*, 71(2):386–402.

[15] Blazek, R., Polacik, M., and Reichard, M. (2013). Rapid growth, early maturation and short generation time in African annual fishes. *Evodevo*, 4(1):24.

[16] Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–20.

[17] Booker, T. R., Jackson, B. C., and Keightley, P. D. (2017). Detecting positive selection in the genome. *BMC Biology*, 15(1):98.

[18] Catchen, J. M., Conery, J. S., and Postlethwait, J. H. (2009). Automated identification of conserved synteny after whole-genome duplication. *Genome Research*, 19(8):1497–1505.

[19] Cellerino, A., Dolfi, L., Ripa, R., Antebi, A., and Valenzano, D. R. (2019). Cell cycle dynamics during diapause entry and exit in an annual killifish revealed by fucci technology. *bioRxiv*.

[20] Cellerino, A., Valenzano, D. R., and Reichard, M. (2016). From the bush to the bench: the annual *Nothobranchius* fishes as a new model system in biology. *Biological Reviews of the Cambridge Philosophical Society*, 91(2):511–33.

[21] Chakraborty, S., Panda, A., and Ghosh, T. C. (2016). Exploring the evolutionary rate differences between human disease and non-disease genes. *Genomics*, 108(1):18–24.

[22] Chang, W. (2014). *extrafont: Tools for using fonts*. R package version 0.17.

[23] Charlesworth, B. (1980). *Evolution in age structured populations, 1st*. Cambridge Univ. Press, Cambridge.

[24] Charlesworth, B. (1994). *Evolution in age structured populations, 2nd*. Cambridge Univ. Press, Cambridge.

[25] Charlesworth, B. (2000). Fisher, Medawar, Hamilton and the evolution of aging. *Genetics*, 156(3):927–931.

[26] Charlesworth, B. (2001). Patterns of age-specific means and genetic variances of mortality rates predicted by the mutation-accumulation theory of ageing. *Journal of Theoretical Biology*, 210(1):47–65.

[27] Charlesworth, B. (2009). Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10(3):195–205.

[28] Charlesworth, B. and Charlesworth, D. (2010). *Elements of Evolutionary Genetics*. Roberts and Company Publishers.

[29] Charlesworth, B. and Hughes, K. A. (1996). Age-specific inbreeding depression and components of genetic variance in relation to the evolution of senescence. *PNAS*, 93(12):6140–6145.

[30] Charlesworth, B., Morgan, M. T., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–303.

[31] Chen, H. Y. and Maklakov, A. A. (2012). Longer life span evolves under high rates of condition-dependent mortality. *Current Biology*, 22(22):2140–2143.

[32] Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92.

[33] Clancy, D. J., Gems, D., Harshman, L. G., Oldham, S., Stocker, H., Hafen, E., Leevers, S. J., and Partridge, L. (2001). Extension of life-span by loss of CHICO, a *Drosophila* insulin receptor substrate protein. *Science*, 292(5514):104–6.

[34] Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2015). Genbank. *Nucleic Acids Research*, 44(D1):D67–D72.

[35] Comfort, A. (1979). *The Biology of Senescence, 3rd Edn*. Churchill Livingstone, Edinburgh and London.

[36] Coombe, L., Zhang, J., Vandervalk, B. P., Chu, J., Jackman, S. D., Birol, I., and Warren, R. L. (2018). Arks: chromosome-scale scaffolding of human genome drafts with linked read kmers. *BMC Bioinformatics*, 19(1):234.

[37] Crick, F. H. (1968). The origin of the genetic code. *Journal of Molecular Biology*, 38(3):367–79.

[38] Cui, R., Medeiros, T., Willemsen, D., Iasi, L., Collier, G. E., Graef, M., Reichard, M., and Valenzano, D. R. (Submitted). Relaxed selection limits lifespan by increasing mutation load.

[39] Di Cicco, E., Tozzini, E. T., Rossi, G., and Cellerino, A. (2011). The short-lived annual fish *Nothobranchius furzeri* shows a typical teleost aging process reinforced by high incidence of age-dependent neoplasias. *Experimental Gerontology*, 46(4):249–56.

[40] Dobzhansky, T. (1973). Nothing in biology makes sense except in light of evolution. *American Biology Teacher*, 35(3):125–129.

[41] Dorn, A., Musilova, Z., Platzer, M., Reichwald, K., and Cellerino, A. (2014). The strange case of East African annual fishes: aridification correlates with diversification for a savannah aquatic group? *BMC Evolutionary Biology*, 14:210.

[42] Edney, E. B. and Gill, R. W. (1968). Evolution of senescence and specific longevity. *Nature*, 220(5164):281–2.

[43] Escobar, J. S., Jarne, P., Charmantier, A., and David, P. (2008). Outbreeding alleviates senescence in hermaphroditic snails as expected from the mutation-accumulation theory. *Current Biology*, 18(12):906–910.

[44] Eyre-Walker, A. and Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8(8):610–8.

[45] Fabian, D. and Flatt, T. (2011). The evolution of aging. *Nature Education Knowledge*, 3(10):9.

[46] Ferretti, L., Ramos-Onsins, S. E., and Perez-Enciso, M. (2013). Population genomics from pool sequencing. *Molecular Ecology*, 22(22):5561–76.

[47] Finch, C. E. (1994). *Longevity, Senescence, and the Genome*. University of Chicago Press.

[48] Fischer, M. C., Rellstab, C., Leuzinger, M., Roumet, M., Gugerli, F., Shimizu, K. K., Holderegger, R., and Widmer, A. (2017). Estimating genomic diversity and population differentiation - an empirical comparison of microsatellite and SNP variation in *Arabidopsis halleri*. *BMC Genomics*, 18(1):69.

[49] Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

[50] Flatt, T. and Partridge, L. (2018). Horizons in the evolution of aging. *BMC Biology*, 16(1):93.

[51] Fontana, L., Partridge, L., and Longo, V. D. (2010). Extending healthy life span-from yeast to humans. *Science*, 328(5976):321–326.

[52] Frick, N. T. and Wright, P. A. (2002). Nitrogen metabolism and excretion in the mangrove killifish *Rivulus marmoratus* I. The influence of environmental salinity and external ammonia. *Journal of Experimental Biology*, 205(Pt 1):79–89.

[53] Friedman, D. B. and Johnson, T. E. (1988). A mutation in the *age-1* gene in *Caenorhabditis elegans* lengthens life and reduces hermaphrodite fertility. *Genetics*, 118(1):75–86.

[54] Futuyma, D. J. and Kirkpatrick, M. (2018). *Evolution, 4th ed.* Oxford University Press Inc.

[55] Gavrilov, L. A. and Gavrilova, N. S. (2002). Evolutionary theories of aging and longevity. *Scientific World Journal*, 2:339–56.

[56] Gems, D., Sutton, A. J., Sundermeyer, M. L., Albert, P. S., King, K. V., Edgley, M. L., Larsen, P. L., and Riddle, D. L. (1998). Two pleiotropic classes of *daf-2* mutation affect larval arrest, adult behavior, reproduction and longevity in *Caenorhabditis elegans*. *Genetics*, 150(1):129–155.

[57] Genade, T., Benedetti, M., Terzibasi, E., Roncaglia, P., Valenzano, D. R., Cattaneo, A., and Cellerino, A. (2005). Annual fishes of the genus *Nothobranchius* as a model system for aging research. *Aging Cell*, 4(5):223–33.

[58] Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P., Sykes, S., Berlin, A. M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E. S., and Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *PNAS*, 108(4):1513–8.

[59] Gonskikh, Y. and Polacek, N. (2017). Alterations of the translation apparatus during aging and stress response. *Mechanisms of Ageing and Development*, 168:30–36.

[60] Gruman, G. J. (2003). *A History of Ideas About the Prolongation of Life*. Springer, Berlin.

[61] Guo, B., Li, Z., and Merila, J. (2016). Population genomic evidence for adaptive differentiation in the baltic sea herring. *Molecular Ecology*, 25(12):2833–52.

[62] Guy, L., Roat Kultima, J., and Andersson, S. G. E. (2010). genoPlotR: comparative gene and genome visualization in R. *Bioinformatics*, 26(18):2334–2335.

[63] Haldane, J. B. S. (1927). A mathematical theory of natural and artificial selection, Part V: Selection and mutation. *Proceedings of the Cambridge Philosophical Society*, 23:838–844.

[64] Haldane, J. B. S. (1932). *The Causes of Evolution*. Longmans, London.

[65] Haldane, J. B. S. (1941). *New Paths in Genetics*. George Allen and Unwin, London.

[66] Haller, B. C. and Messer, P. W. (2017). asymptoticMK: A web-based tool for the asymptotic McDonald-Kreitman test. *G3 (Bethesda)*, 7(5):1569–1575.

[67] Halligan, D. L. and Keightley, P. D. (2009). Spontaneous mutation accumulation studies in evolutionary genetics. *Annual Review of Ecology Evolution and Systematics*, 40:151–172.

[68] Hamilton, M. (2009). *Population Genetics*. Wiley-Blackwell, West Sussex.

[69] Hamilton, W. D. (1966). Moulding of senescence by natural selection. *Journal of Theoretical Biology*, 12(1):12–45.

[70] Harel, I., Valenzano, D. R., and Brunet, A. (2016). Efficient genome engineering approaches for the short-lived African turquoise killifish. *Nature Protocols*, 11(10):2010–2028.

[71] Hartl, D. and Clark, G. (2007). *Principles of Population Genetics, 4th ed.* Sinauer Associates, Sunderland.

[72] Henn, B. M., Botigue, L. R., Bustamante, C. D., Clark, A. G., and Gravel, S. (2015). Estimating the mutation load in human genomes. *Nature Reviews Genetics*, 16(6):333–343.

[73] Herwig, R., Hardt, C., Lienhard, M., and Kamburov, A. (2016). Analyzing and interpreting genome data at the network level with ConsensusPathDB. *Nature Protocols*, 11(10):1889.

[74] Hu, C. K. and Brunet, A. (2018). The African turquoise killifish: A research organism to study vertebrate aging and diapause. *Aging Cell*, 17(3):e12757.

[75] Huff, M. W. and Telford, D. E. (2005). Lord of the rings–the mechanism for oxidosqualene:lanosterol cyclase becomes crystal clear. *Trends in Pharmacological Sciences*, 26(7):335–40.

[76] Hughes, K. A. (2010). Mutation and the evolution of ageing: from biometrics to system genetics. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 365(1544):1273–1279.

[77] Hughes, K. A. and Reynolds, R. M. (2005). Evolutionary and mechanistic theories of aging. *Annual Review of Entomology*, 50:421–445.

[78] Hurst, L. D. (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. *TRENDS in Genetics*, 18(9):486–487.

[79] Jenkins, N. L., McColl, G., and Lithgow, G. J. (2004). Fitness cost of extended lifespan in *Caenorhabditis elegans*. *Proceedings of the Royal Society B-Biological Sciences*, 271(1556):2523–2526.

[80] Jensen, J. D., Payseur, B. A., Stephan, W., Aquadro, C. F., Lynch, M., Charlesworth, D., and Charlesworth, B. (2019). The importance of the neutral theory in 1968 and 50 years on: A response to Kern and Hahn 2018. *Evolution*, 73(1):111–114.

[81] Jubb, R. A. (1971). A new *Nothobranchius* (Pisces, Cyprinodontidae) from Southeastern Rhodesia. *Journal of the American Killifish Association*, 8:12–19.

[82] Kassambara, A. (2018). *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.1.8.

[83] Katju, V., Packard, L. B., Bu, L., Keightley, P. D., and Bergthorsson, U. (2015). Fitness decline in spontaneous mutation accumulation lines of *Caenorhabditis elegans* with varying effective population sizes. *Evolution*, 69(1):104–16.

[84] Katju, V., Packard, L. B., and Keightley, P. D. (2018). Fitness decline under osmotic stress in *Caenorhabditis elegans* populations subjected to spontaneous mutation accumulation at varying population sizes. *Evolution*, 72(4):1000–1008.

[85] Kenyon, C. (2011). The first long-lived mutants: discovery of the insulin/IGF-1 pathway for ageing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1561):9–16.

[86] Kenyon, C., Chang, J., Gensch, E., Rudner, A., and Tabtiang, R. (1993). A *C. elegans* mutant that lives twice as long as wild-type. *Nature*, 366(6454):461–464.

[87] Kim, Y., Nam, H. G., and Valenzano, D. R. (2016). The short-lived African turquoise killifish: an emerging experimental model for ageing. *Disease Models & Mechanisms*, 9(2):115–29.

[88] Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217(5129):624–6.

[89] Kimura, M. (1969). The rate of molecular evolution considered from the standpoint of population genetics. *PNAS*, 63(4):1181–8.

[90] Kimura, M. (1987). *A Stochastic Model of Compensatory Neutral Evolution. In Stochastic Methods in Biology, pp. 2-18*. Springer-Verlag, Berlin.

[91] King, J. L. and Jukes, T. H. (1969). Non-darwinian evolution. *Science*, 164(3881):788–98.

[92] Kirkwood, T. B. (2005). Understanding the odd science of aging. *Cell*, 120(4):437–47.

[93] Kirschner, J., Weber, D., Neuschl, C., Franke, A., Bottger, M., Zielke, L., Powalsky, E., Groth, M., Shagin, D., Petzold, A., Hartmann, N., Englert, C., Brockmann, G. A., Platzer, M., Cellerino, A., and Reichwald, K. (2012). Mapping of quantitative trait loci controlling lifespan in the short-lived fish *Nothobranchius furzeri*–a new vertebrate model for age research. *Aging Cell*, 11(2):252–61.

[94] Klass, M. R. (1983). A method for the isolation of longevity mutants in the nematode *Caenorhabditis elegans* and initial results. *Mechanisms of Ageing and Development*, 22(3-4):279–86.

