Development and application of RNA-seq bioinformatic tools to explore non-model organisms in ageing research

DISSERTATION

zur Erlangung des akademischen Grades
„doctor rerum naturalium“ (Dr. rer. nat.)

vorgelegt

dem Rat der Biologisch-Pharmazeutischen Fakultät
der Friedrich-Schiller-Universität Jena

von

Diplom-Bioinformatiker Martin Bens

geboren am 25.07.1987 in Bad Muskau

seit 1558

Leibniz Institute on Aging –
Fritz Lipmann Institute

Die vorliegende Arbeit wurde in der Zeit von August 2012 bis Juni 2017 am Leibniz-Institut für Alternsforschung – Fritz-Lipmann-Institut in Jena angefertigt und am 03.11.2017 verteidigt.

Gutachter:

1. PD Dr. Matthias Platzer, Jena, Germany

   Platzer Research Group,

   Leibniz Institute on Aging – Fritz Lipmann Institute, Jena, Germany
2. Prof. Dr. Dr. Steve Hoffmann,

   Hoffmann Research Group,

   Leibniz Institute on Aging – Fritz Lipmann Institute, Jena, Germany
3. Prof. Dr. Ingo Ebersberger,

   Department for Applied Bioinformatics,

   Goethe-University, Frankfurt am Main, Germany

„Essentially, all models are wrong, but some are useful."

George Edward Pelham Box
Empirical Model-Building and Response Surfaces,
John Wiley & Son, 1987

Abstract

Advances in sequencing technologies enabled data generation from both genomes and transcriptomes at an unprecedented depth and accuracy. This progress changed the way researchers can approach biological questions, as hypotheses can be generated and verified using virtually any species apart from traditional laboratory organisms. However, genome-wide investigations in such non-model organisms are still hampered by the complexity and costs associated with sequencing and reconstruction of large genomes. Second-generation transcriptome sequencing (RNA-seq) and computational reconstruction of transcripts provides an alternative and cost-effective approach to gain insights into the protein-coding portion of genomes. Yet, lack of publicly available software that performs necessary steps to get from RNA-seq data obtained from non-model organisms to reasonable mRNA catalogues for downstream analyses limited its application to experts.

My thesis comprises the development of `FRAMA`, a software pipeline that delivers mRNA catalogues in the absence of a genomic reference, based on so-called *de novo* transcriptome assembly. Besides efficiently connecting publicly available software, `FRAMA` incorporates custom-build tools to attenuate frequent *de novo* assembly issues. `FRAMA's` competitiveness with genome-based transcript reconstruction approaches was demonstrated by application to RNA-seq data obtained from the naked mole-rat (NMR). This non-model organism gains increasing popularity in ageing research due to its extreme lifespan of >30 years in captivity accompanied by lifelong fertility and an extreme resistance to ageing-related deterioration. Its lifelong fertility is exceptional, considering that NMRs are socially organized in colonies and one breeding female carries the metabolic load of reproduction and still lives at least as long as its non-breeding siblings. As part of my thesis, I investigated this trait by analysing gene expression profiles between reproductively active and inactive NMRs based on an mRNA catalogue delivered by FRAMA. I further applied `FRAMA` to non-model organisms at the other end of the vertebrate lifespan continuum – annual fishes (*Nothobranchius*). This fish genus, comprising one of the shortest-lived vertebrates, naturally evolved short lifespans presumably in adaptation to the transient availablility of water in their habitat. Based on positive selection analyses, we investigated when and how evolution shaped protein-coding genes that are potentially involved in such short lifespans.

Both types of analyses, gene expression in NMRs and positive selection in annual fishes, revealed interesting candidate genes in respect to ageing. Moreover, it shed light on sexual maturation in NMRs and on the evolution of short lifespans in annual fishes.

Zusammenfassung

Technologische Fortschritte haben es ermöglicht, ohne jegliches Vorwissen und in noch nie zuvor dagewesener Tiefe und Genauigkeit, Informationen über Genome und Transkriptome zu gewinnen. Diese Entwicklung hat die Herangehensweise an biologische Fragestellungen verändert, denn nun können Forscher Hypothesen in nahezu jeder Spezies entwickeln und überprüfen, ohne sich auf traditionelle Laborорganismen beschränken zu müssen. Allerdings sind genomweite Untersuchungen in solchen Nicht-Modellorganismen immer noch eingeschränkt, denn die Sequenzierung und rechnergestütze Rekonstruktion von großen Genomen ist ein aufwendiges und kostspieliges Unterfangen. Die Sequenzierung und Rekonstruktion von Transkriptomen bietet hingegen eine kostengünstige Alternative und Einblicke in den proteinkodierenden Teil des Genoms. Der Mangel an Programmen, welche ausgehend von Sequenzierungsdaten passende Transkriptkataloge für weitere Analysen anfertigen, begrenzte allerdings auch diese Anwendung auf Experten.

In meiner Dissertation habe ich eine Software (`FRAMA`) entwickelt, die alle notwendigen Schritte vollzieht um von Sequenzierungsdaten zu Transkriptkatalogen zu gelangen. Dazu werden sowohl öffentlich verfügbare als auch selbst angefertigte Programme effizient miteinander verbunden, um die Transkriptomassemblierung und -annotation vorzunehmen, sowie Schwachstellen der Assemblierung auszubessern. Die Wettbewerbsfähigkeit von `FRAMA` im Vergleich zu genombasierten Verfahren wurde durch Anwendung auf Sequenzierungsdaten des Nacktmulls demonstriert. Der Nacktmull rückt immer stärker in den Fokus der Alternsforschung, durch seine außergewöhnlich Lebensspanne von >30 Jahren in Gefangenschaft, die von einer extremen Widerstandskraft gegen altersbedingten Zerfall und einer lebenslangen Fruchtbarkeit begleitet wird. In Anbetracht seiner eusozialen Lebensweise, ist eine lebenslange Fruchtbarkeit bemerkenswert, denn in einer Kolonie trägt ein einzelnes Weibchen die metabolische Last der Fortpflanzung und wird dabei mindestens genauso alt wie ihre fortpflanzungsinaktiven Geschwister. Als Teil meiner Doktorarbeit habe ich mit Hilfe von Genexpressionsanalysen und basierend auf dem zuvor angefertigten Transkriptkatalog, fortpflanzungaktive und -inaktive Nacktmulle verglichen. Weiterhin habe ich `FRAMA` auf Sequenzierungsdaten von Saisonfischen der Gattung *Nothobranchius* angewandt. Diese Gattung zeigt, vermutlich in Anpassung an die kurze Verfügbarkeit von Wasser in ihrem Habitat, eine sehr kurze Lebensspanne und umfasst eines der kurzlebigstes bekannten Wirbeltiere. Die Transkriptkataloge dienten als Grundlage zur Analyse positiver Selektion um genetische Determinanten kurzer Lebensspannen zu identifizieren.

Beide Analysen haben interessante Genkandidaten in Bezug zur Alterung aufgedeckt und sowohl Einblicke in die sexuelle Reifung von Nacktmullen als auch der Evolution kurzer Lebenspannen in Saisonfischen gegeben.

Table of Contents

1 Introduction

A primary ingredient of life for a wide variety of species is the ageing process [1]. The biological phenomenon that we call ageing is defined as the progressive and irreversible decline in physiological function over time that inevitably ends in death [2]. This process of deterioration usually leads to impairments of health and reproduction, and an increased mortality risk, even if external sources of mortality, such as predators, are absent. Life expectancy, defined as the mean age a population of a certain age will reach, for humans has steadily increased in developed countries [3–5]. This increase is mainly the result of postponed mortality driven by advances in standard of living, education and health-care, including medicine, public health and nutrient supply [5,6]. However, the pace of deterioration with ageing is not slowing down and advanced age is a major risk factor for multiple diseases, including heart disease, cancer, diabetes and Alzheimer's disease. As a consequence of this development, more people reach advanced age, but suffer from these age-associated diseases [5,7]. Thus, despite investigating the molecular causes and effects of ageing to satisfy the human need to understand it [6], ageing research has become crucial to potentially identify strategies that minimize the lifetime spend with age-associated diseases [8].

Understanding ageing is challenging, because it is a complex trait that arises from interactions between multiple biological layers, as well as environmental (*e.g.* radiation) and behavioural factors (*e.g.* physical activity, diet) [9,10]. Recently, a conserved set of biological processes, so-called hallmarks, involved in ageing across different mammalian species have been grouped into three categories: (i) primary (genomic instability, telomere attrition, epigenetic changes, loss of proteostasis), (ii) antagonistic (mitochondrial dysfunction, cellular senescence, deregulated nutrient sensing) and (iii) integrative hallmarks (stem cell exhaustion, altered intracellular communication) [9]. Despite their interactions, the primary cause of ageing is the accumulation of cellular damages (primary hallmarks). Antagonistic hallmarks protect from or compensate these damages initially, but eventually contribute to ageing. Finally, integrative hallmarks arise from the previous two categories and cause the actual deterioration of physiological function [9]. Although understanding the exact molecular mechanisms behind the scenes is an ongoing quest, it is already clear that genes play a major role in these processes. This is shown by hundreds of gene manipulations in laboratory organisms capable of altering these hallmarks and thereby altering (*e.g.* accelerating or delaying) the ageing process [9,11].

Advances in technologies facilitate studying such complex traits by enabling data generation from multiple layers of biological systems, such as genomes, epigenetic patterns, transcriptomes and proteomes. These data can be obtained from biological samples, including tissue samples, whole-blood samples and even single cells. Over the last decade, in particular the development of high-throughput DNA sequencing methods (commonly referred to as

next-generation sequencing but more specifically 2nd-generation sequencing (2GS)) enabled in-depth analyses of biological systems. Among others, 2GS allows inference of sequence composition (sequencing) of whole genomes or targeted features of genomes, such as exons, DNA damage or DNA methylation [12–14]. Nevertheless, whole genome sequencing of large genomes is still costly and their full-length representation require computational reconstruction, which in turn requires sufficient computational resources, time and expertise. Also, additional efforts are needed to reach chromosome-scale resolution (*e.g.* optical mapping, linkage maps) and to annotate genomic features (*e.g.* repeats, genes) [15–19]. Targeted approaches selectively capture a subset of genomic regions and are usually cheaper, but require knowledge of the genomic reference sequence. In contrast, transcriptome sequencing (RNA-seq) is a less complex and, in all above mentioned practical aspects, more effective approach. RNA-seq operates without prior knowledge and provides expression level as well as sequence of transcripts [20]. Quantifying transcript levels helps to investigate which genes are expressed, *e.g.* in a specific organ, tissue or cell of any species under particular conditions and/or at defined time points. In respect to ageing, the comparison of expression levels between different phenotypes can reveal gene candidates that can serve as predictors for the pace of deterioration [21] or as targets to delay ageing [22]. Knowledge of transcript sequence allows to perform genome-wide computational analyses, *e.g.* to reveal sequence adaptations involved in biological traits and the evolutionary forces involved in shaping them [23,24]. In contrast to genome sequencing, RNA-seq can be applied by a broad range of researchers, as it requires lower financial budgets and computational resources, enabling the investigation of a wide range of phenotypes. As an additional benefit, even non-specialists in bioinformatics can analyse previously uncharacterized transcriptomes if user-friendly assembly pipelines are available.

In this thesis, I focused on the development and application of RNA-seq bioinformatic tools to study gene expression and protein-coding sequences (CDSs). Both factors are known to influence lifespan and ageing. In the following two sections, I will provide a general introduction to the transcriptome, the principles of RNA-seq and bioinformatics steps involved in analysing RNA-seq data as well as the organisms that are relevant for ageing research.

1.1 Transcriptome Bioinformatic Analysis

The transcriptome represents the complete set and abundance of transcripts (ribonucleic acids; RNAs) that are expressed in a biological sample at a certain time [20]. RNAs are transcribed from genomic templates and are generally categorized in (i) protein-coding messenger RNAs (mRNAs), (ii) non-protein-coding RNAs (ncRNAs) and (iii) spurious transcripts [25]. This work focuses on mRNAs, which are an important transmitter of biological information between the genome, providing the templates, and proteins, carrying out cellular functions. Nevertheless, all RNA categories are vital to understand the connections between

genotype and phenotype, and especially ncRNAs are increasingly recognized as key regulators of gene expression and are involved in ageing [9,26–28].

In the past, different approaches were developed to infer sequence composition (*e.g.* Sanger sequencing) and expression level (*e.g.* real-time polymerase chain reaction) of mRNAs, but only a small number of mRNAs could be studied simultaneously this way. The first technology that enabled measuring several thousand mRNAs in parallel were DNA microarrays introduced in 1995 [29]. Yet, DNA microarrays are only able to measure the relative expression of mRNAs and, based on hybridisation, require prior knowledge of mRNA sequences that are supposed to be analysed. Since 2005, the introduction of 2GS technologies (454 Life Sciences/Roche) allows to identify both sequence and relative expression at an unprecedented depth and accuracy [30,31]. These methods offer the possibility to study near-complete snapshots of transcriptomes in any species and without prior knowledge. In the following sections, I will describe the basic principle of RNA-seq technology and RNA-seq data analysis.

1.1.1 **RNA-seq principle**

RNA-seq, or whole-transcriptome shotgun sequencing, is the application of 2GS to transcriptomes. 2GS is characterized by massive parallelization of sequencing reactions, amplification of DNA fragments on a solid surface and "sequencing by synthesis" [31]. In principle, total RNA is isolated from biological samples, optionally filtered for target RNA species, and subsequently converted into libraries of complementary DNA fragments. These libraries are then partially sequenced from one end (single-end) or both ends (paired-end). The resulting short nucleotide sequences are called "reads" and have, depending on the sequencing platform, lengths ranging typically from 30-400 bp [12]. During this process, orientation of the original RNAs is lost, but can be preserved by using respective protocols [32]. However, if full-length mRNA sequences are required, these must be computationally reconstructed from reads.

Following this procedure (and sequencing errors and biases apart), an RNA-seq data set consists of reads obtained by random sampling of small sections from RNA molecules that are present in a biological sample. The sampling process is influenced by the abundance and length of RNAs – the higher the expression of an RNA and the longer the molecule, the more reads will be obtained [33]. This allows the quantification of gene/transcript expression levels relative to each other. However, a robust and direct relationship between the number of reads and the actual number of RNA molecules in the biological sample is not possible, unless control RNAs in known amounts are used to calibrate measurements [34]. The detection of low expressed RNAs is limited by the sample size, which is reflected by the number of reads. To investigate low expressed RNAs, the sequencing depth (*i.e.* sample size) needs to be adjusted accordingly and in dependency of the size of the transcriptome being assayed as well as the biological question (*e.g.* RNA discovery, RNA quantification) [35–38].

Within the framework of my thesis, I investigated RNA-seq data by two different routes of downstream analyses: (i) the quantitative route - studying mRNA expression changes between phenotypes - and (ii) the qualitative route - studying positive selection in CDSs. Both routes rely on knowledge of mRNA sequences that can be obtained from the RNA-seq data itself.

1.1.2 *De novo* transcriptome assembly from RNA-seq data

The ultimate goal of transcriptome assemblies is the full-length reconstruction of all transcript species of a transcriptome. Therefore, assembly algorithms were developed that computationally reconstruct full-length sequences *de novo*, *i.e.* without any reference sequence, from the collection of short sequencing reads [37,39–41]. Unfortunately, the number of reads is usually not sufficient to reconstruct all transcripts completely. Thus, assemblers are only able to reconstruct contiguous sequences (contigs) and many contigs represent only a small continuous section of a full-length transcript [37]. Even if a transcript is sufficiently covered by reads, complex loci with multiple exons and extensive alternative splicing events can present too many possible solutions to assembly algorithms and assemblers at best deliver most likely solutions [37].

The reduction of raw assemblies to an evolutionary reasonable set of transcript contigs relies on the identification of homologous transcript counterparts in another species (orthologs). Homology is the relationship between two genes that have descended from a common ancestral gene [42]. Homologs in different species with a last common ancestor before speciation are referred to as orthologs. Because descendants usually rely on the function of the ancestral gene, orthologs are expected to have retained a certain degree of sequence similarity and their function. Thereby, orthologs usually allow the transfer of functional information between species. Paralogs are another group of homologs with the last common ancestor before a gene duplication and are thought to be a major source of functional innovation (neofunctionalization, subfunctionalization) [43]. Homology by this definition is an evolutionary term, but in practice is inferred using sequence similarity [44]. Proper ortholog assignment would require evolutionary analysis, including alignment of transcripts from multiple species and phylogenetic tree reconstruction. However, genome-wide application of such methods is too computationally intensive given tens of thousands of genes (21,243 protein-coding genes in the human genome) [45], represented usually by an order of magnitude more transcript contigs. Therefore, heuristic approaches are commonly used to identify two genes that are mutually most similar between gene sets from two different species (best bidirectional hit) [44]. Besides reducing the assembly to an evolutionary meaningful subset of RNAs, orthologs can be used to determine structural information, such as CDS of mRNAs, and optimize assemblies. This optimization includes the correction of misassembled contigs that arise from falsely connected transcripts originating from adjacent genes, or fragmentation of contigs corresponding to single transcripts [37].

After such a functional and structural annotation of raw assemblies, the subset of evolutionary reasonable contigs, representing ideally full-length RNA sequences, provide the basis for further downstream analysis.

1.1.3 Studying gene expression profiles using RNA-seq data

Gene expression profiling investigates the expression of thousands of genes simultaneously. One of the main goals is to identify statistically significant quantitative changes in gene expression profiles between experimental groups (*e.g.* different conditions, age cohorts). This procedure aims to understand the molecular response, *e.g.* to different stimuli or changes during ageing [46].

To determine expression profiles, reads are usually aligned to a reference genome sequence. If the species' genome sequence is unknown, the genome of a closely-related species can be used or RNA sequences need to be computationally reconstructed and used for alignment. The number of aligned reads (referred to as read count) to a certain region (*e.g.* gene, exon or splice-junction) normalized by the number of sequenced reads can then be used to compare expression levels of genes across samples. To compare expression levels between genes within an RNA-seq sample, read counts need to be additionally normalized by the length of the region [33,46,47]. Different methods utilize read count information to identify a subset of genes showing statistically significant changes across experimental groups (differentially expressed genes; DEGs) [48,49]. Once orthologs are determined, DEGs can also be identified between different species. In order to avoid artefacts in such an analysis, *e.g.* arising from different ortholog lengths, commonly transcribed regions should be determined first [50].

Insights into biological processes are often challenging, even at the level of DEGs. Several tools have been developed to investigate biological processes more broadly, based on predefined gene sets [51–53]. These gene sets categorize genes by common functions or otherwise meaningful criteria (*e.g.* common regulation or location). The Gene Ontology is usually the catalogue of choice for such an analysis [54], but other databases also provide useful gene sets, including "Kyoto Encyclopedia of Genes and Genomes" (KEGG) [55] and Reactome [56] for metabolic pathways, and Digital Ageing Atlas [57] and GenAge [58] for ageing-associated genes. Different methods enable the identification of significantly enriched (over-represented) gene sets, *e.g.* based on DEGs using Fisher's Exact Test or at the level of gene expression (*e.g.* read count, fold change) using `GAGE` to identify gene sets showing altered expression, *e.g.* a common up- or down-regulation [51,59,60].

1.1.4 Studying genome-wide genetic selection by positive selection

One approach to investigate lineage-, species-, or population-specific adaptations via comparative studies is the sequence-based identification of positively selected genes (PSGs). Positive selection describes the selective force that promotes the fixation of advantageous genetic mutations by natural selection [61]. While the concept of positive selection is based on

the simple principle that mutations increasing the chance of survival and reproduction of an organism have a higher chance of being transmitted to the next generation, its detection is complex and requires several steps. In a single gene, detection of positive selection includes (i) identification of orthologs, (ii) accurate multiple sequence alignment (iii) phylogenetic tree reconstruction and (iv) identification of signals of positive selection. A genome-wide application additionally requires robust filtering to remove false positives arising from incomplete sequences, alignment errors or poor conservation [62].

One concept to identify signals of positive selection in two orthologous CDSs relies on the ratio of non-synonymous (amino acid-changing, dN) and synonymous (silent, dS) nucleotide exchanges. Because positive selection promotes changes, non-synonymous exchanges are assumed to be fixed at a higher rate than synonymous exchanges. In consequence, PSGs should show dN/dS ratios significantly greater than 1 [61]. Several statistical methods were developed to test for positive selection in individual phylogenetic branches and in a set of branches [61,63]. Although these methods became increasingly successful in PSG identification, they can be hampered by other selective forces or sequence divergence [63,64]. Nevertheless, comparative studies using this concept proved useful to investigate the links between molecular adaptations and ageing [24,65–67].

1.2 Animal models in research on ageing

As introduced in the beginning, almost all multicellular organisms underlie ageing, a complex time-dependent process that negatively affects the function of multiple organs and increases disease susceptibility. Interestingly, looking at lifespans across different animal species, the maximum lifespan correlates with body mass [68]. However, humans deviate from this relation and are exceptionally long-lived when normalized by their body mass [69]. Nevertheless, the elderly human population becomes increasingly unhealthy with advanced age and suffers from age-associated diseases [5,8]. Fortunately, ageing research has led to remarkable breakthroughs towards the understanding of molecular mechanisms involved in ageing [7,9,70]. Conserved effects of interventions and roles of signalling pathways across different organisms revealed that the pace of ageing is rather flexible and shapeable than fixed [6,8,9]. Among these conserved mechanisms having lifespan modulating effects are: calorie restriction, mitochondrial function, protein turnover, insulin/IGF-signalling, target of rapamycin signalling and sirtuin function [9,70].

Comparative biology has been an essential tool in ageing research for more than a century, starting with the comparison of metabolic rates and their relationship to lifespan in different species (rate-of-living hypothesis) [65–68,71–73]. Yet, most insights into ageing relied on short-lived laboratory organisms to investigate environmental and molecular factors of ageing [9,70,74]. Nowadays, supported by advances in sequencing technologies (2GS, mass spectrometry [75]) and molecular techniques (RNA interference [76], CRISPR/Cas9 [77], induced pluripotent stem cells [78]), research becomes increasingly independent from

traditional laboratory organisms. This progress allows to generate and verify hypotheses about the mechanisms of ageing in a broad range of species. Further, it enables studying particular species having interesting traits for ageing research, such as resistance to age-associated diseases or exceptionally long or short lifespans [79].

The following sections present a selection of organisms that are becoming increasingly popular in ageing research and describe particularly those that I investigated in the course of my thesis. Maximum lifespans are presented according to the "The Animal Ageing and Longevity Database" (AnAge, genomics.senescence.info/species).

1.2.1 **What is a model organism in biology?**

Molecular researchers need experiments to gain insights into complex interactions in biological systems, *e.g.* by investigating the molecular response to external stimuli. However, experiments in the species of interest can be unethical or otherwise unfeasible to perform, *e.g.* because of costly housing or long lifespans. Therefore, model systems that ideally replicate the biological system of interest are used as proxies. Mathematical modelling of multicellular organisms [80] or single organs [81], which would allow computational simulation of experiments, is still in its early stages and living organisms are primarily used as models in biology. The possibility to exploit one particular species to understand molecular mechanisms in another is based on evolutionary ancestries and the conservation of molecular pathways across species. Yet, even slight lineage- or species-specific adaptations can cause altered or novel functions that are only shared between closely related species or limited to a single species [74]. Further, adaptation through the process of natural selection can yield multiple, independent solutions to the same problem [82]. Thus, it is important to choose model organisms wisely to address the research question of interest and transfer gained knowledge to the target species [74,83].

Ideal models are closely related to the target species and share similar traits, *e.g.* in reproduction, diseases and ageing. Still, experiments in animals also require practical and economic factors, including known genes and regulatory regions (ideally a well annotated genome), good health in captivity, inexpensive housing, easy breeding and manipulable genetics and environment [74]. In ageing research, longitudinal and lifespan studies also require comparably short lifespans. Yet, assuming that resistance to ageing must be maintained throughout life, studying young adults of long-lived species should also be rewarding [84]. Traditional model organisms in ageing research are yeasts (*Saccharomyces cerevisiae*, 2 weeks), nematodes (*Caenorhabditis elegans*, 2 months), flies (*Drosophila melanogaster*, 3.5 months) and rodents, in particular mice (*Mus musculus*, 4 years) and rats (*Rattus norvegicus*, 4 years) [9,70,74]. These models fulfil most practical and economic factors mentioned above and are therefore perfectly suitable as experimental systems. Yet, they are only distantly related to humans and on the opposite end of the lifespan continuum (short-lived). Even the closest relatives to humans among these models (mouse/rat) show substantially distinct biological traits, including

differences in reproductive biology (*e.g.* oestrus vs. menstrual cycle; number of pregnancies and offspring) and absence of human age-related diseases (*e.g.* atherosclerosis, Alzheimer's) [74]. This suggests that evolution shaped ageing significantly different in these species [74] and the potential lack of mechanisms to delay ageing might limit research on health- or lifespan extension [79].

1.2.2 **Non-model organisms in ageing research**

Nowadays, mentioned technological advances allow to investigate and manipulate organisms apart from currently established laboratory models, commonly called "non-model" organisms, at the molecular level at similar depth and with a similar efficiency. In respect to ageing, this allows to extend the list of traditional models to more appropriate models for human ageing. Several non-model organisms have been suggested to understand ageing in general, such as biologically immortal hydras and extremely long-lived quahogs (*Arctica islandica*, 507 years), and human ageing in particular, such as nonhuman primates and domesticated species (dogs, cats) that mimic human ageing (particularly brain ageing) [1,74,79,85,86]. Non-model rodent and fish species are of particular interest in my thesis as proxies for human ageing in different aspects.

Rodents are particularly suitable to understand ageing in mammals using comparative biology [85]. Besides well-described laboratory models, this order comprises 2,227 species [87] showing substantially different lifespans even among equally sized animals (*Heterocephalus glaber*, 31 years; *Mus musculus*, 4 years; >7-fold difference). Longevity has presumably evolved independently in different families of rodents, such as mole-rats (Bathyergidae; *Heterocephalus glaber*, 31 years), porcupines (Hystricidae; *Hystrix brachyura*, 27 years), beavers (Castoridae; *Castor canadensis*, 23 years) and squirrels (Sciuridae; *Sciurus carolinensis*, 23 years) [85]. This offers the possibility to discover multiple different mechanisms leading to longevity. In this respect, my work focuses on an extremely long-lived member of the mole-rat clade – the naked mole-rat (detailed in 1.2.3).

Between the comparably closely-related vertebrate models (mouse/rat) and distantly-related invertebrate models (nematode, fly) is another attractive group of animals - fish. Similar to rodents, these vertebrates show a wide variety of different species and lifespans [88]. Ageing studies in guppies (*Poecilia reticulate*, 5 years) demonstrated the potential of fishes as model systems and the zebrafish (*Danio rerio*, 5 years) already emerged as a valuable model particularly in developmental biology (*e.g.* transparent embryos, rapid development, low maintenance cost) [88]. Yet, although its regenerative capabilities are of particular interest, its relatively long lifespan hampers lifelong experimental studies. Due to their short lifespans and rapid life cycles, annual fishes of the genus *Nothobranchius* (detailed in 1.2.4) are particularly interesting for ageing research.

1.2.3 **Rational for the naked mole-rat as a potential model for ageing research**

The naked mole-rat (NMR, *Heterocephalus glaber;* family: Bathyergidae) is a mouse-sized rodent (~35 g) native to hot dry regions of northeast Africa (Somalia, Kenya, Ethiopia) [89]. Although not immune to ageing, these rodents have a remarkable lifespan of >30 years in captivity and, similar to humans, live five times longer than expected by their body mass [69]. Yet, unlike humans (and mice/rats), NMRs are extremely resistant to cancer [90] and show negligible signs of ageing, including no age-related decline in fertility and no increase in mortality rate until sudden death [89]. Still, few signs of ageing can be observed, such as osteoarthritis and skin ageing [69,91]. However, these signs manifest very late in life (~24-27 years), which makes it a perfect model in respect to health span extension in humans [69,89].

Despite showing a healthy and long life, the NMR has another exceptional trait – its eusociality. NMRs form social colonies of up to 300 non-breeding subordinates and one breeding pair of animals [92]. Non-breeders are capable of ascending into breeders even as adults and breeders stay fertile throughout their lifespan. Although currently subject to controversy [93], this eusocial colony structure was thought to be accompanied by different lifespans in NMR breeders and non-breeders, as observed in *Fukomys mechowii* [94], a close relative of the NMR, and in eusocial insects [72]. In *F. mechowii* female and male breeders live 2.2 and 1.5-times, respectively, longer than non-breeders of the same sex [94]. This enables studying different ageing rates within the same genotype. Even if NMR non-breeders and breeders age at equal pace, insights into adaptations that prevent the decline in activity and fertility regardless of the high metabolic demands of pregnancies and lactation are valuable as this contradicts the "disposable soma theory" of ageing [95]. This theory suggests that, due to the organism's limited energy budget, ageing results from a compromise between investment of energy in somatic maintenance and reproduction [95].

In the case of the NMR, two independent draft genome assemblies from 2GS of genomes are available (hetgla1 [96]; hetgla2 [97]) and a third unpublished meta assembly, which combines the previous assemblies, has been performed [67]. However, genome assemblies from 2GS data are often fragmented, have a high gap percentage and homologous genes might have been collapsed, limiting downstream analyses [15]. In addition, finishing a genome, *e.g.* closing assembly gaps, fixing misassembled regions and annotating genes, is still time consuming and expensive [18,19]. As a consequence, the genomic references for the NMR are far from contiguous and complete. So far, transcript sequences have mainly been derived by genome-based gene predictions in hetgla1 and hetgla2. Consequently, their quality largely depends on the quality of the underlying genome assembly. Another important drawback is the inference of untranslated regions (UTR), which cannot be predicted from the genome yet [98]. However, UTRs provide valuable information regarding mRNA stability and mRNA localization [99]. Moreover, UTRs are the landing platform for microRNAs, which gain

increasing attention in ageing research [26,99]. A steadily increasing number of publications on NMR cancer resistance and ageing over the last 8 years [90,100–103] emphasize the efforts that have already been made to establish the NMR as a model organism.

1.2.4 Rational for annual fishes as potential models for ageing research

In contrast to NMRs, annual fishes (genus *Nothobranchius*) offer the possibility to investigate the other end of the lifespan continuum as they show short lifespans. Most of the 43 described species in this genus inhabit ephemeral pools across Africa formed by seasonal rainfalls [104]. Therefore, their life expectancy is dictated by the cessation of their transient habitat and differences in maximum lifespan across different species correlates with the availability of water in their habitat [104]. Among annual fishes, *N. furzeri* shows the most extreme short lifespan of 3-7 months [105], accompanied by a rapid life cycle and early sexual maturation [104]. Such short lifespans enable longitudinal studies and quick, monitorable experiments [21]. The short lifespan of *N. furzeri* in particular is associated with rapid ageing and typical signs of ageing (*e.g.* cancerous lesions, decline in fertility and cognitive function).

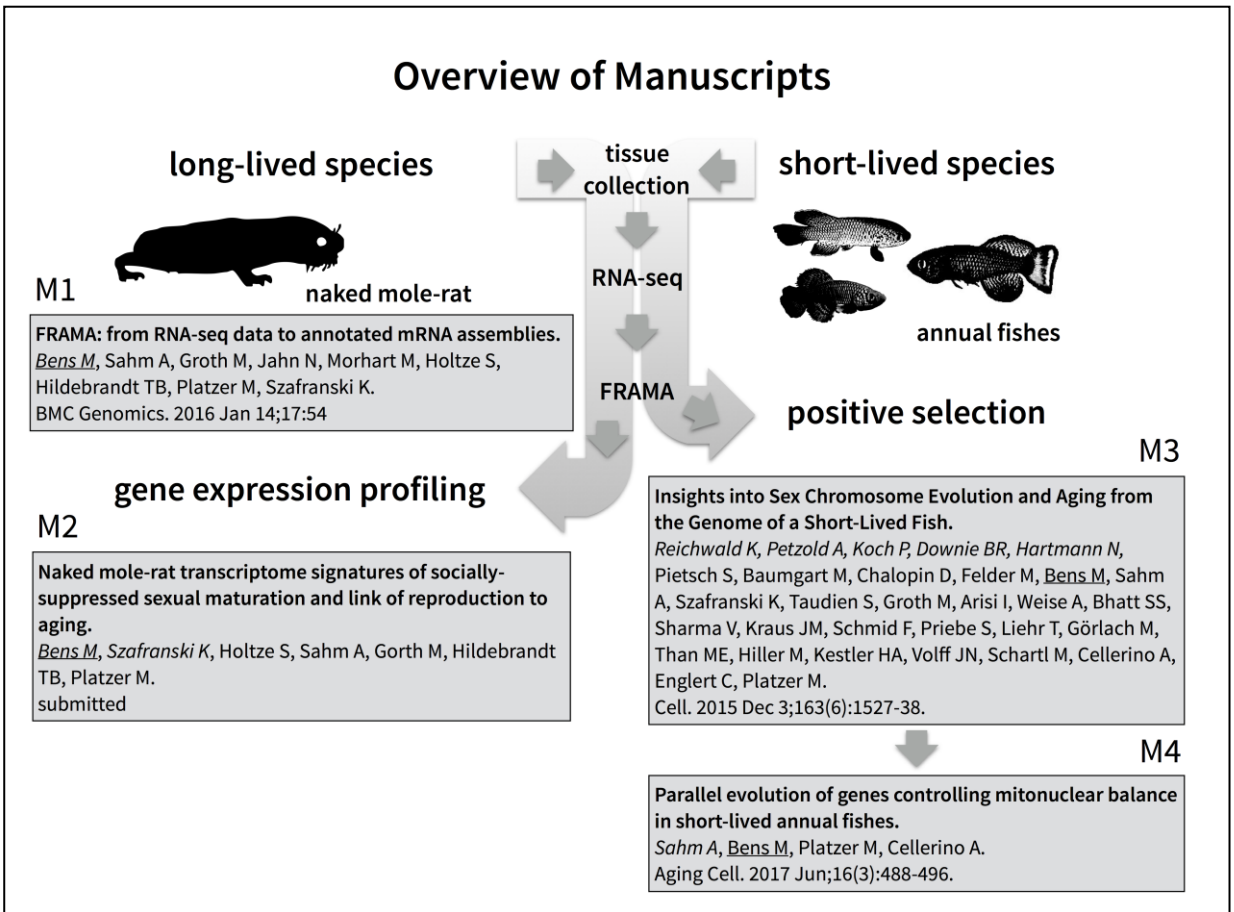


Figure 1: Overview of manuscripts. `FRAMA` (M1) was developed to automate the process from *de novo* transcriptome assembly to a non-redundant set of mRNA sequences suitable for downstream analyses. mRNA catalogues delivered by `FRAMA` provided the basis for two different routes of downstream analyses in two groups of non-model organisms. In long-lived naked mole-rats gene expression was studied (M2), while in short-lived annual fishes positively selected genes were analysed (M3, M4).