[95] Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., Kosiol, C., and Schlotterer, C. (2011a). PoPoolation: A toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLOS ONE*, 6(1):e15925.

[96] Kofler, R., Pandey, R. V., and Schlotterer, C. (2011b). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, 27(24):3435–6.

[97] Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Research*, 27(5):722–736.

[98] Köster, J. and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522.

[99] Kosugi, S., Hirakawa, H., and Tabata, S. (2015). GMcloser: closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments. *Bioinformatics*, 31(23):3733–41.

[100] Kryazhimskiy, S. and Plotkin, J. B. (2008). The population genetics of dN/dS. *PLOS Genetics*, 4(12):e1000304.

[101] Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Research*, 19(9):1639–45.

[102] Kuningas, M., Magi, R., Westendorp, R. G. J., Slagboom, P. E., Remm, M., and van Heemst, D. (2007). Haplotypes in the human *Foxo1a* and *Foxo3a* genes; impact on disease and mortality at old age. *European Journal of Human Genetics*, 15(3):294–301.

[103] Lande, R. (1994). Risk of population extinction from fixation of new deleterious mutations. *Evolution*, 48(5):1460–1469.

[104] Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint*.

[105] Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics*, 26(5):589–95.

[106] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–9.

[107] Lohmueller, K. E., Indap, A. R., Schmidt, S., Boyko, A. R., Hernandez, R. D., Hubisz, M. J., Sninsky, J. J., White, T. J., Sunyaev, S. R., Nielsen, R., Clark, A. G., and Bustamante, C. D. (2008). Proportionally more deleterious genetic variation in European than in African populations. *Nature*, 451(7181):994–U5.

[108] Lohr, J. N., David, P., and Haag, C. R. (2014). Reduced lifespan and increased ageing driven by genetic drift in small populations. *Evolution*, 68(9):2494–508.

[109] Lopez-Otin, C., Blasco, M. A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. *Cell*, 153(6):1194–1217.

[110] Luckinbill, L. S., Arking, R., Clare, M. J., Cirocco, W. C., and Buck, S. A. (1984). Selection for delayed senescence in *Drosophila melanogaster*. *Evolution*, 38(5):996–1003.

[111] Lynch, M. (2006). The origins of eukaryotic gene structure. *Molecular Biology and Evolution*, 23(2):450–68.

[112] Lynch, M. and Gabriel, W. (1990). Mutation load and the survival of small populations. *Evolution*, 44(7):1725–1737.

[113] Mair, W. and University of, L. (2005). *Dietary restriction in Drosophila melanogaster*. University of London, London.

[114] Mank, J. E. and Avise, J. C. (2009). Evolutionary diversity and turn-over of sex determination in teleost fishes. *Sexual Development*, 3(2-3):60–7.

[115] Martinez, D. E. (1998). Mortality patterns suggest lack of senescence in *Hydra*. *Experimental Gerontology*, 33(3):217–225.

[116] Martins, H., Caye, K., Luu, K., Blum, M. G. B., and Francois, O. (2016). Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics. *Molecular Ecology*, 25(20):5029–5042.

[117] Mayr, E. (1997). The objects of selection. *PNAS*, 94(6):2091–2094.

[118] McDonald, J. H. and Kreitman, M. (1991). Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature*, 351(6328):652–4.

[119] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–303.

[120] Medawar, P. B. (1946). Old age and natural death. *The Modern Quarterly*, 2:30–56.

[121] Medawar, P. B. (1952). *An Unsolved Problem of Biology*. H.K. Lewis, London.

[122] Messer, P. W. and Petrov, D. A. (2013). Frequent adaptation and the McDonald-Kreitman test. *PNAS*, 110(21):8615–20.

[123] Mooijaart, S. P., Brandt, B. W., Baldal, E. A., Pijpe, J., Kuningas, M., Beekman, M., Zwaan, B. J., Slagboom, P. E., Westendorp, R. G. J., and van Heemst, D. (2005). *C. elegans* DAF-12, nuclear hormone receptors and human longevity and disease at old age. *Ageing Research Reviews*, 4(3):351–371.

[124] Nadachowska-Brzyska, K., Burri, R., Smeds, L., and Ellegren, H. (2016). PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white *Ficedula* flycatchers. *Molecular Ecology*, 25(5):1058–72.

[125] Nei, M. and Li, W. H. (1979). Mathematical-model for studying genetic-variation in terms of restriction endonucleases. *PNAS*, 76(10):5269–5273.

[126] Neteler, M., Bowman, M. H., Landa, M., and Metz, M. (2012). Grass gis: A multi-purpose open source gis. *Environmental Modelling & Software*, 31:124–130.

[127] Nussey, D. H., Froy, H., Lemaitre, J. F., Gaillard, J. M., and Austad, S. N. (2013). Senescence in natural populations of animals: widespread evidence and its implications for bio-gerontology. *Ageing Research Reviews*, 12(1):214–25.

[128] Oeppen, J. and Vaupel, J. W. (2002). Broken limits to life expectancy. *Science*, 296(5570):1029–31.

[129] Ohta, T. (1972). Population size and rate of evolution. *Journal of Molecular Evolution*, 1(4):305–14.

[130] Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature*, 246(5428):96–8.

[131] Ohta, T. (1974). Mutational pressure as main cause of molecular evolution and polymorphism. *Nature*, 252(5482):351–354.

[132] Ohta, T. (1976). Role of very slightly deleterious mutations in molecular evolution and polymorphism. *Theoretical Population Biology*, 10(3):254–275.

[133] Ohta, T. (1992). The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*, 23:263–286.

[134] Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290.

[135] Partridge, L. and Barton, N. H. (1993). Optimality, mutation and the evolution of ageing. *Nature*, 362(6418):305–11.

[136] Partridge, L. and Barton, N. H. (1996). On measuring the rate of ageing. *Proceedings of the Royal Society B-Biological Sciences*, 263(1375):1365–1371.

[137] Partridge, L. and Fowler, K. (1992). Direct and correlated responses to selection on age at reproduction in *Drosophila melanogaster*. *Evolution*, 46(1):76–91.

[138] Partridge, L. and Gems, D. (2002). Mechanisms of ageing: Public or private? *Nature Reviews Genetics*, 3(3):165–175.

[139] Partridge, L., Thornton, J., and Bates, G. (2015). The new science of ageing. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 370(1676).

[140] Piper, M. D., Selman, C., McElwee, J. J., and Partridge, L. (2008). Separating cause from effect: how does insulin/IGF signalling control lifespan in worms, flies and mice? *J Intern Med*, 263(2):179–91.

[141] Polacik, M., Blazek, R., and Reichard, M. (2016). Laboratory breeding of the short-lived annual killifish *Nothobranchius furzeri*. *Nature Protocols*, 11(8):1396–413.

[142] Pruisscher, P., Nylin, S., Gotthard, K., and Wheat, C. W. (2018). Genetic variation underlying local adaptation of diapause induction along a cline in a butterfly. *Molecular Ecology*.

[143] QGIS Development Team (2009). *QGIS Geographic Information System*. Open Source Geospatial Foundation.

[144] Quinlan, A. R. (2014). BEDTools: The Swiss-army tool for genome feature analysis. *Curr Protoc Bioinformatics*, 47:11 12 1–34.

[145] Quinlan, A. R. and Hall, I. M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–2.

[146] R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[147] Raineri, E., Ferretti, L., Esteve-Codina, A., Nevado, B., Heath, S., and Perez-Enciso, M. (2012). SNP calling by sequencing pooled samples. *BMC Bioinformatics*, 13:239.

[148] Reichard, M., Polacik, M., Blazek, R., and Vrtilek, M. (2014). Female bias in the adult sex ratio of African annual fishes: interspecific differences, seasonal trends and environmental predictors. *Evolutionary Ecology*, 28(6):1105–1120.

[149] Reichwald, K., Lauber, C., Nanda, I., Kirschner, J., Hartmann, N., Schories, S., Gausmann, U., Taudien, S., Schilhabel, M. B., Szafranski, K., Glockner, G., Schmid, M., Cellerino, A., Schartl, M., Englert, C., and Platzer, M. (2009). High tandem repeat content in the genome of the short-lived annual fish *Nothobranchius furzeri*: a new vertebrate model for aging research. *Genome Biology*, 10(2):R16.

[150] Reichwald, K., Petzold, A., Koch, P., Downie, B. R., Hartmann, N., Pietsch, S., Baumgart, M., Chalopin, D., Felder, M., Bens, M., Sahm, A., Szafranski, K., Taudien, S., Groth, M., Arisi, I., Weise, A., Bhatt, S. S., Sharma, V., Kraus, J. M., Schmid, F., Priebe, S., Liehr, T., Gorlach, M., Than, M. E., Hiller, M., Kestler, H. A., Volff, J. N., Schartl, M., Cellerino, A., Englert, C., and Platzer, M. (2015). Insights into sex chromosome evolution and aging from the genome of a short-lived fish. *Cell*, 163(6):1527–38.

[151] Reznick, D. A., Bryga, H., and Endler, J. A. (1990). Experimentally induced life-history evolution in a natural population. *Nature*, 346(6282):357–359.

[152] Reznick, D. N., Bryant, M. J., Roff, D., Ghalambor, C. K., and Ghalambor, D. E. (2004). Effect of extrinsic mortality on the evolution of senescence in guppies. *Nature*, 431(7012):1095–1099.

[153] Ricklefs, R. E. (1998). Evolutionary theories of aging: Confirmation of a fundamental prediction, with implications for the genetic basis and evolution of life span. *American Naturalist*, 152(1):24–44.

[154] Romney, A. L. T., Davis, E. M., Corona, M. M., Wagner, J. T., and Podrabsky, J. E. (2018). Temperature-dependent vitamin d signaling regulates developmental trajectory associated with diapause in an annual killifish. *PNAS*, 115(50):12763–12768.

[155] Rose, M. and Charlesworth, B. (1980). A test of evolutionary theories of senescence. *Nature*, 287(5778):141–2.

[156] Rose, M. R. (1984). Laboratory evolution of postponed senescence in *Drosophila melanogaster*. *Evolution*, 38(5):1004–1010.

[157] Rose, M. R. (1991). *Evolutionary Biology of Aging*. Oxford University Press, New York.

[158] Rose, M. R. and Charlesworth, B. (1981). Genetics of life history in *Drosophila melanogaster*. II. Exploratory selection experiments. *Genetics*, 97(1):187–96.

[159] Rose, M. R., Rauser, C. L., Benford, G., Matos, M., and Mueller, L. D. (2007). Hamilton's forces of natural selection after forty years. *Evolution*, 61(6):1265–1276.

[160] Rowan, B. A., Patel, V., Weigel, D., and Schneeberger, K. (2015). Rapid and inexpensive whole-genome genotyping-by-sequencing for crossover localization and fine-scale genetic mapping. *G3 (Bethesda)*, 5(3):385–98.

[161] RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

[162] Rubin, C. J., Zody, M. C., Eriksson, J., Meadows, J. R., Sherwood, E., Webster, M. T., Jiang, L., Ingman, M., Sharpe, T., Ka, S., Hallbook, F., Besnier, F., Carlborg, O., Bed'hom, B., Tixier-Boichard, M., Jensen, P., Siegel, P., Lindblad-Toh, K., and Andersson, L. (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature*, 464(7288):587–91.

[163] Russell, S. J. and Kahn, C. R. (2007). Endocrine regulation of ageing. *Nature Reviews Molecular Cell Biology*, 8(9):681.

[164] Sahm, A., Bens, M., Platzer, M., and Cellerino, A. (2017). Parallel evolution of genes controlling mitonuclear balance in short-lived annual fishes. *Aging Cell*, 16(3):488–496.

[165] Schiffels, S. and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8):919–25.

[166] Schlotterer, C., Tobler, R., Kofler, R., and Nolte, V. (2014). Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, 15(11):749–63.

[167] Sherman, P. W. and Jarvis, J. U. M. (2002). Extraordinary life spans of naked mole-rats (*Heterocephalus glaber*). *Journal of Zoology*, 258:307–311.

[168] Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–2.

[169] Simons, Y. B., Turchin, M. C., Pritchard, J. K., and Sella, G. (2014). The deleterious mutation load is insensitive to recent population history. *Nature Genetics*, 46(3):220–224.

[170] Slater, G. S. and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6.

[171] Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O., Awedh, M. H., Baldock, R., Barbiera, G., Bardou, P., Beck, T., Blake, A., Bonierbale, M., Brookes, A. J., Bucci, G., Buetti, I., Burge, S., Cabau, C., Carlson, J. W., Chelala, C., Chrysostomou, C., Cittaro, D., Collin, O., Cordova, R., Cutts, R. J., Dassi, E., Di Genova, A., Djari, A., Esposito, A., Estrella, H., Eyras, E., Fernandez-Banet, J., Forbes, S., Free, R. C., Fujisawa, T., Gadaleta, E., Garcia-Manteiga, J. M., Goodstein, D., Gray, K., Guerra-Assuncao, J. A., Haggarty, B., Han, D. J., Han, B. W., Harris, T., Harshbarger, J., Hastings, R. K., Hayes, R. D., Hoede, C., Hu, S., Hu, Z. L., Hutchins, L., Kan, Z., Kawaji, H., Keliet, A., Kerhornou, A., Kim, S., Kinsella, R., Klopp, C., Kong, L., Lawson, D., Lazarevic, D., Lee, J. H., Letellier, T., Li, C. Y., Lio, P., Liu, C. J., Luo, J., Maass, A., Mariette, J., Maurel, T., Merella, S., Mohamed, A. M., Moreews, F., Nabihoudine, I., Ndegwa, N., Noirot, C., Perez-Llamas, C., Primig, M., Quattrone, A., Quesneville, H., Rambaldi, D., Reecy, J., Riba, M., Rosanoff, S., Saddiq, A. A., Salas, E., Sallou, O., Shepherd, R., Simon, R., Sperling, L., Spooner, W., Staines, D. M., Steinbach, D., Stone, K., Stupka, E., Teague, J. W., Dayem Ullah, A. Z., Wang, J., Ware, D., et al. (2015). The biomart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research*, 43(W1):W589–98.

[172] Smith, N. G. and Eyre-Walker, A. (2002). Adaptive protein evolution in *Drosophila*. *Nature*, 415(6875):1022–4.

[173] Stearns, S. C. (2000). Life history evolution: successes, limitations, and prospects. *Naturwissenschaften*, 87(11):476–86.

[174] Stearns, S. C., Ackermann, M., Doebeli, M., and Kaiser, M. (2000). Experimental evolution of aging, growth, and reproduction in fruitflies. *PNAS*, 97(7):3309–13.

[175] Stoletzki, N. and Eyre-Walker, A. (2011). Estimation of the neutrality index. *Molecular Biology and Evolution*, 28(1):63–70.

[176] Taguchi, A. and White, M. F. (2008). Insulin-like signaling, nutrient homeostasis, and life span. *Annual Review of Physiology*, 70:191–212.

[177] Tajima, F. (1989). Statistical-method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595.

[178] Tang, H., Zhang, X., Miao, C., Zhang, J., Ming, R., Schnable, J. C., Schnable, P. S., Lyons, E., and Lu, J. (2015). ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biology*, 16:3.

[179] Tatar, M., Kopelman, A., Epstein, D., Tu, M. P., Yin, C. M., and Garofalo, R. S. (2001). A mutant *Drosophila* insulin receptor homolog that extends life-span and impairs neuroendocrine function. *Science*, 292(5514):107–10.

[180] Tataru, P., Mollion, M., Glemin, S., and Bataillon, T. (2017). Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. *Genetics*, 207(3):1103–1119.

[181] Terzibasi, E., Valenzano, D. R., Benedetti, M., Roncaglia, P., Cattaneo, A., Domenici, L., and Cellerino, A. (2008). Large differences in aging phenotype between strains of the short-lived annual fish *Nothobranchius furzeri*. *PLOS ONE*, 3(12):e3866.

[182] Tiku, V. and Antebi, A. (2018). Nucleolar function in lifespan regulation. *Trends in Cell Biology*, 28(8):662–672.

[183] Valdesalici, S. and Cellerino, A. (2003). Extremely short lifespan in the annual fish *Nothobranchius furzeri*. *Proceedings of the Royal Society B-Biological Sciences*, 270 Suppl 2:S189–91.

[184] Valenzano, D. R., Aboobaker, A., Seluanov, A., and Gorbunova, V. (2017). Non-canonical aging model systems and why we need them. *EMBO J*, 36(8):959–963.

[185] Valenzano, D. R., Benayoun, B. A., Singh, P. P., Zhang, E., Etter, P. D., Hu, C. K., Clement-Ziza, M., Willemsen, D., Cui, R., Harel, I., Machado, B. E., Yee, M. C., Sharp, S. C., Bustamante, C. D., Beyer, A., Johnson, E. A., and Brunet, A. (2015). The African turquoise killifish genome provides insights into evolution and genetic architecture of lifespan. *Cell*, 163(6):1539–54.

[186] Valenzano, D. R., Terzibasi, E., Genade, T., Cattaneo, A., Domenici, L., and Cellerino, A. (2006). Resveratrol prolongs lifespan and retards the onset of age-related markers in a short-lived vertebrate. *Current Biology*, 16(3):296–300.

[187] Van Voorhies, W. A., Fuchs, J., and Thomas, S. (2005). The longevity of caenorhabditis elegans in soil. *Biology Letters*, 1(2):247–249.

[188] Vrtilek, M., Zak, J., Polacik, M., Blazek, R., and Reichard, M. (2018a). Longitudinal demographic study of wild populations of African annual killifish. *Scientific Reports*, 8(1):4774.

[189] Vrtilek, M., Zak, J., Psenicka, M., and Reichard, M. (2018b). Extremely rapid maturation of a wild African annual fish. *Current Biology*, 28(15):R822–R824.

[190] Wagner, J. T., Singh, P. P., Romney, A. L., Riggs, C. L., Minx, P., Woll, S. C., Roush, J., Warren, W. C., Brunet, A., and Podrabsky, J. E. (2018). The genome of *Austrofundulus limnaeus* offers insights into extreme vertebrate stress tolerance and embryonic development. *BMC Genomics*, 19.

[191] Walker, D. W., McColl, G., Jenkins, N. L., Harris, J., and Lithgow, G. J. (2000). Evolution of lifespan in *C. lelegans*. *Nature*, 405(6784):296–297.

[192] Warren, R. L., Yang, C., Vandervalk, B. P., Behsaz, B., Lagman, A., Jones, S. J., and Birol, I. (2015). Links: Scalable, alignment-free scaffolding of draft genomes with long reads. *Gigascience*, 4:35.

[193] Watters, B. (2009). The ecology and distribution of *Nothobranchius* fishes. *Journal of the American Killifish Association*, 42:37–76.

[194] Watterson, G. A. (1975). Number of segregating sites in genetic models without recombination. *Theoretical Population Biology*, 7(2):256–276.

[195] Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., and Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome Research*, 27(5):757–767.

[196] Weismann, A. (1891). *Essays Upon Heredity and Kindred Biological Problems, Vol. 1*. Clarendon Press, Oxford.

[197] Weismann, A. (1892). *Essays Upon Heredity and Kindred Biological Problems, Vol. 2*. Clarendon Press, Oxford.

[198] Wences, A. H. and Schatz, M. C. (2015). Metassembler: merging and optimizing de novo genome assemblies. *Genome Biology*, 16:207.

[199] Westerfield, M. (1995). *The zebrafish book: a guide for the laboratory use of zebrafish (Brachydanio rerio)*. University of Oregon press.

[200] Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

[201] Wildekamp, R. (2004). *A World of Killies. Atlas of the Oviparous Cyprinodontiform Fishes of the World*, volume 4. The American Killifish Association.

[202] Wilicox, B. J., Donlon, T. A., He, Q., Chen, R., Grove, J. S., Yano, K., Masaki, K. H., Willcox, D. C., Rodriguez, B., and Curb, J. D. (2008). FOXO3A genotype is strongly associated with human longevity. *PNAS*, 105(37):13987–13992.

[203] Williams, G. C. (1957). Pleiotropy, natural selection, and the evolution of senescence. *Evolution*, 11(4):398–411.

[204] Wilmoth, J. R. (2000). Demography of longevity: past, present, and future trends. *Experimental Gerontology*, 35(9-10):1111–29.

[205] Wourms, J. P. (1972). The developmental biology of annual fishes. III. Pre-embryonic and embryonic diapause of variable duration in the eggs of annual fishes. *Journal of Experimental Zoology*, 182(3):389–414.

[206] Wright, S. (1921). Systems of mating. *Genetics*, 6(2):111–178.

[207] Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16(2):0097–0159.

[208] Zeileis, A. and Grothendieck, G. (2005). zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, 14(6):1–27.

[209] Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Giron, C. G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., To, J. K., Laird, M. R., Lavidas, I., Liu, Z., Loveland, J. E., Maurel, T., McLaren, W., Moore, B., Mudge, J., Murphy, D. N., Newman, V., Nuhn, M., Ogeh, D., Ong, C. K., Parker, A., Patricio, M., Riat, H. S., Schuilenburg, H., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Zadissa, A., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken, B. L., Cunningham, F., Yates, A., and Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761.

Appendix A

Comparison of `MSMC2` estimates

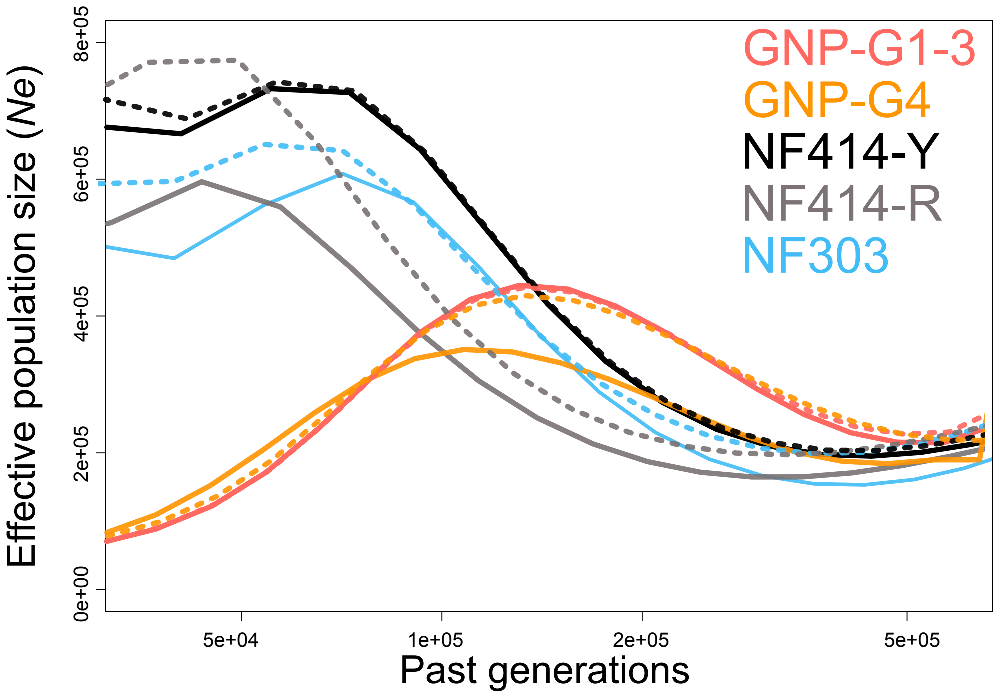

Figure A.1: Comparison of `MSMC2` estimates. Estimated ancestral effective population size *Ne* based on MSMC2 analysis. Time in past generations on the x-axis and the ancestral *Ne* on the y-axis. Comparison between the inference with a coverage threshold of 18 per chromosome (dotted lines) and the population specific mean coverage per chromosome (solid lines). Inference of population NF303 in lightblue; Population GNP in red (GNP-G1-3) and orange (GNP-G4); Population NF414 in black (yellow male, NF414-Y) and grey (red male, NF414-R).

Appendix B

Synteny medaka and platyfish

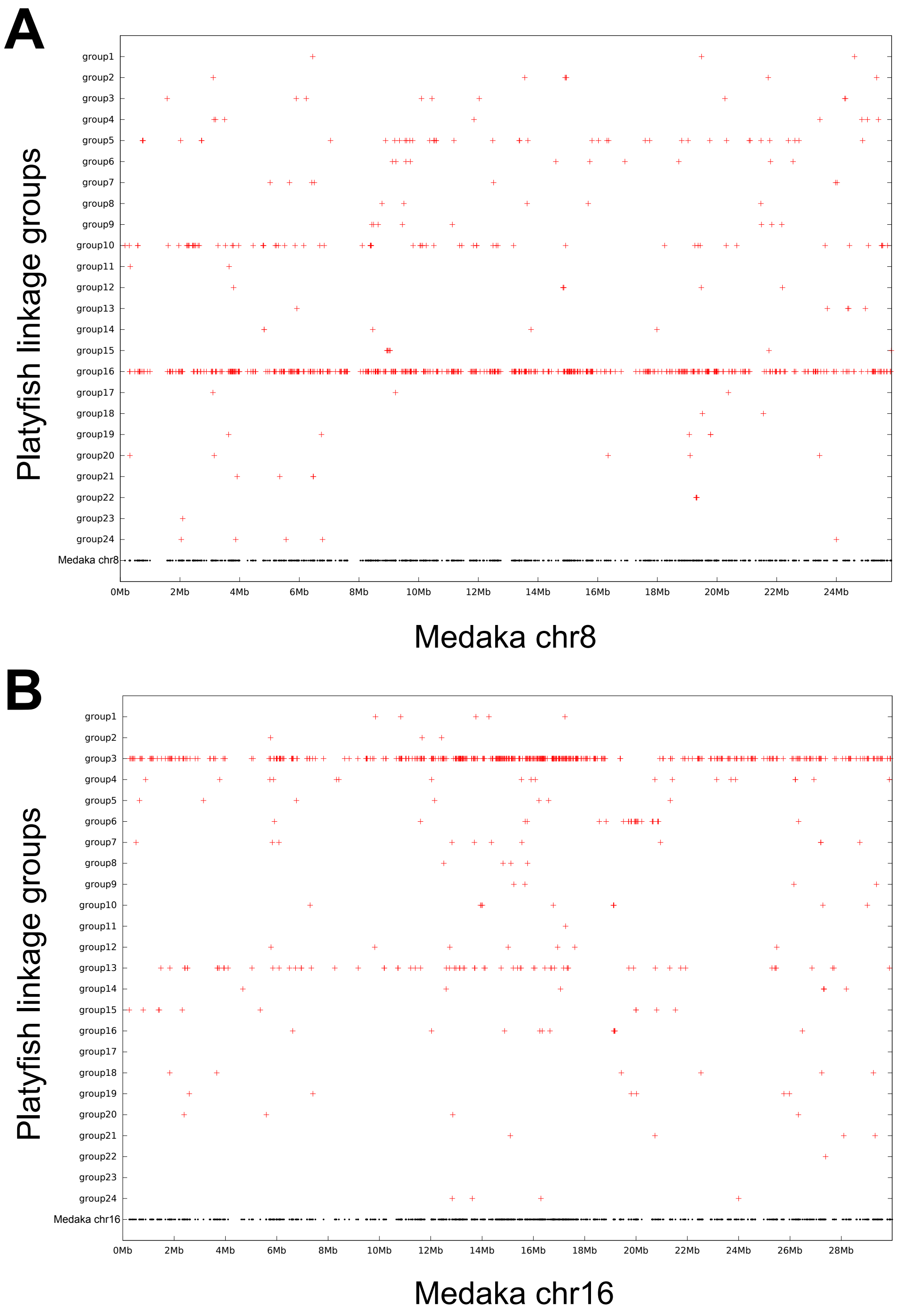

Figure B.1: Synteny medaka and platyfish. Synteny dot plot of A) Medaka chromosome 8 to all platyfish linkage groups; B) Medaka chromosome 16 to all platyfish linkage groups. Plots generated with http://syntenydb.uoregon.edu [18].