Importantly, the involved mechanisms overlap with mammalian hallmarks of ageing [106] and their onset can be modulated, *e.g.* by water temperature, calorie restriction or resveratrol treatment [107,108]. These characteristics, including practical features, such as comparably low requirements in water quality and eggs storable at room temperature for several months [104], make *Nothobranchii* attractive animal models for ageing research.

Since suggested in 2003 [105], *N. furzeri* is rapidly developing as a new model organism in ageing research. Similar to the NMR, two genome assemblies have been performed [109,110], but with greater efforts to reach chromosome-scale resolution [106]. Although this allows thorough investigation of *N. furzeri*, studies on how such short lifespans evolved across different annual fishes are limited. Potential efforts to investigate additional annual fishes could involve genome sequencing as the *N. furzeri* genome should provide a good reference to improve 2GS-based genome assemblies. However, this is still costly even for one species and RNA-seq provides a cost-effective approach to gain insights into protein-coding genes of multiple different species.

1.3 Connections between Manuscripts

Molecular studies in any species greatly benefit from the availability of genetic information. Researchers prefer well-annotated reference genomes. These references provide a plethora of information, from comprehensive templates for mRNAs and ncRNAs, to regulatory motifs and repeat families, and can be studied using comparative genomic approaches that allow to investigate the history of unique and common biological traits across species. Nevertheless, current strategies to assemble reference genomes to chromosome-scale resolution are costly and time intensive. An alternative approach that allows investigating the protein-coding portion of genomes is provided by RNA-seq. Within the scope of my thesis, four scientific papers were published in or submitted to peer-reviewed journals, exemplifying how qualitative and quantitative investigation of transcriptomes can provide insights into molecular mechanisms and adaptation in ageing (Figure 1). Therefore, I developed the software framework `FRAMA` (**F**rom **R**NA-seq to **a**nnotated **m**RNA **a**ssemblies) that automates the assembly, annotation and optimization of protein-coding transcripts from RNA-seq data (M1). `FRAMA` was applied to transcriptomes of non-model species on both ends of the vertebrate lifespan continuum, including the extremely long-lived NMR and a selection of extremely short-lived annual fishes. The delivered mRNA catalogues provided insights into (i) NMR transcriptome signatures of socially-suppressed sexual maturation and links of reproduction to ageing (M2), and (ii) PSGs related to short lifespan in annual fishes (M3, M4).

2 Overview of Manuscripts

2.1 Contribution to each work

Table 1: Overview of manuscripts and proportion of my contribution to each work.

Manuscript	Citation	Contribution
M1	**FRAMA: from RNA-seq data to annotated mRNA assemblies.** Bens M, Sahm A, Groth M, Jahn N, Morhart M, Holtze S, Hildebrandt TB, Platzer M, Szafranski K BMC Genomics. 2016 Jan 14;17:54 published	60%
M2	**Naked mole-rat transcriptome signatures of socially-suppressed sexual maturation and link of reproduction to aging** Bens M, Szafranski K, Holtze S, Sahm A, Groth M, Hildebrandt TB, Platzer M submitted	40%
M3	**Insights into Sex Chromosome Evolution and Aging from the Genome of a Short-Lived Fish.** Reichwald K, Petzold A, Koch P, Downie BR, 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 SS, Sharma V, Kraus JM, Schmid F, Priebe S, Liehr T, Görlach M, Than ME, Hiller M, Kestler HA, Volff JN, Schartl M, Cellerino A, Englert C, Platzer M Cell. 2015 Dec 3;163(6):1527-38. published	5%
M4	**Parallel evolution of genes controlling mitonuclear balance in short-lived annual fishes.** Sahm A, Bens M, Platzer M, Cellerino A. Aging Cell. 2017 Jun;16(3):488-496. published	20%

2.2 Manuscript 1 (M1)

Title: FRAMA: from RNA-seq data to annotated mRNA assemblies

Status: published in *BMC Genomics. 2016 Jan 14;17:54*

Authors: Martin Bens (MB), Arne Sahm (AS), Marco Groth (MG), Niels Jahn (NJ), Michaela Morhart (MM), Susanne Holtze (SH), Thomas B. Hildebrandt (TBH), Matthias Platzer (MP) and Karol Szafranski (KS)

Summary: I developed the software framework `FRAMA`, which assembles, annotates and optimizes transcripts from RNA-seq data. This work describes and assesses every step in `FRAMA`, with special focus on post-assembly tasks, including reduction of redundant transcript contigs, correction of misassembled transcripts, scaffolding of fragmented transcripts and coding sequence identification. `FRAMA` was applied to RNA-seq data obtained from deep sequencing the transcriptome of the naked mole-rat, a promising non-model organism in ageing research.

Authors' contribution: MP, TBH and KS conceived the project. MM, SH, TBH, MB, MG and KS performed the tissue sampling and sequencing experiments. MB, AS, NJ, MP and KS designed and implemented the software. MB, AS, MP and KS performed the validation analysis and discussed the results. MB, MP and KS wrote the paper.

2.3 Manuscript 2 (M2)

Title: Naked mole-rat transcriptome signatures of socially-suppressed sexual maturation and links of reproduction to ageing

Status: submitted to *Genome Research (June 2017)*

Status: **Authors:** Martin Bens (MB)*, Karol Szafranski (KS)*, Susanne Holtze (SH), Arne Sahm (AS), Marco Groth (MG), Thomas B. Hildebrandt (TBH), Matthias Platzer (MP). * shared first authorship

Summary: This work comprises a comparative gene expression study of breeding and non-breeding eusocial, long-lived naked mole-rats and polygynous, not long-lived guinea pigs. Therefore, we accumulated and analysed an extensive set of transcriptome data from ten different tissues of both species. We revealed minor differences between sexes of non-breeding naked mole-rats in contrast to guinea pigs, providing additional evidence for socially suppressed sexual maturation in naked mole-rats. Expression differences between non-breeding and breeding NMRs suggest links to ageing hallmarks by influencing mitochondrial activity and lipid metabolism.

Authors' contribution: TBH, MP and KS conceived the project. SH, TBH, MB, MG and KS performed the animal study, sampling and sequencing experiments. Data analysis and interpretation were performed by MB, KS, SH, MP and TBH. The manuscript was written by MB and KS. TBH and MP are joint senior authors.

2.4 Manuscript 3 (M3)

Title: Insights into Sex Chromosome Evolution and Aging from the Genome of a Short-Lived Fish

Status: published in *Cell. 2015 Dec 3;163(6):1527-38.*

Authors: Kathrin Reichwald (KR)*, Andreas Petzold (AP)*, Philipp Koch (PK)*, Bryan R. Downie (BRD)*, Nils Hartmann (NH)*, Stefan Pietsch (SPe), Mario Baumgart (MBa), Domitille Chalopin (DC, Marius Felder (MF), Martin Bens (MBe), Arne Sahm (AS), Karol Szafranski (KS), Stefan Taudien (ST), Marco Groth (MG), Ivan Arisi (IA), Anja Weise (AW), Samarth S. Bhatt (SSB), Virag Sharma (VS), Johann M. Kraus (JMK), Florian Schmid (FS), Steffen Priebe (SPr), Thomas Liehr (TL), Matthias Görlach (MGo), Manuel E. Than (MET), Michael Hiller (MH), Hans A. Kestler (HAK), Jean-Nicolas Volff (JNV), Manfred Schartl (MS), Alessandro Cellerino (AC), Christoph Englert (CE), Matthias Platzer (MP). * shared first authorship

Summary: Providing a chromosome-scale draft genome sequence, including annotation of protein-coding genes and several classes of non-coding RNAs, for the *N. furzeri*, this publication presents a milestone in establishing this fish as a model organism in molecular research. Despite investigating sex chromosome evolution and reporting *gdf6* gene expression as a marker for sex determination, the age-related analyses included gene expression comparisons between young and old animals in multiple tissues. Positional enrichment revealed non-random organization of temporally regulated DEGs in the genome, suggesting co-regulation of functionally associated genes. I was primarily involved in analysing positive selection in the *N. furzeri*, which indicated also functional adaptation in temporally regulated DEGs.

Authors' contribution: CE, KR, and MP initiated, managed, and drove the genome project. KR, NH, MBa, ST, and MGr prepared the samples. KR, ST, and MGr performed the sequencing AP, PK, BRD, VS, and MH performed the genome assembly and annotation. PK, BRD, DC, and JNV performed the repeat analysis MBa, MGr, AC, and MP performed the mRNA analysis. IA, MBa, AP, and AC performed the miRNA analysis. KR, AW, SSB, and TL performed the chromosome FISH. KR, AP, PK, MF, KS, NH, MS, CE, and MP performed the sex chromosome evolution analysis. MGo and MET performed protein structure modelling. JMK, FS, SPr, PK, HAK, AC, and MP performed the positional gene enrichment analysis. AS, MBe, AP, BRD, AC, and MP performed the positive selection analysis. NH, SPi, and CE performed the diapause analysis. All authors contributed to data interpretation. KR, AP, PK, NH, MS, AC, CE, and MP wrote the manuscript.

2.5 Manuscript 4 (M4)

Title: Parallel evolution of genes controlling mitonuclear balance in short-lived annual fishes

Status: published in *Aging Cell. 2017 Jun;16(3):488-496.*

Authors: Arne Sahm (AS), Martin Bens (MB), Matthias Platzer (MP), Alessandro Cellerino (AC)

Summary: In this work, we present a follow-up study of the positive selection analysis in M3 to investigate the evolution of annual life history in more detail. Again, we performed positive selection analysis, but using a broader range of non-annual outgroup species and by analysing deeper branches of the *N. furzeri* phylogenetic tree. In annual fishes, we identified positively selected genes involved in all steps of mitochondrial biogenesis, a conserved longevity mechanism in model organisms, suggesting a causal link between positively selected genes and annual life history. Further, we identified signs of parallel evolution in two different lineages of annual fishes in a subset of these genes and overlaps with positively selected genes in long-lived mammals.

Authors' contribution: AS and MB performed the analysis; MP and AC supervised the work; and AS, MP, and AC wrote the manuscript.

3 Manuscripts

3.1 Manuscript 1 (M1)

Bens et al. BMC Genomics (2016) 17:54
DOI 10.1186/s12864-015-2349-8

BMC Genomics

SOFTWARE Open Access

FRAMA: from RNA-seq data to annotated mRNA assemblies

Martin Bens[1], Arne Sahm[1], Marco Groth[1], Niels Jahn[1], Michaela Morhart[2], Susanne Holtze[2], Thomas B. Hildebrandt[2], Matthias Platzer[1] and Karol Szafranski[1*]

Abstract

Background: Advances in second-generation sequencing of RNA made a near-complete characterization of transcriptomes affordable. However, the reconstruction of full-length mRNAs via de novo RNA-seq assembly is still difficult due to the complexity of eukaryote transcriptomes with highly similar paralogs and multiple alternative splice variants. Here, we present FRAMA, a genome-independent annotation tool for de novo mRNA assemblies that addresses several post-assembly tasks, such as reduction of contig redundancy, ortholog assignment, correction of misassembled transcripts, scaffolding of fragmented transcripts and coding sequence identification.

Results: We applied FRAMA to assemble and annotate the transcriptome of the naked mole-rat and assess the quality of the obtained compilation of transcripts with the aid of publicy available naked mole-rat gene annotations. Based on a de novo transcriptome assembly (Trinity), FRAMA annotated 21,984 naked mole-rat mRNAs (12,100 full-length CDSs), corresponding to 16,887 genes. The scaffolding of 3488 genes increased the median sequence information 1.27-fold. In total, FRAMA detected and corrected 4774 misassembled genes, which were predominantly caused by fusion of genes. A comparison with three different sources of naked mole-rat transcripts reveals that FRAMA's gene models are better supported by RNA-seq data than any other transcript set. Further, our results demonstrate the competitiveness of FRAMA to state of the art genome-based transcript reconstruction approaches.

Conclusion: FRAMA realizes the *de novo* construction of a low-redundant transcript catalog for eukaryotes, including the extension and refinement of transcripts. Thereby, results delivered by FRAMA provide the basis for comprehensive downstream analyses like gene expression studies or comparative transcriptomics. FRAMA is available at https://github.com/gengit/FRAMA.

Keywords: RNA-seq, Transcriptome assembly, Full-length mRNA, Naked mole-rat

Background

Since decades, characterization of transcriptomes by random sequencing of cDNA has been practiced to decipher the gene repertoire for a large number of organisms [1–4]. The resulting compilation of mRNA sequences, a so-called transcript catalog, is an important fraction of the functional genetic information and serves as a basis for multiple downstream analyses including gene expression studies, using either microarray techniques or tag sequencing, as well as comparative sequence analyses [5, 6]. Particularly, the full-length protein-coding sequence (CDS) represents a crucial entity forming a knowledge base in genetics research [7]. Fragmentary information will lead to incomplete, ambiguous, or even mislead conclusions in downstream analyses. While in principle, a genome-wide catalog of CDSs can also be derived from a genome sequence using gene prediction programs, it is nowadays a standard to support gene predictions with mRNA sequence evidence [8–11]. Transcriptome sequencing is also able to characterize untranslated regions (UTRs) [12], which cannot be predicted from the genome ab initio. UTRs include the landing platforms for potential regulatory interactions with micro-RNAs and, in combination with genomic sequence, also allow definition of promoter regions, both of which are important for functional gene analysis.

* Correspondence: szafrans@fli-leibniz.de
[1]Leibniz Institute on Ageing - Fritz Lipmann Institute, Beutenbergstr. 11, 07745 Jena, Germany
Full list of author information is available at the end of the article

While the introduction of second-generation sequencing of RNA (RNA-seq) made the characterization of transcriptomes very affordable, the short-read RNA-seq data cannot display mRNA molecules in their entirety. Therefore, assembly programs were designed to reconstruct, as good as possible, full-length mRNA sequences from short RNA-seq reads [13, 14]. While these assembly programs have reached an accepted level of quality, they still face severe difficulties. The sequence depth of RNA-seq may be sufficient to detect rare mRNAs but, often, is still too low to allow reconstruction of their entire structure, which results in fragmented transcript contigs. In addition, eukaryotic transcriptomes are very complex by showing several alternative splice variants per gene, multiple gene copies, single nucleotide polymorphisms and transcribed pseudogenes. It is noteworthy that, for protein-coding genes, even the most highly expressed transcript is not necessarily protein-coding [15].

Functionally relevant signatures of non-model organisms in comparison to related organisms, such as gene content and transcript structures, can be read out most conveniently using a low redundancy subset of the transcript assembly. Identification of this representative assembly subset is possible by orthologous inference. In the past, complex algorithms have been developed for genome-wide identification of orthologous and homologous groups between different species [16]. Nevertheless, best available contigs may still show peculiarities, such as incompleteness, retained introns or splicing variants with premature stop codons. Additionally, overlapping genes may result in fusion contigs [17]. Thus, starting from de novo transcriptome assembly, strategies are required to scaffold fragmented contigs, to isolate single transcripts from fusion contigs, and to select or correct contigs in order to show the likely protein-coding transcript variant. Several of these illustrated tasks have been previously addressed in the course of project-specific assembly/annotation projects [18–21], but were not yet incorporated into re-useable software concepts.

Here, we present a genome-independent software tool (FRAMA) that specifically addresses post transcript assembly tasks for eukaryote transcriptomes. These tasks include reduction of assembly redundancy, ortholog-based gene symbol assignment, correction of fusion transcript contigs and scaffolding of fragmented transcript contigs, CDS identification and clipping of weakly supported sequence termini. We applied this pipeline to de novo assembly and annotation of the transcriptome of the naked mole-rat (NMR; *Heterocephalus glaber*), the longest-living rodent known and a promising non-model organism in ageing research [22, 23]. Two independent NMR genome assemblies and associated gene annotations are available [24, 25] and were used for a validation of our pipeline results. The comparison of the different approaches for gene model construction indicates that FRAMA is competitive and fulfills accepted quality standards.

Implementation

FRAMA is a novel software suite that calls components written in Perl and external software (Additional file 1: Table S1), applicable on UNIX/Linux and MacOS computer systems. Mandatory required input are RNA-seq read data, either paired-end or single-end, strand-specific or non strand-specific, and a comprehensively annotated transcriptome of a related species. FRAMA executes in 8 successive steps: (i) assembly, (ii) primary processing, (iii) gene symbol assignment, (iv) fusion detection, (v) scaffolding, (vi) identification of CDS, (vii) identification of mRNA boundaries, and (viii) descriptive assembly statistics (Fig. 1). Software parameters for each step can easily be edited in a parameter file. FRAMA produces a representative compilation of transcripts, a so-called transcript catalog, with CDSs and mRNA boundaries annotated. In the transcript catalog, each transcript will have a one-to-one relationship to an orthologous transcript in the reference transcriptome.

Assembly and primary processing

A variety of de novo transcriptome assembly tools are available, which perform differently well on separate subsets of transcripts [14]. FRAMA currently utilizes Trinity, an allrounder that performs well across different species

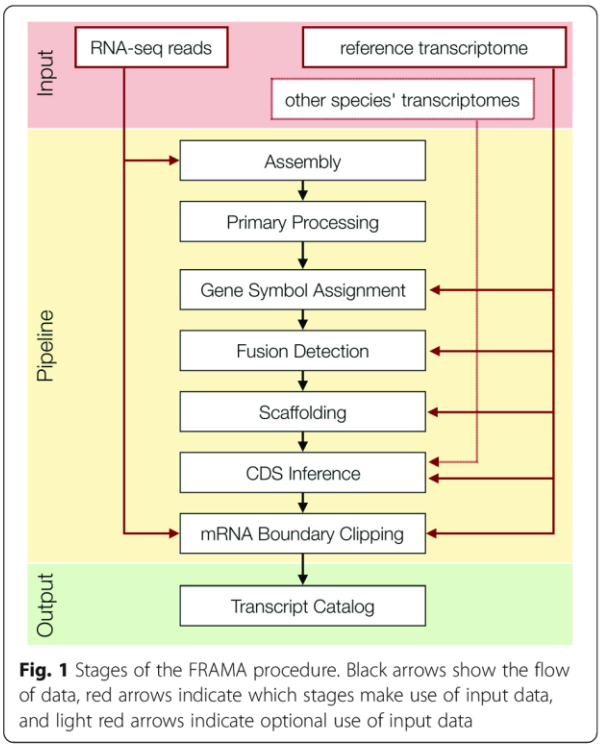


Fig. 1 Stages of the FRAMA procedure. Black arrows show the flow of data, red arrows indicate which stages make use of input data, and light red arrows indicate optional use of input data

and library properties [13, 18, 19]. Trinity starts with a greedy assembly of linear contigs based on the most frequent k-mers to reconstruct one full-length isoform per locus and additional unique regions partially. Then, overlapping contigs are clustered and connected into a de Bruijn graph, which represents different alternative splice variants for one locus or highly similar homologs. Finally, Trinity reports contig sequences that represent probable paths through each graph [13].

NCBI recommends scanning of transcript assembly data for adapter, vector and other cross-project contaminations that might occur. Accordingly, FRAMA examines the final scaffolded and annotated transcriptome for vector contamination using NCBIs VecScreen criteria [26], and match regions are annotated with match score and topological category.

Redundancy among transcript contigs can arise from shorter transcript contigs which are fully embedded in longer contigs or from local differences arising from sequencing errors or allelic variations. In order to reduce redundancy, in an optional step, transcript contigs are clustered using CD-HIT-EST. The cluster will then be replaced by the longest representative contig. Additionally or alternatively, TGICL can be used to combine overlapping transcript contigs into single longer contigs. Order of execution of both software programs can be chosen arbitrarily.

Assignment of gene symbols

Gene symbol assignment to transcript contigs is performed on the nucleotide level, based on best bidirectional BLASTN hits (BBH) against CDSs of an orthologous reference transcriptome. This enables the most sensitive differentiation of paralogous proteins. For example, the genes *CALM1*, *CALM2* and *CALM3* express identical proteins, in the NMR and other mammals, but differ in their CDS (Additional file 2: Figure S1). As an additional advantage of the nucleotide-level search, the identification of CDS for BLASTP or more time-consuming BLASTX searches are not necessary. Following the gene symbol assignment based on BBHs, remaining unassigned transcript contigs that show a single best hit (SBH) to an unassigned reference transcript are labeled and added to the transcript catalog. Annotated transcript contigs become oriented according to its assigned ortholog, which is essential if unoriented read data are used for assembly.

Finally, all annotated transcript contigs are examined for further BLAST hits, which may overlap with the initially identified orthologous region. This identifies "misassembled" contigs, which presumably originate from chimeric cDNA as well as neighboring or overlapping genes. The contigs that contain multiple genes are copied to represent each gene separately, which allows independent processing of the genes in subsequent processing steps.

Scaffolding

FRAMA performs an ortholog-based scaffolding of fragmented transcript contigs (Fig. 2). To achieve this, FRAMA uses transcript contigs without an assigned gene symbol, but with BLASTN hits to previously identified orthologous counterparts. These candidate transcript contigs are then aligned to the orthologous counterpart using MAFFT. Next, the minimum number of fragments spanning most of the reference transcript is determined using a greedy algorithm. Finally, the core contig sequence is extended by the series of winning candidates. Any gap between non-overlapping contigs is filled with an N stretch, whose size corresponds to the size of the orthologous transcript region.

Identification of CDS

In order to identify the CDS, each FRAMA transcript is aligned with orthologous CDSs from the reference transcriptome and, optionally, other species as provided by an ortholog table (Fig. 1). Coordinates of each CDS are transferred to the transcript contig and examined for a valid CDS among all reading frames (Fig. 2). In the first course, a candidate reading frame should fit this orthologous window without premature stop codon or, in case of selenoproteins, without non-UGA stop codons. In the presence of multiple valid coding regions, the most complete one in respect to its corresponding ortholog is chosen. If the described approach fails, the CDS prediction (GENSCAN) that is most similar to that of the assigned ortholog is annotated. As a last resort, the longest ORF computed by EMBOSS GETORF is assigned.

Identification of mRNA boundaries

As mentioned above, neighboring or overlapping genes could result in a single long contig and consequently need to be shortened to obtain one transcript contig corresponding to the assigned gene only. Furthermore, Trinity has difficulties determining the precise end of 3' ends, in particular due to the imprecise cellular mechanism of 3' end cleavage, alternative poly(A) sites or possible genomic contamination. Fortunately, mRNA 3' termini share significant sequence conservation between species, e.g., human and mouse [27], and further evidence like poly(A) signal motifs and poly(A)-containing reads are used to infer more precise 3' ends. Specifically, FRAMA scores potential 3' ends according to the occurrence of poly(A) signals. Additionally, informative drops in read coverage as well as reads that contain protruding poly(A) sequence are identified via re-alignment of the RNA-seq data. Finally, a local alignment with 50 bp of the orthologous mRNA terminus is computed with EMBOSS needle.

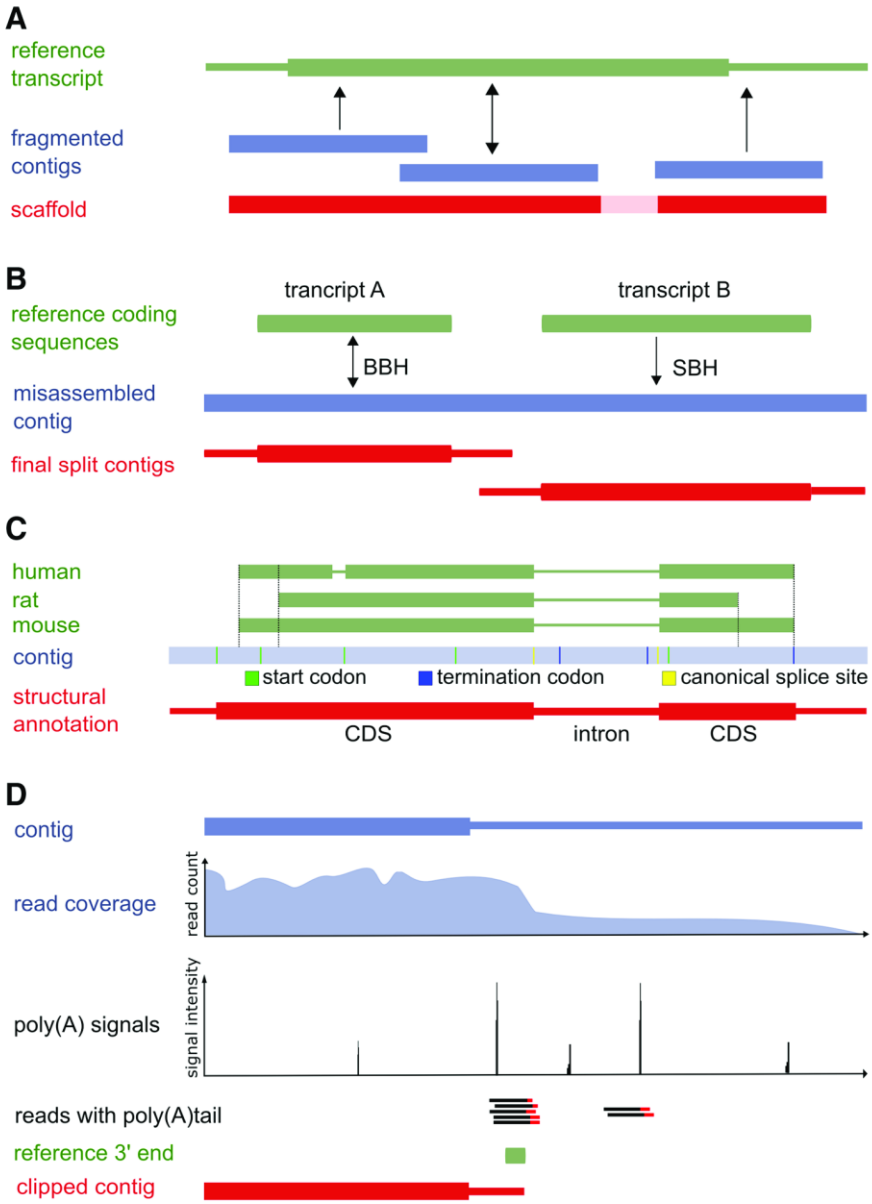


Fig. 2 Schematic illustration of complex processing stages in FRAMA: **a** inference of CDS using orthologous transcripts from related species; **b** ortholog-based detection of fusion contigs; **c** scaffolding; **d** clipping of transcript 3′ termini by the use of weighted scores for indicative features. Horizontal bars indicate contigs and mRNAs, thicker regions indicate CDS. Colors code the origin of sequence data: Trinity contig (*blue*), orthologous transcript (*green*), final FRAMA transcript (*red*)

Each contig position is assigned a weighted score based on all four features using fuzzy logics, and clipping is applied at the most reliable position, using an empirically validated threshold. If GENSCAN predicts a promoter sequence, 5′ ends are clipped as well. In case of extra CDS regions that are predicted by GENSCAN and supported by a BLAST hit, clipping is always applied, either according to the scoring scheme or, if no reliable position was identified, at the center of intercoding regions.

Results

Sequencing

A limited overview of a tissue's mRNA content could be obtained from assembly of 20 million RNA-seq reads preferably 100 nt or longer [28]. For a near-complete picture of a multi-cellular eukaryote, well over 100 million RNA-seq reads and a diversified tissue sampling are desirable, in order to recover tissue-specific genes and genes which are generally low in expression. For an

application of FRAMA, we chose the latter concept and obtained strand-specific Illumina RNA-seq data from ten different tissues of the NMR (Additional file 1: Table S3). After quality filtering and joining of overlapping paired-end reads, the data consisted of 352.5 million single-end fragments with an average length of 194 bp (67.9 Gb in total). For quality control, reads were aligned to the NMR genome sequence, resulting in 90.9–96.2 % mapped reads per sample. Mapping rates above 90 % are comparably high and indicate good base quality of the RNA-seq data and good correspondence between RNA-seq data and the genome sequence [29]. Taking a curated set of NMR transcripts (TCUR), we could further validate that the dUTP protocol for RNA-seq is highly strand-specific. At least 99.85 % of mapped reads had the correct orientation.

Assembly and primary processing

Read data from the ten tissue samples were used as pooled input to Trinity/FRAMA. The use of pooled samples was shown to improve the completeness of transcript contigs in contrast to merging of sample-specific assemblies [18]. The resulting raw assembly comprised 660,649 individual graphs, which, theoretically, reflect the number of assembled gene loci, and 1,042,649 transcript contigs. The length of contigs ranged from 200 bp, the default threshold of Trinity, up to 32,980 bp, with an N50 of 2687 bp (Additional file 1: Table S5).

Trials on meta-assembly indicate that both, CD-HIT-EST and TGICL do minor reductions (8.6 and 11.4 %, respectively) of the transcript contig set while an impact on the final transcript catalog is undetectable. Intending most conservative processing of the NMR data, we chose to continue with the primary Trinity assembly and in order to avoid false assemblies, e.g., collapsing of paralogs or joining of neighboring genes.

One step of sequence post-processing is the clipping of putative sequencing adapters from contig ends, which may show up even if adapter clipping was performed on the input RNA-seq data (0.04 % of contigs). Moreover, FRAMA scans transcript contigs for putative vector contamination, as recommended by the NCBI. As might be expected for the in vitro-cloned RNA-seq libraries, the sequence data is free of cloning vectors. However, NCBI VecScreen indicated 8 strong and 26 moderate vector hits, which we all classified as false positives upon thorough inspection. For example, vector pSOS (acc. no. AF102576.1) contains a fragment of human *SOS1* which produces a strong hit to the *SOS1* transcript of the NMR. Unfortunately, masking of these regions is required for submission to the NCBI Transcript Shotgun Assembly archive.

Assignment of gene symbols

We chose human as the reference organism since the human gene annotation has superior quality and, in terms of sequence similarity, it is closer to the naked mole-rat than mouse, which has a gene annotation of similar quality (Additional file 1: Table S4). Using 34,655 human protein-coding reference transcripts (19,178 genes), FRAMA was able to identify 21,984 NMR counterparts, corresponding to 16,887 genes in total (88.0 % of human genes). The longest NMR transcript contig (32,980 bp) corresponds to the longest human gene, titin.

In general, transcripts that could not be identified in the NMR have much lower expression levels in human tissues, compared to those which could be identified (Additional file 2: Figure S2). For example, reconstructed versus non-reconstructed genes show 1301-fold higher median expression in human liver, and 396-fold higher expression in human kidney (both $p < <0.001$, Mann-Whitney *U* test). On the other hand, some highly expressed genes in human liver lack orthologs in the NMR. However, several of these were identified as primate-specific genes. For example, the top-expressed orphan human genes comprise three metallothionein genes (*MT1F*, *MT1H*, *MT1M*) which are part of the primate-specific expansion of the metallothionein-1 and -2 family [30]; four cytochrome P450 genes (*CYP2C8*, *CYP2C9*, *CYP2C19* and *CYP4F11*) which are primate-specific paralogs at multiple branches of the large family tree [31]; and factors of the major histocompatibility complex, *HLA-B* and *HLA-E*, which underwent fast evolution in primate populations [32].

Scaffolding

Scaffolding was applied to 3684 FRAMA transcripts (3488 genes) and added 3.29 Mb sequence, resulting in a median information increase of 1.27-fold. We manually inspected 31 scaffolded FRAMA transcripts comprising 81 fragments in comparison to a curated set of NMR transcripts (TCUR) and determined errors in 5 scaffold fragments (6.2 %). Further, of all scaffolded FRAMA transcripts we identified only 111 (3.0 %) that show non-overlapping hits to multiple genome contigs in both genome assemblies. These failure rates likely represent the upper bound of errors since some of the non-validated scaffolds may result from fragmented genome data.

Following a series of physical processing steps from the initial Trinity assembly to pre-final transcript sequences, we sought to assess the completeness of the transcript catalog produced by FRAMA. For this we used CEGMA (Additional file 1: Table S6), a tool that identifies 248 eukaryotic core protein-coding genes and diagnoses their completeness. Since 245 genes scored "CDS complete" (98.8 %), the transcript sequence set produced by FRAMA appeared almost complete, within the performance range of other, genome-based transcript catalogs (TGNOMON 247, equivalent to 99.6 %; TKIM 237, 95.6 %; see Methods for definition of reference transcript

sets). Interestingly, the initial Trinity transcriptome assembly contained even slightly less CEGMA genes (243 complete scores) than that of FRAMA, indicating that the final FRAMA output essentially encompasses all relevant genes contained in the initial assembly, and that subsequent processing steps even improved the recovery of the core gene set.

Identification of CDS

The majority of coding regions (13,841 genes; 82.0 %) were assigned with evidence from orthologous sequences. GENSCAN additionally identified CDS of 2765 genes, of which 26.4 % contained introns with canonical splice sites. Taken together, most resulting NMR genes had a full-length ORF including start and stop codon (12,100; 71.1 %; Fig. 3a). This is further supported by 12,583 genes (74.5 %) that had their CDS reconstructed over >90 % of the orthologous length (Fig. 3b). Correctness of the inferred CDS and the assigned gene symbol was validated by BLASTP searches against the human proteome, revealing 96.3 % of transcript contigs that hit proteins with the correct gene symbol, plus 2.9 % that gave hits to the same gene family.

Identification of mRNA boundaries

During gene symbol assignment, FRAMA identified 12 fusion transcript contigs that arose mostly from neighboring genes (Fig. 4). This does not reflect the total number of misassembled transcript contigs, because different misassembled variants have been assigned to different orthologous genes by the BBH/SBH strategy. In total, GENSCAN predicted multiple CDS for 1127 FRAMA NMR transcripts (5.1 %; 1069 genes). This is a higher proportion than seen on human and mouse RefSeq transcripts (3.5 and 2.6 %, respectively), which we consider as the background level of false positive GENSCAN predictions. Consistently, 52.4 % of the NMR transcripts with extra CDS predictions are supported by cross-species BLAST hits (591 transcripts, 516 genes) and thus likely result from correct CDS predictions. The remaining proportion of spurious predictions is comparable to the level in human and mouse transcripts. In total, summing the effect of all clipping procedures, FRAMA removed 5.13 Mb sequence from 5556 transcripts (4774 genes).

Genome-based validation of transcript catalog

A recurring problem in the validation of de novo assemblies is the absence of a reference or gold standard. We chose to compare transcripts computed by FRAMA (TFRAMA) with publicly available NMR transcripts and gene annotations (Additional file 1: Table S7). We considered in-house curated transcripts (TCUR) that were reconstructed using a genome-independent approach as the gold standard in this comparison of NMR sequences. Two previous efforts provided NMR transcript catalogs based on a combination of ab-initio gene prediction, orthologous matching and RNA-seq evidence - one by Kim et al. reported transcript models (TKIM) [24] based on genome assembly hetgla1, and one computed RefSeq transcripts using NCBI's GNOMON pipeline (TGNOMON) based on both available genome assemblies (hetgla1, hetgla2). Further, our validation included transcripts obtained only from ab initio prediction (TGENSCAN).