Appendix C

F_{ST} overrepresentation analysis

Table C.1: Low F_{ST} overrepresentation: GNP and NF414 (q-value < 0.05)

***P*-value**	**q-value**	**GO ID**	**Category**	**Level**	**Name**
7.3e-14	6.2e-12	GO:0044391	c	3	ribosomal subunit
6.7e-13	4.6e-11	GO:1990904	c	2	ribonucleoprotein complex
2.6e-12	1.1e-10	GO:0005840	c	3	ribosome
1.4e-11	4.8e-10	GO:0003735	m	2	structural constituent of ribosome
2.4e-10	2.2e-08	GO:0022626	c	4	cytosolic ribosome
2.5e-09	9.3e-07	GO:0043604	b	5	amide biosynthetic process
3.7e-09	1.7e-07	GO:0044445	c	4	cytosolic part
6.2e-09	1.9e-07	GO:0015934	c	4	large ribosomal subunit
7.4e-09	1.9e-06	GO:0006413	b	3	translational initiation
1.0e-08	1.9e-06	GO:0006401	b	5	RNA catabolic process
2.5e-08	1.2e-05	GO:0043603	b	4	cellular amide metabolic process
4.1e-08	5.1e-06	GO:0019083	b	5	viral transcription
8.4e-08	1.1e-05	GO:0022613	b	3	ribonucleoprotein complex biogenesis
1.2e-07	1.1e-05	GO:0006412	b	5	translation
1.6e-07	1.2e-05	GO:0019080	b	5	viral gene expression
1.9e-07	1.2e-05	GO:0043043	b	5	peptide biosynthetic process
2.1e-07	4.8e-05	GO:0042254	b	4	ribosome biogenesis
4.8e-07	7.5e-05	GO:0006518	b	4	peptide metabolic process

7.0e-07	2.4e-05	GO:0098798	c	2	mitochondrial protein complex
1.9e-06	2.3e-04	GO:0090150	b	4	establishment of protein localization to membrane
2.6e-06	7.3e-05	GO:0022625	c	5	cytosolic large ribosomal subunit
3.6e-06	3.1e-04	GO:0034641	b	3	cellular nitrogen compound metabolic process
4.4e-06	1.0e-04	GO:0015935	c	4	small ribosomal subunit
5.5e-06	5.1e-04	GO:0034655	b	4	nucleobase-containing compound catabolic process
7.8e-06	5.7e-04	GO:0044270	b	4	cellular nitrogen compound catabolic process
8.5e-06	5.7e-04	GO:0046700	b	4	heterocycle catabolic process
9.6e-06	1.3e-04	GO:0022627	c	5	cytosolic small ribosomal subunit
1.3e-05	7.4e-04	GO:0019439	b	4	aromatic compound catabolic process
1.5e-05	7.4e-04	GO:1901566	b	4	organonitrogen compound biosynthetic process
1.6e-05	7.4e-04	GO:0044271	b	4	cellular nitrogen compound biosynthetic process
2.4e-05	1.0e-03	GO:1901361	b	4	organic cyclic compound catabolic process
5.8e-05	4.4e-03	GO:0003723	m	4	RNA binding
7.3e-05	2.8e-03	GO:0044265	b	4	cellular macromolecule catabolic process
7.9e-05	2.2e-03	GO:0098800	c	3	inner mitochondrial membrane protein complex
1.0e-04	5.5e-03	GO:0006414	b	5	translational elongation
2.3e-04	4.2e-03	GO:0005743	c	4	mitochondrial inner membrane
2.7e-04	1.3e-02	GO:0042773	b	5	ATP synthesis coupled electron transport
3.5e-04	1.2e-02	GO:0034645	b	4	cellular macromolecule biosynthetic process
3.7e-04	8.4e-03	GO:0000313	c	2	organellar ribosome
3.7e-04	7.8e-03	GO:0005761	c	3	mitochondrial ribosome
4.0e-04	1.3e-02	GO:0006119	b	4	oxidative phosphorylation
4.4e-04	2.9e-02	GO:0006139	b	3	nucleobase-containing compound metabolic process
4.4e-04	1.4e-02	GO:0016032	b	4	viral process
5.7e-04	8.7e-03	GO:0005740	c	4	mitochondrial envelope
6.1e-04	1.0e-02	GO:0005762	c	3	mitochondrial large ribosomal subunit
6.1e-04	5.7e-03	GO:0000315	c	5	organellar large ribosomal subunit
7.1e-04	1.0e-02	GO:0019866	c	3	organelle inner membrane

7.7e-04	2.2e-02	GO:0009059	b	4	macromolecule biosynthetic process
7.8e-04	3.3e-02	GO:0042255	b	5	ribosome assembly
7.9e-04	2.2e-02	GO:0010467	b	4	gene expression
8.3e-04	2.2e-02	GO:0022904	b	4	respiratory electron transport chain
8.9e-04	3.3e-02	GO:0022618	b	5	ribonucleoprotein complex assembly
9.3e-04	4.7e-02	GO:0046483	b	3	heterocycle metabolic process
1.1e-03	2.7e-02	GO:0072594	b	4	establishment of protein localization to organelle
1.1e-03	4.7e-02	GO:0044403	b	3	symbiont process
1.2e-03	2.8e-02	GO:0071826	b	4	ribonucleoprotein complex subunit organization
1.2e-03	1.5e-02	GO:0043233	c	2	organelle lumen
1.2e-03	1.3e-02	GO:0070013	c	3	intracellular organelle lumen
1.2e-03	3.9e-02	GO:0043097	b	5	pyrimidine nucleoside salvage
1.2e-03	1.5e-02	GO:0070469	c	2	respiratory chain
1.2e-03	1.3e-02	GO:1903561	c	3	extracellular vesicle
1.2e-03	2.8e-02	GO:0045333	b	4	cellular respiration
1.3e-03	3.9e-02	GO:0043624	b	5	cellular protein complex disassembly
1.3e-03	1.5e-02	GO:0043230	c	2	extracellular organelle
1.4e-03	4.7e-02	GO:0006725	b	3	cellular aromatic compound metabolic process
1.5e-03	1.4e-02	GO:1905368	c	3	peptidase complex
1.5e-03	4.7e-02	GO:1901360	b	3	organic cyclic compound metabolic process
1.6e-03	4.7e-02	GO:0044249	b	3	cellular biosynthetic process
1.7e-03	4.8e-02	GO:0033365	b	5	protein localization to organelle
1.7e-03	3.6e-02	GO:0008655	b	4	pyrimidine-containing compound salvage
1.7e-03	1.5e-02	GO:0071013	c	3	catalytic step 2 spliceosome
1.8e-03	2.3e-02	GO:0070062	c	4	extracellular exosome
1.9e-03	3.8e-02	GO:0009057	b	4	macromolecule catabolic process
2.0e-03	2.0e-02	GO:0044455	c	2	mitochondrial membrane part
2.0e-03	2.3e-02	GO:0005689	c	4	U12-type spliceosomal complex
2.2e-03	1.7e-02	GO:0044444	c	3	cytoplasmic part
2.8e-03	1.9e-02	GO:0098803	c	3	respiratory chain complex
2.9e-03	2.5e-02	GO:0044446	c	2	intracellular organelle part

3.0e-03	1.9e-02	GO:0031966	c	3	mitochondrial membrane
3.0e-03	3.0e-02	GO:0005753	c	4	mitochondrial proton-transporting ATP synthase complex
3.8e-03	2.3e-02	GO:0045259	c	3	proton-transporting ATP synthase complex
4.3e-03	2.4e-02	GO:0005746	c	3	mitochondrial respiratory chain
5.1e-03	3.5e-02	GO:0016592	c	2	mediator complex
5.4e-03	2.9e-02	GO:0005681	c	3	spliceosomal complex
7.1e-03	3.3e-02	GO:0043231	c	3	intracellular membrane-bounded organelle
7.4e-03	3.3e-02	GO:0016469	c	3	proton-transporting two-sector ATPase complex
7.6e-03	4.4e-02	GO:0005615	c	2	extracellular space
9.5e-03	5.0e-02	GO:0031975	c	2	envelope
9.5e-03	4.1e-02	GO:0031967	c	3	organelle envelope

Table C.2: Low F_{ST} overrepresentation: GNP and NF303 (q-value < 0.05)

P-value	q-value	GO ID	Category	Level	Name
1.2e-07	7.6e-06	GO:0098798	c	2	mitochondrial protein complex
2.5e-06	7.9e-05	GO:0003735	m	2	structural constituent of ribosome
2.5e-06	1.9e-04	GO:0005840	c	3	ribosome
1.3e-05	1.1e-03	GO:0005743	c	4	mitochondrial inner membrane
1.5e-05	5.7e-04	GO:0044391	c	3	ribosomal subunit
2.9e-05	7.3e-04	GO:0098800	c	3	inner mitochondrial membrane protein complex
4.8e-05	9.1e-04	GO:0019866	c	3	organelle inner membrane
1.1e-04	3.8e-03	GO:0022626	c	4	cytosolic ribosome
1.3e-04	3.8e-03	GO:0015935	c	4	small ribosomal subunit
1.6e-04	3.2e-03	GO:0000313	c	2	organellar ribosome
1.6e-04	2.4e-03	GO:0005761	c	3	mitochondrial ribosome
1.8e-04	3.2e-03	GO:1990204	c	2	oxidoreductase complex
1.9e-04	2.4e-03	GO:0044429	c	3	mitochondrial part
1.9e-04	3.2e-03	GO:0044455	c	2	mitochondrial membrane part
2.3e-04	4.9e-03	GO:0005740	c	4	mitochondrial envelope
2.5e-04	2.7e-03	GO:0031966	c	3	mitochondrial membrane
5.8e-04	1.9e-02	GO:0022627	c	5	cytosolic small ribosomal subunit
1.0e-03	9.7e-03	GO:0030964	c	3	NADH dehydrogenase complex
1.0e-03	1.3e-02	GO:0005747	c	4	mitochondrial respiratory chain complex I
1.0e-03	1.3e-02	GO:0045271	c	4	respiratory chain complex I
1.1e-03	1.3e-02	GO:0005739	c	4	mitochondrion
1.5e-03	1.3e-02	GO:0098803	c	3	respiratory chain complex
2.4e-03	1.8e-02	GO:0005746	c	3	mitochondrial respiratory chain
3.3e-03	4.3e-02	GO:0070469	c	2	respiratory chain
3.4e-03	3.6e-02	GO:0044445	c	4	cytosolic part

Table C.3: Low F_{ST} overrepresentation: NF303 and NF414 (q-value < 0.05)

***P*-value**	**q-value**	**GO ID**	**Category**	**Level**	**Name**
1.3e-04	1.1e-02	GO:0005739	c	4	mitochondrion
1.6e-04	6.7e-03	GO:0008121	m	4	ubiquinol-cytochrome-c reductase activity
1.6e-04	6.7e-03	GO:0016681	m	4	oxidoreductase activity, acting on diphenols and related substances as donors, cytochrome as acceptor
2.7e-04	2.0e-02	GO:0016679	m	3	oxidoreductase activity, acting on diphenols and related substances as donors
2.9e-04	8.3e-03	GO:0003723	m	4	RNA binding
4.3e-04	3.5e-02	GO:0045275	c	3	respiratory chain complex III
4.3e-04	1.8e-02	GO:0005750	c	4	mitochondrial respiratory chain complex III
5.7e-04	2.7e-02	GO:0031625	m	5	ubiquitin protein ligase binding
6.2e-04	4.3e-02	GO:0070069	c	2	cytochrome complex
7.0e-04	2.4e-02	GO:0003735	m	2	structural constituent of ribosome
7.4e-04	1.6e-02	GO:0044389	m	4	ubiquitin-like protein ligase binding
1.4e-03	4.8e-02	GO:1990904	c	2	ribonucleoprotein complex
1.4e-03	2.4e-02	GO:0001076	m	4	transcription factor activity, RNA polymerase II transcription factor binding

Table C.4: High F_{ST} overrepresentation: NF303 and GNP (q-value < 0.05)

***P*-value**	**q-value**	**GO ID**	**Category**	**Level**	**Name**
7.1e-05	2.3e-02	GO:0034641	b	3	cellular nitrogen compound metabolic process
3.5e-04	2.8e-02	GO:0016779	m	4	nucleotidyltransferase activity
7.5e-04	2.8e-02	GO:0036041	m	4	long-chain fatty acid binding
9.2e-04	3.6e-02	GO:1901567	m	2	fatty acid derivative binding
2.2e-03	4.3e-02	GO:0003735	m	2	structural constituent of ribosome

Appendix D

Rate of molecular evolution

Table D.1: Substitutions in synonymous and non-synonymous sites

Outgroup	**Type**	**GNP**	**NF414**	**NF303**
N. orthonotus	synonymous	89766	80918	36671
N. orthonotus	non-synonymous	83419	74857	34404
N. rachovii	synonymous	130571	122396	91308
N. rachovii	non-synonymous	141305	134409	106112
Total number of sites	synonymous	5270947	5313481	5306770
Total number of sites	non-synonymous	22086652	22267506	22233200

Table D.2: Efficiency of natural selection in turquoise killifish populations

Index	**Species**	**GNP**	**NF414**	**NF303**
DoS median	***N. orthonotus***	-0.167	-0.021	-0.014
classical MK-α	***N. orthonotus***	-0.385	-0.123	-0.059
asymptotic MK-α	***N. orthonotus***	-0.211	-0.027	0.042
DoS median	***N. rachovii***	-0.143	0.000	0.000
classical MK-α	***N. rachovii***	-0.174	0.063	0.151
asymptotic MK-α	***N. rachovii***	-0.057	0.149	0.227