In transcript-genome alignments 96.8 % of TFRAMA could be aligned (92.7 % of sequence), but only 78.7 % of these transcripts were aligned over their entire length (>99 %). Since a realignment of TGENSCAN to its source genome gives 98.9 % of transcripts matching over their entire length (99.9 % of sequence), the technical error rate appears negligible. Interestingly, TCUR showed non-matching and mismatching regions with a rate depending on the genome sequence, 4.1 % exons on hetgla1, 1.0 % on hetgla2 (Additional file 1: Tables S8 and S9). However, 92.0 % of conflicting regions were validated by one genome version,

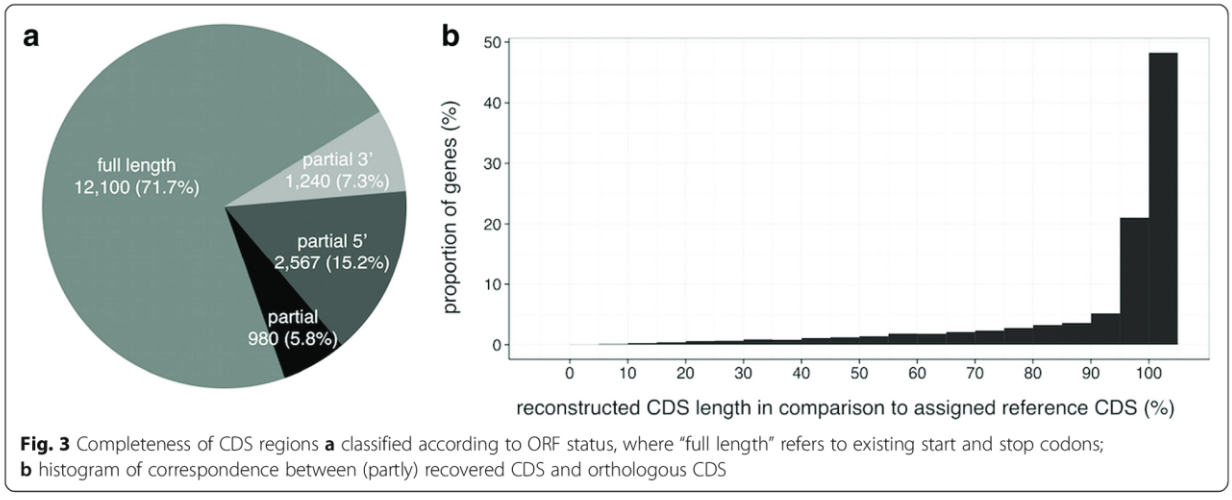


Fig. 3 Completeness of CDS regions **a** classified according to ORF status, where "full length" refers to existing start and stop codons; **b** histogram of correspondence between (partly) recovered CDS and orthologous CDS

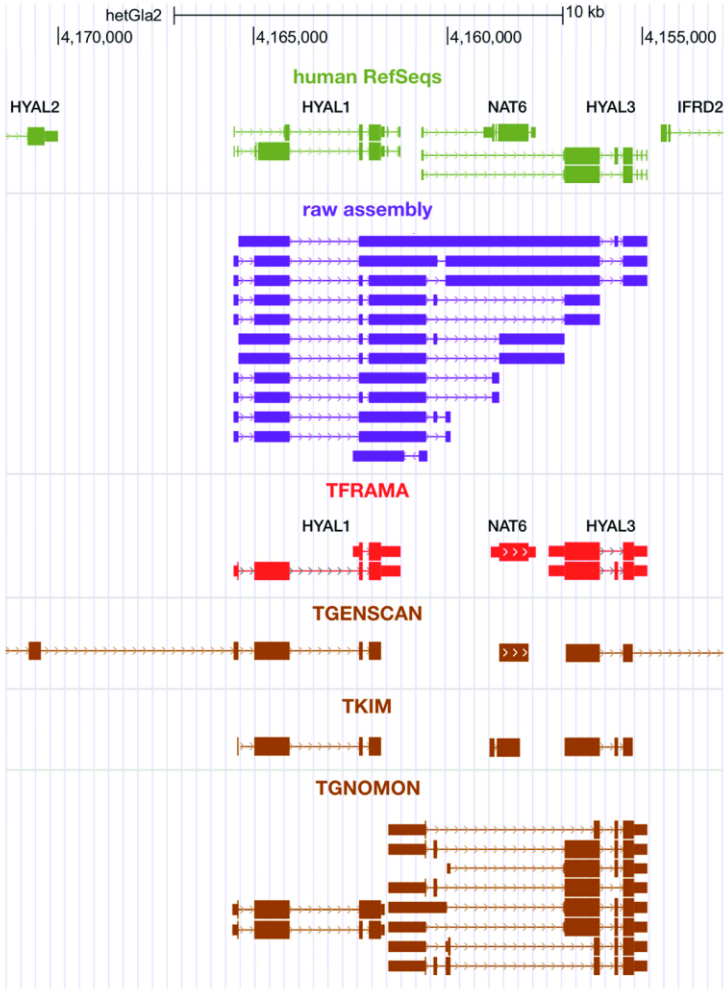


Fig. 4 A genome-based transcript map showing misassembled Trinity contigs (*purple track*) and improvements made by FRAMA's mRNA boundary clipping (*red track*). Human RefSeq counterparts to FRAMA transcripts are shown in green. Trinity provides a plethora of (putative) transcript isoforms (63 contigs) for the *HYAL1-NAT6-HYAL3* locus, many of them being read-through variants that join neighboring genes (informative subset in purple track). Although FRAMA is not able to resolve the shared first exon of the *NAT6-HYAL3* locus correctly, mRNA boundary clipping improved the raw assembly substantially by separating the gene loci. Genome-based methods (brown tracks) struggle in predicting the correct gene loci, too: TKIM shows the best performance, separating each gene locus correctly. GENSCAN correctly separates *HYAL1*, *NAT6* and *HYAL3* loci, but joins neighboring loci (*HYAL1* with *HYAL2* and *HYAL3* with *IFRD2*). GNOMON correctly provides several different *HYAL3* variants, but misses *NAT6* completely. Throughout the figure, thick bars represent coding regions, thin bars untranslated regions and lines introns. Arrows on lines or bars indicate the direction of transcription. Accession numbers of external gene models are listed in Additional file 1: Table S11

which indicates that missing or discontinuous genome sequence are the source of conflicts with TCUR transcript models. We reject the possibility that genetic differences of the underlying NMR material explain the genome-transcriptome differences since well-aligned regions have very high sequence similarity, 99.9 % between TCUR and both genome versions and 99.9 % between TFRAMA and hetgla2. In conclusion, TFRAMA consistently fills missing and weak genome sequence. Effectively, TFRAMA-genome alignments spanned 1695 sequence gaps within scaffolds of hetgla2 and added 408,293 bp novel sequence.

We also validated the consistency of transcript sets, using the RNA-seq data produced in this study, by calculating the proportion of transcript-genome alignments covered by reads (coverage breadth). As expected, the majority of TFRAMA (98.1 %) is completely supported by RNA-seq reads (transcripts with >95 % coverage breadth). In contrast, only 18.7 % of TGENSCAN are completely supported by reads, while 22.4 % are sparsely covered (<5 % coverage breadth). Evidence-based methods show better agreement with our experimental data (TGNOMON 87.6 %, TKIM 71.5 % completely supported).

We compared the transcript-genome alignments of TGNOMON, TKIM, TGENSCAN and TFRAMA with those of our gold standard data set, TCUR (Table 1, Additional file 2: Figure S3). All methods achieved a similar recovery rate of TCUR gene loci (TGNOMON 135, 99.3 %; TKIM 122, 89.7 %; TGENSCAN 133, 97.8 %; TFRAMA 129, 94.9 %). The assigned gene symbols, if present, were consistent with the TCUR annotation (Additional file 1: Table S10).

Next, we investigated the structural agreement between transcripts of the different transcript cataloging methods. Overlapping transcripts from different sources were classified based on the number and type of shared exons (Additional file 2: Figure S4): (i) identical transcripts have all exons exactly corresponding, (ii) matching transcripts share all exons, but not necessarily all exon boundaries, and (iii) others. Application of this classification scheme on TCUR loci showed that the proportion of identical and matching transcript models differed largely between genome-dependent methods (TGNOMON 122 of 135, 90.4 %; TKIM 66 of 122, 54.1 %; TGENSCAN: 19 of 133, 14.3 %). TFRAMA showed results close to TGNOMON (identical/matching 115; 89.1 %) and outperformed TKIM and TGENSCAN. Given that these primary results indicated superior quality of TGNOMON in respect to curated transcripts, we used it as a reference for a second, genome-wide quality assessment. According to this, TFRAMA resembles TGNOMON transcript models by showing the highest number of identical and matching loci (10,590; 73.6 %), in contrast to TKIM (8029; 53.8 %) and TGENSCAN (2628; 16.3 %). More specifically, TFRAMA also shows more transcript models identical to a TGNOMON counterpart (8463; 58.8 %) than TKIM (5382; 36.0 %). Together, this demonstrates a quality ranking of TGNOMON > TFRAMA > TKIM > TGENSCAN.

Performance evaluation

The runtime of FRAMA mainly depends on the number of input reads, the resulting number of assembled transcript contigs and the size of the reference transcriptome. For the complete NMR dataset and 34,655 reference transcripts as input, FRAMA had a total runtime of 338 h on an 8-CPUs Linux workstation (Intel Xeon, 2.83 GHz, Model E5440) and a memory size of 32 GByte. The major computational load was due to de novo assembly and BLAST searches, each taking about 40 % of the total runtime. Using a smaller input subset of 40 million reads, the total run time of FRAMA decreased to 48 h, indicating that the total runtime linearly depends on the volume of the read data.

Discussion

Though whole-genome sequencing and assembly is an essential prerequisite for genome-wide analyses, providing a plethora of information, it is still quite labor-intensive, time-consuming and costly. For example, three groups have independently worked on NMR genome assemblies and associated gene annotations, over the last four years [24, 25, 33]. In contrast, transcriptome sequencing and de novo transcriptome assembly is an affordable approach for first-pass sequence analysis of novel organisms, given automated concepts for extraction of transcripts from RNA-seq data. Towards this goal, we present FRAMA, an mRNA assembly and annotation pipeline for eukaryotes, which is designed to transform a primary transcriptome assembly into a comprehensive, but low-redundant, catalog of reconstructed mRNA sequences.

FRAMA is extensively guided by orthologous transcripts of a reference organism. Orthologs are used (i) for assignment of gene symbols to anonymous transcript contigs, (ii) for identification of representative transcripts from a complicated mixture of mRNA isoforms, and (iii) for

Table 1 Results of structural agreement of overlapping loci in the hetgla2 genome sequence

	Recovered[a]	Identical[b]	Matching[b]	Other[b]
TCUR (loci: 136)				
TFRAMA	129; 94.9 %	100; 77.5 %	15; 11.6 %	14; 10.9 %
TGNOMON	135; 99.3 %	114; 84.4 %	8; 5.9 %	13; 9.6 %
TKIM	122; 89.7 %	50; 41.0 %	16; 13.1 %	56; 45.9 %
TGENSCAN	133; 97.8 %	13; 9.8 %	6; 4.5 %	114; 85.7 %
TGNOMON (loci: 19,746)				
TFRAMA	14,387; 72.9 %	8463; 58.8 %	2127; 14.8 %	3797; 26.4 %
TKIM	14,933; 75.6 %	5382; 36.0 %	2647; 17.7 %	6904; 46.2 %
TGENSCAN	16,082; 81.4 %	1584; 9.8 %	1044; 6.5 %	13,454; 83.7 %

Each orthologous set of transcripts was compared to TCUR and TGNOMON, after filtering of alignments with perfectly aligned CDS (>99 % recovered in genome). CDSs are considered overlapping if they share nucleotides on the same strand. CDS overlap cases were classified to the following categories: identical (identical exons), matching (shared exons), or 'other' (unequal number of exons)

[a]Number of overlapping loci and their proportion of the loci in reference

[b]Number of identical, matching and other transcript models and their proportion of the loci in overlap

refinement of representative transcripts, including scaffolding of fragmented transcript contigs, removal of likely intron contamination, and clipping of weakly supported 3' ends. Given the high relevance of the reference organism, the primary question is what species should be used. Often, there will be a tradeoff between closely related species that have a relatively weak gene annotation on one hand, and more distantly related species with a more comprehensive annotation on the other hand. Applied to the NMR case, the closest-related model organism is the guinea pig (CDS similarity NMR/guinea pig 92.3 %, NMR/human 89.1 %, Additional file 1: Table S4), with an estimated divergence time of 41 Mya [33]. However, the guinea pig genome sequence is rather fragmentary, and the gene annotation is largely confined to the results of Ensembl and NCBI annotation pipelines, which are driven by gene prediction and homology inference. Human, with a divergence time of ca. 88 Mya [34], seems more challenging with regard to sequence similarity searches, but is outstanding in its extensive and experimentally based gene annotation. In fact, human as a homology reference for the NMR gave very satisfying results in this study (88.0 % recovered orthologs), which suggests that even organisms as distant as 100 Mya or more could serve as a reliable basis for ortholog inference. Consistent with this, a methodological survey showed that ortholog inference using a BBH scheme performs well in comparison to other assignment methods, irrespective of species distance [16].

The simplification of gene content via orthologous inference is to some extent artificial, since the ortholog-driven approach fails to identify species-specific paralogs - at best, they are misclassified as orthologs. However, the low-redundant transcript catalog is a comfortable starting point for identification of such species-specific paralogs. It is also clear that a transcript catalog based on RNA-seq will remain incomplete with respect to the total gene content of an organism. Since, even after sampling of multiple tissues and developmental stages, mRNAs with highly specific and restricted expression profiles will not be sufficiently covered. A good example that illustrates both, tissue-specific expression as well as species-specific paralogy, is the family of olfactory receptors (ORs). Humans have 388 functional OR genes, predominantly expressed in sensory neurons of the nasal mucosa, whereas rats have 1259 OR genes. Consistently, the subterranean NMR, which has an outstanding olfactory capacity, show signs of ongoing positive selection and expansion of the OR family, according to targeted genome resequencing [35]. An incompleteness of such tissue-specific transcripts may be acceptable if a limited set of tissues will be analyzed in subsequent studies, and the established gene catalog contains all the genes expressed in those addressed tissues. Furthermore, tissue-specific expression patterns are typically known from related organisms and rarely change during evolution [36]. Thus, even a limited gene catalog from selected tissues can be expected to be conclusive with respect to gene content.

A clear advantage of FRAMA is that it does not require genome data, allowing the study of non-model organisms with yet unknown genome sequence. When we analyzed the FRAMA results for the NMR, we obtained quality measures for the two available genome sequences, which further illustrate the independence of the transcriptome approach. Given a good correspondence on the sequence level (99.9 %), the NMR transcriptome provided exon sequences that filled genomic gap regions estimated to make up 1.0 % of the latest available genome sequence [24]. In addition, reconstructed mRNAs spanned 1695 gaps within genomic scaffolds, thereby driving genome assembly towards higher contiguity. Together, curated as well as FRAMA transcripts provided independent support for improvements made in NMR genome assemblies through the past years [24].

Modern genome annotation strategies incorporate RNA-seq data as experimental evidence for genes. As it had to be expected, FRAMA based on RNA-seq alone does not outperform qualified genome-based annotation strategies, like NCBI's GNOMON pipeline, that use multiple sources of gene support in addition to transcriptome sequencing [11]. On the other hand, the FRAMA transcript catalog outperformed the ab initio gene prediction using GENSCAN and the annotation of the first NMR genome. Moreover, the FRAMA transcript catalog was close to the result of GNOMON with respect to structurally identical or matching transcript models (Table 1, Additional file 2: Figure S4). The latter can be considered as the currently best NMR genome annotation and is also well supported by an independent set of scientist-curated NMR transcripts (Table 1, dataset TCUR). Striking heterogeneities were found between different genome-based annotations, especially if one assumes that the same experimental evidence of RNA-seq data was used. The compared methods have similar sensitivity in recovery of gene loci, measured on the TCUR dataset, but the results differ largely on the gene structure level. However, such heterogeneities are in agreement with a recent benchmark study on genome-based RNA-seq transcript reconstruction [37].

Conclusions

FRAMA realizes the *de novo* construction of a low-redundant transcript catalog for eukaryotes, including the extension and refinement of transcripts. Thereby, it delivers a compilation of transcripts which we regard suitable for comprehensive downstream analyses performed by biologists without bioinformatics expert support.

Methods

For a full list of external software including versions and references see Additional file 1: Table S1.

Tissue sampling

Samples from cerebellum, pituitary, thyroid, adrenal gland, kidney, skin, liver and ovary were collected from one female naked mole-rat from a previously established colony, kept at the Leibniz Institute for Zoo and Wildlife Research (IZW, Berlin) [38]. Hypothalamus and testis samples were obtained from a male animal of the same colony. Animal housing and tissue sampling was compliant with national and state legislation (breeding allowance #ZH 156; ethics approval G 0221/12 "Exploring long health span", Landesamt für Gesundheit und Soziales, Berlin).

RNA-seq

Prior to RNA isolation, tissue was disrupted in the homogenization buffer of the RNA extraction protocol using a Tissue Lyser instrument (Qiagen). RNA was isolated using the RNeasy Mini kit (Qiagen), performing specialized protocols for brain and muscle tissues as recommended by the manufacturer. The RNA was treated with DNase I on the affinity column before elution. Strand specific RNA-seq libraries, including poly-A(+) mRNA selection and RNA fragmentation, were prepared using the TruSeq Stranded RNA LT Kit (Illumina) according to the supplier's instructions, with 2 μg total RNA as input. The resulting libraries had insert sizes of ca. 100–400 bp as indicated by DNA 7500 Chips run on an Agilent Bioanalyzer 2100 instrument (Agilent). All ten libraries were combined into a single pool. Sequencing of 200-nt paired-end reads was performed using an Illumina HiSeq 2500 apparatus in Rapid mode with TruSeq Rapid SBS chemistry on two lanes (Illumina). Read data for each library were extracted in FastQ format using the CASAVA software v1.8.4 (Illumina) using default settings.

Read preprocessing

Quality of RNA-seq reads was inspected using FastQC. Raw data was screened for potential cross-contamination with foreign species, including human, pig, mouse and guinea pig. Overlapping paired-end reads were joined into single longer reads (93.8 %), and adapter sequences of these and remaining reads were clipped using SeqPrep (parameters: –A < adaptrev > –B < adaptfwd>). Non-overlapping reads were quality-trimmed at the 3' end using sickle (parameters: –x -q 23 -l 35), and reads shorter than 35 bp were discarded. Reverse-complemented antisense reads and sense reads were pooled with joined long reads to generate a set of stranded single reads (simply "reads" in the following).

Reference sequence sets

Human transcripts, used as the reference for transcriptome reconstruction, were part of the human genome annotation release 105 obtained from the National Center for Biotechnology Information (NCBI). Selection for known protein-coding Reference Sequences (RefSeqs; NM-style accessions) resulted in 34,655 transcripts. Public human RNA-seq data (Illumina Body Map 2.0, Illumina Corp., unpublished) were used to assess mRNA expression. Mouse protein-coding RefSeqs were part of the mouse genome annotation release 104 obtained from NCBI (77,610 transcripts). NMR genome assemblies were previously reported by Kim et al. [24] (Bioproject: PRJNA68323; hetgla1) and Keane et al. [25] (Bioproject: PRJNA72441; hetgla2). The most recent hetgla2 genome sequence was used as the reference unless stated otherwise. Four sets of NMR transcripts from different sources were used for comparison: 76,826 Reference Sequence mRNAs modeled by NCBI's eukaryotic genome annotation pipeline, GNOMON (NCBI Heterocephalus glaber Annotation Release 100; abbreviated as TGNOMON); 21,771 CDSs published by Kim et al. [24] (Bioproject: PRJNA68323; abbreviated as TKIM); 55,730 GENSCAN predictions obtained from UCSC (abbreviated as TGENSCAN); and 142 curated mRNA sequences obtained from GenBank (Additional file 1: Table S2; abbreviated as TCUR).

Read alignment

Spliced alignment of the RNA-seq reads against the genome sequence was performed with STAR allowing 2 % mismatches within the aligned region and a maximum of 5 multiple hits per read (parameters: –outSAMstrandField intronMotif –outFilterMultimapNmax 5 –outFilterMismatchNoverLmax 0.02). RNA-seq read counts per gene were obtained via mapping with BOWTIE; per gene, the longest transcript was used as mapping template, and unique hits for each read were required. A comparison of human samples, based on expression values scaled to fragments per kb transcript per million fragments (FPKM) [39], was done using the Mann–Whitney *U*-test (two-sided), and p-values were obtained via a Monte Carlo-based approximation implemented in the R package COIN.

Multiple sequence alignment

For orthologous assignment of CDS we created a resource of multi-species mRNA alignments. Starting with the reference mRNAs of human, dog, mouse, and rat (NCBI RefSeq, release 61), orthologous clusters were identified using the HomoloGene database (release 67) [40]. Multiple protein sequence alignments for each cluster were computed using CLUSTALW (parameter: gapext = –2). For each human isoform, a sub-alignment

was extracted from the orthologous cluster, such that the one most similar isoform from each of the other species was contained.

Analysis of transcript-to-genome alignments

Quality of transcript sequence sets was assessed from transcript-to-genome alignments. The following approach was applied to all transcript sets to ensure equal conditions. Transcript sequences were mapped with BLAT (parameter: –extendThroughN) and filtered for one global best hit using the BLAT utility pslCDnaFilter (parameters: –globalNearBest = 0.0 -minAlnSize = 100 -minId = 0.9). Spliced alignment was determined with SPLIGN (parameters: –gap_extension_score –520 -type est -direction sense -min_exon_idty 0.85 -min_compartment_idty 0.6) within the best BLAT hit region including 1 kb up- and downstream. Poorly aligned regions were determined with an in-house implemented hidden Markov model, which identifies regions of significantly high mismatch density due to lack of appropriately aligning genome regions.

An all-against-all comparison between gene annotations was used to determine shared genes and transcripts. Briefly, within a gene annotation, genes are defined either by single-transcript loci or by multiple transcripts overlapping on the same strand. One-to-one relationships between transcripts from different annotations were calculated with EVALUATOR.pl, which utilizes a stable marriage algorithm to pair transcripts for each gene locus. The number of overlapping, missing or wrong exons was determined with in-house software. The structural agreement was investigated for the CDS of transcripts with perfectly aligned CDS (>99 % aligned).

Data access

RNA-seq data and assembled transcripts with full-length CDS were deposited at NCBI databases (linked to Bioproject PRJNA283581). FRAMA is available for download at https://github.com/gengit/FRAMA.

Availability and requirements

Project name: FRAMA (from RNA-seq to annotated mRNA assembly)

Project home page: https://github.com/gengit/FRAMA

Operating System: UNIX/Linux

Programming language: Perl, R

Other requirements: Additional file 1: Table S1 and https://github.com/gengit/FRAMA.

License: FLI-Licence

Availability of supporting data

Additional file 1: Supplementary Tables.
Additional file 2: Supplementary Figures.

Additional files

Additional file 1: Table S1. List of external software. **Table S2**: NMR transcript data set TCUR, and orthologous transcripts from human, mouse and guinea pig. Multi-species mRNA alignments were constructed independently from those described in the main text, using the sequence database entries as listed. **Table S3**: Naked mole-rat samples for strand-specific RNA-seq, and produced RNA-seq data. **Table S4**: Pairwise transcript sequence identities between NMR and related mammals. The analysis is based on 142 multiple sequence alignments of the CDSs of NMR, guinea pig, human and mouse (as listed in Additional file 1: Table S2). Identity values were computed based on gap-masked alignments. **Table S5**: Statistics of the transcriptome data produced by Trinity (column "transcript assembly") and subsequently processed using FRAMA (column "transcript catalog"). **Table S6**: CEGMA results on transcriptome datasets. As defined by CEGMA, 'complete proteins' are recovered with >70 % in comparison to CEGMA's core proteins. 'Partial proteins' additionally include proteins, which exceed a certain alignment score threshold. CEGMAs software components were used as suggested: geneid (v1.4), genewise (wise2.2.3-rc7), hmmer (HMMER 3.0), NCBI BLAST+ (2.2.25). **Table S7**: Source of transcript sequence sets and underlying input data. **Table S8**: Transcript-genome alignment statistics of curated dataset (TCUR) in hetgla1. The alignments comprise 1473 well-aligned blocks and 81 unaligned or mismatching blocks. Transcripts show 99.9 % average identity within well-aligned blocks. **Table S9**: Transcript-genome alignment of curated dataset (TCUR) in hetgla2. The alignments comprise 1525 well-aligned blocks and 16 unaligned or mismatching blocks. Transcripts show 99.9 % average identity within well-aligned blocks. **Table S10**: Correspondence of gene symbols between transcript sets. The evaluation considered gene loci overlapping in the hetgla2 genome sequence, where all transcript-genome alignments of a gene were considered to define the gene locus. Only genes with ascertained function (non-LOC gene symbol) were compared. **Table S11**: Accession numbers of sequences that are shown in the genome-based transcript map (hetgla2, scaffold JH602043; Fig. 4). Accession numbers for each sequence are listed in the same order as shown in Fig. 4 (from top to bottom). (XLSX 77 kb)

Additional file 2: Figure S1. Multiple sequence alignments of CALM1, CALM2 and CALM3 in human and NMR. **(A)** protein coding sequence **(B)** protein sequence. All protein coding sequences encode for the same protein sequence. The nucleotide identity between human and NMR orthologs is higher (97 % CALM1, 98 % CALM2, 95 % CALM3) than the intra-species paralog identity (e.g., human CALM1/CALM2 highest identity with 85 %). **Figure S2**. Recovery of transcripts is predicted by the expression level in the reference organism - **(A)** human liver, **(B)** human kidney. Public human Illumina RNA-seq data were obtained from the Short Read Archive at the EBI (accessions ERR030895 and ERR030893, respectively). Box plots show the human expression levels in log-scale FPKM; zero FPKM values were initially transformed to 0.80 times the lowest finite value. Human genes are displayed in three groups: all genes ("all"), genes recovered as orthologous NMR transcripts ("recovered"), and genes missing in the NMR transcript catalog ("missing"). Boxes enclose the data ranges of the central two-third quantiles, and central bars indicate the data medians. Note that the group-wise medians are significantly influenced by the fraction of zero-expression genes; these are 12 % in the liver-recovered group, 56 % in the liver-missing group, 7 % in the kidney-recovered group, and 49 % in the kidney-missing group. **Figure S3**. Results of structural agreement between transcript sets. The evaluation considered gene loci overlapping in the hetgla2 genome. Each transcript set was compared to TCUR. **Figure S4**. Classification of exons into four categories (exact, overlapping, missing and wrong) based on the reference transcript model. Exact exons share the same boundaries. Overlapping exons share base pairs, but not necessarily any boundary. Exons only present in the predicted transcript model are classified as wrong. Exons only present in the reference transcript model are classified as missing. (DOCX 841 kb)

Abbreviations

BBH: best bidirectional blast hit; CDS: protein-coding sequence; MSA: multiple sequence alignment; NMR: naked mole-rat; RNA-seq: second-generation sequencing of RNA; SBH: single best blast hit; UTR: untranslated regions.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
MP, TBH and KS conceived the project. MM, SH, TBH, MB, MG and KS performed the tissue sampling and sequencing experiments. MB, AS, NJ, MP and KS designed and implemented the software. MB, AS, MP and KS performed the validation analysis and discussed the results. MB, MP and KS wrote the paper, and all authors read and approved the final manuscript.

Acknowledgments
We thank Ivonne Görlich for excellent technical assistance and Andreas Petzold for helpful discussions. This work was funded by the Bundesministerium für Bildung und Forschung (BMBF) and the Leibniz-Gesellschaft through the Senatsausschusswettbewerb (SAW), grant SAW-2012-FLI-2, as well as the Deutsche Forschungsgemeinschaft (DFG), grant PL 173/8-1.

Author details
[1]Leibniz Institute on Ageing - Fritz Lipmann Institute, Beutenbergstr. 11, 07745 Jena, Germany. [2]Leibniz Institute for Zoo and Wildlife Research, Alfred-Kowalke-Straße 17, 10315 Berlin, Germany.

Received: 1 August 2015 Accepted: 22 December 2015
Published online: 14 January 2016

References

1. Martin KJ, Pardee AB. Identifying expressed genes. Proc Natl Acad Sci U S A. 2000;97:3789–91.
2. Hillier LD, Lennon G, Becker M, Bonaldo MF, Chiapelli B, Chissoe S, et al. Generation and analysis of 280,000 human expressed sequence tags. Genome Res. 1996;6:807–28.
3. Bonaldo MF, Lennon G, Soares MB. Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res. 1996;6:791–806.
4. McCombie WR, Adams MD, Kelley JM, FitzGerald MG, Utterback TR, Khan M, et al. Caenorhabditis elegans expressed sequence tags identify gene families and potential disease gene homologues. Nat Genet. 1992;1:124–31.
5. Duncan RP, Husnik F, Van Leuven JT, Gilbert DG, Dávalos LM, McCutcheon JP, et al. Dynamic recruitment of amino acid transporters to the insect/symbiont interface. Mol Ecol. 2014;23:1608–23.
6. Agarwal P, Parida SK, Mahto A, Das S, Mathew IE, Malik N, et al. Expanding frontiers in plant transcriptomics in aid of functional genomics and molecular breeding. Biotechnol J. 2014;9:1480–92.
7. Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, Irie R, et al. Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet. 2004;36:40–5.
8. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 2012;13:329–42.
9. Stanke M, Tzvetkova A, Morgenstern B. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. Genome Biol. 2006;7(1):S11.
10. Korf I, Flicek P, Duan D, Brent MR. Integrating genomic homology into gene structure prediction. Bioinformatics. 2001;17 Suppl 1:S140–8.
11. Gnomon–NCBI eukaryotic gene prediction tool [http://www.ncbi.nlm.nih.gov/projects/genome/guide/gnomon.shtml]. Accessed 2014-10-27.
12. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011;12:671–82.
13. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.
14. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28:1086–92.
15. Gonzàlez-Porta M, Frankish A, Rung J, Harrow J, Brazma A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol. 2013;14:R70.
16. Altenhoff AM, Dessimoz C. Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol. 2009;5, e1000262.
17. Schliesky S, Gowik U, Weber APM, Bräutigam A. RNA-Seq Assembly – Are We There Yet? Front Plant Sci. 2012;3:220.
18. Zhao Q-Y, Wang Y, Kong Y-M, Luo D, Li X, Hao P. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics. 2011;12(14):S2.
19. Duan J, Xia C, Zhao G, Jia J, Kong X. Optimizing de novo common wheat transcriptome assembly using short-read RNA-Seq data. BMC Genomics. 2012;13:392.
20. Krasileva KV, Buffalo V, Bailey P, Pearce S, Ayling S, Tabbita F, et al. Separating homeologs by phasing in the tetraploid wheat transcriptome. Genome Biol. 2013;14:R66.
21. Surget-Groba Y, Montoya-Burgos JI. Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Res. 2010;20:1432–40.
22. Buffenstein R. Negligible senescence in the longest living rodent, the naked mole-rat: Insights from a successfully aging species. J Comp Physiol B Biochem Syst Environ Physiol. 2008;178:439–45.
23. Austad SN. Comparative biology of aging. J Gerontol Ser A Biol Sci Med Sci. 2009;64:199–201.
24. Kim EB, Fang X, Fushan AA, Huang Z, Lobanov AV, Han L, et al. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. Nature. 2011;479:223–7.
25. Keane M, Craig T, Alfoldi J, Berlin AM, Johnson J, Seluanov A, et al. The Naked Mole Rat Genome Resource: facilitating analyses of cancer and longevity-related adaptations. Bioinformatics. 2014;30:3558–60.
26. VecScreen [http://www.ncbi.nlm.nih.gov/tools/vecscreen/about/]. Accessed 2015-07-13.
27. Morgan M, Iaconcig A, Muro AF. Identification of 3′ gene ends using transcriptional and genomic conservation across vertebrates. BMC Genomics. 2012;13:708.
28. Francis WR, Christianson LM, Kiko R, Powers ML, Shaner NC, Haddock DSH. A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. BMC Genomics. 2013;14:167.
29. Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Rätsch G, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods. 2013;10:1185–91.
30. Moleirinho A, Carneiro J, Matthiesen R, Silva RM, Amorim A, Azevedo L. Gains, losses and changes of function after gene duplication: Study of the metallothionein family. PLoS One. 2011;6, e18487.
31. Uno Y, Iwasaki K, Yamazaki H, Nelson DR. Macaque cytochromes P450: nomenclature, transcript, gene, genomic structure, and function. Drug Metab Rev. 2011;43:346–61.
32. Parham P, Abi-Rached L, Matevosyan L, Moesta AK, Norman PJ, Older Aguilar AM, et al. Primate-specific regulation of natural killer cells. J Med Primatol. 2010;39:194–212.
33. Fang X, Seim I, Huang Z, Gerashchenko MV, Xiong Z, Turanov AA, et al. Adaptations to a subterranean environment and longevity revealed by the analysis of mole rat genomes. Cell Rep. 2014;8:1354–64.
34. Adkins RM, Walton AH, Honeycutt RL. Higher-level systematics of rodents and divergence time estimates based on two congruent nuclear genes. Mol Phylogenet Evol. 2003;26:409–20.
35. Stathopoulos S, Bishop JM, O'Ryan C. Genetic signatures for enhanced olfaction in the African mole-rats. PLoS One. 2014;9, e93336.
36. Liao BY, Zhang J. Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. Mol Biol Evol. 2006;23:1119–28.
37. Steijger T, Abril JF, Engström PG, Kokocinski F. The RGASP Consortium, Hubbard TJ, et al.: Assessment of transcript reconstruction methods for RNA-seq. Nat Methods. 2013;10:1177–84.
38. Roellig K, Drews B, Goeritz F, Hildebrandt TB. The long gestation of the small naked mole-rat (Heterocephalus glaber Rüppell, 1842) studied with ultrasound biomicroscopy and 3D-ultrasonography. PLoS One. 2011;6, e17744.
39. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5:621–8.
40. Acland A, Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2014;42:D7–17.

3.2 Manuscript 2 (M2)

Draft document.