Appendix E

Distribution of fitness effects

Table E.1: DFE NF303 model choice

Model	**< -100**	**(-100, -10)**	**(-10, -1)**	**(-1, 0)**	**(0, 1)**	**1 <**	**df**	**log lk**	**AIC**
A del - r - eps	0.626	0.087	0.067	0.221	0	0	4	-99329.4782	198666.9564
A del + r - eps	0.607	0.105	0.077	0.212	0	0	13	-134.9742	295.9485
A del - r + eps	0.234	0.430	0.206	0.129	0	0	5	-30401.6677	60813.3355
A del + r + eps	0.607	0.105	0.077	0.212	0	0	14	-135.1071	298.2142
B del - r - eps	0.626	0.087	0.067	0.221	0	0	4	-99328.6250	198665.2500
B del + r - eps	0.607	0.105	0.077	0.212	0	0	13	-134.9473	295.8946
B del - r + eps	0.234	0.430	0.206	0.129	0	0	5	-30401.6677	60813.3355
B del + r + eps	0.606	0.105	0.077	0.212	0	0	14	-135.2926	298.5853
C del - r - eps	0.626	0.087	0.067	0.221	0	0	4	-99328.6250	198665.2500
C del + r - eps	0.607	0.105	0.077	0.212	0	0	13	-134.9473	295.8946
C del - r + eps	0.234	0.430	0.206	0.129	0	0	5	-30401.6677	60813.3355
C del + r + eps	0.606	0.105	0.077	0.212	0	0	14	-135.2926	298.5853

Table E.2: DFE NF414 model choice

Model	**< -100**	**(-100, -10)**	**(-10, -1)**	**(-1, 0)**	**(0, 1)**	**1 <**	**df**	**log lk**	**AIC**
A del - r - eps	0.359	0.285	0.163	0.192	0	0	4	-3331.5714	6671.1429
A del + r - eps	0.566	0.129	0.091	0.214	0	0	13	-152.4884	330.9769
A del - r + eps	0.359	0.285	0.163	0.192	0	0	5	-3331.5725	6673.1450
A del + r + eps	0.566	0.129	0.091	0.214	0	0	14	-152.6699	333.3397
B del - r - eps	0.359	0.285	0.163	0.192	0	0	4	-3331.5714	6671.1429
B del + r - eps	0.566	0.129	0.091	0.214	0	0	13	-152.5069	331.0139
B del - r + eps	0.359	0.285	0.163	0.192	0	0	5	-3331.5773	6673.1545
B del + r + eps	0.565	0.130	0.091	0.213	0	0	14	-150.5594	329.1189
C del - r - eps	0.359	0.285	0.163	0.192	0	0	4	-3331.5714	6671.1429
C del + r - eps	0.566	0.129	0.091	0.214	0	0	13	-152.5069	331.0139
C del - r + eps	0.359	0.285	0.163	0.192	0	0	5	-3331.5773	6673.1545
C del + r + eps	0.565	0.130	0.091	0.213	0	0	14	-150.5594	329.1189

Table E.3: DFE GNP model choice

Model	**< -100**	**(-100, -10)**	**(-10, -1)**	**(-1, 0)**	**(0, 1)**	**1 <**	**df**	**log lk**	**AIC**
A del - r - eps	5.7e-02	4.4e-01	3.0e-01	2.0e-01	0.0e+00	0.0e+00	4	-1780.7702	3569.5404
A del + r - eps	1.2e-02	4.2e-01	3.8e-01	1.9e-01	0.0e+00	0.0e+00	13	-170.9231	367.8461
A del - r + eps	1.6e-02	4.3e-01	3.6e-01	1.9e-01	0.0e+00	0.0e+00	5	-1344.2411	2698.4822
A del + r + eps	1.2e-02	4.2e-01	3.8e-01	1.9e-01	0.0e+00	0.0e+00	14	-164.4842	356.9684
B del - r - eps	5.7e-02	4.4e-01	3.0e-01	2.0e-01	0.0e+00	0.0e+00	4	-1780.7702	3569.5404
B del + r - eps	1.2e-02	4.2e-01	3.8e-01	1.9e-01	0.0e+00	0.0e+00	13	-170.9247	367.8494
B del - r + eps	1.6e-02	4.3e-01	3.6e-01	1.9e-01	0.0e+00	0.0e+00	5	-1344.2411	2698.4822
B del + r + eps	1.2e-02	4.2e-01	3.8e-01	1.9e-01	0.0e+00	0.0e+00	14	-164.4814	356.9627
C del - r - eps	5.7e-02	4.4e-01	3.0e-01	2.0e-01	0.0e+00	0.0e+00	4	-1780.7702	3569.5404
C del + r - eps	1.2e-02	4.2e-01	3.8e-01	1.9e-01	0.0e+00	0.0e+00	13	-170.9247	367.8494
C del - r + eps	1.6e-02	4.3e-01	3.6e-01	1.9e-01	0.0e+00	0.0e+00	5	-1344.2411	2698.4822
C del + r + eps	1.2e-02	4.2e-01	3.8e-01	1.9e-01	0.0e+00	0.0e+00	14	-164.4814	356.9627
A - r - eps	9.1e-01	4.6e-02	1.0e-02	1.4e-03	1.5e-03	3.4e-02	5	-20421.4981	40852.9961
A + r - eps	1.2e-02	4.2e-01	3.8e-01	1.8e-01	1.2e-02	0.0e+00	14	-170.8333	369.6666
A - r + eps	1.6e-02	4.3e-01	3.6e-01	1.9e-01	5.4e-03	0.0e+00	6	-1344.1347	2700.2694
A + r + eps	1.2e-02	4.2e-01	3.8e-01	1.8e-01	1.2e-02	0.0e+00	15	-164.3833	358.7667
B - r - eps	5.7e-02	4.4e-01	3.0e-01	2.0e-01	0.0e+00	0.0e+00	6	-1780.7702	3573.5404
B + r - eps	1.2e-02	4.2e-01	3.8e-01	1.9e-01	0.0e+00	2.2e-04	15	-170.9107	371.8214
B - r + eps	3.2e-22	6.2e-01	1.8e-01	6.8e-09	0.0e+00	2.0e-01	7	-1104.7426	2223.4851
B + r + eps	1.2e-02	4.2e-01	3.8e-01	1.9e-01	0.0e+00	1.2e-03	16	-164.3789	360.7579
C - r - eps	4.2e-15	7.3e-01	4.8e-02	4.3e-10	2.3e-01	0.0e+00	6	-1515.0998	3042.1996
C + r - eps	1.2e-02	4.2e-01	3.8e-01	1.9e-01	4.3e-04	0.0e+00	15	-170.8975	371.7949
C - r + eps	1.6e-02	4.3e-01	3.6e-01	1.9e-01	6.1e-06	0.0e+00	7	-1344.2388	2702.4776
C + r + eps	1.2e-02	4.2e-01	3.8e-01	1.9e-01	6.9e-04	8.1e-11	16	-164.4277	360.8554

Appendix F

DoS GO slim families

Table F.1: DoS Gene Ontology slim families: GNP *N. rachovii*

GO slim term	Median DoS GO-term	Median DoS overall	*P*-value	FDR
aging	-2.5e-01	-1.7e-01	4.8e-02	7.7e-01
anatomical structure development	-1.6e-01	-1.7e-01	7.2e-01	9.9e-01
anatomical structure formation involved in morphogenesis	-1.3e-01	-1.7e-01	4.2e-01	9.2e-01
ATPase activity	-3.0e-01	-1.7e-01	3.0e-02	7.7e-01
autophagy	-1.4e-01	-1.7e-01	8.8e-01	9.9e-01
biological process	-1.6e-01	-1.7e-01	8.5e-01	9.9e-01
biosynthetic process	-1.4e-01	-1.7e-01	9.9e-01	1.0e+00
carbohydrate metabolic process	-2.2e-01	-1.7e-01	2.5e-01	8.6e-01
catabolic process	-1.3e-01	-1.7e-01	6.3e-01	9.9e-01
cell	-1.7e-01	-1.7e-01	9.0e-01	9.9e-01
cell-cell signaling	-2.1e-01	-1.7e-01	1.8e-01	8.6e-01
cell adhesion	-1.7e-01	-1.7e-01	4.2e-01	9.2e-01
cell cycle	-1.1e-01	-1.7e-01	1.6e-01	8.6e-01
cell death	-2.0e-01	-1.7e-01	4.9e-02	7.7e-01
cell differentiation	-1.7e-01	-1.7e-01	6.1e-01	9.9e-01
cell division	-4.0e-02	-1.7e-01	1.3e-01	8.6e-01

cell junction organization	-2.2e-01	-1.7e-01	2.6e-01	8.6e-01
cell morphogenesis	-1.7e-01	-1.7e-01	6.4e-01	9.9e-01
cell motility	-1.8e-01	-1.7e-01	4.3e-01	9.2e-01
cell proliferation	-1.4e-01	-1.7e-01	6.4e-01	9.9e-01
cellular amino acid metabolic process	2.9e-02	-1.7e-01	1.2e-01	8.6e-01
cellular component assembly	-1.5e-01	-1.7e-01	8.4e-01	9.9e-01
cellular nitrogen compound metabolic process	-1.4e-01	-1.7e-01	8.0e-01	9.9e-01
cellular protein modification process	-1.7e-01	-1.7e-01	5.0e-01	9.6e-01
cellular component	-1.6e-01	-1.7e-01	9.6e-01	9.9e-01
chromosome	-8.7e-02	-1.7e-01	6.1e-01	9.9e-01
chromosome organization	-1.7e-01	-1.7e-01	9.7e-01	9.9e-01
chromosome segregation	-1.0e-01	-1.7e-01	4.6e-01	9.2e-01
cilium	-1.4e-01	-1.7e-01	2.3e-01	8.6e-01
circulatory system process	-1.4e-01	-1.7e-01	6.0e-01	9.9e-01
cofactor metabolic process	-1.5e-01	-1.7e-01	4.0e-01	9.2e-01
cytoplasm	-1.7e-01	-1.7e-01	1.0e+00	1.0e+00
cytoplasmic, membrane-bounded vesicle	-1.7e-01	-1.7e-01	5.1e-01	9.6e-01
cytoskeletal protein binding	-1.5e-01	-1.7e-01	9.5e-01	9.9e-01
cytoskeleton	-1.5e-01	-1.7e-01	8.1e-01	9.9e-01
cytoskeleton-dependent intracellular transport	-2.0e-01	-1.7e-01	4.4e-01	9.2e-01
cytoskeleton organization	-1.9e-01	-1.7e-01	3.3e-01	9.2e-01
cytosol	-1.7e-01	-1.7e-01	7.2e-01	9.9e-01
developmental maturation	-1.7e-01	-1.7e-01	1.3e-01	8.6e-01
DNA binding	-1.5e-01	-1.7e-01	7.6e-01	9.9e-01
DNA metabolic process	-1.9e-01	-1.7e-01	9.0e-01	9.9e-01
embryo development	-1.7e-01	-1.7e-01	4.5e-01	9.2e-01
endoplasmic reticulum	-1.7e-01	-1.7e-01	4.0e-01	9.2e-01
endosome	-3.8e-02	-1.7e-01	1.7e-01	8.6e-01
enzyme binding	-1.4e-01	-1.7e-01	7.7e-01	9.9e-01
enzyme regulator activity	-2.0e-01	-1.7e-01	4.0e-01	9.2e-01
extracellular matrix organization	-1.1e-01	-1.7e-01	9.6e-01	9.9e-01
extracellular region	-1.8e-01	-1.7e-01	3.1e-01	9.2e-01

extracellular space	-1.4e-01	-1.7e-01	6.5e-01	9.9e-01
generation of precursor metabolites and energy	-3.3e-02	-1.7e-01	7.2e-01	9.9e-01
Golgi apparatus	-2.0e-01	-1.7e-01	1.8e-01	8.6e-01
growth	-2.5e-01	-1.7e-01	1.1e-02	7.7e-01
GTPase activity	0.0e+00	-1.7e-01	9.1e-01	9.9e-01
helicase activity	-3.4e-01	-1.7e-01	8.9e-02	8.6e-01
histone binding	0.0e+00	-1.7e-01	4.1e-02	7.7e-01
homeostatic process	-1.7e-01	-1.7e-01	9.7e-01	9.9e-01
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	-5.5e-02	-1.7e-01	7.9e-01	9.9e-01
hydrolase activity, acting on glycosyl bonds	-2.3e-01	-1.7e-01	6.4e-01	9.9e-01
immune system process	-1.7e-01	-1.7e-01	8.0e-01	9.9e-01
intracellular	-1.7e-01	-1.7e-01	9.0e-01	9.9e-01
ion binding	-1.3e-01	-1.7e-01	5.4e-01	9.6e-01
isomerase activity	2.8e-02	-1.7e-01	8.3e-01	9.9e-01
kinase activity	-9.4e-02	-1.7e-01	3.7e-01	9.2e-01
ligase activity	-7.1e-02	-1.7e-01	2.0e-01	8.6e-01
lipid binding	-8.5e-02	-1.7e-01	4.3e-01	9.2e-01
lipid metabolic process	-1.0e-01	-1.7e-01	6.3e-01	9.9e-01
lipid particle	-3.1e-02	-1.7e-01	9.5e-01	9.9e-01
locomotion	-1.7e-01	-1.7e-01	5.1e-01	9.6e-01
lyase activity	-2.8e-01	-1.7e-01	3.4e-01	9.2e-01
lysosome	-1.6e-01	-1.7e-01	7.0e-01	9.9e-01
macromolecular complex assembly	-1.5e-01	-1.7e-01	2.8e-01	8.8e-01
membrane organization	-2.7e-01	-1.7e-01	8.1e-03	7.7e-01
methyltransferase activity	-1.5e-01	-1.7e-01	4.1e-01	9.2e-01
microtubule organizing center	-1.3e-01	-1.7e-01	2.4e-01	8.6e-01
mitochondrion	-1.3e-01	-1.7e-01	1.5e-01	8.6e-01
mitochondrion organization	-1.4e-01	-1.7e-01	5.7e-02	7.7e-01
mitotic nuclear division	-6.1e-02	-1.7e-01	1.6e-01	8.6e-01
molecular function	-1.7e-01	-1.7e-01	5.2e-01	9.6e-01
mRNA binding	-1.4e-01	-1.7e-01	7.1e-01	9.9e-01