Manuscript Title

Naked mole-rat transcriptome signatures of socially-suppressed sexual maturation and links of reproduction to ageing

Authors

Martin Bens 1*, Karol Szafranski 1*, Susanne Holtze 2, Arne Sahm 1, Marco Groth 1, Hans A. Kestler 1,3, Thomas B. Hildebrandt 2**, Matthias Platzer 1**

Affiliations:

1 Leibniz Institute on Aging - Fritz Lipmann Institute, Germany

2 Leibniz Institute for Zoo and Wildlife Research, Germany

3 Institute of Medical Systems Biology, Ulm University, Ulm, Germany

* authors contributed equally to this work

** shared senior authors

Corresponding author: Martin Bens, Martin.Bens@leibniz-fli.de

Email addresses: Martin Bens, Martin.Bens@leibniz-fli.de; Karol Szafranski, Karol.Szafranski@leibniz-fli.de; Susanne Holtze, holtze@izw-berlin.de; Arne Sahm, Arne.Sahm@leibniz-fli.de; Marco Groth, Marco.Groth@leibniz-fli.de; Thomas B. Hildebrandt, hildebrandt@izw-berlin.de; Matthias Platzer, Matthias.Platzer@leibniz-fli.de

Keywords

reproduction, sexual maturation, ageing, RNA-seq, naked-mole rat, eusociality

Abbreviations

GO	Gene Ontology
NMR	naked mole-rat
GP	guinea pig
DEG	differentially expressed gene
FET	Fisher's exact test
FDR	false discovery rate
NMR-BvN	comparison of NMR breeders vs. non-breaders (both sexes)
NMR-N-FvM	comparison of NMR Non-breeder Females vs. Males
NMR-B-FvM	comparison of NMR Breeder Females vs. Males
NMR-M-BvN	comparison of NMR Male Breeders vs. Non-breeders
NMR-F-BvN	comparison of NMR Female Breeders vs. Non-breeders
GP-N-FvM	comparison of GP Non-breeder Females vs. Males
GP-B-FvM	comparison of GP Breeder Females vs. Males
GP-M-BvN	comparison of GP Male Breeders vs. Non-breeders
GP-F-BvN	comparison of GP Female Breeders vs. Non-breeders

Abstract

Naked mole-rats (NMRs) are eusocially organized in colonies with an extreme reproductive skew towards one pair of breeding animals. Although breeders carry the additional metabolic load of reproduction, laboratory animals remain fertile and healthy throughout their extremely long lifespan of >30 years. Here, we present a comparative transcriptome analysis of breeders versus non-breeders of the eusocial, long-lived NMR versus the polygynous and shorter-lived guinea pig (GP). It gives insights into interspecies differences in sexual maturation and a naturally evolved case of positive correlation between reproduction and longevity. We found low levels of transcriptional differentiation between sexes in adult NMR non-breeders, providing molecular evidence that sexual maturation in NMRs is socially suppressed. After transition into breeders, both NMR sexes show pronounced feedback signalling via gonadal steroids and expression changes similar to socially-regulated reproductive phenotypes in fish. Remarkably, genes which are higher or lower expressed in NMR compared to GP are also preferentially up- or downregulated in NMR breeders, suggesting a gradual expression strategy related to fitness maintenance in reproductive NMRs. Moreover, status-related expression differences show significant enrichment for ageing-associated genes only in NMRs, indicating differences in the genetic impact on lifespan between species. This is further supported by opposing expression changes in status-related genes between species, such as fibroblast growth factor receptor 2. In addition, tissue-specific changes of mitochondrial activity in skin of male NMR breeder indicate delayed organ ageing.

Introduction

The naked mole-rat (NMR, *Heterocephalus glaber*) has become increasingly popular as an animal model in a variety of research fields due to its unique biology. This includes an exceptionally long lifespan and resistance to cancer (Kim et al. 2011; Tian et al. 2013). According to The Animal Ageing and Longevity Database (AnAge, genomics.senescence.info/species) the maximum recorded lifespan is 31 years, i.e. 368% of the prediction based on body mass. NMRs stay fertile throughout their long and healthy life, i.e. show an extraordinary long life- and healthspan (Buffenstein and Jarvis 2002). This lifelong fertility becomes even more astonishing, considering the extreme reproductive skew in NMR colonies. Like eusocial insects, NMRs are socially organized in colonies consisting of a pair of reproducing animals (breeders, queen and pasha(s)) and up to 300 subordinates (non-breeders, female and male workers) (Jarvis 1981). However, although workers are in principle capable of reproduction (Faulkes et al. 1990a, 1991), sexual maturation is suppressed through the behavior of the dominating queen (Smith et al. 1997). Non-breeding animals of both sexes are the backbone of the social organisation of the colony and take care of foraging, brood care, colony defence and digging (Burda et al. 2000).

Naturally, new NMR colonies originate from fissioning of existing colonies or formation by dispersers. Dispersers are animals that leave their natal colony and migrate into other colonies or form new colonies (Braude 2000; O'Riain et al. 1996). When under laboratory conditions non-breeders are removed from the colony and paired with the opposite sex, they have the capability to ascend into breeders. This process is accompanied with physiological and behavioural changes, and results in the formation of a new colony (Jarvis 1981; Faulkes et al. 1990a). Remarkably, despite the queen's enormous metabolic load of producing a large litter every three months and being exclusively in charge of lactation (Orr et al. 2016), preliminary data from the wild NMR colonies indicate that breeders live longer than their non-breeding counterparts (Hochberg et al. 2016). Under laboratory conditions it seems there is no evidence of a significant difference in the life expectancy between breeders and workers (Buffenstein 2005). In closely related eusocial Fukomys mole-rats the breeding animals show extended longevity in captivity (Dammann and Burda 2006; Dammann et al. 2011). This contrasts the disposable soma theory of ageing, which hypothesizes that energy is either allocated to body maintenance or reproduction (Kirkwood 1977).

In NMR, the reproductive suppression in female non-breeders is mediated through inhibition of gonadotropin-releasing hormone (GnRH) secretion from the hypothalamus (Faulkes et al. 1990a). This in turn leads to an inhibition of follicle stimulating hormone (FSH) and luteinizing hormone (LH) released by the pituitary gland and causes a block of ovulation. Also for male non-breeders, reproductive suppression is caused by inhibition of GnRH secretion, which results in lower levels of urinary testosterone and plasma LH (Faulkes et al. 1991). The impact, however, is less profound compared to females as spermatogenesis is attenuated, but not entirely suppressed (van der Horst et al. 2011). Nevertheless, weight of testis and number of active spermatozoa is higher in breeders (Faulkes et al. 1994, 1991). The role of GnRH in mediating environmental cues to allow or block reproduction is well described in a variety of species (Abbott et al. 1988; White et al. 2002).

The NMR can be regarded as a neotenic species and the prolonged retention of juvenile features has been linked to its longevity (Skulachev et al. 2017). In comparison to mice, e.g. postnatal NMR brain maturation occurs at slower rate (Orr et al. 2016) and puberty is delayed. Female and male NMRs may reach sexual maturity at 7.5 to 12 months of age (Sherman et al. 1991b). In the colony, however, the queen suppresses sexual maturation in both non-breeding males and females by aggressive social behaviour (Smith et al. 1997) and can delay – independently of neoteny - the puberty of female workers for life (Dengler-Crish and Catania 2007). Thus, sexual dimorphism is almost absent among non-breeding NMRs. Both sexes show almost no difference in morphology, including body mass, body size and even external genitalia, as well as no behavioural differences, in the sense that non-breeders participate and behave equally in all colony labours (Sherman et al. 1991a). Nevertheless, these features are correlated with colony rank. Most profound differences can be observed comparing NMR queens vs. non-breeders, reflected in morphological differences, such as an elongated spine and higher body mass of queens, as well as behavioural differences, such as increased aggressiveness, copulation and genital nuzzling (Sherman et al. 1991a).

In this work, we characterize the molecular signature of reproductive status (breeder vs. non-breeder) in tissue samples of ten organs or their substructures (hereinafter called for simplicity "tissues") from both sexes using RNA-seq. We contrast the NMR results with the transcript profiles of corresponding samples of guinea pig (GP, *Cavia porcellus*), a closely related, polygynous, not long-lived rodent species (AnAge: 12 years maximal longevity, 89% of the prediction based on body mass). We specifically focused our analyses on transcriptome signatures of the socially-suppressed sexual maturation in NMR as well as on differentially expressed genes (DEGs) that may contribute to the exceptional long and healthy lifespan of NMR breeders.

Results

To gain molecular insights into the fascinating combination of NMR phenotypes, in particular their eusocial reproduction, lifelong fertility, extraordinary healthspan and longevity, we aimed to collect a comprehensive set of tissue samples for male and female breeders and non-breeders of NMR and GP - six biological replicates each. Towards this, NMR non-breeders were removed from their natal colony, paired with an unrelated animal of the opposite sex from a second colony and thereby turned into breeders. Respective male and female litter siblings remained in the two colonies as non-breeder controls. Time to first litters averaged in 6.5±4.9 (s.d.) months and duration of pregnancies was approximately 70 days. GP breeders and non-breeders were housed as pairs of opposite or same sex, respectively. For this species, the time to first litters was 4.1±0.8 month and the pregnancies lasted about 68 days.

Female breeders gave birth to two litters each, with two exceptions. One NMR female was pregnant at least twice (ultrasonographically verified), but never gave birth to live offspring, and another gave birth to three litters, due to a pregnancy fathered by one of her sons. At time of sampling, NMRs and GPs reached an age of 3.4±0.5 and 0.9±0.1 years, respectively (for more details see Supplementary Table S1).

Tissue and species are the major determinants of transcriptomes

To compare gene expression between reproductive statuses (breeder vs. non-breeder) in NMR and GP, we performed RNA-seq of ten different tissue samples (heart - Hrt, skin - Skn, liver - Lvr, kidney - Kid, cerebellum - Cer, hypothalamus - Hyp, pituitary - Pit, thyroid - Thy, adrenal - Adr, and gonads – Gon, represented by either ovary – Ova or testis – Tes) from 24 animals for each species (six males, six females per status; Supplement Figure S1). Seven of the 480 samples (1.5%) had to be excluded for different reasons (Supplement Table S2, S3). On average (±s.d.), we obtained per sample 27.6±3.6 million high-quality reads with 84.1±16.1% unique mapping rate (Supplement Tables S4). The grand mean of pairwise Pearson correlation within the 40 replicate groups (2 statuses x 2 sexes x 10 tissues per sex) was 0.981±0.013 and 0.984±0.01 for NMR and GP, respectively, indicating high consistency between replicate samples (Supplement Table S5).

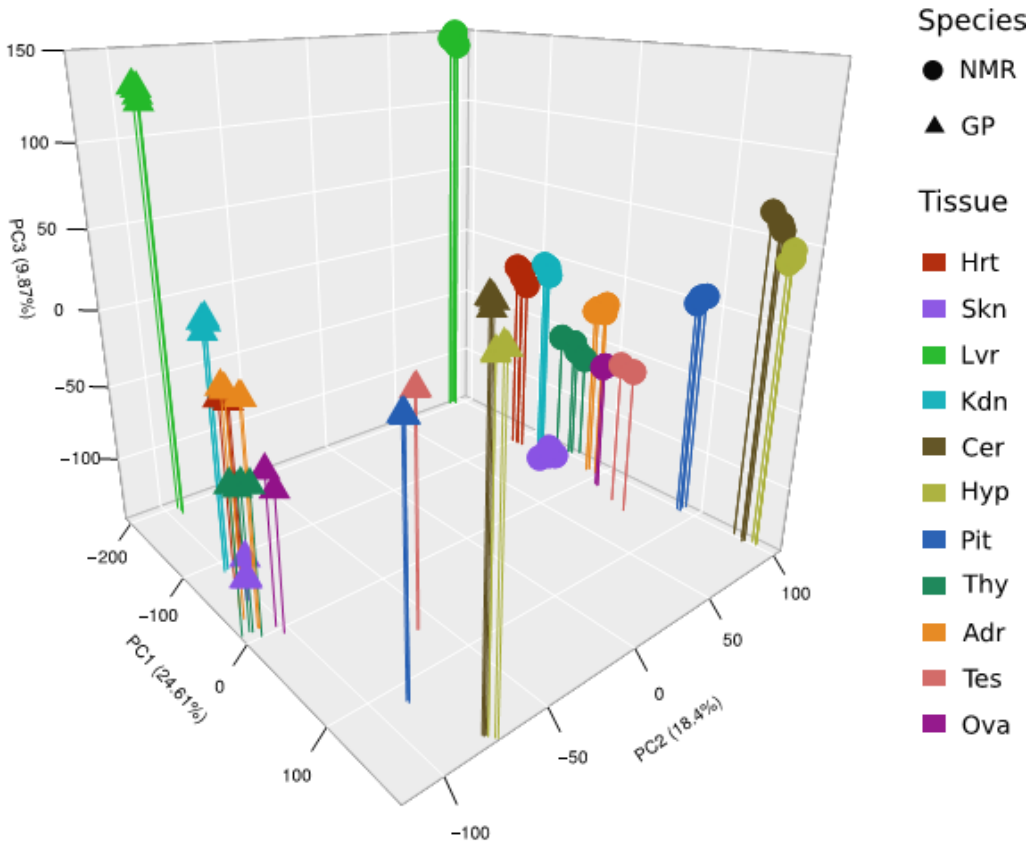


Figure 1: Principle component (PC) analysis based on logarithm of mean expression level between replicates (four datasets per tissue and species: 2 sexes x 2 statuses, except gonads). Tissues are separated by PC1 and PC3, species by PC2.

Based on these data, unsupervised hierarchical clustering gave a similar cluster hierarchy of tissues for both species (Supplement Figure 2). Brain tissues are grouped (Pit as a sister group to Cer and Hyp); Kid and Thy are sister groups to the cluster of Adr and Ova. The results are confirmed by principle component analysis, separating tissues by the first and species by the second component (Figure 1). At this level of

analysis, ovary was the only tissue, which showed a separation of samples with respect to breeding status. Together, this indicates that: (i) tissue source is dominant over other biological variables such as species, sex and status and (ii) the impact of sex and status on transcriptome profiles is subtle.

Cross-species DEGs are enriched in ageing-related genes

To further characterize the species differences between the long-lived NMR and the shorter-lived GP, we next determined gene expression differences based on orthologous transcribed regions that show high sequence similarity. This filtering method avoided potentially misleading signals that may arise from assembly artefacts or the comparison of different transcript isoforms and identified 9,593 genes suitable for further analyses.

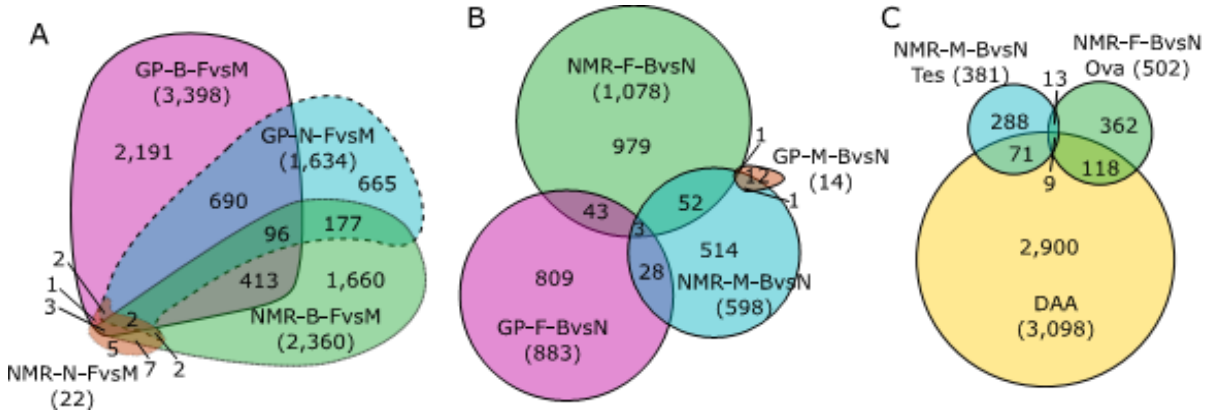


Figure 2: Euler diagrams of analyzed gene sets. (A) DEGs of females vs. males within non-breeders and breeders for each species. (B) DEGs of breeders vs. non-breeders within the same sex for each species. (C) NMR gonad DEGs of breeders vs.

Across all tissues, we identified 16,692 significant expression differences (EDs; 8,920/7,772 higher/lower expressed in NMR) in 5,601 genes (FDR<0.01, $|\log_2 FC|>2$; Supplement Table S6, Supplement Data 1). To assess the association of cross-species DEGs with ageing, we examined their overlap with ageing-related genes of human and mouse obtained from the Digital Ageing Atlas (DAA) (Craig et al. 2014). This test revealed a significant overlap with DAA containing 999 genes (17.8% of DEGs; p=0.008, Fisher's exact test (FET); Supplement Table S7). The enrichment analysis of shared ageing-related genes reveals 212 significant GO terms (FDR<0.05, Supplement Table S8). Further summarization in 80 non-redundant GO term sets by REVIGO reveals that the top 15 ranked sets are associated with lipid biosynthetic process (GO:0008610), growth (neuron projects development, GO:0031175; chemotaxis, GO:0006935; system development, GO:0048731; blood vessel development, GO:0001568, wound healing, GO:0042060; hippocampus development, GO:0021766) and response to glucocorticoids (GO:0051384) (Supplement Figure S3).

Sexual differentiation and maturation in NMR are delayed until transition from worker to breeder

Next, we determined DEGs between sexes within the groups of non-breeders and breeders for each tissue and species (Table 1; Supplement Table S9, Supplement Data 2). GP non-breeder females vs. males (GP-N-FvM) show over all tissues except gonads 1,713 significant expression differences (EDs) in 1,634 genes (FDR<0.01), primarily in Adr (858), Lvr (383), Thy (347) and Kid (109). Between female and male GP breeders (GP-B-FvM), we observed 3,654/3,398 EDs/DEGs. These transcriptome data confirm a prominent level of sexual differentiation among sexually mature GPs that further increases after onset of breeding. Breeders have 790 DEGs in common with non-breeders ($p<2.2x10^{-16}$, FET; Figure 2A). Functional enrichment analysis of these shared genes reveals 158 significant GO terms and 67 sets (Supplement Table S10). Among highest ranked sets we find immune system (lymphocyte activiation, GO:0046649; leucocyte migration GO:0050900; B cell receptor signalling pathway, GO:0050853), steroid metabolic process (GO:0008202) and sets related to oxidative stress (respiratory electron transport chain, GO:0022904; superoxide anion generation, GO:0042554) (Supplement Figure S5).

Similarly to GP-B-FvM, NMR-B-FvM shows 2,456/2,360 sex-related EDs/DEGs (FDR<0.01), mostly in Thy (1,791) and Adr (533). The overlap with GP-B-FvM is with 514 DEGs considerable but does not reach significance (p=0.062, FET; Figure 2A). Nevertheless, these data indicate basic similarities in sexual differentiation among breeders of both species. Functional enrichment analysis of NMR-B-FvM DEGs revealed 74 terms and 35 sets related to general biological processes (Supplement Table S11 and Supplement Figure S6).

Table 1: Numbers of DEGS identified in the different comparisons (FDR<0.01).

	female vs. male				breeder vs. non-breeder			
	non-breeder		breeder		females		males	
Organ	GP	NMR	GP	NMR	GP	NMR	GP	NMR
Hrt	4	10	89	13	2	3	0	0
Skn	6	6	3	5	5	1	0	223
Lvr	383	4	235	10	71	0	9	1
Kid	109	6	106	21	1	4	0	0
Cer	1	5	2	25	0	15	0	0
Hyp	2	8	1	11	0	5	2	0
Pit	347	9	307	47	114	114	2	1
Thy	3	0	2,087	1,791	675	285	1	0
Adr	858	4	824	533	0	201	1	4
Gon	-	-	-	-	18	502	3	381
ED*	1,713	52	3,654	2,456	886	1,130	18	610
NR**	1,634	22	3,398	2,360	883	1,078	14	598

* significant expression differences across tissues
** non-redundant set of significant expression differences across tissues

Surprisingly, only 22 NMR-N-FvM EDs/DEGs were detected across all tissues (Supplement Table S12), indicating an only minor sex differentiation between NMR non-breeding females and males. Nonetheless, we observed a significant overlap with GP-N-FvM DEGs (p=0.017, FET; Figure 2A) as four of the six common DEGs are located on the X chromosome where they presumably escape X inactivation in both species (Supplement Table S12).

Status change of NMRs is accompanied by major changes in the endocrine system

Then, we determined DEGs between breeders and non-breeders within the same sex for each species (Table 1; Supplement Table S13, Supplement Data 3). Females show a similar amount of EDs/DEGs in both species (GP-F-BvN: 886/883, NMR-F-BvN: 1,130/1,078) but have only 46 DEGs in common (Figure 2B). This is less than expected by chance, although not reaching significance (p=0.075, FET for depletion) and indicates that the molecular signature of the transition from female non-breeder to breeder is different in both species. E.g. in GP-F-BvN, only 18 DEGs are observed in Ova and none in Adr while in NMR-F-BvN, these tissues show most of the differences with 502 and 201 DEGs, respectively. Functional enrichment analysis of DEGs in NMR Ova reveals 283 GO terms that can be grouped into 104 sets (Supplement Table S14). Highest ranked categories are related to reproductive structure development (GO:0048608), steroid metabolism (GO:0008202), Ser/Thr kinase signalling (GO:0007178) and lipid homeostasis (GO:0055088) (Supplement Figure S7). This indicates substantial physiological and endocrine effects in NMR females in adaption to their role as breeders. Functional enrichment analysis in Adr revealed an obvious directionality in expression changes. DEGs are preferentially upregulated in reproduction (GO:0000003, 24 of 28) and endocrine system development (GO:0035270, 17 of 18) among highest ranked GO term sets (Supplement Figure S8), altogether 20 sets composed of 39 terms (Supplement Table S15), consistent with its role in hormone secretion and changes in Ova. Tallying with this, Cer DEGs are enriched and upregulated in steroid metabolic process (GO:0008202, 6 of 6 upregulated) and response to hormones (GO:0009725, 7 of 7) (Supplement Figure S9, altogether 38 sets composed of 93 terms, Supplement Table S16).

In male GPs, status-related differences (GP-M-BvN) are almost absent across all tissues (only 18 EDs/14 DEGs). In contrast, NMR-M-BvN shows 610/598 EDs/DEGs, predominantly in Tes (381) and Skn (223). NMRs share 55 status-related DEGs in both sexes (p=0.008, FET; Supplement Table S17), while the few status-related changes in male GPs show no overlap with those in females (Figure 2B). Among shared DEGs in NMRs, we identified 10 genes involved in endocrine signaling, including *SSTR3* (somatostatin receptor), *TAC4* (tachykinin), *PRDX1* (peroxiredoxin 1) and ACPP (acid phosphatase, prostate), as well as in general signaling via cAMP signaling (three genes) and through G-protein coupled receptors (four genes).

Mitochondrial genes show opposed expression changes in Tes and Skn after status change of NMR males

In NMR-M-BvN Tes, functional enrichment analysis reveals 128 GO terms that group into 39 sets (Supplement Table S18). Highest ranked GO sets are metabolism- and energy-

related, and the DEGs enriched therein are mostly upregulated, e.g. lipid biosynthetic process (GO:0008610, 75 of 82 genes), coenzyme metabolic process (GO:0006732, 20 of 21), energy derivation by oxidation of organic compounds (GO:0015980, 25 of 25), oxidation-reduction process (GO:0055114, 64 of 64) and glutathione metabolic process (GO:0006749, 13 of 14) (Supplement Figure S10). Further, we observe an upregulation of response to stimulus (GO:0050896, 79 of 106), in line with an upregulation of steroid metabolic process (GO:0008202, 28 of 28) included in lipid biosynthetic process set. In accordance with the dominance of energy-related processes, DEGs are enriched and preferentially upregulated in the top GO cellular component terms mitochondria (GO:0005739, 61 of 62) and peroxisomes (GO:0005777, 13 of 13) (Supplement Table S19). Together, this indicates increased demands of energy, e.g. to produce steroid hormones in Tes of NMR breeders.

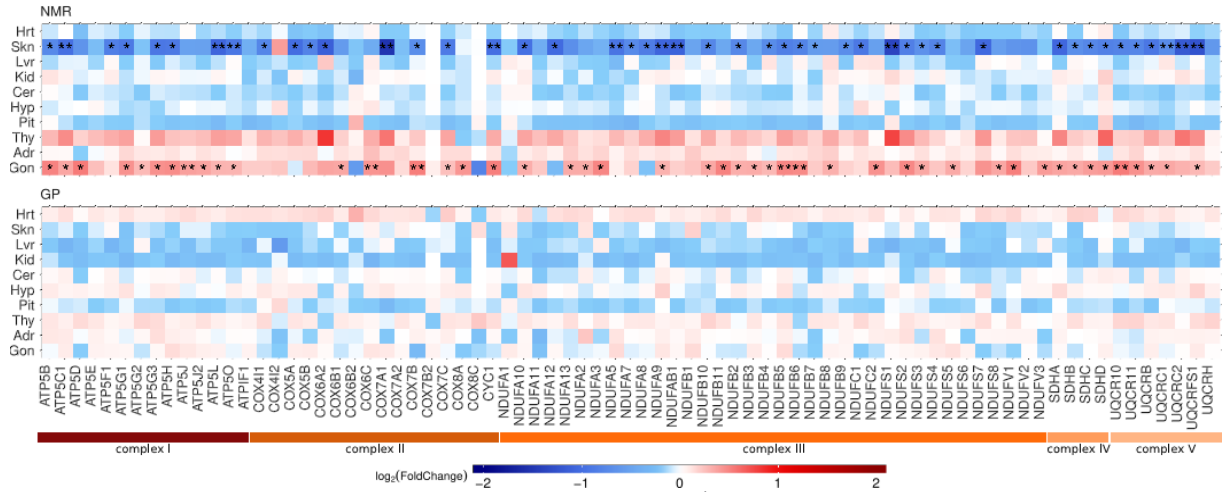


Figure 3: Heatmap represents log_2 fold-changes in the male breeder vs. non-breeder comparison of all 77 nuclear genes encoding for mitochondrial respiratory chain complexes. Significant expression changes were observed in NMR Skn (46 genes, mean fold-change 0.76) and Tes (46, 1.89) and are indicated by asterisks *: FDR<0.05, **: FDR<0.01.

In Skn, NMR-M-BvN DEGs are enriched in 121 GO terms that can be summarized in 29 sets (Supplemental Table S20, Supplemental Figure S11). Similar to Tes, Skn shows enrichment of energy-related processes. However, GO term sets in Skn are mostly downregulated, including energy derivation by organic compounds (GO:0015980, 36 of 38 genes) and oxidation-reduction process (GO:0055114, 42 of 46). Consistently, this includes genes associated with mitochondria (GO:0044429, 56 of 57) and respiratory chain (GO:0070469, 11 of 11).

The overlap between mitochondrial DEGs in Tes and Skn comprises 6 genes ($p=7.27x10^{-8}$, FET; Figure 3). Among common genes *PINK1* (PTEN induced putative kinase 1) is 1.5-fold upregulated in Tes and 2.5-fold downregulated in Skn, indicating a role in regulation of mitophagy (Narendra et al. 2010) in both tissues.

To follow up the mitochondria-related findings, we determined the 'mitonuclear transcript ratio' as the read count ratio of mitochondrial-encoded genes versus nuclear-encoded genes. It differs largely between tissues and species (Supplement Table S21). Hrt shows the highest mitonuclear ratio, with a minor difference between species (NMR 30.1%, GP 30.4%). Tes shows the lowest ratio, particularly in GPs (NMR

5.5%, GP 0.9%), and a 43.1% increase in NMR-M-BvN (p=0.003, t-test; Supplement Figure S12). This increase is accompanied with an upregulation of nuclear genes encoding for mitochondrial respiratory chain complexes (Figure 3; Supplement Table S22). The expected increase in ROS is compensated by an on average 1.59-fold upregulation of eight antioxidant DEGs (Supplement Table S23). Consistently with functional enrichment analysis mentioned above, we observe an opposing effect in Skn of NMR male breeders, which show a decline in mitonuclear ratio together with a downregulation of nuclear genes of the respiratory complexes (Figure 3). In line with the downregulation of OXPHOS, we observe a downregulation of antioxidant enzymes SOD2 (superoxide dismutase 2, 2.64-fold) and PRDX3 (peroxiredoxin 3, 2.14-fold).

NMR status-related DEGs are enriched in ageing-related genes

Driven by the exceptional long and healthy lifespan of NMR breeders, we searched our transcriptome data for molecular evidences supporting this phenomenon. First, we assessed the connection of cross-species DEGs with expression changes that are associated with status change in each species. DEGs shared in both contrasts (NMR vs. GP and breeder vs. non-breeder, FDR<0.05) provided the basis for a correlation analyses of fold changes in each species (Supplement Figure S4). Hypothesizing, in agreement with the disposable soma theory of ageing, a negative impact of reproduction on lifespan for GP and, in contradiction to this theory, an inverse effect for NMR, we confirmed this by opposing correlations (combined $p=3.8x10^{-9}$ (Lancaster procedure (Dai et al. 2014)), negative correlation for GP (1,380 shared DEGs; Spearman correlation -0.1; $p=1.6x10^{-4}$) and positive correlation for NMR (3,695; 0.17, $p=1.7x10^{-24}$)). This means that DEGs with higher expression in NMR than GP are preferentially up-regulated in NMR breeders compared to non-breeders, and vice versa.

Second, we found that only NMR status-related DEGs show significant enrichments of ageing-related genes from DAA (Supplement Table S24): males in Skn (55 genes; q=0.0012, FET) and Tes (80 genes, q=0.01), and females in Ova (127 genes; $q=1.2x10^{-7}$), Thy (59 genes, q=0.033) and Adr (43 genes, q=0.038). The significant overlap of 22 DEGs between NMR-F-BvN and NMR-M-BvN in Gon (p=0.0035, FET) contains nine ageing-related genes (p=0.004, Figure 2C). In GP, only the non-redundant set of DEGs in GP-F-BvN shows a tendency of enrichment (160 genes, q=0.051), in contrast to NMRs, which show enrichment in males (134 genes, $q=4.5x10^{-5}$) and females (245 genes, $q=7x10^{-9}$).

Third, we expect that reproduction impacts the life expectancy of NMR and GP differently. Therefore, we searched for status-related DEGs that are shared in both species, but show opposing direction of expression. These genes might mark different coping mechanisms with the metabolic load of reproduction. As described above, the overlap of DEGs between species is very low (Supplemental Table S25). Nevertheless, opposing direction of expression change can be observed in Ova (1 of 2 shared DEGs; *FGFR2*: fibroblast growth factor receptor 2) and female thyroid (8/8) and testis (1/1).

Fourth, and based on a similar assumption, we determined enrichment of ageing-related genes in the 20%-quantile of genes, which show the greatest interspecies difference in status-related expression changes (FDR<0.05). In males, we found significant enrichments of DAA genes in Skn ($q=7.04*10^{-7}$) and Tes (q=0.0025). In the latter, only

genes upregulated in NMR breeders show enrichment regardless of the direction of expression change in GPs. In contrast, genes in Skn that have the same direction of expression change in both species exhibit significance. Like males, female Skn (q=0.0064) is enriched for ageing-related genes if one considers downregulated expression in breeders in both species. Further, genes with opposing expression change in females show enrichment in Hrt (q=0.0024), Pit (q=0.0053) and Ova (0.0013). Further functional enrichment analysis of these ageing-related gene sets reveals differences between sexes. In males, the non-redundant set of genes shows enrichment in 268 GO terms and 89 sets (Supplement Table S27, Supplement Figure S13). Highest ranked sets are associated with lipid metabolism (GO:0006629), energy metabolism (energy derivation by oxidation of organic compounds, GO:0015980; mitochondrial ATP synthesis coupled proton transport, GO:0042776), glutathione metabolic process (GO:0006749) and immune system (innate immune response-activating signal transduction, GO:0002758). Females, show enrichment in 528 GO terms summarized in 129 sets (Supplement Table S27). Notably, highest ranked sets reveal enrichment in positive regulation of tumour necrosis factor production (GO:0032760) and negative regulation of programmed cell death (GO:0043069) (Supplement Figure S14).

Discussion

In our comparative study of breeders vs. non-breeders of the eusocial long-lived NMR and the colonial, polygynous and shorter-lived GP, we accumulated a comprehensive set of transcriptome data to gain insights into naturally evolved interspecies differences in sexual maturation and links between reproduction and ageing. Both species are able to breed year-round and produce four to five litters per year (Roellig et al. 2011; Peaker and Taylor 1996). Both species have a similar average gestation period of 70 days, which is long compared to similarly sized species. Notably, NMRs produce on average 10.5 offspring per litter, which is twice the number of offspring produced by similarly sized rodents and more than three times higher compared to GPs (average 3.2 offspring) (Roellig et al. 2011; Peaker and Taylor 1996). This further underscores the apparent contradiction of the NMR queen’s enormous metabolic load and extraordinary long life/healthspan (Buffenstein and Jarvis 2002) to the disposable soma theory of ageing (Kirkwood 1977), indicating that natural ways to extended healthspan remain to be uncovered in NMRs.

A first study to identify adaptations to unique NMR traits at the transcriptome level compared liver gene expression of young adult non-breeding male NMRs and mice (Yu et al. 2011). Higher NMR transcript levels were observed for genes associated with oxidoreduction and mitochondria as well as highlighted epithelial cell adhesion molecule (Epcam), alpha2-macroglobulin (A2m), and the mitochondrial complex II subunit (Sdhc) as candidates for specifying species differences in ageing and cancer. Our present study, more comprehensive in several aspects (sex, breeding status, numbers of animals and tissue samples), is based on a comparison of NMR vs. GP, which are phylogenetically closer than NMR and mouse. It revealed that between NMR and GP most analysed genes are differentially expressed and that these DEGs are significantly enriched for ageing-related genes in DAA. Among the latter, the main functional commonality is their association with lipid metabolism. Among cross-species DEGs that

are consistently differentially expressed across all tissues, we identified further ageing-related candidates. *RRAGB* (Ras related GTP binding B), overexpressed in NMRs, interacts with mTORC1 complex (Sancak et al. 2008) whose dysregulation has been linked to ageing and diseases such as cancer and diabetes (Zoncu et al. 2011). Further, *TMEM8C* (transmembrane protein 8C) shows overexpression in NMRs, is essential for muscle regeneration (Millay et al. 2013) and might be linked to the resistance to muscle loss in aged NMRs (Stoll et al. 2016; Holtze et al. 2016).

In respect to molecular signature of reproductive status (breeder vs. non-breeder) in NMR and GP, our main finding is the nearly complete absence of significant transcriptional differences between sexes in adult non-breeding NMRs. This fits their grossly identical morphology and identical behaviour in stable colonies (Sherman et al. 1991a). This is in stark contrast to non-breeding adult GPs of an even younger age where we observed thousands of DEGs. GP non-breeders and breeders share a large and highly significant number of sex-related DEGs. Among others, these DEGs are enriched in GO terms related to steroid metabolism and immune system. The effect of gonadal steroids on the immune system is well described in GPs and other mammals (Grossman 1985). After separation of NMR non-breeders from their colony, sexes became not only distinguishable by morphology and behaviour (Dengler-Crish and Catania 2007; Sherman et al. 1991a), but also by gene expression. This differentiation on transcriptional level provides further molecular support for the previously described suppression of sexual maturation in non-breeding adult NMRs by their colony environment (Smith et al. 1997; Dengler-Crish and Catania 2007).

This is in line with the observed major changes in the endocrine system after status change in NMRs but not in GPs. The upregulation (i) of genes related to steroid metabolism in Ova and Tes, (ii) of endocrine system development in female Adr as well as (iii) of steroid metabolic process and response to hormones in female Cer, indicate increased feedback signalling via gonadal steroids after NMRs transition into breeders. Gonadal hormones (estrogens, androgens, progesterogens) are an essential part of the hypothalamus-pituitary-gonads axis by mediating growth, secondary sex characteristics and reproduction (Clarke et al. 2012). They mediate feedback signals to hypothalamus and pituitary, which in turn regulate the secretion of gonadotropic hormones and neuropeptides. A selection of respective significantly upregulated genes is provided as Supplement Text S1.

In testis of immature chicken, expression of *ADIPOR1/2* (adiponectin receptor 1/2) is significantly less compared to mature animals (Ocón-Grove et al. 2008). It has been hypothesised that these genes are involved in supporting the higher metabolic activity related to spermatogenesis, testicular steroid hormone production, and transport of spermatozoa and testicular fluid. In line with these observations, our results show a significant upregulation of *ADIPOR2* in Tes of NMR breeders (1.7-fold).

The renin-angiotensin system predominantly involved in cardiovascular control has also been associated with reproduction in mice and human (Pan et al. 2013). In particular, signalling through receptors coded by *AGTR1* (angiotensin II receptor type 1) in males, and *AGTR1/2* in females has been associated with fertility and stimulation of

reproduction. In accordance, we observe significant upregulation of *AGTR1* and *AGTR2* in gonads of NMR male and female breeders, respectively.

Notably, although well expressed, we found no significant differences for gonadotropin-related genes (*GNRH1*, *FSHB*, *CGA*, *LHB)*. This indicates a similar intracellular transcript turnover in non-breeders and breeders, and is consistent with previous results indicating that LH is stored in non-breeders, ready to be released upon GnRH signalling (Faulkes et al. 1990b). Previously, elevated diencephalon mRNA levels were observed in NMR for *AR* (androgen receptor) in male NMR breeders and of *CYP19A1* (aromatase), *ESR1* (estrogen receptor 1) and *PGR* (progesterone receptor) in female breeders (Swift-Gallant et al. 2015). Our corresponding Hyp data do not support those increases (log2FC<0.1). Nevertheless, we identified significantly elevated transcript levels of these genes both in male (Tes: *AR*) and female breeders (Pit: *PGR*; Ova: *AR*, *CYP19A1*, *ESR1*, *PGR*) supporting a complex status-related function of these genes. Further, steroid feedback and GnRH secretion are integrated by brain GABA and glutamate signalling in mammals and cichlids (Renn et al. 2008). Although, we do not observe equivalent significant differences in the NMR brain, the data show an upregulation of genes coding for three GABA receptor subunits (*GABRB3*, *GABRG1*, *GABRP*) and one glutamate receptor subunit (*GRIK2*) in Tes, as well as two of three differentially expressed GABA (*GABRB2*, *GABRG3*) and three glutamate receptor subunits (*GRIK1*, *GRIK2*, *GRIK4*) in Ova and female Pit of NMR breeders. Together, this suggests increased neuronal plasticity and/or activity predominantly in gonads of NMRs after becoming breeders.

Another important group of regulators of the hypothalamus-pituitary-gonad axis are glucocorticoids. They have been linked to stress, reproduction and social behaviour in a variety of species, including members of muroidae, primates and cichlids (Gesquiere et al. 2011; O'Connor et al. 2013; Crump and Chevins 1989). However, in NMRs, correlation between social status and urinary cortisol is not clear and seems to depend on colony stability (Clarke and Faulkes 1997, 1998). Here, we observe a significant upregulation of the glucocorticoid receptor gene (*NR3C1*) in Tes of NMR male and Thy of female breeders as well as of *HSD11B1* (hydroxysteroid 11-beta dehydrogenase 1) in Skn of male breeders, indicating increased conversion of inactive cortisone to receptor-active cortisol. Interestingly, this is in line with elevated expression of glucocorticoid receptor in Tes in African cichlid breeders. Males can reversibly change between dominant and subordinate phenotypes (Maruska and Fernald 2011). Similar to NMRs, only dominant phenotypes are reproductively active.

In male African cichlids, moreover, *SST* (somatostatin) coding for an important peptide hormone that is involved in the inhibition of growth hormone, and its receptor gene (*SSTR3*) have recently been implicated in regulating aggressive behaviour (Trainor and Hofmann 2006). In particular, aggression and androgen level are negatively correlated with expression of *SSTR3* in fish Tes. Similarly, we observe in NMR breeder Tes a significant upregulation of *AR* (androgen receptor) as well as an upregulation of genes involved in the conversion of testosterone (*HSD17B1*, hydroxysteroid 17-beta dehydrogenase 1, q=0.0123) and dihydrotestosterone (*SRD5A1*, steroid 5 alpha-reductase 1, q=0.045), together with significant downregulation of *SSTR3*. This indicates that *SSTR3* may be associated with social dominance in NMRs as well.

As our study was primarily motivated by the exceptional long and healthy lifespan of NMR breeders, we performed a correlation analysis between species (NMR vs. GP) and status (breeder vs. non-breeder) EDs confirming the basic hypothesis of the present work: the status-change into reproductive animals results in a gradual expression strategy related to fitness maintenance in reproductive NMRs in contrast to GPs. Genes which are higher or lower expressed in NMR compared to GP are also preferentially up- or downregulated in NMR breeders (positive correlation) opposite to GPs (negative correlaton). In other words, the positive correlation in NMR contradicts the disposable soma theory of ageing, as EDs contributing to a long lifespan (higher/lower expression in NMR than GP) are tendentially increased in NMR breeders compared to non-breeders, while diminished in GPs as expected by this theory.

Furthermore, several significant evidences indicate that NMR status change has an impact on genes involved in ageing: enrichment of ageing-related genes in the non-redundant DEG sets of males and females, as well as enrichment in most tissues with at least 50 DEGs (male Skn and Tes; female Ova, Thy and Adr). This contrasts with our observations in GPs, which show only a tendency of ageing-relation for the non-redundant set of status-related DEGs in females.

Further, we observe significant tissue-specific changes in mitochondrial activity of male NMR breeders. While Tes shows an upregulation of nuclear-encoded mitochondrial genes and a respective increase in mitonuclear transcript ratio, Skn shows the opposite. Moreover, we observe significant enrichments of genes involved in fatty acid metabolism among status-related DEGs in both NMR tissues. Consistent with the role of mitochondria in lipid homeostasis and the observed directionality of changes in mitochondrial activity, fatty acid metabolism DEGs in Tes were preferentially up- and in Skn downregulated. While the increased mitochondrial activity in Tes probably complies with demands of energy for the production of sex steroids and their anabolic effect on physiology, such as growth of testis (Faulkes et al. 1994), the observed changes in Skn may indicate a link to the extraordinary healthspan of NMR breeders. Previously, it was observed that inhibition of complex I activity during adult life prolongs lifespan and rejuvenates the tailfin transcriptome in short-lived fish (Baumgart et al. 2016).

Although only a small subset of status-related DEGs is shared between species, among them are ageing-related genes having opposing expression changes. We identified *FGFR2*, which is in Ova downregulated in NMRs, but upregulated in GPs. Members of this receptor family bind growth factors and thereby influence mitogenesis, differentiation and cancer (Jackson et al. 1997; Jang et al. 2001; Hunter et al. 2007; Cerliani et al. 2011). *FGFR2* is indirectly linked to ageing according to AgeFactDB, and FGFR2 blockers have been suggested as a therapeutic target for cancer treatment (Turner and Grose 2010). The downregulation of FGFR2 in NMR female breeders might suggest intrinsic anti-cancer and longevity mechanisms. Among status-related DEGs in NMR, two are related to senescence. *TOM1* (target of myb1 membrane trafficking protein), which is involved in the protein-degradation system (Makioka et al. 2016), is downregulated in female Thy. Increased expression of *TOM1* has been observed in cellular senescence in human fibroblast cell lines (Guo et al. 2004). In Tes, we observe upregulation of *PIR*

(Pirin), which inhibits cellular senescence in melanocytic cells (Licciulli et al. 2011).

Taken together, our comparative transcriptome analysis of breeders versus non-breeders of the eusocial, long-lived NMR versus the polygynous and shorter-lived GP provides novel insights into socially regulated sexual maturation and natural ways to extended life/healthspan that encourage further functional and mechanistic investigations of these extraordinary NMR phenotypes.

Materials & Methods

Animals

Naked mole-rats. NMR colonies were kept inside a climatized box (2x1x1 m) in artificial burrow systems, consisting of eight cylindrical acrylic glass containers (diameter: 240 mm height: 285 or 205 mm). The latter functioned as variable nest boxes, food chambers or toilet areas, and were interconnected with acrylic tubes having an inner diameter of 60 mm. Husbandry conditions were stabile during the entire experimental period of 22 months. Temperature and humidity were adjusted to 27.0□2.0°C and 85.0±5.0%, respectively. In general, the NMR colonies were kept in darkness except for 2 to 4 hours of daily husbandry activities. Fresh vegetable food was provided daily and *ad libitum*. In addition, commercial rat pellets (Vita special, Vitakraft GmbH, Bremen, Germany) were fed as an additional source of protein and trace elements.

To turn them into breeders, randomly selected non-breeding animals derived from two long-term (>4 years) established colonies of more than 50 individuals were separated and paired with the opposite sex. As non-breeder controls, litter siblings of paired animals remained in their colonies as workers. After the lactation period of the second set of live offspring the tissue sampling was scheduled. To avoid further pregnancies in the females, male partners were removed and euthanized 8-10 days postpartum. The tissue collection in the females took place 40-50 days after the end of last pregnancy.

Guinea pigs. GPs (breed: Dunkin Hartley HsdDhl:DH, Harlan Laboratories, AN Venray, Netherlands) were housed in standardized GP cages (length: 850 mm, width: 470 mm, height: 450 mm) in breeding pairs plus offspring or in same-gender pairs of two. Commercial guinea pig pellets and commercial pet food hay (Hellweg Zooland GmbH, Berlin, Germany) were provided together with vitamin C enriched water *ad libitum*. Housing temperature and humidity were 18.0±2.0°C and 45.0±5.0%, respectively. A 12h light/dark regime was provided.

After an initial adaption period of 6 to 8 weeks the GPs were randomly divided in breeding pairs or in same-gender pairs of two. The offspring were separated from their parents after weaning (~3 weeks postpartum). Tissue collection was scheduled after the lactation period of the second set of live offspring. To avoid further pregnancies in the females, male partners were removed between eleven days before and seven days after birth of the second litter. The tissue collection in the females took place 42-83 days after the end of last pregnancy.

For tissue collection, all animals were anaesthetized by 3% isoflurane inhalation anaesthesia (Isofluran CP, CP-Pharma, Burgdorf, Germany) and euthanized by surgical decapitation. Animal housing and tissue collection at the Leibniz Institute for Zoo and Wildlife Research was compliant with national and state legislation (breeding allowance #ZH 156; ethics approval G 0221/12 "Exploring long health span", Landesamt für Gesundheit und Soziales, Berlin).

Sample collection, RNA Isolation and Sequencing

For *de novo* transcriptome assembly, animals were euthanized and ten tissue samples (heart – Hrt (NMR only), skin - Skn, liver - Lvr, kidney - Kid, cerebellum - Cer, hypothalamus - Hyp, pituitary - Pit, thyroid - Thy, adrenal - Adr, and gonads - Gon (testis – Tes /ovaries - Ova)) were collected from NMR and GP individuals, as described previously (Bens et al. 2016). Strand-specific RNA-seq were prepared using the TruSeq Stranded RNA LT Kit (Illumina), and 200-nt reads were obtained using a HiSeq2500 (Illumina), as described previously (Bens et al. 2016).

For expression analysis, the same ten tissues were collected from NMR and GP breeders and non-breeders. RNA was purified as described above. Library preparation was done using Illumina's TruSeq RNA Library Prep Kit v2 kit following the manufacturer's description. Quantification and quality check of the libraries was done using Agilent's Bioanalyzer 2100 in combination with a DNA 7500 Kit (both Agilent Technologies). Sequencing was done on a HiSeq 2500 running the machine in 51 cycle, single-end, high-output mode by multiplexing seven samples per lane. Demultiplexing and extraction of read information in FastQ format was done using the tool bcl2astq v1.8.4 (provided by Illumina).

Data Analysis

De novo transcriptome assembly and annotation for GP was performed as described in (Bens et al. 2016). Briefly, overlapping paired-end reads were joined into single fragments and then assembled by Trinity (Grabherr et al. 2011). Gene symbols were assigned to the assembled transcripts by similarity to human transcripts using FRAMA (Bens et al. 2016).

As a reference for RNA-seq data mapping the public NMR (Bioproject PRJNA72441) (Keane et al. 2014) and GP genomes (UCSC, cavpor3) were used. Reference transcript sets of NMR and GP were mapped to the corresponding genome in two steps: BLAT (v36) (Kent 2002) was used to identify the locus and then SPLIGN (v1.39.8) (Kapustin et al. 2008) was applied to splice align the transcript sequence within BLAT locus. RNA-seq data were aligned to the corresponding reference genome utilizing STAR (v2.4.1d) (Dobin et al. 2013) with a maximum mismatch of 6% and a minimum aligned length of 90%. Reads mapped to multiple loci were discarded. Gene expression was quantified using HTSEQ (v0.6.1p1) (Anders et al. 2015) based on the aligned reference transcripts. The pairwise Pearson correlation between biological replicates was calculated based on 16,339 and 16,009 genes in NMR and GP, respectively (Supplement Table S5).

The 'mitonuclear transcript ratio' was calculated as the read count ratio of 13 mitochondrial-encoded genes versus all nuclear-encoded genes.

PosiGene was applied to the transcriptome of human (**), NMR and GP with the parameter '-prank=0 -max_anchor_gaps_hard=100 -rs=NMR' to determine orthologous transcribed regions in NMR and GP having a protein identity >70%. RNA-seq data were aligned to the corresponding transcriptomes utilizing bowtie2 (2.2.9) (Langmead and Salzberg 2012) with the parameter '--very-sensitive-local'.

DESeq2 (v1.6.3) (Love et al. 2014) was used to identify DEGs. A false discovery rate <0.01 (FDR; Benjamini Hochberg corrected p values, (Kasen et al. 1990)) was used for the identification of significant DEGs.

Gene Ontology analysis was performed using the web interface of GoMiner (Database build 2011-01) based on the functional annotation of human genes (UniProt) (Zeeberg et al. 2003). A FDR<0.05 was used for the identification of significant GO terms that were summarized by REVIGO (parameter SimRel=0.5) into non-redundant GO term sets (Supek et al. 2011). GO term sets were then ranked by number of summarized GO terms and number of changed genes.

Overlap between gene sets was determined with Fisher's exact test (FET) using the one-sided option. Generally, we tested for enrichment if not otherwise stated.

We obtained 3,009 ageing-related genes in human and mouse from the Digital Ageing Atlas (DAA) (Craig et al. 2014). The corresponding counterparts in the NMR (2,588) and GP (2,539) were used for enrichment analysis and results were corrected for multiple testing (FDR). P-values corrected for multiple testing are indicated by q and nominal p-values by p.

To examine the connection between reproduction and ageing in both species, we determined the difference in log2-fold-change (breeders vs. non-breeders) of NMR and GP. For fold-changes moving in opposite directions between species, we calculated the absolute difference $(|\log_2 NMR_{BvsN} - \log_2 GP_{BvsN}|)$, and for fold-changes moving in the same direction, higher fold-changes in NMR-BvsN were rewarded $(|\log_2 NMR_{BvsN}| - |\log_2 GP_{BvsN}|)$. The 20%-quantile of genes having the greatest difference was determined separately for (i) the complete gene set and for genes showing (ii) opposing and (iii) unidirectional fold-changes. All sets were tested for enrichment of ageing-related genes.

Statistical analyses were performed in R (version v3.1.2).

Data Accession

RNA-seq data for gene expression profiling were deposited at Gene Expression Omnibus (GSE98719). RNA-seq data for *de novo* assembly were deposited at Sequence Read Archive (SRP104222, SRP061363) and the corresponding gene collection is available as a gff3-file (ftp://genome.leibniz-fli.de/pub/nmr2017/).

Acknowledgements

We like to acknowledge Angelika Kissmann and Jette Ziep for their professional help with the animal husbandry. We also thank Michaela Wetzel for her logistic help during tissue sampling and storage as well as for her support with the animal husbandry. We are grateful to Ivonne Görlich for excellent technical assistance and Debbra Weih for

critical reading of the manuscript. This research was supported by the Bundesministerium für Bildung und Forschung (BMBF) and the Leibniz-Gesellschaft through the Senatsausschusswettbewerb (SAW) 2012, as well as the Deutsche Forschungsgemeinschaft (DFG), grant PL 173/8-1.

Competing interests

None declared.

Author Contributions

TBH, MP and KS conceived the project. SH, TBH, MB, MG and KS performed the animal study, sampling and sequencing experiments. Data analysis and interpretation were performed by MB, KS, SH, MP and TBH. The manuscript was written by MB and KS, and revised by all authors. The manuscript submitted by MB; all authors reviewed and approved the submitted manuscript. TBH and MP are joint senior authors.

References

Abbott DH, Hodges JK, George LM. 1988. Social status controls LH secretion and ovulation in female marmoset monkeys (Callithrix jacchus). J Endocrinol 117: 329–339.

Anders S, Pyl PT, Huber W. 2015. HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics 31: 166–169.

Baumgart M, Priebe S, Groth M, Hartmann N, Menzel U, Pandolfini L, Koch P, Felder M, Ristow M, Englert C, et al. 2016. Longitudinal RNA-Seq Analysis of Vertebrate Aging Identifies Mitochondrial Complex I as a Small-Molecule-Sensitive Modifier of Lifespan. Cell Syst 2: 122–32.

Bens M, Sahm A, Groth M, Jahn N, Morhart M, Holtze S, Hildebrandt TB, Platzer M, Szafranski K. 2016. FRAMA: from RNA-seq data to annotated mRNA assemblies. BMC Genomics 17: 54.

Braude S. 2000. Dispersal and new colony formation in wild naked mole-rats: evidence against inbreeding as the system of mating. Behav Ecol 11: 7–12.

Buffenstein R. 2005. The naked mole-rat: a new long-living model for human aging research. J Gerontol A Biol Sci Med Sci 60: 1369–77.

Buffenstein R, Jarvis JUM. 2002. The naked mole rat--a new record for the oldest living rodent. Sci aging Knowl Environ 2002: pe7.

Burda H, Honeycutt RL, Begall S, Locker-Grutjen O, Scharff A. 2000. Are naked mole-rats eusocial and if so, why? Behav Ecol Sociobiol 47: 293–303.

Cerliani JP, Guillardoy T, Giulianelli S, Vaque JP, Gutkind JS, Vanzulli SI, Martins R, Zeitlin E, Lamb CA, Lanari C. 2011. Interaction between FGFR-2, STAT5, and progesterone receptors in breast cancer. Cancer Res 71: 3720–31.

Clarke FM, Faulkes CG. 1997. Dominance and queen succession in captive colonies of the eusocial naked mole-rat, Heterocephalus glaber. Proc R Soc London B Biol Sci 264: 993–1000.

Clarke FM, Faulkes CG. 1998. Hormonal and behavioural correlates of male dominance and reproductive status in captive colonies of the naked mole-rat,

Heterocephalus glaber. Proc Biol Sci 265: 1391–1399.

Clarke IJ, Campbell R, Smith JT, Prevot V, Wray S. 2012. Neuroendocrine control of reproduction. In Handbook of Neuroendocrinology, pp. 197–235, Elsevier Inc.

Craig T, Smelick C, Tacutu R, Wuttke D, Wood SH, Stanley H, Janssens G, Savitskaya E, Moskalev a., Arking R, et al. 2014. The Digital Ageing Atlas: integrating the diversity of age-related changes into a unified resource. Nucleic Acids Res 43: D873--D878.

Crump CJ, Chevins PF. 1989. Prenatal stress reduces fertility of male offspring in mice, without affecting their adult testosterone levels. Horm Behav 23: 333–43.

Dai H, Leeder JS, Cui Y. 2014. A modified generalized Fisher method for combining probabilities from dependent tests. Front Genet 5: 32.

Dammann P, Burda H. 2006. Sexual activity and reproduction delay ageing in a mammal. Curr Biol 16: 117–118.

Dammann P, Šumbera R, Maßmann C, Scherag A, Burda H. 2011. Extended longevity of reproductives appears to be common in Fukomys mole-rats (Rodentia, Bathyergidae). PLoS One 6: 2–8.

Dengler-Crish CM, Catania KC. 2007. Phenotypic plasticity in female naked mole-rats after removal from reproductive suppression. J Exp Biol 210: 4351–4358.

Dobin A, Davis C a., Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21.

Faulkes CG, Abbott DH, Jarvis JUM. 1990a. Social suppression of ovarian cyclicity in captive and wild colonies of naked mole-rats, Heterocephalus glaber. Reproduction 88: 559–568.

Faulkes CG, Abbott DH, Jarvis JUM. 1991. Social suppression of reproduction in male naked mole-rats, Heterocephalus glaber. J Reprod Fertil 91: 593–604.

Faulkes CG, Abbott DH, Jarvis JUM, Sherriff FE. 1990b. LH responses of female naked mole-rats, Heterocephalus glaber, to single and multiple doses of exogenous GnRH. Reproduction 89: 317–323.

Faulkes CG, Trowell SN, Jarvis JU, Bennett NC. 1994. Investigation of numbers and motility of spermatozoa in reproductively active and socially suppressed males of two eusocial African mole-rats, the naked mole-rat (Heterocephalus glaber) and the Damaraland mole-rat (Cryptomys damarensis). J Reprod Fertil 100: 411–6.

Gesquiere LR, Learn NH, Simao MCM, Onyango PO, Alberts SC, Altmann J. 2011. Life at the Top: Rank and Stress in Wild Male Baboons. Science 333: 357 LP-360.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29: 644–652.

Grossman CJ. 1985. Interactions between the gonadal steroids and the immune system. Science.

Guo S, Zhang Z, Tong T. 2004. Cloning and characterization of cellular senescence-associated genes in human fibroblasts by suppression subtractive hybridization. Exp Cell Res 298: 465–72.

Hochberg ME, Noble RJ, Braude S. 2016. A Hypothesis to Explain Cancers in Confined Colonies of Naked Mole Rats. bioRxiv.

Holtze S, Eldarov CM, Vays VB, Vangeli IM, Vysokikh MY, Bakeeva LE, Skulachev

VP, Hildebrandt TB. 2016. Study of age-dependent structural and functional changes of mitochondria in skeletal muscles and heart of naked mole rats (Heterocephalus glaber). Biochem 81: 1429–1437.

Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, et al. 2007. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 39: 870–4.

Jackson D, Bresnick J, Rosewell I, Crafton T, Poulsom R, Stamp G, Dickson C. 1997. Fibroblast growth factor receptor signalling has a role in lobuloalveolar development of the mammary gland. J Cell Sci 1261–8.

Jang JH, Shin KH, Park JG. 2001. Mutations in fibroblast growth factor receptor 2 and fibroblast growth factor receptor 3 genes associated with human gastric and colorectal cancers. Cancer Res 61: 3541–3.

Jarvis JU. 1981. Eusociality in a mammal: cooperative breeding in naked mole-rat colonies. Science 212: 571–3.

Kapustin Y, Souvorov A, Tatusova T, Lipman D. 2008. Splign: algorithms for computing spliced alignments with identification of paralogs. Biol Direct 3: 20.

Kasen S, Ouellette R, Cohen P. 1990. Mainstreaming and postsecondary educational and employment status of a rubella cohort. Am Ann Deaf 135: 22–6.

Keane M, Craig T, Alfoldi J, Berlin a. M, Johnson J, Seluanov a., Gorbunova V, Di Palma F, Lindblad-Toh K, Church GM, et al. 2014. The Naked Mole Rat Genome Resource: facilitating analyses of cancer and longevity-related adaptations. Bioinformatics 30: 3558–3560.

Kent WJ. 2002. BLAT - The BLAST-like alignment tool. Genome Res 12: 656–664.

Kim EB, Fang X, Fushan A a., Huang Z, Lobanov A V, Han L, Marino SM, Sun X, Turanov A a., Yang P, et al. 2011. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. Nature 479: 223–227.

Kirkwood TB. 1977. Evolution of ageing. Nature 270: 301–304.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–359.

Licciulli S, Luise C, Scafetta G, Capra M, Giardina G, Nuciforo P, Bosari S, Viale G, Mazzarol G, Tonelli C, et al. 2011. Pirin inhibits cellular senescence in melanocytic cells. Am J Pathol 178: 2397–406.

Love MI, Huber WW, Anders S, Lönnstedt I, Speed T, Robinson M, Smyth G, McCarthy D, Chen Y, Smyth G, et al. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15: 550.

Makioka K, Yamazaki T, Takatama M, Ikeda M, Murayama S, Okamoto K, Ikeda Y. 2016. Immunolocalization of Tom1 in relation to protein degradation systems in Alzheimer's disease. J Neurol Sci 365: 101–7.

Maruska KP, Fernald RD. 2011. Plasticity of the reproductive axis caused by social status change in an african cichlid fish: II. Testicular gene expression and spermatogenesis. Endocrinology 152: 291–302.

Millay DP, O/'Rourke JR, Sutherland LB, Bezprozvannaya S, Shelton JM, Bassel-Duby R, Olson EN. 2013. Myomaker is a membrane activator of myoblast fusion and muscle formation. Nature 499: 301–305.

Narendra DP, Jin SM, Tanaka A, Suen D-F, Gautier CA, Shen J, Cookson MR, Youle

RJ. 2010. PINK1 Is Selectively Stabilized on Impaired Mitochondria to Activate Parkin. PLOS Biol 8: e1000298.

O'Connor CM, Rodela TM, Mileva VR, Balshine S, Gilmour KM. 2013. Corticosteroid receptor gene expression is related to sex and social behaviour in a social fish. Comp Biochem Physiol A Mol Integr Physiol 164: 438–46.

O'Riain MJ, Jarvis JU, Faulkes CG. 1996. A dispersive morph in the naked mole-rat. Nature 380: 619–621.

Ocón-Grove OM, Krzysik-Walker SM, Maddineni SR, Hendricks GL, Ramachandran R. 2008. Adiponectin and its receptors are expressed in the chicken testis: influence of sexual maturation on testicular ADIPOR1 and ADIPOR2 mRNA abundance. Reproduction 136: 627–38.

Orr ME, Garbarino VR, Salinas A, Buffenstein R. 2016. Extended Postnatal Brain Development in the Longest-Lived Rodent: Prolonged Maintenance of Neotenous Traits in the Naked Mole-Rat Brain. Front Neurosci 10: 504.

Pan PP, Zhan QT, Le F, Zheng YM, Jin F. 2013. Angiotensin-converting enzymes play a dominant role in fertility. Int J Mol Sci 14: 21071–21086.

Peaker M, Taylor E. 1996. Sex ratio and litter size in the guinea-pig. Reproduction 108: 63–67.

Renn SCP, Aubin-Horth N, Hofmann H a. 2008. Fish and chips: functional genomics of social plasticity in an African cichlid fish. J Exp Biol 211: 3041–3056.

Roellig K, Drews B, Goeritz F, Hildebrandt TB. 2011. The long gestation of the small naked mole-rat (Heterocephalus glaber Rüppell, 1842) studied with ultrasound biomicroscopy and 3D-ultrasonography. PLoS One 6: e17744.

Sancak Y, Peterson TR, Shaul YD, Lindquist RA, Thoreen CC, Bar-Peled L, Sabatini DM. 2008. The Rag GTPases Bind Raptor and Mediate Amino Acid Signaling to mTORC1. Science 320: 1496 LP-1501.

Sherman PW, Jarvis JUM, Alexander RD. 1991a. An Ethogram for the Naked Mole-Rat: Nonvocal Behaviors. In The Biology of the Naked Mole-rat, Monographs in behavior and ecology, pp. 209–242, Princeton University Press.

Sherman PW, Jarvis JUM, Alexander RD. 1991b. Reproduction in Naked Mole-Rats. In The Biology of the Naked Mole-rat, Monographs in behavior and ecology, pp. 384–425, Princeton University Press.

Skulachev VP, Holtze S, Vyssokikh MY, Bakeeva LE, Skulachev M V., Markov A V., Hildebrandt TB, Sadovnichii VA. 2017. Neoteny, Prolongation of Youth: From Naked Mole Rats to "Naked Apes" (Humans). Physiol Rev 97: 699–720.

Smith TE, Faulkes CG, Abbott DH. 1997. Combined olfactory contact with the parent colony and direct contact with nonbreeding animals does not maintain suppression of ovulation in female naked mole-rats (Heterocephalus glaber). Horm Behav 31: 277–288.

Stoll EA, Karapavlovic N, Rosa H, Woodmass M, Rygiel K, White K, Turnbull DM, Faulkes CG. 2016. Naked mole-rats maintain healthy skeletal muscle and Complex IV mitochondrial enzyme function into old age. Aging (Albany NY).

Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. Revigo summarizes and visualizes long lists of gene ontology terms. PLoS One 6.

Swift-Gallant A, Mo K, Peragine DE, Monks DA, Holmes MM. 2015. Removal of reproductive suppression reveals latent sex differences in brain steroid hormone

receptors in naked mole-rats, Heterocephalus glaber. Biol Sex Differ 6: 31.

Tian X, Azpurua J, Hine C, Vaidya A, Myakishev-Rempel M, Ablaeva J, Mao Z, Nevo E, Gorbunova V, Seluanov A. 2013. High-molecular-mass hyaluronan mediates the cancer resistance of the naked mole rat. Nature 499: 346–349.

Trainor BC, Hofmann HA. 2006. Somatostatin Regulates aggressive behavior in an african cichlid fish. Endocrinology 147: 5119–5125.

Turner N, Grose R. 2010. Fibroblast growth factor signalling: from development to cancer. Nat Rev Cancer 10: 116–129.

van der Horst G, Maree L, Kotzé SH, O'Riain MJ. 2011. Sperm structure and motility in the eusocial naked mole-rat, Heterocephalus glaber: a case of degenerative orthogenesis in the absence of sperm competition? BMC Evol Biol 11: 351.

White S a, Nguyen T, Fernald RD. 2002. Social regulation of gonadotropin-releasing hormone. J Exp Biol 205: 2567–2581.

Yu C, Li Y, Holmes A, Szafranski K, Faulkes CG, Coen CW, Buffenstein R, Platzer M, de Magalhães JP, Church GM. 2011. RNA sequencing reveals differential expression of mitochondrial and oxidation reduction genes in the long-lived naked mole-rat when compared to mice. PLoS One 6: e26729.

Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, et al. 2003. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol 4: R28.

Zoncu R, Efeyan A, Sabatini DM. 2011. mTOR: from growth signal integration to cancer, diabetes and ageing. Nat Rev Mol Cell Biol 12: 21–35.

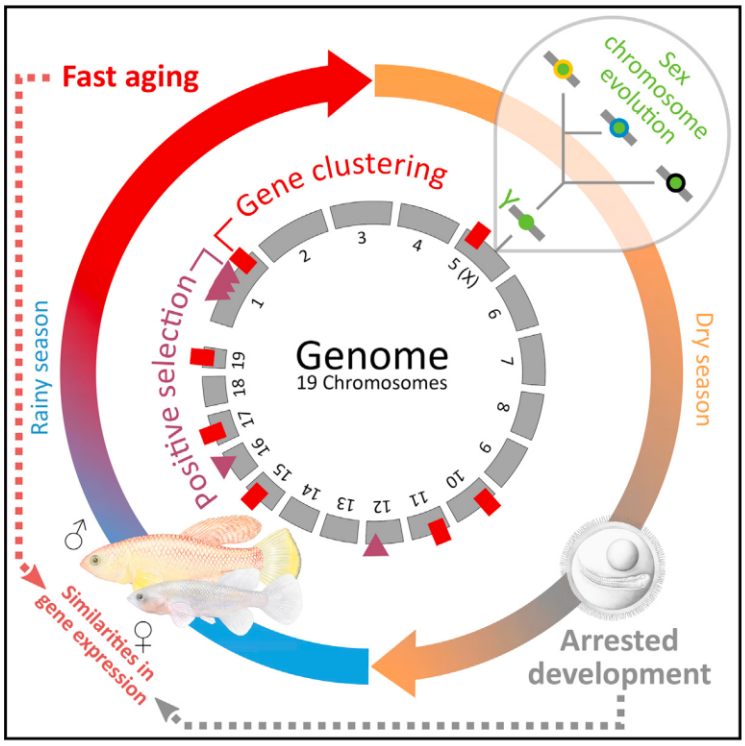
3.3 Manuscript 3 (M3)

Resource

Cell

Insights into Sex Chromosome Evolution and Aging from the Genome of a Short-Lived Fish

Graphical Abstract



Highlights

- The genome sequence of a very short-lived fish is a resource for aging research
- The sex chromosomes display features of early mammalian XY evolution
- Aging-related genes are clustered in specific genomic regions
- Transcriptional profiles show similarities between developmental arrest and aging

Authors

Kathrin Reichwald, Andreas Petzold, Philipp Koch, ..., Alessandro Cellerino, Christoph Englert, Matthias Platzer

Correspondence

matthias.platzer@leibniz-fli.de

In Brief

The turquoise killifish has a lifespan of only 4–12 months and yet its aging shares many similarities with that of humans. We sequenced and analyzed the killifish genome and provide insights into its biology. We detected very early stages of sex chromosome evolution, identified the sex-determining master gene, found clustering of aging-related genes in the genome, identified genes under positive selection, and discovered that similar gene sets are regulated during developmental arrest of embryos and aging.