mRNA processing	-1.7e-01	-1.7e-01	7.5e-02	7.9e-01
neurological system process	-1.5e-01	-1.7e-01	2.1e-01	8.6e-01
nuclear chromosome	-8.1e-02	-1.7e-01	9.6e-01	9.9e-01
nuclear envelope	-1.3e-01	-1.7e-01	7.4e-01	9.9e-01
nuclease activity	-1.9e-01	-1.7e-01	7.2e-01	9.9e-01
nucleic acid binding transcription factor activity	-5.3e-02	-1.7e-01	5.1e-01	9.6e-01
nucleobase-containing compound catabolic process	-1.3e-01	-1.7e-01	4.9e-01	9.6e-01
nucleocytoplasmic transport	-1.1e-01	-1.7e-01	6.9e-01	9.9e-01
nucleolus	-1.3e-01	-1.7e-01	6.6e-01	9.9e-01
nucleoplasm	-1.5e-01	-1.7e-01	5.3e-01	9.6e-01
nucleotidyltransferase activity	-1.2e-01	-1.7e-01	8.0e-01	9.9e-01
nucleus	-1.4e-01	-1.7e-01	8.6e-01	9.9e-01
organelle	-1.7e-01	-1.7e-01	8.0e-01	9.9e-01
oxidoreductase activity	-2.4e-01	-1.7e-01	9.6e-01	9.9e-01
peptidase activity	-1.7e-01	-1.7e-01	3.9e-01	9.2e-01
peroxisome	-4.2e-01	-1.7e-01	5.0e-02	7.7e-01
phosphatase activity	-3.9e-02	-1.7e-01	2.2e-01	8.6e-01
pigmentation	0.0e+00	-1.7e-01	5.6e-01	9.8e-01
plasma membrane	-1.7e-01	-1.7e-01	3.1e-01	9.2e-01
plasma membrane organization	-2.3e-01	-1.7e-01	2.5e-01	8.6e-01
protein binding, bridging	6.2e-02	-1.7e-01	9.7e-02	8.6e-01
protein complex	-1.6e-01	-1.7e-01	4.1e-01	9.2e-01
protein complex assembly	-1.7e-01	-1.7e-01	2.4e-01	8.6e-01
protein folding	-1.3e-01	-1.7e-01	4.5e-01	9.2e-01
protein maturation	-1.2e-01	-1.7e-01	3.4e-01	9.2e-01
protein targeting	-6.3e-02	-1.7e-01	5.4e-01	9.6e-01
protein transporter activity	-1.4e-02	-1.7e-01	9.5e-01	9.9e-01
proteinaceous extracellular matrix	-9.7e-02	-1.7e-01	7.8e-01	9.9e-01
reproduction	-1.4e-01	-1.7e-01	7.2e-01	9.9e-01
response to stress	-1.7e-01	-1.7e-01	8.5e-01	9.9e-01
ribonucleoprotein complex assembly	-1.9e-01	-1.7e-01	6.3e-01	9.9e-01
ribosome	-1.4e-01	-1.7e-01	4.5e-01	9.2e-01

ribosome biogenesis	-1.7e-01	-1.7e-01	8.9e-01	9.9e-01
RNA binding	-1.4e-01	-1.7e-01	8.6e-01	9.9e-01
rRNA binding	0.0e+00	-1.7e-01	2.3e-01	8.6e-01
signal transducer activity	-1.7e-01	-1.7e-01	3.2e-01	9.2e-01
signal transduction	-1.9e-01	-1.7e-01	5.9e-02	7.7e-01
small molecule metabolic process	-1.2e-01	-1.7e-01	9.3e-01	9.9e-01
structural constituent of ribosome	-8.3e-02	-1.7e-01	2.2e-01	8.6e-01
structural molecule activity	-1.7e-01	-1.7e-01	4.1e-01	9.2e-01
sulfur compound metabolic process	-2.1e-01	-1.7e-01	1.3e-01	8.6e-01
symbiosis, encompassing mutualism through parasitism	-1.8e-01	-1.7e-01	3.0e-01	9.2e-01
transcription factor activity, protein binding	-2.1e-01	-1.7e-01	6.2e-02	7.7e-01
transcription factor binding	-1.7e-01	-1.7e-01	7.4e-02	7.9e-01
transferase activity, transferring acyl groups	-1.7e-01	-1.7e-01	9.2e-01	9.9e-01
transferase activity, transferring alkyl or aryl (other than methyl) groups	-2.4e-01	-1.7e-01	8.0e-01	9.9e-01
transferase activity, transferring glycosyl groups	-3.0e-01	-1.7e-01	3.0e-02	7.7e-01
translation	-8.6e-02	-1.7e-01	2.6e-01	8.6e-01
translation factor activity, RNA binding	0.0e+00	-1.7e-01	1.0e-01	8.6e-01
transmembrane transport	-1.9e-01	-1.7e-01	1.3e-01	8.6e-01
transmembrane transporter activity	-2.0e-01	-1.7e-01	1.8e-01	8.6e-01
transport	-1.3e-01	-1.7e-01	9.9e-01	1.0e+00
tRNA metabolic process	-2.3e-02	-1.7e-01	2.6e-01	8.6e-01
ubiquitin-like protein binding	-1.5e-01	-1.7e-01	8.9e-01	9.9e-01
unfolded protein binding	-2.5e-01	-1.7e-01	4.5e-01	9.2e-01
vacuolar transport	0.0e+00	-1.7e-01	2.0e-01	8.6e-01
vacuole	-1.1e-01	-1.7e-01	5.6e-01	9.8e-01
vesicle-mediated transport	-1.1e-01	-1.7e-01	8.2e-01	9.9e-01

Table F.2: DoS Gene Ontology slim families: NF414 *N. rachovii*

GO slim term	**Median DoS GO-term**	**Median DoS overall**	***P*-value**	**FDR**
aging	0.0e+00	0	7.2e-01	9.6e-01
anatomical structure development	0.0e+00	0	8.6e-01	9.6e-01
anatomical structure formation involved in morphogenesis	0.0e+00	0	3.4e-01	9.0e-01
ATPase activity	0.0e+00	0	1.6e-01	9.0e-01
autophagy	0.0e+00	0	2.3e-01	9.0e-01
biological process	0.0e+00	0	8.5e-01	9.6e-01
biosynthetic process	0.0e+00	0	8.9e-01	9.6e-01
carbohydrate metabolic process	-1.5e-02	0	1.8e-01	9.0e-01
catabolic process	0.0e+00	0	7.8e-01	9.6e-01
cell	0.0e+00	0	4.7e-01	9.2e-01
cell-cell signaling	0.0e+00	0	7.3e-01	9.6e-01
cell adhesion	0.0e+00	0	3.0e-01	9.0e-01
cell cycle	0.0e+00	0	3.0e-02	6.9e-01
cell death	0.0e+00	0	9.8e-01	9.9e-01
cell differentiation	0.0e+00	0	2.8e-01	9.0e-01
cell division	0.0e+00	0	1.0e-01	8.9e-01
cell junction organization	-1.7e-02	0	6.3e-01	9.6e-01
cell morphogenesis	-9.2e-04	0	4.6e-01	9.2e-01
cell motility	0.0e+00	0	6.3e-01	9.6e-01
cell proliferation	0.0e+00	0	6.0e-01	9.5e-01
cellular amino acid metabolic process	0.0e+00	0	6.2e-01	9.6e-01
cellular component assembly	0.0e+00	0	9.2e-01	9.7e-01
cellular nitrogen compound metabolic process	0.0e+00	0	5.4e-01	9.3e-01
cellular protein modification process	0.0e+00	0	8.4e-01	9.6e-01
cellular component	0.0e+00	0	7.9e-01	9.6e-01
chromosome	0.0e+00	0	9.0e-01	9.6e-01
chromosome organization	0.0e+00	0	1.2e-01	8.9e-01
chromosome segregation	0.0e+00	0	2.9e-01	9.0e-01

cilium	0.0e+00	0	5.3e-01	9.3e-01
circulatory system process	0.0e+00	0	4.1e-01	9.0e-01
cofactor metabolic process	0.0e+00	0	8.9e-01	9.6e-01
cytoplasm	0.0e+00	0	4.5e-02	7.7e-01
cytoplasmic, membrane-bounded vesicle	0.0e+00	0	8.9e-01	9.6e-01
cytoskeletal protein binding	0.0e+00	0	5.4e-01	9.3e-01
cytoskeleton	0.0e+00	0	5.3e-01	9.3e-01
cytoskeleton-dependent intracellular transport	0.0e+00	0	8.7e-01	9.6e-01
cytoskeleton organization	0.0e+00	0	4.4e-01	9.2e-01
cytosol	0.0e+00	0	1.3e-01	8.9e-01
developmental maturation	0.0e+00	0	9.4e-01	9.8e-01
DNA binding	0.0e+00	0	2.6e-01	9.0e-01
DNA metabolic process	0.0e+00	0	4.6e-01	9.2e-01
embryo development	0.0e+00	0	4.3e-01	9.1e-01
endoplasmic reticulum	0.0e+00	0	8.4e-01	9.6e-01
endosome	0.0e+00	0	9.0e-01	9.6e-01
enzyme binding	0.0e+00	0	3.1e-01	9.0e-01
enzyme regulator activity	0.0e+00	0	3.1e-01	9.0e-01
extracellular matrix organization	0.0e+00	0	7.5e-01	9.6e-01
extracellular region	0.0e+00	0	4.2e-01	9.0e-01
extracellular space	0.0e+00	0	5.1e-01	9.3e-01
generation of precursor metabolites and energy	0.0e+00	0	1.8e-01	9.0e-01
Golgi apparatus	0.0e+00	0	4.2e-01	9.0e-01
growth	-3.2e-02	0	7.9e-02	8.4e-01
GTPase activity	0.0e+00	0	1.4e-01	8.9e-01
helicase activity	0.0e+00	0	5.5e-01	9.3e-01
histone binding	0.0e+00	0	9.2e-01	9.7e-01
homeostatic process	0.0e+00	0	2.6e-01	9.0e-01
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	-5.0e-02	0	3.7e-01	9.0e-01
hydrolase activity, acting on glycosyl bonds	-1.8e-02	0	1.0e+00	1.0e+00
immune system process	0.0e+00	0	5.8e-01	9.5e-01

intracellular	0.0e+00	0	3.1e-01	9.0e-01
ion binding	0.0e+00	0	8.2e-01	9.6e-01
isomerase activity	-2.2e-02	0	5.8e-01	9.5e-01
kinase activity	0.0e+00	0	3.8e-01	9.0e-01
ligase activity	0.0e+00	0	3.0e-01	9.0e-01
lipid binding	0.0e+00	0	1.6e-01	9.0e-01
lipid metabolic process	0.0e+00	0	3.1e-01	9.0e-01
lipid particle	2.7e-02	0	2.3e-01	9.0e-01
locomotion	0.0e+00	0	2.5e-01	9.0e-01
lyase activity	-4.1e-02	0	3.6e-01	9.0e-01
lysosome	0.0e+00	0	5.4e-01	9.3e-01
macromolecular complex assembly	0.0e+00	0	8.4e-01	9.6e-01
membrane organization	0.0e+00	0	5.4e-01	9.3e-01
methyltransferase activity	0.0e+00	0	5.9e-01	9.5e-01
microtubule organizing center	0.0e+00	0	4.1e-01	9.0e-01
mitochondrion	0.0e+00	0	2.0e-02	5.8e-01
mitochondrion organization	0.0e+00	0	6.0e-02	8.2e-01
mitotic nuclear division	0.0e+00	0	8.2e-02	8.4e-01
molecular function	0.0e+00	0	7.3e-01	9.6e-01
mRNA binding	4.2e-02	0	4.3e-02	7.7e-01
mRNA processing	0.0e+00	0	7.2e-01	9.6e-01
neurological system process	0.0e+00	0	7.3e-01	9.6e-01
nuclear chromosome	0.0e+00	0	3.3e-01	9.0e-01
nuclear envelope	0.0e+00	0	7.9e-01	9.6e-01
nuclease activity	-1.0e-02	0	9.7e-01	9.9e-01
nucleic acid binding transcription factor activity	0.0e+00	0	7.1e-01	9.6e-01
nucleobase-containing compound catabolic process	0.0e+00	0	8.9e-01	9.6e-01
nucleocytoplasmic transport	0.0e+00	0	9.7e-01	9.9e-01
nucleolus	0.0e+00	0	2.9e-01	9.0e-01
nucleoplasm	0.0e+00	0	7.8e-02	8.4e-01
nucleotidyltransferase activity	0.0e+00	0	3.5e-01	9.0e-01
nucleus	0.0e+00	0	2.7e-01	9.0e-01