Accession Numbers

KG817100
KG959958

Reichwald et al., 2015, Cell *163*, 1527–1538
December 3, 2015 ©2015 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2015.10.071

CellPress

Resource

Cell

Insights into Sex Chromosome Evolution and Aging from the Genome of a Short-Lived Fish

Kathrin Reichwald,[1,14] Andreas Petzold,[1,14,16] Philipp Koch,[1,14] Bryan R. Downie,[1,14] Nils Hartmann,[1,14] Stefan Pietsch,[1] Mario Baumgart,[1] Domitille Chalopin,[2,17] Marius Felder,[1] Martin Bens,[1] Arne Sahm,[1] Karol Szafranski,[1] Stefan Taudien,[1] Marco Groth,[1] Ivan Arisi,[3] Anja Weise,[4] Samarth S. Bhatt,[4] Virag Sharma,[5,6] Johann M. Kraus,[7] Florian Schmid,[7,8] Steffen Priebe,[9] Thomas Liehr,[4] Matthias Görlach,[1] Manuel E. Than,[1] Michael Hiller,[5,6] Hans A. Kestler,[1,7,10] Jean-Nicolas Volff,[2] Manfred Schartl,[11,12] Alessandro Cellerino,[1,13,15] Christoph Englert,[1,10,15] and Matthias Platzer[1,15,*]

[1]Leibniz Institute on Aging-Fritz Lipmann Institute (FLI), Jena 07745, Germany
[2]Institut de Génomique Fonctionnelle de Lyon, Ecole Normale Supérieure de Lyon, CNRS UMR5242, Université Claude Bernard Lyon 1, 69364 Lyon Cedex, France
[3]Genomics Facility, European Brain Research Institute (EBRI) Rita Levi-Montalcini, Rome 00143, Italy
[4]Jena University Hospital, Institute of Human Genetics, Friedrich Schiller University, Jena 07743, Germany
[5]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden 01307, Germany
[6]Max Planck Institute for the Physics of Complex Systems, Dresden 01307, Germany
[7]Medical Systems Biology, Ulm University, Ulm 89069, Germany
[8]International Graduate School in Molecular Medicine at Ulm University (GSC270), Ulm 89069, Germany
[9]Leibniz Institute for Natural Product Research and Infection Biology-Hans-Knoell-Institute (HKI), Jena 07745, Germany
[10]Faculty of Biology and Pharmacy, Friedrich Schiller University Jena, Jena 07743, Germany
[11]Department of Physiological Chemistry, Biocenter, University of Wuerzburg, Wuerzburg 97074, Germany
[12]Comprehensive Cancer Center Mainfranken, University Hospital Wuerzburg, Wuerzburg 97074, Germany
[13]Laboratory of Biology, Scuola Normale Superiore, Pisa 56126, Italy
[14]Co-first author
[15]Co-senior author
[16]Present address: Deep Sequencing Group SFB 655, Biotechnology Center, Dresden University of Technology, Dresden 01307, Germany
[17]Present address: Department of Genetics, University of Georgia, Athens, GA 30602, USA
*Correspondence: matthias.platzer@leibniz-fli.de
http://dx.doi.org/10.1016/j.cell.2015.10.071

SUMMARY

The killifish *Nothobranchius furzeri* is the shortest-lived vertebrate that can be bred in the laboratory. Its rapid growth, early sexual maturation, fast aging, and arrested embryonic development (diapause) make it an attractive model organism in biomedical research. Here, we report a draft sequence of its genome that allowed us to uncover an intra-species Y chromosome polymorphism representing—in real time—different stages of sex chromosome formation that display features of early mammalian XY evolution "in action." Our data suggest that *gdf6Y*, encoding a TGF-β family growth factor, is the master sex-determining gene in *N. furzeri*. Moreover, we observed genomic clustering of aging-related genes, identified genes under positive selection, and revealed significant similarities of gene expression profiles between diapause and aging, particularly for genes controlling cell cycle and translation. The annotated genome sequence is provided as an online resource (http://www.nothobranchius.info/NFINgb).

INTRODUCTION

The turquoise killifish *Nothobranchius furzeri* (Jubb, 1971) is an annual fish that inhabits seasonal freshwater ponds in the southeast of Africa. It is characterized by rapid growth, early sexual maturation, and an exceptionally short lifespan reflecting the adaptation to the ephemeral nature of the habitat (Blažek et al., 2013; Cellerino et al., 2015; Genade et al., 2005). Several laboratory strains exist differing in their origin and lifespan. The GRZ strain comes from a semi-arid habitat in Zimbabwe (Figures 1A, 1C, and 2A) (Jubb, 1971), where its founders were collected in 1969 and have a maximum lifespan of 4–6 months. To date, this is the shortest maximum lifespan reported for a vertebrate bred in captivity (Valdesalici and Cellerino, 2003). Strains from semi-arid or more humid regions in Mozambique (e.g., MZM-0403 and MZM-0410) (Figure 1B) and the borderland between Mozambique and Zimbabwe (MZZW-0701) have a longer maximum lifespan of ~1 year (Terzibasi et al., 2008; Tozzini et al., 2013). These strains were established only several years ago and are genetically heterogeneous, whereas GRZ is highly inbred (Reichwald et al., 2009). In spite of the short lifespan, both GRZ and MZM strains show typical signs of aging, i.e., a decline in cognitive and behavioral capacity accompanied by aging-related histological changes (Di Cicco et al., 2011; Terzibasi et al., 2007) as well as aging-related telomere shortening and impairment of mitochondrial function (Hartmann et al.,

Cell

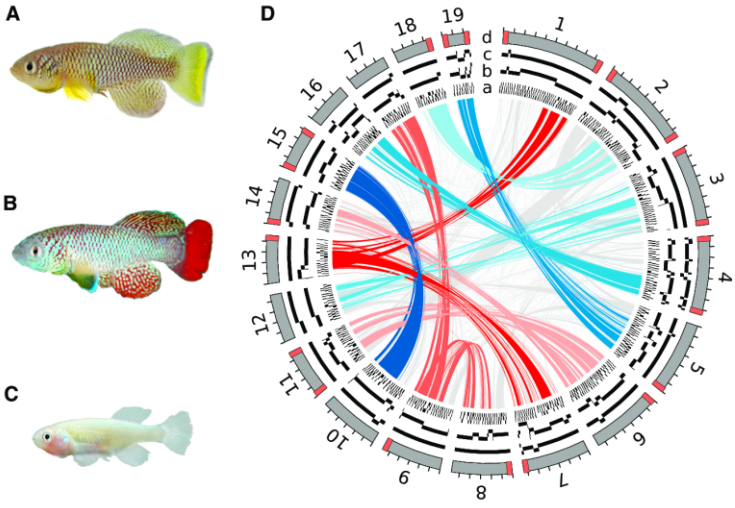


Figure 1. The Turquoise Killifish and the Genome Assembly

(A) Adult GRZ male.

(B) Adult MZM-0403 male.

(C) Adult GRZ female.

(D) Circles: the stepwise assembly of the reference sequence is represented from the inner to the outer circle. a: scaffolds obtained by applying programs ALLPATHS-LG and KILAPE. b: super-scaffolds built upon integration of optical mapping data. c: genetic scaffolds generated by linkage map integration. d: synteny groups defined upon analyses of synteny in medaka and stickleback. Synteny groups are sorted by length and numbered accordingly. Chromosome ends identified by optical mapping are marked in orange. The distance between two ticks is 10 Mb. Center: pairs of paralogous genes for synteny groups with a 1:1 (1:2) relation are connected by blue (red) lines; different hues define different chromosomal pairs (trios). Grey lines indicate gene pairs that do not follow our classification of chromosomal paralogy.

See also Figures S1 and S2 and Data S1.

2009, 2011). Lifespan determination in *N. furzeri* is polygenic; four quantitative loci relevant for lifespan are presently known (Kirschner et al., 2012).

Due to its fast development, *N. furzeri* can reach sexual maturity in <3 weeks and first signs of sexual dimorphism are apparent at 2 weeks after hatching (Blažek et al., 2013). In vertebrates, the gonads are usually the last organ system to develop into the functional adult structure. In fish, gonad differentiation commences only at late larval stages or even after metamorphosis and full functionality is reached at puberty (Devlin and Nagahama, 2002). Also, the sex determination system that provides the decision whether the undifferentiated gonad anlage of the embryo will develop later into a testis or an ovary is very plastic and can differ between closely related fish species or even within species (Volff et al., 2007). The fact that sex determination systems can change easily or arise rapidly during fish evolution together with the necessary rapid development of the reproductive system observed in *N. furzeri*, raises the question whether fast lifecycle and short lifespan influenced the evolution of the primary sex-determining (SD) gene and the sex chromosomes. Thus far, the segregation analyses of four sex-linked markers in crosses of GRZ and MZM-0403 is concordant with an XY sex determination system (Kirschner et al., 2012; Valenzano et al., 2009). The identical morphology (homomorphy) of the putative sex chromosomes (Reichwald et al., 2009) pointed to their young age and possibly to a situation of "sex chromosome evolution in action."

To survive the dry season, embryos of *N. furzeri* are protected from dehydration by a desiccation-resistant chorion and can enter into a state of developmental arrest termed diapause; the latter being a well-known adaptation in animal species to overcome unfavorable conditions. In *N. furzeri*, the arrest may occur at three distinct developmental stages (diapause I, II, and III) and can last for more than a year. Also in the nematode *Caenorhabditis elegans*, a larval arrest is observed (dauer larvae), and genes relevant for entering and maintaining the dauer state affect lifespan (Kenyon et al., 1993). We therefore analyzed whether differentially expressed genes (DEGs) in *N. furzeri* diapause versus non-diapause embryos are regulated in aging.

Recently, protocols for transgenesis (Hartmann and Englert, 2012; Valenzano et al., 2011) and CRISPR/Cas9-mediated mutagenesis have been established for *N. furzeri* (Harel et al., 2015). These tools, together with the short lifespan, make *N. furzeri* a very attractive vertebrate model to study aging, developmental arrest, and the interrelationship between both phenotypes. Here, we report a high-quality draft sequence of the *N. furzeri* genome. We provide insights into the very early evolutionary stages of an XY sex determination system, reveal clustering of aging-related genes in specific genomic regions, identify genes under positive selection and detect common expression profiles in diapause and aging.

RESULTS AND DISCUSSION

Assembly and Annotation of a High-Quality Draft Genome Sequence with Long-Range Contiguity

Today's challenge in genome analysis is generating a reference sequence of high quality and long-range contiguity. The *N. furzeri* project required special efforts because the genome is large and repeat-rich (Reichwald et al., 2009). In these two aspects, it resembles the zebrafish genome, for which a high-quality reference sequence was published only recently (Howe et al., 2013). We sequenced genomic DNA from *N. furzeri* females of the highly inbred GRZ strain (Figures 1A and 1C) in which all autosomes and the X chromosome are nearly homozygous. Using Illumina and Roche next-generation sequencing (NGS) technologies, we obtained whole-genome shotgun (WGS) data from 17 paired-end and mate-pair libraries amounting to 236 Gb (158-fold coverage, based on a genome-size estimate of 1.5 Gb; Figure S1B; Data S1A and S1B). Further, we sequenced

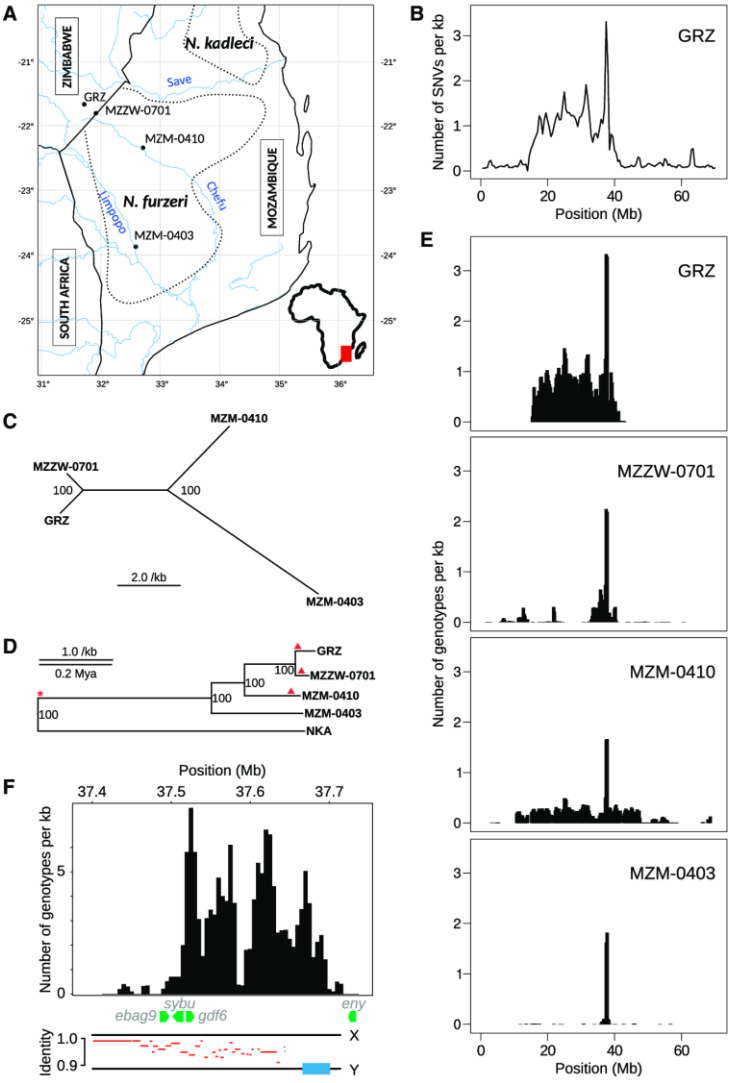


Figure 2. Phylogeny and Sex Chromosome Analyses of *N. furzeri* Strains GRZ, MZZW-0701, MZM-0410, and MZM-0403

(A) Geographic origin of *N. furzeri* strains is indicated by dots. The distribution range of *N. furzeri* and *N. kadleci* is marked by dotted lines.

(B) SNV density profile for synteny group (sgr) 05 obtained by aligning GRZ male WGS reads to the female reference sequence (sliding window: 1 Mb, step size: 500 kb).

(C) Phylogenetic relationship between strains based on WGS variation data.

(D) Phylogenetic tree of *N. furzeri* strains rooted by their sister species *N. kadleci* (NKA) based on exonic variations obtained by RNA-seq (Data S4I). The divergence time of *N. furzeri* and *kadleci* was estimated as 0.75 Mya (Dorn et al., 2014) and used for scaling. Red marks indicate the primary (asterisk) and secondary (triangle) events leading to the suppression of recombination shown in (E) and (F).

(E) Genotype density profiles for sgr05 of the strains. Genotype data were filtered for SNV positions in a given strain where all females are homozygous and all males are heterozygous (sliding window: 500 kb, step size: 250 kb).

(F) Top: zoom into the genotype density profile of the sex-determining region (SDR) in MZM-0403 (sliding window: 10 kb, step size: 5 kb). Genes annotated in the SDR of the GRZ female reference sequence are shown as green arrows. Bottom: identity plot (red lines) of BAC-based X and Y chromosome-specific sequences (black lines). The blue box represents a Y-specific 35 kb tandem repeat cluster composed of repeat units of 634 nt and 150 nt.

See also Figure S3 and Data S1I and S2A–S2C.

genomic insert ends of 81,393 BACs and fosmids to assist in the assembly and to provide a physical resource of the *N. furzeri* genome (5.3-fold clone coverage; Data S1C and S1D). To build the assembly, a five-step strategy was applied: we started with ALLPATHS-LG (Gnerre et al., 2011), continued with scaffolding, integrated optical and three genetic linkage maps, and finished with comparative synteny mapping in two closely related fish species (Table 1). The incorporation of optical mapping data remarkably improved the assembly contiguity (30-fold, Figure S1C). The genome assembly, in the following referred to as reference sequence, comprises 1.24 Gb (scaffold N50: ~0.5 Mb, optical N50: ~16 Mb, synteny N50: ~57 Mb), of which 87% were assigned to 19 synteny groups (sgrs) (Figure 1D). For 15 sgrs, we identified the corresponding *N. furzeri* chromosomes by fluorescence in situ hybridization using BAC probes (Figure S2A; Data S1E). Further, optical mapping data indicate that the assembly reached 22 of 38 chromosome ends (Figures 1D and S1D).

We built a comprehensive catalog of repetitive elements using Sanger/Illumina WGS reads and the genome assembly. Based on the Sanger data, we determined a repeat content of 64.6%, comprising 42.1% dispersed and 22.5% tandem repeats. This was confirmed by non-assembled NGS WGS data (Figure S2B). The *N. furzeri* reference sequence, however, contains only 35% repeats. In particular, tandem repeats are underrepresented (2% instead of 22.5%). This is most likely caused by the short NGS reads that collapse during the assembly process. Dispersed repeats amount to 33%, with LINEs being most abundant (8.4%) attributable to a recent expansion in this class of retrotransposons (Figure S2C; Data S1F). Finally, we confirmed the high quality of the reference sequence by PacBio WGS- and BAC sequencing (Data S1H and S1I) showing that gaps

Cell

Table 1. Statistics of the Stepwise Assembly

Assembly Step	Number of Scaffolds	Total Length (bp)	Fraction of N[a] (%)	Longest Assembly Unit (bp)	N50 (bp)
A ALLPATHS-LG	15,930	900,823,930	9.9	1,451,049	132,538
B Scaffolding + gap filling	7,675	943,595,854	9.2	3,869,209	494,454
C Optical map integration	6,012	1,230,898,532	30.4	44,272,285	15,858,201
D Genetic map integration	5,924	1,239,698,532	30.9	96,068,516	48,234,189
E Synteny integration	5,896	1,242,498,532	31.0	98,476,147	57,367,160
Anchoring within the Final Assembly					
Chromosomes/synteny groups	19	1,078,719,814	33.64	98,476,147	63,666,967
Autosomes	18	1,008,464,687	33.42	98,476,147	57,680,405
X chromosome	1	70,255,127	36.78	70,255,127	70,255,127
Unassigned	5,877	163,778,718	13.94	1,706,182	81,864

See also Data S1.

[a]Unresolved nucleotide positions, stands for A, C, G, or T.

contained in the genome assembly are almost entirely composed of repeats (83.1%).

We performed gene annotation using comprehensive RNA sequencing (RNA-seq) and microRNA sequencing (miRNA-seq) datasets as well as protein homology and in silico prediction tools (Figure S2D; Data S1K–S1N). We annotated 26,141 protein-coding genes with 59,154 transcripts, and 59 rRNA, 453 tRNA, 184 small nucleolar RNA (snoRNA), 598 miRNA, and 117 other non-protein coding RNA (ncRNA) genes (a detailed description of the miRNome will be reported elsewhere; M. Baumgart, I.A., and A.P., unpublished data). The teleost genome duplication (TGD) is reflected by the presence of 2,229 paralogous gene pairs, representing 17% of the *N. furzeri* protein-coding genes; further, we identified five pairs of putatively paralogous chromosomes with a 1:1 and three triads with a 1:2 relationship (Figure 1D; Data S1O).

To assess the completeness of the reference sequence with respect to the non-repetitive fraction of the genome, we used the Core Eukaryotic Genes Mapping Approach (CEGMA) (Parra et al., 2007) and searched in *N. furzeri* for orthologs of 248 highly conserved genes present in most eukaryotic genomes. Of these, we detected 98% in the reference sequence with 95% being completely covered (Data S1P). Furthermore, we could align 91% of the *N. furzeri* transcript catalog (Petzold et al., 2013) with the reference sequence strongly suggesting a highly complete representation of the genic fraction of the genome. Moreover, the PacBio-based estimate of the repeat content in gaps confirms that ~90% of the non-repetitive genome fraction is represented in the assembly. The annotated genome reference sequence is accessible at the *N. furzeri* Information Network Genome Browser (NFINgb, http://www.nothobranchius.info/NFINgb). Its long-range contiguity, chromosomal scale assembly, and completeness of genic regions allow studying the biology of the *N. furzeri* genome.

Insights into Early Events of XY Sex Chromosome Evolution

To map the SD region (SDR) in the reference sequence, which represents a GRZ female genome, we performed additional WGS sequencing of four GRZ males (Data S2B). Because the GRZ strain is highly inbred, we expected genomic variations predominantly in the region of suppressed recombination between male and female sex chromosomes. Accordingly, male single nucleotide variations (SNVs) were mainly confined to a region on sgr05 (Figures 2B and S3A) that bears the only four sex-linked markers identified so far (Kirschner et al., 2012; Valenzano et al., 2009). This male-specific region of the Y chromosome (MSY) encompasses 26.1 Mb (sgr05: 15,031,832–41,162,746) and exhibits a distinct peak in variation density at position 37.6 Mb. PCR/Sanger sequencing-based validation of sex-linkage for selected SNVs pointed to an intra-species sex chromosome polymorphism between *N. furzeri* strains. For example, variations in the syntabulin gene (*sybu*) are associated with sex in GRZ, MZZW-0701, and MZM-0410 but not in MZM-0403, whereas SNVs up to 42 kb upstream of *sybu* show sex-linkage in all strains (Data S2A).

By analyzing the intra-species variations by additional WGS data from males and females of MZZW-0701, MZM-0410, and MZM-0403 in more detail (Data S2B), we identified ~3.3 million SNVs (accessible at NFINgb). Using those SNVs to determine the phylogenetic relationship between strains, we found a good agreement with the geographic location of collection sites (Figures 2A and 2C). Rooting of the phylogenetic tree revealed that MZM-0403 belongs to a different lineage than the three other strains (Figure 2D), thus confirming the deep geographic structuring of the species (Bartáková et al., 2013; Dorn et al., 2011). We next searched genome-wide for signs of suppressed recombination and identified the most prominent region in all strains on sgr05 (Figures 2E and S3A). In GRZ, the SNV and genotype density profiles coincide (Figures 2B and 2E) suggesting that the same genetic signal of suppressed sex chromosomal recombination was detected with both approaches. While the size of the MSY differs considerably between strains, ranging from 196 kb to 37 Mb (Data S2C), the position of the variation peak is identical. To date, intra-species sex chromosome polymorphisms have been observed only in exceptional cases and only by using cytogenetic methods, e.g., in guppy (Nanda et al., 2014).

Cell

Comparative variation analyses of this remarkable strain-specific Y chromosome polymorphism indicate a two-step scenario for its evolution. First, an ancient event in the common ancestor of all strains led to suppressed recombination in a 196-kb region and the emergence and/or fixation of a SD signal. This stage of early sex-chromosome evolution is conserved in MZM-0403 (Figure 2F, top). In all four strains, the highest number of sex chromosome-specific SNVs was accumulated in the 196-kb region, indicating that recombination suppression shielded in the ancestral state the newly evolving SD gene from cross-over and the proto-Y from losing its identity. To shed light on the mechanism of recombination suppression, we sequenced X- and Y-specific BACs harboring this region using PacBio technology (Data S1I). The BAC-based X-specific assemblies confirmed the reference sequence. In addition, we obtained a corresponding Y-specific region encompassing a 35 kb tandem-repeat cluster (Figure 2F, bottom) that similarly to the MSY of the medaka fish (Kondo et al., 2006) may prevent recombination in flanking regions.

Secondary events encompassing larger regions (7–37 Mb), yet containing the primary SDR, occurred independently in each of the three northern strains. By applying FISH analysis, we identified an inversion as the secondary cross-over barrier in MZM-0410 (Figure S3B). Thus, the individually structured *N. furzeri* Y chromosomes seem to reflect the first stages of the mammalian XY evolution that has shaped these chromosomes by consecutive inversions into evolutionary strata over 320 million years (Lahn and Page, 1999). Also, the sex chromosomes of the flatfish *Cynoglossus semilaevis* estimated to be ~30 million years old, have most likely diverged due to suppression recombination by a large inversion (Chen et al., 2014). For *N. furzeri* we estimate the occurrence of the secondary recombination suppression in GRZ around 70 thousand years ago (kya), in MZZW-0701 50 kya and in MZM-0410 38 kya by dating the primary event to the species split between *N. furzeri* and *N. kadleci* at 750 kya (Dorn et al., 2014) (Figure 2D; Data S2D). Although this is a rough estimate, we conclude that the secondary events are very young in evolutionary terms compared to previously studied SD systems.

Our data demonstrate that during early sex chromosome evolution, a whole set of different Ys can be created. In-depth analyses of Y polymorphisms in species with older Y chromosomes will allow studying whether in a second phase the most successful Y might make a sweep through the species. Such a sweep would then lead to a situation noticed for mammalian Ys where only minor sequence variations mark the Y haplotypes in a later phase of Y chromosome evolution (Ellegren, 2003). Future studies will clarify whether population-genetic fragmentation (Bartáková et al., 2013), short lifespan, annualism, and/or the multiple specific adaptations of *N. furzeri* facilitated its unprecedented Y chromosome polymorphism.

Tracing the Emergence of a Novel Sex-Determining Gene: *gdf6Y*

We next attempted to identify the SD gene in *N. furzeri*. The minimal MSY was observed in MZM-0403 encompassing 196 kb and coinciding with the peak of Y-specific sequence variation at position 37.6 Mb in sgr05 (Figures 2E and 2F). This region contains only one annotated gene, *gdf6*, encoding growth differentiation factor 6, a member of the TGF-β family. We propose *gdf6Y* as symbol for the gene in the MSY. In GRZ, the *gdf6Y* coding sequence (CDS) differs from *gdf6* on the X chromosome in 22 SNVs and a 9-bp deletion, resulting in 15 amino acid (aa) exchanges and a 3 aa deletion (Figure S4A). All non-synonymous SNVs and the deletion are conserved between strains (Data S3A and S3B). Remarkably, the part of *gdf6Y* coding for the C-terminal 120 aa and homologous to the mature human GDF6, contains five non-synonymous but no synonymous substitutions indicating that positive selection acted on this part of the protein. The mature growth factor is highly conserved between vertebrates, and all male substitutions affect aa conserved between the *N. furzeri* X-chromosomal Gdf6 and its human ortholog (Figure 3A). Scanning all 339 genes in the 26.1 Mb MSY of GRZ confirms the sequence variations in *gdf6Y* as by far strongest signal of local positive selection (Data S3C).

To evaluate the impact of these aa changes, we performed homology modeling using the structure of the human receptor-bound GDF5 dimer (Kotzsch et al., 2009). Four of the five aa differing between mature Gdf6 and Gdf6Y reside in the modeled region (Figure 3A). Two of them (Gdf6Y/Gdf6: R405/Q408 and V407/E410) point outward into the solvent and reside at the edge of a β sheet (Figures 3B and S4B) that undergoes an induced fit upon formation of the GDF5:receptor complex (Kotzsch et al., 2009). The other two (T364/I367 and V372/M375) are located in a helix being part of the protomer interface but also contacting the receptor (Kotzsch et al., 2009). Hence, all four X/Y variable aa might have a bearing on protein interactions, either during dimerization or in the process of forming complexes with receptor(s).

Comparative analyses of *gdf6*/*gdf6Y* transcript levels revealed biallelic expression in early developmental stages of male and female GRZ and a significantly higher overall expression in males starting at day 3 post-hatching (Figures S4C and S4D; Data S3D). In RNA-seq data of adult ovaries, we found few *gdf6* reads, whereas in testes only *gdf6Y* mRNAs were detected at a considerable level. A possible explanation for the male-specific expression from the Y-chromosomal locus is a *gdf6Y*-specific deletion of 241 bp (sgr05: 37,526,406–37,526,646) in the 3′UTR including a potential mir-430 binding site (Figures S4E–S4G). In fish, mir-430 is an important regulator of germline-specific gene expression (Mishima et al., 2006). It is tempting to speculate that this deletion was the primary event marking the inception of the XY differentiation.

Gdf6Y expression peaks shortly after hatching; this is a time period when sex determination occurs in many fish species. Gdf6 is a member of the TGF-β family known to play a predominant role in developmental processes. Other members of the TGF-β family, e.g., the anti-Mullerian hormone (AMH) and the gonadal soma-derived growth factor (GSDF), as well as their receptors are important factors in sexual development of mammals and other vertebrates and function as master male sex determinants in several fish species (Josso and Clemente, 2003; Kikuchi and Hamaguchi, 2013; Morrish and Sinclair, 2002; Myosho et al., 2012; Rondeau et al., 2013). Gdf9 and Bmp15 are important players in ovarian development of mammals (Otsuka et al., 2011) and fish (Clelland and Kelly, 2011). Gdf6 has not

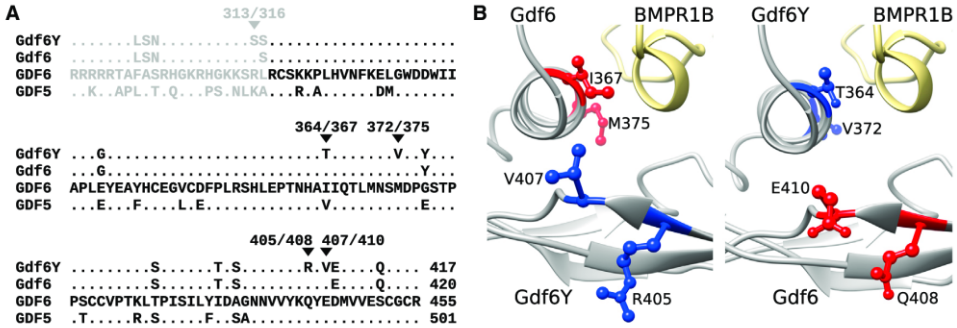


Figure 3. Gdf6Y/Gdf6 Homology Modeling

(A) ClustalW alignment of C-terminal, highly conserved 125 aa of *N. furzeri* Gdf6Y and Gdf6 as well as human GDF6 and GDF5. Amino acids (aa) identical to GDF6 are shown as dots. Amino acids varying between Gdf6Y and Gdf6 are highlighted by filled triangles and their numbers. The first 22 aa depicted in gray were not included in the modeling because they are missing in the reference structure.

(B) Detailed ribbon representations of two regions (left, right) of the modeled Gdf6Y/Gdf6 hetero-dimer (gray) receptor (yellow) complex given in Figure S4B. The four Gdf6Y/Gdf6 variable aa covered by the model are shown with side chains in blue for Gdf6Y and red for Gdf6. In the dimer, these aa are located spatially close to each other in the two regions shown.

See also Figure S4 and Data S3.

been described in the context of gonad development so far; how it acts as a master sex regulator in *N. furzeri* warrants further investigation.

Genomic Positional Enrichment of Aging-Related Genes

Recently, data have accumulated suggesting that eukaryotic genes located in physical proximity may be co-regulated and/or have similar functions. Correlations between chromosomal position and membership of functional gene sets were identified for yeast (Santoni et al., 2013) and human (Thévenin et al., 2014) genomes. Hence, chromosomal and spatial co-localization in the nucleus may indicate co-regulation. It was previously shown that 3D chromatin structure couples nuclear compartmentalization of chromatin domains with the control of gene activity (Guelen et al., 2008) and thus contributes to cell-specific gene expression (Zullo et al., 2012). In this context, it is noteworthy that cellular senescence is associated with modifications of the global chromatin interaction network (Chandra et al., 2015). To our knowledge, it has not yet been investigated whether genes relevant for organismal aging are clustered in genomic regions.

Taking advantage of the long-range contiguity of the *N. furzeri* reference sequence, we set out to study whether aging-related genes show positional gene enrichment (PGE) in sgrs. To this end, we identified aging-related DEGs in three tissues (brain, liver, and skin) by applying two different approaches: (1) we compared young versus old MZM-0410 (5 weeks versus 39 weeks, corresponding to 10% versus 75% of maximum lifespan), and (2) we compared GRZ versus MZM-0410 at 12 weeks. As aging rates differ between these strains (Terzibasi et al., 2008), the same chronological age in the second approach corresponds to 50% of the maximum lifespan in GRZ and 24% in MZM-0410 (Data S4A–S4G).

In total, we detected ten PGE regions. Four of those are based on DEGs obtained by the first approach and six were identified by the second approach (false discovery rate [FDR] < 0.05, scan statistics; Data S4H). These regions are located on seven sgrs, extend over 2.6–9.2 Mb, and contain 11–23 DEGs. On three sgrs, two PGE regions each overlap non-randomly ($p = 0.0012$, resampling test) indicating that the same genomic features were detected by different approaches and in samples from different organs. One of the latter PGE regions located on sgr10 and detected based on DEGs in skin aging (Figure 4A) is enriched for the GO term "response to wounding" (FDR < 0.05, Fisher's exact test). The genes are downregulated in aging (Figure 4B) thus suggesting their co-regulation and providing a link to the well-accepted aging-related phenotype of decreased regenerative capacity (Conboy et al., 2005). These findings demonstrate that *N. furzeri* genes related to aging are distributed non-randomly in the genome and that positional clustering may allow their co-regulation.

Positively Selected Genes in *N. furzeri*

The availability of high-quality genomic reference sequences facilitates the identification of genes under positive selection. To identify genes potentially relevant for adaptation of life-history traits we analyzed *N. furzeri* in comparison with *N. pienaari* because these sympatric species show convergent evolution of short lifespan (Tozzini et al., 2013). Therefore, we generated CDS data for *N. pienaari* and, additionally, for four longer-lived Nothobranchius species as well as the non-annual killifish *Aphyosemion striatum* as outgroup by RNA-seq of brain samples (Data S4I). The consensus tree based on multi-species CDS alignments matched well their reported phylogeny (Dorn et al., 2014) (Figure 4C). To avoid assembly errors, only de novo assembled *N. furzeri* transcripts that show 100% identity to the reference sequence (n = 23,108; corresponding to 11,748 genes) were analyzed. Accordingly, for *N. pienaari* we included transcripts showing at least 99% coverage and 98% identity to the *N. furzeri* reference sequence (n = 5,576; corresponding to 5,363 genes). We identified seven genes under

Cell

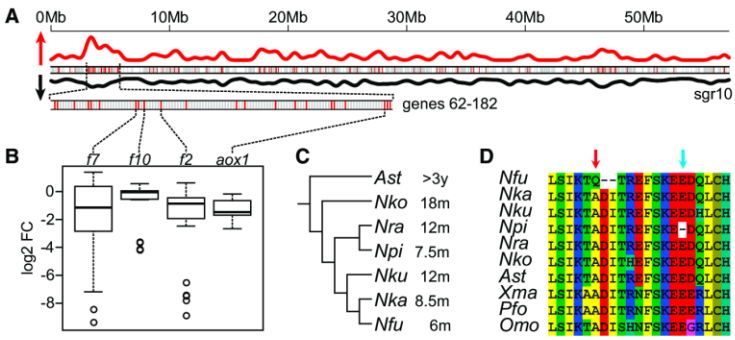


Figure 4. Positional Gene Enrichment and Positive Selection

(A) Schematic representation of synteny group (sgr) 10 and a region of positional gene enrichment. The genes in the sgr are represented by vertical bars: red, differentially expressed; gray, not differentially expressed. The density of all genes on the sgr (black line) and those differentially expressed (red line) is shown (kernel density estimation, Gaussian kernel). Arrows indicate the direction of increasing values.