organelle	0.0e+00	0	3.0e-01	9.0e-01
oxidoreductase activity	0.0e+00	0	7.4e-01	9.6e-01
peptidase activity	0.0e+00	0	7.3e-01	9.6e-01
peroxisome	-2.6e-02	0	8.1e-01	9.6e-01
phosphatase activity	0.0e+00	0	5.3e-01	9.3e-01
pigmentation	-5.3e-02	0	1.1e-01	8.9e-01
plasma membrane	0.0e+00	0	1.3e-01	8.9e-01
plasma membrane organization	0.0e+00	0	5.6e-01	9.4e-01
protein binding, bridging	-5.4e-02	0	1.6e-01	9.0e-01
protein complex	0.0e+00	0	3.5e-01	9.0e-01
protein complex assembly	0.0e+00	0	8.3e-01	9.6e-01
protein folding	0.0e+00	0	8.8e-01	9.6e-01
protein maturation	2.9e-02	0	2.1e-02	5.8e-01
protein targeting	0.0e+00	0	7.5e-01	9.6e-01
protein transporter activity	5.1e-02	0	8.5e-02	8.4e-01
proteinaceous extracellular matrix	1.3e-02	0	6.7e-01	9.6e-01
reproduction	0.0e+00	0	9.0e-01	9.6e-01
response to stress	0.0e+00	0	6.2e-01	9.6e-01
ribonucleoprotein complex assembly	0.0e+00	0	9.8e-01	9.9e-01
ribosome	2.9e-02	0	1.9e-03	1.3e-01
ribosome biogenesis	0.0e+00	0	4.6e-01	9.2e-01
RNA binding	0.0e+00	0	3.8e-01	9.0e-01
rRNA binding	0.0e+00	0	1.2e-01	8.9e-01
secondary metabolic process	0.0e+00	0	8.1e-01	9.6e-01
signal transducer activity	0.0e+00	0	7.5e-01	9.6e-01
signal transduction	0.0e+00	0	5.9e-01	9.5e-01
small molecule metabolic process	0.0e+00	0	9.1e-01	9.6e-01
structural constituent of ribosome	7.1e-02	0	7.4e-04	1.0e-01
structural molecule activity	0.0e+00	0	3.7e-01	9.0e-01
sulfur compound metabolic process	-3.3e-02	0	3.4e-01	9.0e-01
symbiosis, encompassing mutualism through parasitism	0.0e+00	0	6.8e-01	9.6e-01

transcription factor activity, protein binding	0.0e+00	0	2.4e-01	9.0e-01
transcription factor binding	0.0e+00	0	6.9e-01	9.6e-01
transferase activity, transferring acyl groups	-4.9e-02	0	3.8e-01	9.0e-01
transferase activity, transferring alkyl or aryl (other than methyl) groups	-8.7e-02	0	3.4e-01	9.0e-01
transferase activity, transferring glycosyl groups	0.0e+00	0	6.9e-01	9.6e-01
translation	0.0e+00	0	2.8e-03	1.3e-01
translation factor activity, RNA binding	0.0e+00	0	4.0e-01	9.0e-01
transmembrane transport	0.0e+00	0	2.3e-01	9.0e-01
transmembrane transporter activity	0.0e+00	0	4.8e-01	9.3e-01
transport	0.0e+00	0	5.1e-02	7.8e-01
tRNA metabolic process	4.5e-02	0	1.8e-01	9.0e-01
ubiquitin-like protein binding	3.5e-02	0	1.5e-01	9.0e-01
unfolded protein binding	0.0e+00	0	8.6e-01	9.6e-01
vacuolar transport	0.0e+00	0	2.7e-01	9.0e-01
vacuole	0.0e+00	0	8.6e-01	9.6e-01
vesicle-mediated transport	0.0e+00	0	4.8e-01	9.3e-01

Table F.3: DoS Gene Ontology slim families: NF303 *N. rachovii*

GO slim term	**Median DoS GO-term**	**Median DoS overall**	***P*-value**	**FDR**
aging	1.9e-03	0	8.4e-01	9.5e-01
anatomical structure development	0.0e+00	0	7.8e-01	9.2e-01
anatomical structure formation involved in morphogenesis	0.0e+00	0	8.7e-01	9.5e-01
ATPase activity	3.3e-02	0	1.1e-01	4.6e-01
autophagy	0.0e+00	0	6.9e-01	9.0e-01
biological process	0.0e+00	0	7.6e-01	9.1e-01
biosynthetic process	0.0e+00	0	9.7e-01	9.8e-01
carbohydrate metabolic process	-2.6e-02	0	1.4e-04	1.4e-02
catabolic process	0.0e+00	0	8.6e-01	9.5e-01
cell	0.0e+00	0	2.8e-01	6.4e-01
cell-cell signaling	0.0e+00	0	3.6e-01	7.4e-01
cell adhesion	6.9e-03	0	2.5e-01	6.3e-01
cell cycle	4.9e-03	0	6.5e-02	3.3e-01
cell death	0.0e+00	0	3.1e-01	6.6e-01
cell differentiation	0.0e+00	0	7.8e-01	9.2e-01
cell division	3.6e-02	0	5.3e-03	9.1e-02
cell junction organization	0.0e+00	0	4.6e-01	8.2e-01
cell morphogenesis	0.0e+00	0	2.1e-02	1.9e-01
cell motility	1.7e-02	0	2.0e-01	5.7e-01
cell proliferation	0.0e+00	0	8.1e-01	9.4e-01
cellular amino acid metabolic process	0.0e+00	0	6.8e-01	9.0e-01
cellular component assembly	0.0e+00	0	6.3e-01	8.9e-01
cellular nitrogen compound metabolic process	0.0e+00	0	6.8e-01	9.0e-01
cellular protein modification process	0.0e+00	0	7.5e-01	9.0e-01
cellular component	0.0e+00	0	5.2e-01	8.3e-01
chromosome	1.6e-02	0	2.2e-02	1.9e-01
chromosome organization	2.9e-02	0	1.0e-02	1.5e-01
chromosome segregation	4.8e-02	0	1.6e-02	1.7e-01

cilium	4.0e-02	0	1.2e-01	4.8e-01
circulatory system process	0.0e+00	0	2.8e-01	6.4e-01
cofactor metabolic process	-2.9e-02	0	4.3e-02	3.0e-01
cytoplasm	0.0e+00	0	8.6e-02	3.9e-01
cytoplasmic, membrane-bounded vesicle	0.0e+00	0	5.7e-01	8.6e-01
cytoskeletal protein binding	0.0e+00	0	2.9e-01	6.4e-01
cytoskeleton	0.0e+00	0	4.2e-01	8.0e-01
cytoskeleton-dependent intracellular transport	0.0e+00	0	9.3e-01	9.7e-01
cytoskeleton organization	3.9e-03	0	5.3e-01	8.3e-01
cytosol	0.0e+00	0	9.9e-01	9.9e-01
developmental maturation	0.0e+00	0	4.5e-01	8.2e-01
DNA binding	0.0e+00	0	1.8e-01	5.6e-01
DNA metabolic process	0.0e+00	0	7.1e-01	9.0e-01
embryo development	0.0e+00	0	1.3e-01	4.8e-01
endoplasmic reticulum	0.0e+00	0	6.8e-01	9.0e-01
endosome	0.0e+00	0	5.0e-01	8.3e-01
enzyme binding	9.4e-03	0	1.7e-01	5.6e-01
enzyme regulator activity	2.5e-02	0	5.1e-01	8.3e-01
extracellular matrix organization	1.6e-02	0	4.2e-01	8.0e-01
extracellular region	0.0e+00	0	1.3e-01	4.8e-01
extracellular space	2.4e-02	0	2.7e-01	6.4e-01
generation of precursor metabolites and energy	0.0e+00	0	2.6e-01	6.3e-01
Golgi apparatus	0.0e+00	0	5.9e-01	8.7e-01
growth	0.0e+00	0	7.2e-01	9.0e-01
GTPase activity	0.0e+00	0	2.8e-01	6.4e-01
helicase activity	0.0e+00	0	8.7e-01	9.5e-01
histone binding	0.0e+00	0	4.3e-01	8.0e-01
homeostatic process	0.0e+00	0	9.7e-01	9.8e-01
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	9.4e-03	0	7.3e-01	9.0e-01
hydrolase activity, acting on glycosyl bonds	0.0e+00	0	4.2e-01	8.0e-01
immune system process	1.7e-02	0	3.3e-02	2.5e-01

intracellular	0.0e+00	0	4.9e-01	8.2e-01
ion binding	0.0e+00	0	9.1e-01	9.6e-01
isomerase activity	2.2e-02	0	2.0e-01	5.7e-01
kinase activity	0.0e+00	0	9.1e-01	9.6e-01
ligase activity	2.8e-02	0	3.0e-01	6.4e-01
lipid binding	0.0e+00	0	5.5e-01	8.4e-01
lipid metabolic process	1.6e-02	0	8.1e-02	3.8e-01
lipid particle	1.0e-01	0	6.8e-02	3.4e-01
locomotion	0.0e+00	0	8.7e-01	9.5e-01
lyase activity	-5.1e-02	0	2.0e-04	1.4e-02
lysosome	0.0e+00	0	7.2e-01	9.0e-01
macromolecular complex assembly	0.0e+00	0	1.8e-01	5.6e-01
membrane organization	0.0e+00	0	6.2e-01	8.8e-01
methyltransferase activity	0.0e+00	0	3.4e-01	7.1e-01
microtubule organizing center	8.4e-03	0	2.0e-01	5.7e-01
mitochondrion	0.0e+00	0	2.0e-01	5.7e-01
mitochondrion organization	1.0e-02	0	4.6e-02	3.0e-01
mitotic nuclear division	5.0e-02	0	5.0e-03	9.1e-02
molecular function	0.0e+00	0	9.3e-01	9.7e-01
mRNA binding	4.4e-02	0	3.0e-01	6.4e-01
mRNA processing	0.0e+00	0	4.9e-01	8.2e-01
neurological system process	0.0e+00	0	5.8e-01	8.6e-01
nitrogen cycle metabolic process	3.8e-01	0	3.5e-03	8.1e-02
nuclear chromosome	0.0e+00	0	8.5e-01	9.5e-01
nuclear envelope	7.8e-03	0	4.7e-01	8.2e-01
nuclease activity	4.8e-02	0	1.8e-01	5.6e-01
nucleic acid binding transcription factor activity	0.0e+00	0	3.7e-01	7.4e-01
nucleobase-containing compound catabolic process	0.0e+00	0	5.8e-02	3.3e-01
nucleocytoplasmic transport	7.0e-03	0	1.4e-01	4.8e-01
nucleolus	1.7e-02	0	2.8e-02	2.2e-01
nucleoplasm	0.0e+00	0	2.2e-01	6.1e-01
nucleotidyltransferase activity	0.0e+00	0	4.2e-01	8.0e-01

nucleus	0.0e+00	0	3.7e-01	7.4e-01
organelle	0.0e+00	0	5.1e-01	8.3e-01
oxidoreductase activity	0.0e+00	0	8.8e-01	9.5e-01
peptidase activity	0.0e+00	0	7.4e-01	9.0e-01
peroxisome	-2.6e-02	0	2.4e-01	6.2e-01
phosphatase activity	0.0e+00	0	4.7e-01	8.2e-01
pigmentation	0.0e+00	0	4.7e-01	8.2e-01
plasma membrane	1.3e-02	0	4.0e-02	2.9e-01
plasma membrane organization	3.3e-02	0	6.5e-02	3.3e-01
protein binding, bridging	0.0e+00	0	9.0e-01	9.6e-01
protein complex	0.0e+00	0	5.2e-02	3.3e-01
protein complex assembly	3.9e-03	0	1.1e-01	4.8e-01
protein folding	0.0e+00	0	1.8e-01	5.6e-01
protein maturation	3.6e-02	0	5.5e-01	8.4e-01
protein targeting	0.0e+00	0	6.7e-01	9.0e-01
protein transporter activity	0.0e+00	0	7.2e-01	9.0e-01
proteinaceous extracellular matrix	0.0e+00	0	8.6e-01	9.5e-01
reproduction	0.0e+00	0	7.3e-01	9.0e-01
response to stress	0.0e+00	0	6.7e-01	9.0e-01
ribonucleoprotein complex assembly	0.0e+00	0	9.9e-01	9.9e-01
ribosome	4.5e-02	0	3.5e-04	1.6e-02
ribosome biogenesis	1.9e-02	0	2.2e-02	1.9e-01
RNA binding	2.0e-02	0	1.3e-02	1.6e-01
rRNA binding	1.2e-01	0	3.0e-03	8.1e-02
secondary metabolic process	8.2e-02	0	1.8e-01	5.6e-01
signal transducer activity	0.0e+00	0	8.2e-01	9.4e-01
signal transduction	0.0e+00	0	4.8e-01	8.2e-01
small molecule metabolic process	0.0e+00	0	6.5e-01	9.0e-01
structural constituent of ribosome	1.9e-02	0	2.0e-03	6.9e-02
structural molecule activity	0.0e+00	0	6.0e-02	3.3e-01
sulfur compound metabolic process	0.0e+00	0	7.5e-02	3.6e-01

symbiosis, encompassing mutualism through parasitism	0.0e+00	0	1.2e-01	4.8e-01
transcription factor activity, protein binding	0.0e+00	0	6.1e-01	8.8e-01
transcription factor binding	0.0e+00	0	9.6e-01	9.8e-01
transferase activity, transferring acyl groups	1.7e-02	0	7.1e-01	9.0e-01
transferase activity, transferring alkyl or aryl (other than methyl) groups	5.1e-02	0	2.9e-01	6.4e-01
transferase activity, transferring glycosyl groups	-4.8e-02	0	5.9e-02	3.3e-01
translation	8.6e-03	0	1.3e-02	1.6e-01
translation factor activity, RNA binding	4.3e-03	0	6.1e-01	8.8e-01
transmembrane transport	2.2e-02	0	1.2e-01	4.8e-01
transmembrane transporter activity	1.4e-02	0	2.0e-01	5.7e-01
transport	0.0e+00	0	1.2e-02	1.6e-01
tRNA metabolic process	0.0e+00	0	5.4e-01	8.4e-01
ubiquitin-like protein binding	0.0e+00	0	2.4e-01	6.2e-01
unfolded protein binding	0.0e+00	0	7.0e-01	9.0e-01
vacuolar transport	0.0e+00	0	2.4e-01	6.2e-01
vacuole	0.0e+00	0	8.1e-01	9.4e-01
vesicle-mediated transport	0.0e+00	0	2.4e-01	6.2e-01