(B) Relative downregulation of four DEGs with the GO annotation "response to wounding" in aging skin. Gene symbols *f2*, *f7*, and *f10* stand for coagulation factors II, VII, and X. *aox1*, aldehyde oxidase 1. Boxes, first and third quartiles; horizontal line, median; whiskers, most extreme value within 1.5× of inter-quartile range; dots, outliers. Expression differences were calculated by pairwise comparisons (n = 25) between the samples.

(C) Phylogram of the species used for transcriptome sequencing based on Dorn et al. (2014). For each species, the captive median lifespan is reported: *A. striatum* (unpublished), *N. korthause* (Baumgart et al., 2015), *N. rachovii*, *N. pienaari*, *N. kuhntae*, *N. furzeri* (Tozzini et al., 2013), and *N. kadleci* (Ng'oma et al., 2014).

(D) Alignment of the Id3 C terminus. The red arrow indicates aa under positive selection in *N. furzeri* followed by a two aa deletion. The blue arrow indicates the *N. pienaari*-specific deletion. The background color of each aa relates to the chemical nature of its side chain.

See also Data S4A–S4M.

positive selection in *N. furzeri* and one in *N. pienaari* (FDR < 0.05, Data S4J) highlighting the importance of a reference sequence for evolutionary analyses. Remarkably, five of these genes are either up- or downregulated in aging in at least one of three MZM-0410 organs (brain, liver, skin at 39 versus 5 weeks; Data S4A and S4K–S4M).

The signature of selection for *id3* (inhibitor of DNA binding 3, dominant negative helix-loop-helix protein) is particularly interesting. *Id3* is upregulated during aging in brain and skin and is also a key component of TGF-β signaling. TGF-β regulates inflammation, is involved in aging-related diseases such as tumorigenesis, fibrosis, glaucoma, and osteoarthritis (Krieglstein et al., 2012), and regulates life-history traits in *C. elegans* (Luo et al., 2010; Shaw et al., 2007). In *N. furzeri*, the gene shows signs of positive selection; i.e., a radical substitution of a non-polar by a charged aa followed by a 2-aa deletion (Figure 4D). Interestingly, at 10-aa distance in *N. pienaari* one evolutionarily conserved aa is deleted suggesting convergent evolution.

Another interesting gene under positive selection is *ikbip* (I Kappa B Kinase Interacting Protein), a pro-apoptotic gene (Hofer-Warbinek et al., 2004) downregulated in skin aging. Apoptosis is relevant for both diapause and aging. Diapausing killifish embryos are resistant to apoptosis (Meller and Podrabsky, 2013), but apoptosis is induced in aging *N. furzeri* (Di Cicco et al., 2011; Ng'oma et al., 2014). Apoptosis-related genes were shown to be age-regulated across tissues in a meta-analysis of mammalian aging (de Magalhães et al., 2009). Studies of larger taxonomical samples, including genomic and transcriptomic sequence datasets, are needed for further investigation of positive selection and convergent evolution in Nothobranchius species.

Overlap of Transcriptional Changes in Developmental Arrest and Aging

Last, we assessed the potential relation between developmental arrest (diapause) and aging in *N. furzeri*. Focusing on diapause II at the somite stage, we determined gene expression changes between arrested and non-arrested embryos at a comparable morphological stage using RNA-seq (Figure 5A). We identified 1,256 down- and 971 upregulated genes in arrested GRZ and MZM-0403 embryos (FDR < 0.05, DEseq and edgeR; Data S4O). In the set of downregulated genes, pattern specification processes including embryonic development of different organs and processes associated with cell proliferation were enriched (p < 0.05, hypergeometric test). Processes enriched in upregulated genes were more diverse and included translational elongation, ribosome biogenesis, metabolism and regulation of cellular component movement (Figure S5A). Decreased rates of cell proliferation and changes in the metabolic status have also been observed in diapause embryos of the South American killifish *Austrofundulus limnaeus* (Podrabsky and Culpepper, 2012). Upregulation of genes involved in translational and ribosomal processes, however, was unexpected. A possible explanation is the need for immediate cellular activity once environmental conditions trigger the exit from diapause.

We then analyzed whether there were similar gene expression changes in diapause and aging. To this end, we again employed MZM-0410 RNA-seq data (brain, liver, skin; 5/12/20/27/39 weeks; Data S4A) and focused on genes showing a monotonic increase or decrease of transcript levels in aging (Data S4P–S4R). In brain, the number of genes that were either down- or upregulated in both aging and diapause was significantly higher than the number of genes downregulated in brain aging and upregulated in diapause or vice versa (p < 0.001, chi-square test; Figure 5B). We therefore concentrated on the first two groups with highest DEG numbers and found that all significantly enriched processes in the group of downregulated genes were associated with cell-cycle progression and DNA replication (Figure 5C). Previous work suggests that brain aging in *N. furzeri* is associated with reduced mitotic activity of adult neuronal stem cells (Tozzini et al., 2012). Unexpectedly, the

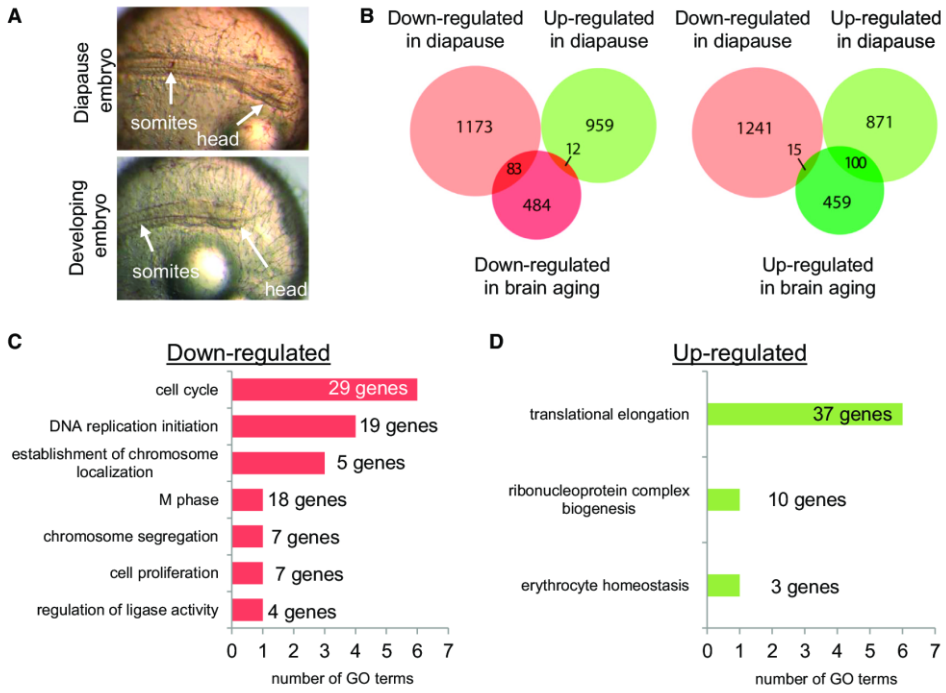


Figure 5. RNA-Seq Analyses of Diapause Embryos and Brain Aging

(A) The embryo (upper picture) has arrested in diapause II for 9 months, whereas the non-arrested embryo (lower picture) exhibiting a comparable morphological stage has an age of 6 days post fertilization.

(B) Venn-analyses of genes downregulated (light red) and upregulated (light green) in diapause as well as monotonic downregulated (dark red) and upregulated (dark green) in brain aging.

(C) Enrichment analyses of genes downregulated in diapause and brain aging. Numbers of involved genes and GO terms are shown for each biological process.

(D) Enrichment analyses of genes upregulated in diapause and brain aging.

See also Figure S5 and Data S4A and S4N–S4U.

two major processes enriched in upregulated genes in diapause and brain aging were translational elongation and ribonucleoprotein complex biogenesis (Figure 5D). The small number of overlapping DEGs between diapause and liver aging prevented further analysis (Figure S5B). Similar to brain, we identified in skin a significantly higher number of genes that were either up- or downregulated both in diapause and aging than genes regulated in opposite ways ($p < 0.001$, chi-square test, Figure S5B). Analysis of consistently downregulated genes showed enrichment of diverse processes. In the respective set of upregulated genes, however, again translational elongation and ribosome biogenesis were enriched (Figure S5C). Previously, aging-related upregulation of genes encoding translational and ribosomal proteins has been reported for human brain, muscle, and kidney suggesting a compensatory mechanism for aging-related increase in protein damage (Zahn et al., 2006). To our knowledge, a common expression profile for vertebrate developmental arrest and aging has not been described before.

In the nematode *C. elegans*, a link between developmental arrest, the so-called dauer larvae, and longevity has been identified. When mutated, some genes affecting dauer formation such as *daf-2* (a homolog of the insulin and IGF-1 receptor) increase lifespan (Kenyon et al., 1993; Lin et al., 1997; Ogg et al., 1997; Shaw et al., 2007). Moreover, the gene expression profile of dauer larvae shows similarities to the expression profile of long-lived adult mutants (McElwee et al., 2004). At present, the absence of long-lived mutants prevents such kinds of analysis in *N. furzeri*. Our comparison of gene expression changes between *N. furzeri* diapause embryos and *C. elegans* dauer larvae (Wang and Kim, 2003) revealed little overlap (Data S4V). This does not seem surprising, given the long evolutionary distance between the two species and their different habitats. The identification of e.g., *daf-16*/*FoxO4* being upregulated in embryonic arrest of both species, however, indicates commonalities between the two processes and calls for further analyses, e.g., genomic manipulation of the *FoxO4* locus in *N. furzeri*.

In conclusion, the high-quality draft sequence of the *N. furzeri* genome provided here and the availability of several *N. furzeri* strains that differ in lifespan represent excellent resources for studying and identifying genes involved in aging and longevity.

Cell

Furthermore, the novel genomic engineering tools now available in *N. furzeri* such as the CRISPR/Cas system (Harel et al., 2015) will allow the generation of mutant lines at a large scale providing a platform for drug screening and sophisticated models to study aging as well as aging-related and other diseases and to develop novel therapies.

EXPERIMENTAL PROCEDURES

Additional details are provided in the Supplemental Experimental Procedures.

Animal Material

Sample acquisition was carried out in accordance with the "principles of laboratory animal care" and the current version of the German Law on the Protection of Animals.

De Novo Genome Sequencing and Assembly

Two adult female GRZ were sequenced using Illumina technology and assembled with ALLPATHS-LG. In parallel, two adult male GRZ were sequenced using Roche technology; these data served for long-range scaffolding and gap filling. Further, optical mapping (OpGen; http://www.opgen.com) was performed in one adult female GRZ. By combining restriction maps obtained with this procedure and sequence scaffolds, superscaffolds were formed. These were manually ordered in genetic scaffolds based on own genetic maps (Kirschner et al., 2012; Ng'oma et al., 2014). Finally, by synteny analyses in medaka and stickleback, genetic scaffolds were arranged in sgrs.

Repeat Annotation

Repeats are identified by (1) RepeatModeler in the reference sequence, (2) RepeatMasker, RepeatScout (Price et al., 2005) for assembled Sanger sequences generated by whole-genome sample sequencing, and (3) RepARK (Koch et al., 2014) for WGS Illumina reads. Subsequently, libraries were merged in a *N. furzeri*-specific repeat library and finally used to annotate the reference sequence by RepeatMasker and TandemRepeatFinder (Benson, 1999).

Gene Annotation and Identification of Paralogs

Protein-coding genes were annotated based on (1) ab initio gene prediction, (2) protein sequence similarity, and (3) Illumina RNA-seq data. Results were combined into CDS models with EVM and UTRs, and transcripts were constructed with PASA (Haas et al., 2008). Gene symbols and functions were annotated using homologous proteins of medaka, platyfish, stickleback, tetraodon, and zebrafish obtained from Ensembl (Cunningham et al., 2015). InterProScan75 (Zdobnov and Apweiler, 2001) was used to identify protein domains and to retrieve Gene Ontology annotations.

MiRNA genes were identified from Illumina miRNA-seq data. To detect rRNA genes, BLAT searches using known *N. furzeri* rRNA sequences (Reichwald et al., 2009) as queries were performed. In addition, miRNA, tRNA, rRNA, and other non-protein-coding genes were identified using ab initio gene prediction tools.

TGD-derived paralogs were identified with Ensembl Compara. First, *N. furzeri* genes were used to find orthologs in medaka, platyfish, stickleback, tetraodon, and zebrafish. Next, Ensembl gene IDs served as queries in Ensembl Compara to detect pairwise paralogous relationships. Any pair of duplicated genes originating before the teleost split was discarded. Finally, *N. furzeri* genes related to the same orthologous gene were also included.

Genomic Resequencing of *N. furzeri* Strains and Variation Calling

Illumina WGS reads generated for all strains were mapped to the reference sequence with Bowtie2 (Langmead and Salzberg, 2012) (minimum mapping quality score of 11). Regions with alignment gaps were realigned with GATK (McKenna et al., 2010) and duplicate reads marked with Picard Tools (http://picard.sourceforge.net). Sequence variations and genotypes were called with GATK. Selected genomic regions were resequenced in additional specimens by PCR and Sanger technology as described (Reichwald et al., 2009).

Overrepresentation Analysis

Zebrafish orthologs of *N. furzeri* genes were retrieved using BLAST. Human orthologs were fetched with R package orthology. GO enrichment analysis was done using DAVID (Huang et al., 2009) and summarized by REViGO (Supek et al., 2011).

Positional Gene Enrichment

Aging-related DEGs were identified by Illumina RNA-seq. Scan statistics (Glaz et al., 2001) were used to test if an observed accumulation of k DEGs on a sgr containing N genes is likely to happen by chance. The scan statistic S is the maximal k in any interval W of fixed size w ($w = 0.1 \times N$). Subsequently, an overrepresentation analysis for each detected genomic region was performed.

Positive Selection

Protein-coding sequences of *N. kadleci*, *N. korthausae*, *N. kuhntae*, *N. pienaari*, *N. rachovii*, *A. striatum*, and *N. furzeri* were assembled de novo using Illumina RNA-seq data. Prank (Löytynoja and Goldman, 2008) alignments of orthologous CDS were filtered by Gblocks (Talavera and Castresana, 2007) and in-house software. Then, the improved branch-site test of positive selection was applied as described (Zhang et al., 2005). Ka/Ks ratios were calculated for all CDS pairs in the SDR both in total and in 333 nt windows sampled using a step size of 99 nt.

Gene Expression Analysis in Diapause Embryos

In total, 287 diapause and 239 non-diapause embryos were collected at the somite stage. Approximately 30 embryos per state were pooled resulting in eight diapause and eight non-diapause samples. Total RNA was extracted and sequenced by Illumina RNA-seq. Significant DEGs were identified and an overrepresentation analysis was performed.

ACCESSION NUMBERS

The accession number for the *N. furzeri* genome project including genome assembly and NGS data (WGS, RNA-seq, and BAC-seq) reported in this paper is BioProject: PRJEB5837. The accession numbers for the *N. furzeri* GRZ genomic insert end sequences of BACs and fosmids are GenBank: KG817100 to KG959958. The accession number for assembled Sanger WGS sequences is BioProject: PRJNA29535. Accession numbers of individual datasets are given in Data S1, S2, S3, and S4.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and four data sets and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.10.071.

AUTHOR CONTRIBUTIONS

C.E., K.R., and M.P. initiated, managed, and drove the genome project. K.R., N.H., M.Ba., S.T., and M.Gr. prepared the samples. K.R., S.T., and M.Gr. performed the sequencing. A.P., P.K., B.R.D., V.S., and M.H. performed the genome assembly and annotation. P.K., B.R.D., D.C., and J.N.V. performed the repeat analysis. M.Ba., M.Gr., A.C., and M.P. performed the mRNA analysis. I.A., M.Ba., A.P., and A.C. performed the miRNA analysis. K.R., A.W., S.S.B., and T.L. performed the chromosome FISH. K.R., A.P., P.K., M.F., K.S., N.H., M.S., C.E., and M.P. performed the sex chromosome evolution analysis. M.Go. and M.E.T. performed protein structure modeling. J.M.K., F.S., S.Pr., P.K., H.A.K., A.C., and M.P. performed the PGE analysis. A.S., M.Be., A.P., B.R.D., A.C., and M.P. performed the positive selection analysis. N.H., S.Pi., and C.E. performed the diapause analysis. All authors contributed to data interpretation. K.R., A.P., P.K., N.H., M.S., A.C., C.E., and M.P. wrote the manuscript.

ACKNOWLEDGMENTS

We thank Silke Foerste, Ivonne Goerlich, Ivonne Heinze, Christin Hahn, Cornelia Luge, Sabine Matz, Martin Neumann, and Bernd Senf for technical

assistance. We thank Karl Lenhard Rudolph for discussions and Cornelia Platzer for critical reading of the manuscript. This work was supported by the Leibniz Association (WGL: PAKT-2006-FLI to C.E. and M.P., and SAW-2012-FLI to M.P.), the German Research Foundation (DFG: RE 3505/1-1 to K.R., HA 6214/2-1 to N.H., and SFB 1074 project Z1 to H.A.K.), the German Federal Ministry of Education and Research (BMBF: JenAge 0315581A/C to A.C., C.E., and M.P.; 031A099 to M.H.; Gerontosys II, Forschungskern SySTaR, project ID 0315894A to H.A.K.), the European Community's Seventh Framework Program (FP7/2007-2013 under grant agreement 602783 to H.A.K.), and the Italian Ministry of Higher Education (FIRB: RBAP10L8TY to I.A.).

Received: June 3, 2015
Revised: August 11, 2015
Accepted: October 21, 2015
Published: December 3, 2015

REFERENCES

Bartáková, V., Reichard, M., Janko, K., Polačik, M., Blažek, 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 Evol. Biol. *13*, 196.

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*, 63–69.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. *27*, 573–580.

Blažek, R., Polačik, M., and Reichard, M. (2013). Rapid growth, early maturation and short generation time in African annual fishes. Evodevo *4*, 24.

Cellerino, A., Valenzano, D.R., and Reichard, M. (2015). From the bush to the bench: the annual Nothobranchius fishes as a new model system in biology. Biol. Rev. Camb. Philos. Soc. Published online April 28, 2015. http://dx.doi.org/10.1111/brv.12183.

Chandra, T., Ewels, P.A., Schoenfelder, S., Furlan-Magaril, M., Wingett, S.W., Kirschner, K., Thuret, J.Y., Andrews, S., Fraser, P., and Reik, W. (2015). Global reorganization of the nuclear landscape in senescent cells. Cell Rep. *10*, 471–483.

Chen, S., Zhang, G., Shao, C., Huang, Q., Liu, G., Zhang, P., Song, W., An, N., Chalopin, D., Volff, J.N., et al. (2014). Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. Nat. Genet. *46*, 253–260.

Clelland, E.S., and Kelly, S.P. (2011). Exogenous GDF9 but not Activin A, BMP15 or TGFβ alters tight junction protein transcript abundance in zebrafish ovarian follicles. Gen. Comp. Endocrinol. *171*, 211–217.

Conboy, I.M., Conboy, M.J., Wagers, A.J., Girma, E.R., Weissman, I.L., and Rando, T.A. (2005). Rejuvenation of aged progenitor cells by exposure to a young systemic environment. Nature *433*, 760–764.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2015). Ensembl 2015. Nucleic Acids Res. *43*, D662–D669.

de Magalhães, J.P., Curado, J., and Church, G.M. (2009). Meta-analysis of age-related gene expression profiles identifies common signatures of aging. Bioinformatics *25*, 875–881.

Devlin, R.H., and Nagahama, Y. (2002). Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences. Aquaculture *208*, 191–364.

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. Exp. Gerontol. *46*, 249–256.

Dorn, A., Ng'oma, E., Janko, K., Reichwald, K., Polačik, M., Platzer, M., Cellerino, A., and Reichard, M. (2011). Phylogeny, genetic variability and colour polymorphism of an emerging animal model: the short-lived annual Nothobranchius fishes from southern Mozambique. Mol. Phylogenet. Evol. *61*, 739–749.

Dorn, A., Musilová, 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 Evol. Biol. *14*, 210.

Ellegren, H. (2003). Levels of polymorphism on the sex-limited chromosome: a clue to Y from W? BioEssays *25*, 163–167.

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*, 223–233.

Glaz, J., Naus, J., and Wallenstein, S. (2001). Scan Statistics (New York: Springer).

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. USA *108*, 1513–1518.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W., and van Steensel, B. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature *453*, 948–951.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. *9*, R7.

Harel, I., Benayoun, B.A., Machado, B., Singh, P.P., Hu, C.K., Pech, M.F., Valenzano, D.R., Zhang, E., Sharp, S.C., Artandi, S.E., and Brunet, A. (2015). A platform for rapid exploration of aging and diseases in a naturally short-lived vertebrate. Cell *160*, 1013–1026.

Hartmann, N., and Englert, C. (2012). A microinjection protocol for the generation of transgenic killifish (Species: Nothobranchius furzeri). Dev. Dyn. *241*, 1133–1141.

Hartmann, N., Reichwald, K., Lechel, A., Graf, M., Kirschner, J., Dorn, A., Terzibasi, E., Wellner, J., Platzer, M., Rudolph, K.L., et al. (2009). Telomeres shorten while Tert expression increases during ageing of the short-lived fish Nothobranchius furzeri. Mech. Ageing Dev. *130*, 290–296.

Hartmann, N., Reichwald, K., Wittig, I., Dröse, S., Schmeisser, S., Lück, C., Hahn, C., Graf, M., Gausmann, U., Terzibasi, E., et al. (2011). Mitochondrial DNA copy number and function decrease with age in the short-lived fish Nothobranchius furzeri. Aging Cell *10*, 824–831.

Hofer-Warbinek, R., Schmid, J.A., Mayer, H., Winsauer, G., Orel, L., Mueller, B., Wiesner, Ch., Binder, B.R., and de Martin, R. (2004). A highly conserved proapoptotic gene, IKIP, located next to the APAF1 gene locus, is regulated by p53. Cell Death Differ. *11*, 1317–1325.

Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L., et al. (2013). The zebrafish reference genome sequence and its relationship to the human genome. Nature *496*, 498–503.

Huang, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Josso, N., and Clemente, Nd. (2003). Transduction pathway of anti-Müllerian hormone, a sex-specific member of the TGF-beta family. Trends Endocrinol. Metab. *14*, 91–97.

Jubb, R.A. (1971). A new Nothobranchius (Pisces, Cyprinodontidae) from Southeastern Rhodesia. J Am Killifish Association *8*, 12–19.

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

Kikuchi, K., and Hamaguchi, S. (2013). Novel sex-determining genes in fish and sex chromosome evolution. Dev. Dyn. *242*, 339–353.

Kirschner, J., Weber, D., Neuschl, C., Franke, A., Böttger, M., Zielke, L., Powalsky, E., Groth, M., Shagin, D., Petzold, A., et al. (2012). Mapping of quantitative trait loci controlling lifespan in the short-lived fish Nothobranchius furzeri–a new vertebrate model for age research. Aging Cell *11*, 252–261.

Cell

Koch, P., Platzer, M., and Downie, B.R. (2014). RepARK–de novo creation of repeat libraries from whole-genome NGS reads. Nucleic Acids Res. *42*, e80.

Kondo, M., Hornung, U., Nanda, I., Imai, S., Sasaki, T., Shimizu, A., Asakawa, S., Hori, H., Schmid, M., Shimizu, N., and Schartl, M. (2006). Genomic organization of the sex-determining and adjacent regions of the sex chromosomes of medaka. Genome Res. *16*, 815–826.

Kotzsch, A., Nickel, J., Seher, A., Sebald, W., and Müller, T.D. (2009). Crystal structure analysis reveals a spring-loaded latch as molecular mechanism for GDF-5-type I receptor specificity. EMBO J. *28*, 937–947.

Krieglstein, K., Miyazono, K., ten Dijke, P., and Unsicker, K. (2012). TGF-β in aging and disease. Cell Tissue Res. *347*, 5–9.

Lahn, B.T., and Page, D.C. (1999). Four evolutionary strata on the human X chromosome. Science *286*, 964–967.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lin, K., Dorman, J.B., Rodan, A., and Kenyon, C. (1997). daf-16: An HNF-3/forkhead family member that can function to double the life-span of Caenorhabditis elegans. Science *278*, 1319–1322.

Löytynoja, A., and Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science *320*, 1632–1635.

Luo, S., Kleemann, G.A., Ashraf, J.M., Shaw, W.M., and Murphy, C.T. (2010). TGF-β and insulin signaling regulate reproductive aging via oocyte and germline quality maintenance. Cell *143*, 299–312.

McElwee, J.J., Schuster, E., Blanc, E., Thomas, J.H., and Gems, D. (2004). Shared transcriptional signature in Caenorhabditis elegans Dauer larvae and long-lived daf-2 mutants implicates detoxification system in longevity assurance. J. Biol. Chem. *279*, 44533–44543.

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 Res. *20*, 1297–1303.

Meller, C.L., and Podrabsky, J.E. (2013). Avoidance of apoptosis in embryonic cells of the annual killifish Austrofundulus limnaeus exposed to anoxia. PLoS ONE *8*, e75837.

Mishima, Y., Giraldez, A.J., Takeda, Y., Fujiwara, T., Sakamoto, H., Schier, A.F., and Inoue, K. (2006). Differential regulation of germline mRNAs in soma and germ cells by zebrafish miR-430. Curr. Biol. *16*, 2135–2142.

Morrish, B.C., and Sinclair, A.H. (2002). Vertebrate sex determination: many means to an end. Reproduction *124*, 447–457.

Myosho, T., Otake, H., Masuyama, H., Matsuda, M., Kuroki, Y., Fujiyama, A., Naruse, K., Hamaguchi, S., and Sakaizumi, M. (2012). Tracing the emergence of a novel sex-determining gene in medaka, Oryzias luzonensis. Genetics *191*, 163–170.

Nanda, I., Schories, S., Tripathi, N., Dreyer, C., Haaf, T., Schmid, M., and Schartl, M. (2014). Sex chromosome polymorphism in guppies. Chromosoma *123*, 373–383.

Ng'oma, E., Reichwald, K., Dorn, A., Wittig, M., Balschun, T., Franke, A., Platzer, M., and Cellerino, A. (2014). The age related markers lipofuscin and apoptosis show different genetic architecture by QTL mapping in short-lived Nothobranchius fish. Aging (Albany, N.Y.) *6*, 468–480.

Ogg, S., Paradis, S., Gottlieb, S., Patterson, G.I., Lee, L., Tissenbaum, H.A., and Ruvkun, G. (1997). The Fork head transcription factor DAF-16 transduces insulin-like metabolic and longevity signals in C. elegans. Nature *389*, 994–999.

Otsuka, F., McTavish, K.J., and Shimasaki, S. (2011). Integral role of GDF-9 and BMP-15 in ovarian function. Mol. Reprod. Dev. *78*, 9–21.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics *23*, 1061–1067.

Petzold, A., Reichwald, K., Groth, M., Taudien, S., Hartmann, N., Priebe, S., Shagin, D., Englert, C., and Platzer, M. (2013). The transcript catalogue of the short-lived fish Nothobranchius furzeri provides insights into age-dependent changes of mRNA levels. BMC Genomics *14*, 185.

Podrabsky, J.E., and Culpepper, K.M. (2012). Cell cycle regulation during development and dormancy in embryos of the annual killifish Austrofundulus limnaeus. Cell Cycle *11*, 1697–1704.

Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. Bioinformatics *21* (*Suppl 1*), i351–i358.

Reichwald, K., Lauber, C., Nanda, I., Kirschner, J., Hartmann, N., Schories, S., Gausmann, U., Taudien, S., Schilhabel, M.B., Szafranski, K., et al. (2009). High tandem repeat content in the genome of the short-lived annual fish Nothobranchius furzeri: a new vertebrate model for aging research. Genome Biol. *10*, R16.

Rondeau, E.B., Messmer, A.M., Sanderson, D.S., Jantzen, S.G., von Schalburg, K.R., Minkley, D.R., Leong, J.S., Macdonald, G.M., Davidsen, A.E., Parker, W.A., et al. (2013). Genomics of sablefish (Anoplopoma fimbria): expressed genes, mitochondrial phylogeny, linkage map and identification of a putative sex gene. BMC Genomics *14*, 452.

Santoni, D., Castiglione, F., and Paci, P. (2013). Identifying correlations between chromosomal proximity of genes and distance of their products in protein-protein interaction networks of yeast. PLoS ONE *8*, e57707.

Shaw, W.M., Luo, S., Landis, J., Ashraf, J., and Murphy, C.T. (2007). The C. elegans TGF-beta Dauer pathway regulates longevity via insulin signaling. Curr. Biol. *17*, 1635–1645.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS ONE *6*, e21800.

Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. *56*, 564–577.

Terzibasi, E., Valenzano, D.R., and Cellerino, A. (2007). The short-lived fish Nothobranchius furzeri as a new model system for aging studies. Exp. Gerontol. *42*, 81–89.

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*, e3866.

Thévenin, A., Ein-Dor, L., Ozery-Flato, M., and Shamir, R. (2014). Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome. Nucleic Acids Res. *42*, 9854–9861.

Tozzini, E.T., Baumgart, M., Battistoni, G., and Cellerino, A. (2012). Adult neurogenesis in the short-lived teleost Nothobranchius furzeri: localization of neurogenic niches, molecular characterization and effects of aging. Aging Cell *11*, 241–251.

Tozzini, E.T., Dorn, A., Ng'oma, E., Polačik, M., Blažek, R., Reichwald, K., Petzold, A., Watters, B., Reichard, M., and Cellerino, A. (2013). Parallel evolution of senescence in annual fishes in response to extrinsic mortality. BMC Evol. Biol. *13*, 77.

Valdesalici, S., and Cellerino, A. (2003). Extremely short lifespan in the annual fish Nothobranchius furzeri. Proc. Biol. Sci. *270* (*Suppl 2*), S189–S191.

Valenzano, D.R., Kirschner, J., Kamber, R.A., Zhang, E., Weber, D., Cellerino, A., Englert, C., Platzer, M., Reichwald, K., and Brunet, A. (2009). Mapping loci associated with tail color and sex determination in the short-lived fish Nothobranchius furzeri. Genetics *183*, 1385–1395.

Valenzano, D.R., Sharp, S., and Brunet, A. (2011). Transposon-Mediated Transgenesis in the Short-Lived African Killifish Nothobranchius furzeri, a Vertebrate Model for Aging. G3 (Bethesda) *1*, 531–538.

Volff, J.N., Nanda, I., Schmid, M., and Schartl, M. (2007). Governing sex determination in fish: regulatory putsches and ephemeral dictators. Sex Dev. *1*, 85–99.

Wang, J., and Kim, S.K. (2003). Global analysis of dauer gene expression in Caenorhabditis elegans. Development *130*, 1621–1634.

Zahn, J.M., Sonu, R., Vogel, H., Crane, E., Mazan-Mamczarz, K., Rabkin, R., Davis, R.W., Becker, K.G., Owen, A.B., and Kim, S.K. (2006). Transcriptional profiling of aging in human muscle reveals a common aging signature. PLoS Genet. *2*, e115.

Zdobnov, E.M., and Apweiler, R. (2001). InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics *17*, 847–848.

Zhang, J., Nielsen, R., and Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. *22*, 2472–2479.

Zullo, J.M., Demarco, I.A., Piqué-Regi, R., Gaffney, D.J., Epstein, C.B., Spooner, C.J., Luperchio, T.R., Bernstein, B.E., Pritchard, J.K., Reddy, K.L., and Singh, H. (2012). DNA sequence-dependent compartmentalization and silencing of chromatin at the nuclear lamina. Cell *149*, 1474–1487.

3.4 Manuscript 4 (M4)

Aging Cell (2017) **16**, pp488–496 Doi: 10.1111/acel.12577

Parallel evolution of genes controlling mitonuclear balance in short-lived annual fishes

Arne Sahm,[1] Martin Bens,[1] Matthias Platzer[1] and Alessandro Cellerino[1,2]

[1]Leibniz Insitute on Ageing, Fritz-Lipmann Institute, Jena 07745, Germany
[2]Bio@SNS, Scuola Normale Superiore, Pisa 56124, Italy

Summary

The current molecular understanding of the aging process derives almost exclusively from the study of random or targeted single-gene mutations in highly inbred laboratory species, mostly invertebrates. Little information is available as to the genetic mechanisms responsible for natural lifespan variation and the evolution of lifespan, especially in vertebrates. Here, we investigated the pattern of positive selection in annual (i.e., short-lived) and nonannual (i.e., longer-lived) African killifishes to identify a genomic substrate for evolution of annual life history (and reduced lifespan). We identified genes under positive selection in all steps of mitochondrial biogenesis: mitochondrial (mt) DNA replication, transcription from mt promoters, processing and stabilization of mt RNAs, mt translation, assembly of respiratory chain complexes, and electron transport chain. Signs of paralleled evolution (i.e., evolution in more than one branch of *Nothobranchius* phylogeny) are observed in four out of five steps. Moreover, some genes under positive selection in *Nothobranchius* are under positive selection also in long-lived mammals such as bats and mole-rats. Complexes of the respiratory chain are formed in a coordinates multistep process where nuclearly and mitochondrially encoded components are assembled and inserted into the inner mitochondrial membrane. The coordination of this process is named mitonuclear balance, and experimental manipulations of mitonuclear balance can increase longevity of laboratory species. Our data strongly indicate that these genes are also casually linked to evolution lifespan in vertebrates.

Key words: evolution; gerontogenes; lifespan; longevity regulation; longevity gene; molecular biology of aging; mortality.

Introduction

The current molecular understanding of the aging process derives almost exclusively from the study of random or targeted single-gene mutations in highly inbred laboratory species, mostly invertebrates. Little information is available as to the genetic mechanisms responsible for natural lifespan variation and the evolution of longevity, especially in vertebrates. Yet, natural variability in lifespan across vertebrate species greatly exceeds the magnitude of life extension that has been obtained by single-gene manipulations, and a comparative approach may reveal novel genetic pathways that are responsible for evolution of lifespan.

Correspondence
Alessandro Cellerino, Leibniz Insitute on Ageing, Fritz-Lipmann Institute, Jena 07745, Germany. Tel.: +39-050-3152756; fax: +39-050-3152760; e-mail: alessandro.cellerino@sns.it

Accepted for publication *23 December 2016*

The increasing availability of sequenced genomes and transcriptomes of related species with differing lifespans can facilitate the identification of putative aging-related genes by analysis of positive selection. Positive selection is the evolutionary process by which a mutation becomes fixed in a population because it increases fitness. If two branches of an evolutionary tree differ in a key phenotype (lifespan, in this case), the genes under positive selection likely played a role in the evolution of that phenotype. In interspecies comparisons, positive selection on protein-coding sequences results in an increase in the rate of non-synonymous substitutions as compared with random genetic drift. Statistical models based on the ratio of non-synonymous to synonymous substitution rates (d_N/d_S) can identify specific amino acid codons within a given gene that evolved due to positive selection and are widely used in comparative genomics (Kosiol *et al.*, 2008; Roux *et al.*, 2014; Davies *et al.*, 2015).