Appendix G

Pathway overrepresentation DoS

Table G.1: DoS <2.5^{th} percentile pathway overrepresentation with outgroup *N. orthonotus*

Population	**Pathway**	***P*-value**	**q-value**	**Source**
GNP	WNT ligand biogenesis and trafficking	5.1e-06	3.6e-03	Reactome
GNP	TCF dependent signaling in response to WNT	1.0e-05	3.6e-03	Reactome
GNP	Signaling by WNT	4.1e-05	9.8e-03	Reactome
GNP	Class B/2 (Secretin family receptors)	5.8e-05	1.0e-02	Reactome
GNP	DNA Damage Recognition in GG-NER	1.1e-04	1.5e-02	Reactome
GNP	Dopaminergic synapse - Homo sapiens (human)	1.2e-04	1.5e-02	KEGG
GNP	DARPP-32 events	4.4e-04	4.6e-02	Reactome
GNP	Wnt Mammals	6.0e-04	4.8e-02	INOH
GNP	Wnt Canonical	6.0e-04	4.8e-02	INOH
GNP	Disassembly of the destruction complex and recruitment of AXIN to the membrane	8.4e-04	6.1e-02	Reactome
GNP	Cocaine addiction - Homo sapiens (human)	1.0e-03	6.6e-02	KEGG
GNP	Deregulated CDK5 triggers multiple neurodegenerative pathways in Alzheimer´s disease models	2.2e-03	8.7e-02	Reactome
GNP	Neurodegenerative Diseases	2.2e-03	8.7e-02	Reactome
GNP	CTLA4 inhibitory signaling	2.2e-03	8.7e-02	Reactome
GNP	Breast cancer - Homo sapiens (human)	2.3e-03	8.7e-02	KEGG

GNP	Hepatocellular carcinoma - Homo sapiens (human)	2.3e-03	8.7e-02	KEGG
GNP	WNT-Core	2.3e-03	8.7e-02	Signalink
GNP	Formation of TC-NER Pre-Incision Complex	2.3e-03	8.7e-02	Reactome
GNP	Neddylation	2.3e-03	8.7e-02	Reactome
GNP	mTOR signaling pathway - Homo sapiens (human)	2.7e-03	9.7e-02	KEGG
GNP	The information processing pathway at the ifn beta enhancer	3.4e-03	9.7e-02	BioCarta
GNP	Wnt signaling pathway - Homo sapiens (human)	3.5e-03	9.7e-02	KEGG
GNP	Wnt Signaling Pathway and Pluripotency	3.5e-03	9.7e-02	Wikipathways
GNP	fig-met-1-last-solution	4.2e-03	9.7e-02	Wikipathways
GNP	Degradation of beta-catenin by the destruction complex	4.2e-03	9.7e-02	Reactome
GNP	Hippo signaling pathway - Homo sapiens (human)	4.3e-03	9.7e-02	KEGG
GNP	Global Genome Nucleotide Excision Repair (GG-NER)	4.4e-03	9.7e-02	Reactome
GNP	Metabolism of carbohydrates	4.8e-03	9.7e-02	Reactome
GNP	Amphetamine addiction - Homo sapiens (human)	5.1e-03	9.7e-02	KEGG
GNP	Costimulation by the CD28 family	6.7e-03	9.7e-02	Reactome
GNP	Regulation of ck1/cdk5 by type 1 glutamate receptors	6.9e-03	9.7e-02	BioCarta
GNP	Nicotine Activity on Dopaminergic Neurons	6.9e-03	9.7e-02	Wikipathways
GNP	Chromatin remodeling by hswi/snf atp-dependent complexes	6.9e-03	9.7e-02	BioCarta
GNP	FCERI mediated MAPK activation	6.9e-03	9.7e-02	Reactome
GNP	C-MYC pathway	6.9e-03	9.7e-02	PID
GNP	Deactivation of the beta-catenin transactivating complex	6.9e-03	9.7e-02	Reactome
GNP	Nicotine Pathway (Dopaminergic Neuron), Pharmacodynamics	6.9e-03	9.7e-02	PharmGKB
GNP	Nucleotide Excision Repair	7.7e-03	9.7e-02	Reactome

GNP	Spliceosome - Homo sapiens (human)	7.7e-03	9.7e-02	KEGG
GNP	mRNA Splicing - Major Pathway	7.9e-03	9.7e-02	Reactome
GNP	Wnt	8.9e-03	9.7e-02	NetPath
GNP	mRNA Splicing	9.2e-03	9.7e-02	Reactome
GNP	Central carbon metabolism in cancer - Homo sapiens (human)	9.9e-03	9.7e-02	KEGG
NF414	Synthesis of 5-eicosatetraenoic acids	5.7e-05	3.5e-02	Reactome
NF414	glutathione redox reactions I	1.9e-04	4.8e-02	HumanCyc
NF414	Cap-dependent Translation Initiation	4.6e-04	4.8e-02	Reactome
NF414	Eukaryotic Translation Initiation	4.6e-04	4.8e-02	Reactome
NF414	Synthesis of 15-eicosatetraenoic acid derivatives	4.6e-04	4.8e-02	Reactome
NF414	Synthesis of 12-eicosatetraenoic acid derivatives	4.6e-04	4.8e-02	Reactome
NF414	Selenium Metabolism and Selenoproteins	6.3e-04	5.6e-02	Wikipathways
NF414	Formation of a pool of free 40S subunits	9.3e-04	7.2e-02	Reactome
NF414	Translation	1.2e-03	8.0e-02	Reactome
NF414	SRP-dependent cotranslational protein targeting to membrane	1.3e-03	8.0e-02	Reactome
NF414	L13a-mediated translational silencing of Ceruloplasmin expression	1.8e-03	9.1e-02	Reactome
NF414	GTP hydrolysis and joining of the 60S ribosomal subunit	1.8e-03	9.1e-02	Reactome

Table G.2: DoS <2.5th percentile pathway overrepresentation with outgroup *N. rachovii*

Population	**Pathway**	***P*-value**	**q-value**	**Source**
GNP	WNT ligand biogenesis and trafficking	1.9e-05	8.6e-03	Reactome
GNP	Signaling by WNT	7.8e-05	1.8e-02	Reactome
GNP	TCF dependent signaling in response to WNT	1.5e-04	2.3e-02	Reactome
GNP	Glucocorticoid receptor regulatory network	3.4e-04	2.5e-02	PID
GNP	Deregulated CDK5 triggers multiple neurodegenerative pathways in Alzheimer,s disease models	3.7e-04	2.5e-02	Reactome
GNP	Neurodegenerative Diseases	3.7e-04	2.5e-02	Reactome
GNP	Proteoglycans in cancer - Homo sapiens (human)	3.7e-04	2.5e-02	KEGG
GNP	Signaling by Nuclear Receptors	5.2e-04	3.0e-02	Reactome
GNP	Degradation of beta-catenin by the destruction complex	7.3e-04	3.7e-02	Reactome
GNP	ESR-mediated signaling	8.9e-04	4.1e-02	Reactome
GNP	Wnt	1.0e-03	4.2e-02	NetPath
GNP	Wnt Signaling Pathway and Pluripotency	1.3e-03	4.7e-02	Wikipathways
GNP	mTOR signaling pathway - Homo sapiens (human)	1.8e-03	5.3e-02	KEGG
GNP	TNF related weak inducer of apoptosis (TWEAK) Signaling Pathway	1.9e-03	5.3e-02	Wikipathways
GNP	Nicotine Pathway (Dopaminergic Neuron), Pharmacodynamics	1.9e-03	5.3e-02	PharmGKB
GNP	Signaling pathways regulating pluripotency of stem cells - Homo sapiens (human)	2.1e-03	5.3e-02	KEGG
GNP	Class B/2 (Secretin family receptors)	2.2e-03	5.3e-02	Reactome
GNP	Signaling by TGF-beta Receptor Complex	2.2e-03	5.3e-02	Reactome
GNP	Breast cancer - Homo sapiens (human)	2.4e-03	5.5e-02	KEGG
GNP	Regulation of TP53 Activity	3.9e-03	6.5e-02	Reactome
GNP	WNT-Core	4.0e-03	6.5e-02	Signalink
GNP	Estrogen-dependent gene expression	5.9e-03	8.5e-02	Reactome
GNP	Presenilin action in Notch and Wnt signaling	6.9e-03	8.5e-02	PID

GNP	ESC Pluripotency Pathways	7.0e-03	8.5e-02	Wikipathways
GNP	PTEN Regulation	7.0e-03	8.5e-02	Reactome
GNP	Signaling by TGF-beta family members	9.5e-03	9.0e-02	Reactome
GNP	Wnt Mammals	9.5e-03	9.0e-02	INOH
GNP	Wnt Canonical	9.5e-03	9.0e-02	INOH
GNP	Gastric cancer - Homo sapiens (human)	9.8e-03	9.0e-02	KEGG
GNP	Hepatocellular carcinoma - Homo sapiens (human)	9.8e-03	9.0e-02	KEGG

Table G.3: DoS >97.5^{th} percentile pathway overrepresentation with outgroup *N. orthonotus*

Population	**Pathway**	***P*-value**	**q-value**	**Source**
GNP	adenosine ribonucleotides *de novo* biosynthesis	5.5e-04	5.6e-02	HumanCyc
GNP	superpathway of purine nucleotide salvage	1.8e-03	5.6e-02	HumanCyc
GNP	purine nucleotides *de novo* biosynthesis	1.8e-03	5.6e-02	HumanCyc
GNP	Allograft Rejection	6.1e-03	7.5e-02	Wikipathways
GNP	Anchoring of the basal body to the plasma membrane	6.9e-03	7.5e-02	Reactome
GNP	JAK-STAT-Core	7.4e-03	7.5e-02	Signalink
GNP	PLK1 signaling events	7.4e-03	7.5e-02	PID
GNP	Organelle biogenesis and maintenance	9.4e-03	8.4e-02	Reactome

Table G.4: DoS >97.5th percentile pathway overrepresentation with outgroup *N. rachovii*

Population	**Pathway**	***P*-value**	**q-value**	**Source**
NF303	Ribosome - Homo sapiens (human)	3.3e-06	2.2e-03	KEGG
NF303	Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.	1.4e-05	4.6e-03	Reactome
NF303	Respiratory electron transport	1.0e-04	2.0e-02	Reactome
NF303	The citric acid (TCA) cycle and respiratory electron transport	1.2e-04	2.0e-02	Reactome
NF303	Complex I biogenesis	2.2e-04	2.8e-02	Reactome
NF303	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	4.8e-04	4.5e-02	Reactome
NF303	Nonsense-Mediated Decay (NMD)	4.8e-04	4.5e-02	Reactome
NF303	Translation	6.7e-04	4.6e-02	Reactome
NF303	Mitochondrial translation initiation	8.0e-04	4.6e-02	Reactome
NF303	Cytoplasmic Ribosomal Proteins	8.4e-04	4.6e-02	Wikipathways
NF303	Peptide chain elongation	8.4e-04	4.6e-02	Reactome
NF303	Mitochondrial translation termination	9.0e-04	4.6e-02	Reactome
NF303	Eukaryotic Translation Termination	9.7e-04	4.6e-02	Reactome
NF303	Mitochondrial translation elongation	1.0e-03	4.6e-02	Reactome
NF303	FGFR2 ligand binding and activation	1.1e-03	4.6e-02	Reactome
NF303	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	1.1e-03	4.6e-02	Reactome
NF303	Selenocysteine synthesis	1.3e-03	5.0e-02	Reactome
NF303	Mitochondrial translation	1.4e-03	5.0e-02	Reactome
NF303	Eukaryotic Translation Elongation	1.5e-03	5.0e-02	Reactome
NF303	FGFR2b ligand binding and activation	1.6e-03	5.0e-02	Reactome
NF303	Oxidative phosphorylation	1.7e-03	5.0e-02	Wikipathways
NF303	Thermogenesis - Homo sapiens (human)	1.8e-03	5.0e-02	KEGG
NF303	Cap-dependent Translation Initiation	1.8e-03	5.0e-02	Reactome
NF303	Eukaryotic Translation Initiation	1.8e-03	5.0e-02	Reactome
NF303	Formation of a pool of free 40S subunits	2.1e-03	5.6e-02	Reactome
NF303	Parkinson,s disease - Homo sapiens (human)	2.4e-03	6.1e-02	KEGG

NF303	Oxidative phosphorylation - Homo sapiens (human)	4.0e-03	9.8e-02	KEGG
NF303	SRP-dependent cotranslational protein targeting to membrane	4.2e-03	9.8e-02	Reactome
NF303	Electron Transport Chain	4.6e-03	9.8e-02	Wikipathways
NF303	L13a-mediated translational silencing of Ceruloplasmin expression	4.6e-03	9.8e-02	Reactome
NF303	GTP hydrolysis and joining of the 60S ribosomal subunit	4.6e-03	9.8e-02	Reactome
GNP	Mitochondrial translation termination	7.7e-04	8.3e-02	Reactome
GNP	Mitochondrial translation	1.2e-03	8.3e-02	Reactome
NF414	Electron Transport Chain	1.1e-04	7.5e-02	Wikipathways

Erklärung zur Dissertation

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt habe, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universtität zur Prüfung vorgelegen hat; dass sie – abgesehen von unten angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde.
Die Bestimungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Dr. Dario Valenzano betreut worden.
Nachfolgend genannte Teilpublikationen liegen vor:

Willemsen, D., Cui, R., Reichard, M., Valenzano, D.R., Intra-species differences in population size shape life history and genome evolution. bioRxiv 2019. https://doi.org/10.1101/852368

David Willemsen
11. February 2020