One of the main limitations in applying this approach to the investigation of the genetic basis for lifespan evolution is the lack of a group of related species that are good laboratory organisms, are genetically tractable, and at the same time show naturally evolved large-scale differences in lifespan. Genome-wide scans for positive selection were performed in several long-lived mammals (bats, the naked mole-rat, the bowhead whale). However, it is not possible to establish a link between positively selected genes (PSGs) and evolution of longevity because the short-lived sister *taxon* (i.e., the most closely related species/clade) may not be available for analysis, making it impossible to exclude that of a codon change was selected before longevity evolved [for a discussion see (Sahm *et al.*, 2016a)] and it is very often impossible to relate a codon change to one of the several traits that distinguish two taxa (e.g., a PSG in *H. sapiens* may be related to longevity, bipedalism, absence of fur, speech, relative brain size, or any other trait that distinguish humans from apes).

Annual fishes of the genus *Nothobranchius* are small teleost fishes from East Africa adapted to the alternation of wet and rainy season. They inhabit ephemeral habitats that last a few months (Tozzini *et al.*, 2013). This short lifespan is retained under captive conditions and is coupled to rapid expression of a host of conserved age-associated phenotypes (Cellerino *et al.*, 2016). In addition, a key adaptation of annual fishes is the ability to enter diapause – a state when development halts – at specific stages during embryonic life, that is necessary to survive the dry season. The genus *Nothobranchius* evolved from a non-annual (therefore longer-lived) ancestor, the non-annual sister genus (Aphyosemion), is clearly identified (Furness *et al.*, 2015), and the two taxa provide a sharp phenotypic contrast. Duration of the habitat (aridity) strictly limits natural lifespan of Nothobranchius fishes.

We specifically tested whether differences in habitat duration led to the evolution of a different rate of senescence in Nothobranchius populations from southern and central Mozambique, a region characterized by a major gradient in aridity. Two independent evolutionary lineages of Nothobranchius are found in this area: *N. furzeri* and *N. kuhntae* belong to one lineage while *N. rachovii* and *N. pienaari* belong to another lineage (Dorn *et al.*, 2014). For each lineage, one species originates from semi-arid habitat (*N. furzeri* and *N. pienaari*, respectively) and another species from the humid habitat (*N. kuhntae* and *N. rachovii*, respectively). In both species pairs, the species from

more arid habitats showed shortened lifespan and accelerated expression of aging markers (Tozzini *et al.*, 2013), thereby providing a clear example of parallel evolution.

We previously sequenced and assembled the genome of *N. furzeri* as well as the transcriptomes of *N. kadleci* (the sister species of *N. furzeri*), *N. pienaari*, *N. rachovii*, and *N. kuhntae* together with *N. korthause* [a long-lived *Nothobranchius*, lifespan 18 months (Baumgart *et al.*, 2015)], and *Aphyosemion striatum* (lifespan > 3 years) as a representative of the non-annual sister genus (Reichwald *et al.*, 2015). We found seven genes under positive selection in *N. furzeri* and one in *N. pienaari*, another very-short-lived species, using the other six species of Nothobranchiidae as outgroups (Reichwald *et al.*, 2015). Here, we use a different selection of outgroups and analyze deeper branches of the *N. furzeri* phylogenetic tree to identify PSGs: (i) in coincidence with the evolution of annual life and (ii) showing parallel evolution in the two clades that are found in southern and central Mozambique.

Some results of this study were published in the form of preprint (Sahm *et al.*, 2016b).

Results

Genomewide scan strategy

In addition to sequence data presented previously (Reichwald *et al.*, 2015), we obtained from GenBank the RefSeq mRNA sequences of the phylogenetically closest outgroups from Ovalentaria (Fig. 1) and analyzed the pattern of positive selection along three internal branches of the tree: The first branch corresponds to the last common ancestor (LCA) of all *Nothobranchius* spp. (N-branch) and it marks the transition to annual life cycle. The other two branches correspond to the LCA of *N. pienaari* and *N. rachovii* (PR-branch) and LCA of *N. furzeri*, *N. kadleci* and *N. kuhntae* (FKK-branch), respectively. These two branches diverged in the Pleistocene, share the same distribution, and species belonging to the two clades can be found sympatric in the same pond (Dorn *et al.*, 2014). They represent therefore independent adaptations to the paleoclimatic changes of that period that was characterized by long-term progressive aridification of East Africa (Dorn *et al.*, 2014) and likely they were both subject to continued selection on adaptations linked to annual life cycle.

In each calculation, the background was the union of all the branches of the tree excluding the respective foreground branch, that is, when studying the N-branch the FKK and PR (and their child branches) are included in the background. In all comparisons, we defined PSGs based on nominal significant *P*-values (i.e., < 0.05, not corrected for multiple testing). This was a deliberate choice because of several reasons. First, we aim primarily at identifying parallel evolution at the level of pathway and not individual genes. Second, the number of genes strongly influences the sensitivity of Fisher's exact test, and it is not meaningful to perform GO analysis on lists containing few genes. Third, *de novo* transcriptome assembly projects inevitably generate incomplete data and a fraction of genes will show incomplete taxon coverage. We specifically tested whether PSGs have higher taxon coverage than the whole set of tested genes. However, this was not the case (Fig. 2) and only one PSG has a taxon coverage smaller than five.

Positive selection acts on mitochondrial and mitonuclear balance proteins

We found 75 PSGs in the N-branch, 106 in the PR-branch, and 88 in the FKK-branch ($P < 0.05$, branch-site test; Tables S1–S3, Supporting information). Among these, four code for components of the mitochondrial respiratory chain complex I in the N-branch (GO:0005747, fold-enrichment = 14, $P = 0.0002$, Fisher's exact test; Fig. 3, Table S1, Supporting information). Therefore, emergence of annual life cycle is coincident with strong positive selection on mitochondrial respiration. This is in line with the evidence that diapause is linked to profound remodeling of mitochondrial physiology (Duerr & Podrabsky 2010). Three further genes of complex I are under positive selection in the PR-branch (fold-enrichment = 8.8, $P = 0.005$, Fisher's exact test) and one further gene in the FKK-branch, indicating parallel and continued positive selection on complex I during the evolutionary history of Nothobranchius (Fig. 3, Tables S2 and S3, Supporting information).

Strikingly, other nine genes were under positive selection in both the PR- and FKK-branches (Table 1). Among these, are *TFB2M* (transcription factor B2, mitochondrial) and *POLRMT* (polymerase (RNA) mitochondrial) that together with *TFAM* (transcription factor A, mitochondrial) form the ternary complex that transcribes the entire mitochondrial genome (Litonin *et al.* 2010) and *FASTKD5* (fast kinase domain 5) that is necessary for processing of mitochondrial mRNAs (Antonicka & Shoubridge 2015). Further signs of parallel positive selection were evident at the level of functional gene groups. In addition to *FASTKD5*, *FASTKD1*

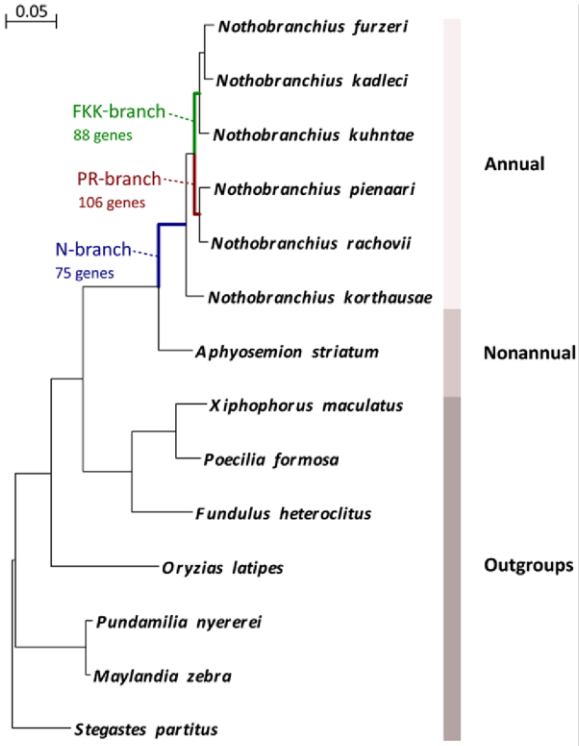


Fig. 1 Phylogeny of the analyzed species. Maximum-likelihood tree figure. Phylogeny of the analyzed species and their life history. Maximum-likelihood nucleotide-based phylogenetic tree of species that were used for genome-scale scans for positively selected genes. Outgroups from Ovalentaria are indicated as well as the three branches (N-, PR-, and FKK-branch) that are reported in the text. The alignment is based on concatenation of 4865 genes. The represented tree is the consensus of 1046 different trees created by splitting the alignment in fragments of 15 knt and calculating a tree for each fragment. The calibration bar refers to substitutions per nucleotidic site.

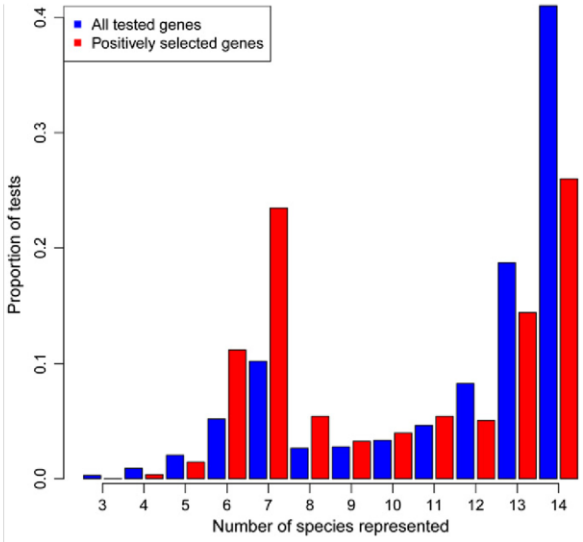


Fig. 2 Distribution of taxon coverage for all tested genes (blue bars) and the positively selected genes (red bars). The *X*-axis reports number of taxa for which a gene sequence is available, and the *Y*-axis reports the fraction of genes falling into each coverage class.

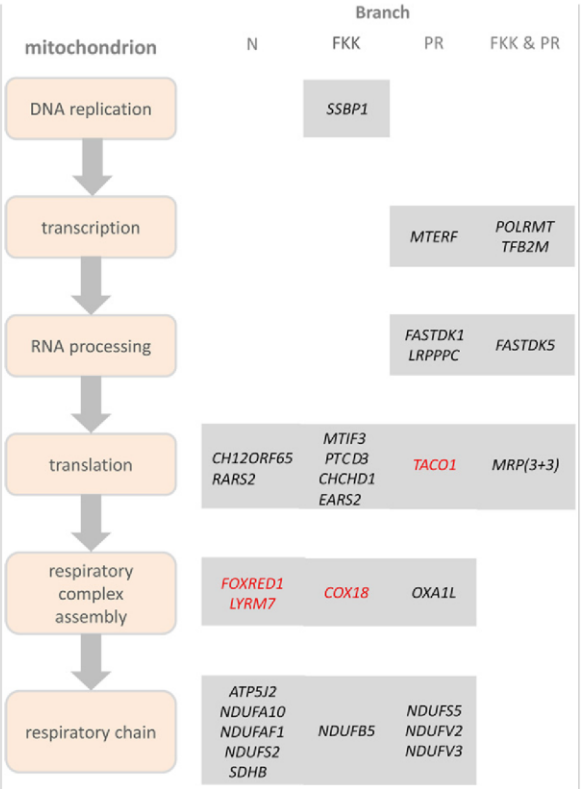


Fig. 3 Genes controlling mitochondrial biogenesis and mitonuclear balance under selection in the three branches. Mitochondrial biogenesis was divided into the following processes: mtDNA replication, transcription from mitochondrial promoters, processing and stabilization of mitochondrial RNAs, translation, assembly of respiratory chain complexes and electron transport chain. Genes in black are classified based on their GO annotation genes in red genes are involved in mitochondrial biogenesis based on literature but not annotated as such in GO (see text for references). The term MRP indicates mitochondrial ribosomal proteins (MRPL53, MRPS31, and MPRS26 in FKK-branch and MRPL23, MRPL3, and MTG2 in the RP-branch, respectively).

and *LRPPPC* (leucine-rich pentatricopeptide repeat containing), that control stability of mitochondrial RNAs (Sasarman *et al.* 2010), were positively selected in PR-branch. Three mitochondrial ribosome proteins (MRPs) were under positive selection in each of the two branches (GO:0005761, fold-enrichment = 9.1 and 14.7, respectively, $P = 0.02$ and 0.01, Fisher's exact test for the PR- and FKK-branch, respectively). In addition, two recently identified MRPs (Koc *et al.* 2013) were positively selected in FKK-branch: *PTCD3* (pentatricopeptide repeat-containing protein 3) and *CHCHD1* (coiled-coil-helix-coiled-coil-helix domain containing protein 1). Two further genes important for translation of mitochondrial RNAs were also positively selected: *MTIF3* (mitochondrial translation initiation factor 3) in FKK-branch and *TACO1* (translational activator of mitochondrially encoded cytochrome C oxidase I) in PR-branch (Fig. 3).

Respiratory chain complexes are large protein complexes that undergo multistep assembly where nuclearly and mitochondrially encoded components are combined and inserted into the mitochondrial inner membrane (Ghezzi & Zeviani 2012). Several genes involved in this process were positively selected: *COX18* (cytochrome C oxidase assembly factor) (Sacconi *et al.* 2009) in FKK-branch, *OXA1*L (oxidase (cytochrome c) assembly 1-like) (Stiburek *et al.* 2007; Haque *et al.* 2010) in PR-branch, *FOXRED1* (FAD-dependent oxidoreductase domain containing 1; Fassone *et al.* 2010) and *LYRM7* (LYR motif containing 7) (Sanchez *et al.* 2013) in N-branch (Fig. 3). Therefore, proteins necessary for mitochondrial biogenesis and more specifically for expression of mitochondrially encoded genes and assembly of respiratory chain complexes show signs of parallel evolution. Altogether, among the observed 269 cases of positive selection along the three branches, 33 could be assigned to mitochondrial proteins and those involved in the mitochondrial biogenesis and mitonuclear balance (Fig. 3 and Tables S1–S3, Supporting information).

We also compared expression levels of genes in mitochondrial biogenesis and mitonuclear balance in two contrasts of a short- and a long-lived species: *N. furzeri* vs. *A. striatum* and mouse vs. naked mole-rat (Yu *et al.*, 2011). For the genes, 1-to-1 orthology relationships based on ENSEMBL IDs could be established for 23. Of these, 12 (*RARS2*, *FASTKD5*, *POLRMT*, *OXA1L*, *NDUFAF1*, *C12orf65*, *NDUFS2*, *MTG2*, *PTCD3*, *MRPS31*, *NDUFS5*, *NDUFB5*) have a lower expression in both long-lived species (P-value = 0.005985, binomial test, ¼, Table S4, Supporting information).

Gene enrichment is not due to expression bias or incomplete lineage sorting

To ensure the statistical significance of our observation and exclude that biases due to the transcriptome assembly process and sequencing biases – in particular toward highly expressed genes – are responsible for the enrichment of mitochondrial proteins, we performed a simulation where we built two gene sets for each of three tested branches: an expression-adjusted background gene set and a "mitochondrial biogenesis" gene set. The later was derived from the union of the GO categories

Table 1 Genes that are positively selected both in PR- and FKK-branch

Gene symbol	Gene name	Function
ETAA1	Ewing tumor-associated antigen 1	DNA damage response
POLRMT	Polymerase (RNA) mitochondrial (DNA directed)	Transcription of mtDNA
PRRC2C	Proline-rich coiled-coil 2C	Poly-A RNA binding
APOA1	Apolipoprotein A-I	High-density lipoprotein particle binding
FASTKD5	FAST kinase domains 5	Regulation of mitochondrial RNA stability
TAF1C	TATA box-binding protein (TBP)-associated factor, RNA polymerase I, C, 110 kDa	Transcription of nuclear DNA
TFB2M	Transcription factor B2, mitochondrial	Transcription of mtDNA
CLIP1A (Nfu_g_1_008997)	CAP-GLY domain containing linker protein 1a	Unknown
SI:DKEYP-77H1.4 (*Nfu_g_1_001190*)	Uncharacterized	

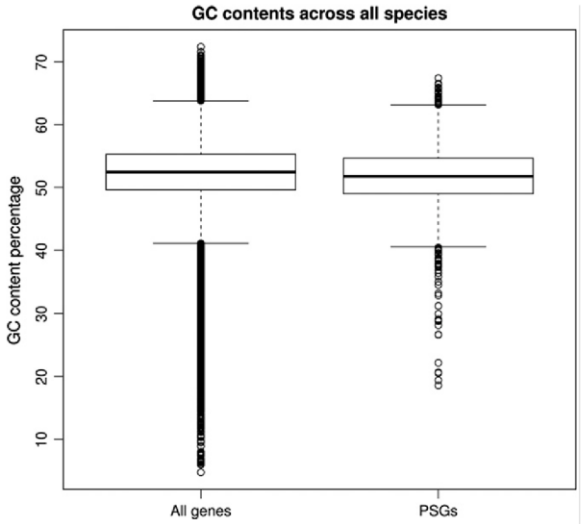


Fig. 4 Distribution of GC content in all the tested transcripts and the positively selected genes. Data from all species and tests are pooled. Box plots indicate 10%, 25%, median, 75% and 90% quantiles, and points represent outliers.

mitochondrial RNA metabolic process (GO:0000959), mitochondrial translation (GO:0032543), cellular respiration (GO: 0045333), mitochondrial respiratory chain complex assembly (GO:0033108), mitochondrial morphogenesis (GO:0070584). Per simulation run, we then randomly draw for each of the three branches from the background set a number of genes that equals the number of PSGs that were identified in the respective branch and calculated the sum of drawn mitochondrial biogenesis genes. In none of 1.000.000 simulation runs, a higher number than 21 was observed (95% quantile: 11). We concluded that our finding of 33 cases of positive selection on mitochondrial biogenesis genes is highly significant (simulated $P < 10^{-6}$) and not caused by an expression or sequencing bias. Analysis of GC content demonstrated that PSGs did not differ from all analyzed genes (Fig. 4).

To exclude that enrichment was due to incomplete lineage sorting (Mendes & Hahn, 2016), we tested all PSGs that were identified initially based on a globally estimated tree again with a tree that was estimated using only the individually tested gene. Among all PSGs, 91% (244/269) were supported by this approach as well. Only one gene involved in mitochondrial biogenesis, namely *SDHB* in the N-branch, was not confirmed to be a PSG by this analysis.

Positively selected genes are enriched among annualism-related genes

To derive independent evidence that the PSGs may be involved in the evolution of annual life style, we compared the union of the PSGs with two sets of differentially expressed genes (DEGs) in *N. furzeri*: (i) DEGs detected during brain aging (Baumgart *et al.*, 2014) and (ii) DEGs detected during diapause (Reichwald *et al.*, 2015). PSGs showed an over-representation among upregulated DEGs during diapause ($P = 0.023$, respectively, Fisher's exact test, Fig. 5) and among these 17 genes, *TFB2M* (PSG in both PR- and FKK-branch) and the assembly factor *NDUFA1* (PSG in the N-branch) are of relevance for mitonuclear balance. Over-representation of PSGs among upregulated DEGs is

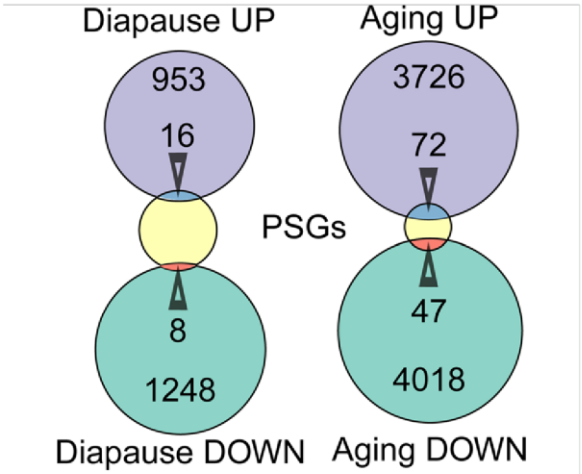


Fig. 5 Overlap of positively selected genes (PSGs) with genes regulated during diapause or aging. Differentially expressed genes were obtained from Baumgart *et al.* (2014) and Reichwald *et al.* (2015) for brain aging and diapause, respectively. The arrowheads point to the intersection of the sets and indicate the number of genes in the respective intersections. The numbers within the circles indicate the number of genes in each set excluded from the intersections. The total number of PSGs in 267 in both cases.

observed also during aging ($P = 0.0093$, respectively, Fisher's exact test, Fig. 5), among these 47 genes, TFB2M is again present. PSGs upregulated during aging were also four genes of the cytokine–cytokine receptor interaction pathway (*CSF1RA*, *FLT1*, *IL2RGA*, *IL2ST*; dre04060 KEGG, $P = 0.0001$, Fisher's exact test).

In addition, we compared PSGs with results of longitudinal gene expression in *N. furzeri*. Gene co-expression network analysis revealed that *ETAA1*, positively selected in both PR- and FKK-branch, and *APOA1BP* (apolipoprotein A1 binding protein), the binding partner of

the PSG *APOA1,* are part of a module of co-regulated genes highly enriched with MRPs and complex I components and negatively correlated with longevity (Baumgart *et al.*, 2016; Fig. 6).

Positively selected genes overlap with those in long-lived mammals

Previous analysis of PSGs in the *N. furzeri* genome suggested that some aging-relevant genes (e.g., the insulin like growth factor 1 receptor) can be positively selected both in short- and long-lived species (Valenzano *et al.*, 2015). We therefore compared *Nothobranchius* PSGs with PSGs detected by others in six species/clades of long-lived mammals (naked mole-rat, mole-rat LCA, blind mole-rat, human, bowhead whale and Brandt's bat). In all species, some PSGs overlap with those detected in *Nothobranchius* (Table S13, Supporting information). Of particular interest because under selection in more than two species/branches are: (i) *POLRMT*, that is a PSG in PR- and FKK-branch as well as in the Brandt's bat and the two extracellular matrix genes (ii) tenascin (*TNC*), a PSG in humans, mole-rats and in the FKK-branch regulated during both aging and diapause and (iii) Collagen type IV alpha 2 (*COL4A2*), a PSG in the naked mole-rat, the mole-rat LCA and in the FKK-branch also regulated during aging.

Discussion

The coordinated synthesis and assembly of mitochondrially and nuclearly encoded components of the respiratory chain is a conserved longevity mechanism that is controlled by MRPs (Dillin *et al.*, 2002; Houtkooper *et al.*, 2013). Knock-down of MRPs during early life in *C. elegans* results in an impaired assembly of respiratory complexes and life extension. Studies in the mouse and *N. furzeri* have shown that MRPs and nuclearly encoded complex I components are tightly co-regulated and expression of these genes during early adult life is predictive of lifespan in vertebrates (Miwa *et al.*, 2014; Baumgart *et al.*, 2016). Further, inhibition of complex I activity during adult life prolongs lifespan and rejuvenates the transcriptome in *N. furzeri* (Baumgart *et al.*, 2016).

Here, we show that these same genes are under positive selection in annual fish strongly suggesting that that evolution of the genes controlling mitonuclear balance is causally linked to evolution of short lifespan and annual life cycle. This is supported by our findings that several of these genes are positively selected along more than one investigated branch and are differentially regulated during diapause and aging of the shortest lived *Nothobranchius* species.

At the single-gene level, of particular interest are PSGs that are detected both in the PR- and FKK-branch that represent examples of parallel evolution. *POLRMT* and *TFB2M* are part of the ternary complex that transcribes mitochondrial DNA (including mitochondrial rRNAs) and *TAF1C* (TATA box-binding protein-associated factor RNA polymerase I subunit C) that is part of the multisubunit SL1 complex, which is required for RNA polymerase I to synthesize ribosomal RNA. Therefore, these three genes are at the core of the process that controls the balance between biogenesis of cytosolic and mitochondrial ribosomes. *APOA1* (apolipoprotein A1) is a component of HDL particles that have an obvious relevance for human age-related diseases. Polymorphisms of *APOA1* are associated with coronary artery disease (Helgadottir *et al.*, 2016) and it is an interactor of the *APOE*, a well-described genetic risk factor for Alzheimer's and cardiovascular diseases (Mahley, 2016) and the locus with the largest statistical support for an association with extreme longevity (Broer *et al.*, 2015). Interestingly, its expression in the liver is correlated with body weight in mice (Pearson's correlation coefficient: −0.84 for females and −0.74 for males, http://phenome.jax.org/). *EAAT1* shows striking similarities with *APOA1*. Its expression in the liver correlates with female body weight in mice (Pearson's correlation coefficient: +0.94). *ETAA1* and *APOA1BP* have central positions in the gene module of co-expressed genes whose expression is negatively correlated with lifespan that also contains MRPs and complex I (Fig. 6), strongly suggesting that they are involved in mitonuclear balance. Interestingly, a function of *ETAA1* in DNA repair was recently demonstrated (Lee *et al.*, 2016) and the gene coding for another protein important in DNA repair, *XRCC5*, was previously shown to be under positive selection in *N. furzeri* (Valenzano *et al.*, 2015; Sahm *et al.*, 2016a). However, it is not possible to determine *in silico* whether the substitutions observed in the two lineages cause similar changes of mitochondrial function and parallel selection on the same genes does not represent a proof of functional convergence.

Are the genes under selection in short-lived species also involved in evolution of longevity? Data supporting this notion come from different

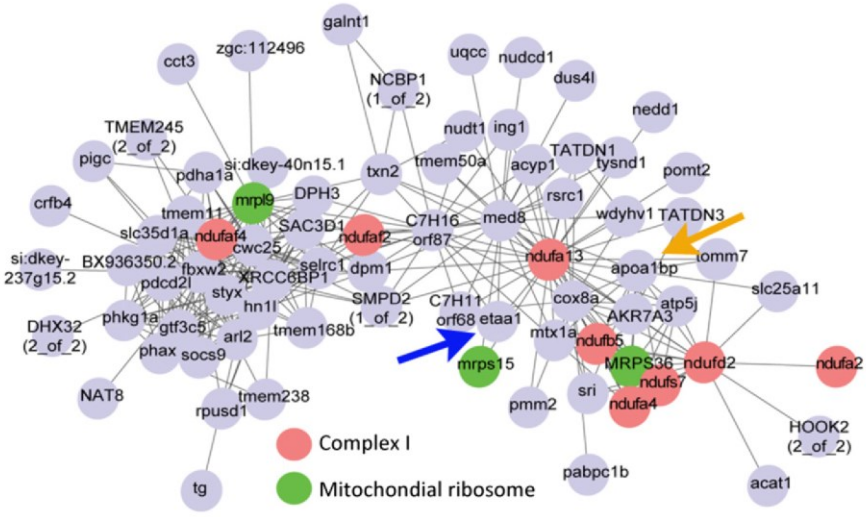


Fig. 6 Position of ETAA1 and APOA1BP in the network of longevity-associated genes described by Baumgart *et al.* (2016). Picture reproduced with permission from Baumgart *et al.* (2016) and the gene annotation conforms to the annotation of transcripts described in Reichwald *et al.* (2015). Red genes code for complex I components and green genes for mitochondrial ribosome components.

studies of positive selection in the genomes of long-lived species. Ant workers can live on average ten times as long as their solitary ancestors and queens with 10 years at average and nearly 30 years at maximum even more than 100 times as long (Jemielity *et al.*, 2005). In an examination of seven ants genomes, highly significant enrichments of PSGs were documented for a series of GO terms that are related to the respiratory chain or mitochondrial biogenesis; especially mitochondrial electron transport (GO:0006120), mitochondrial respiratory chain complex I (GO:0005747), and mitochondrial large/small ribosomal subunit (GO:0005762/GO:0005763) (Roux *et al.*, 2014). The same study reported based on expression data obtained in the fire ant *S. invicata* that PSGs are highly expressed in queens, intermediately expressed in workers and weakest expressed in males which are the shortest lived ant caste. While the expression of PSGs correlates with lifespan of the respective caste, there is no differential expression across mitochondrial genes in general between queens and workers. This means that the association between PSGs and caste biased gene expression cannot be simply explained by higher overall levels of genes that are involved in mitochondrial activity but suggests a relation between queen-specific expression of PSGs and longer lifespan. Notably, consistent with the results of our study, there was no evidence found for positive selection on mitochondrial-encoded genes in ants. Furthermore, respiratory chain genes were found to be under positive selection in the bats *P. poliocephalus* and *M. lucifugus* (Shen *et al.*, 2010). Both are long-lived mammals, while *P. poliocephalus* reaches a maximum age of 23.6 years at a weight of 675 g resulting in a lifespan that is 1.7 times larger than expected based on the body mass, *M. lucifugus* even reaches a maximum age of 34 at a weight of only 10 g resulting in lifespan almost five times longer than expected based on body mass (Tacutu *et al.*, 2013).

Overlaps with long-lived mammals are detectable also at the level of single genes. Of particular interest is *POLRMT* this gene that codes for the mitochondrial RNA polymerase is a PSG in the PR- and FKK-branch and also in the Brandt's bat (Seim *et al.*, 2013) and, as discussed above, it is of key importance for mitonuclear balance. It is tempting to speculate that positive selection in short- and long-lived species modulates mitochondrial function in opposite directions. However, as discussed above, it is not possible to predict the functional impact of molecular evolution and this hypothesis will require experimental tests. Indirect evidence in favor comes from the observation that a significant fraction of mitochondrial biogenesis and mitonuclear balance genes are lower expressed in the longer lived element of two comparisons of long- and short-lived species: *N. furzeri* vs. *A. striatum* and mouse vs. naked mole-rat.

This hypothesis is also supported by direct measurements of complex I activity. Assays of mitochondrial physiology in the bivalve *Arctica islandica* (the longest lived metazoan with maximum lifespan exceeding 500 years) and two taxonomically related species of comparable size revealed lower activity of complex I resulting in reduced production of reactive oxygen species (Munro *et al.*, 2013). Similarly, low activity of complex I and low production of reactive oxygen species were related to longevity in homeotherm vertebrates (Brunet-Rossinni, 2004; Lambert *et al.*, 2010) and, finally, conditions that increase mouse longevity are associated with reduced expression of complex I (Miwa *et al.*, 2014).

Comparison of positive selection at the gene level between *Nothobranchius* and long-living mammals identified *TNC* and *COL4A2* as particularly interesting candidates as they are PSGs in two mammalian clades each and are also both differentially expressed in *Nothobranchius furzeri* aging (Reichwald *et al.*, 2015). These data lend further support to the notion that extracellular matrix genes are regulators of lifespan that derives from meta-analysis of genomewide transcript regulation (de Magalhaes *et al.* 2009), positive selection analysis (Li & de Magalhaes, 2013), and experimental approaches (Ewald *et al.*, 2015).

Finally, it should be noted that PSGs upregulated during aging were enriched for terms related to inflammation that are also known to be a highly conserved hallmark of aging at the transcriptome level (Baumgart *et al.*, 2014).

Experimental procedures

Genome-scale identification of positively selected genes

The basis for this work were protein-coding sequences (CDSs) of six *Nothobranchius* species (*N. furzeri*, *N. kadleci*, *N. kuhntae*, *N. pienaari*, *N. rachovii, and N. korthausae*) and *A. striatum* from transcriptome catalogs that were recently assembled and annotated (Reichwald *et al.*, 2015) with FRAMA (Bens *et al.*, 2016). The reads were adapter clipped with seqprep (https://github.com/jstjohn/SeqPrep) and quality trimmed with sickle (Joshi & Fass, 2011) before assembly [for more information about tissues, read numbers, filtered bases, etc., see Table S12 (Supporting information) or (Reichwald *et al.*, 2015)]. CDSs from seven additional outgroups (*Xiphophorus maculatus*, *Poecilia formosa*, *Fundulus heteroclitus*, *Maylandia zebra*, *Pundamilia nyererei*, *Stegastes partitus*, *Oryzias latipes*) were obtained from NCBI RefSeq (14.12.15) and assigned to ortholog groups by the best bidirectional BLAST hit criterion Camacho *et al.* (2009) against *N. furzeri*.

For each N. furzeri CDS isoform, the most similar isoform of each other species was determined by pairwise comparison. To reduce the risk of aligning nonhomologous codons, these sequences were required to have additionally at least a similarity of 70% with *N. furzeri* and 50% with each other species on protein level. The selected isoforms in each ortholog group were aligned with PRANK (Loytynoja & Goldman, 2008), which is the alignment software of choice for positive selection analysis (Fletcher & Yang, 2010). The alignments were stringently filtered with GBLOCKS (Talavera & Castresana, 2007) to remove gaps and unreliable alignment columns around them that could produce false signals of positive selection (−b2 = total number of sequences in the alignment, b4 = 30, t=c). Then, for each alignment the branch-site test of positive selection (Yang & Nielsen, 2002; Zhang *et al.*, 2005) was applied: The respectively tested branch (LCA, FKK, or PR) was marked as 'foreground', and all other branches were marked as 'background'. The program CODEML from the PAML (Yang, 2007) package was called separately for models M2a0 (model = 2, NSsites = 2; fix_omega = 1, omega = 1) and M2a (model = 2, NSsites = 2; fix_omega = 0, omega = 1) as described in the PAML User Guide (http://abacus.gene.ucl.ac.uk/software/pamlDOC.pdf). To calculate a *P*-value, the chi-square distribution with one degree of freedom was used to compare the likelihoods of both models: $P = \chi^2$ (2* (ln(likelihood(M2a))-ln(likelihood(M2a0))),1).

Sites under positive selection were inferred by the Bayes empirical Bayes method (Yang *et al.*, 2005) provided by CODEML. Sites that were predicted in a two amino acid frame next to a block which was deleted by GBLOCKS were removed and an adjusted *P*-value calculated. For 9017, 12028, and 10976 genes, *P*-values were calculated in the N-, FKK- and PR-branches, respectively (Table S5, Supporting information). We considered all genes with *P*-values ≤ 0.05 as positively selected genes (PSGs).

Since high rates of false positive were detected in some automated genome-scale scans for PSGs in the past (Mallick *et al.*, 2009; Markova-Raina & Petrov, 2011), we demanded our candidates to fulfill further strict filter criteria. Candidates were removed that had: (I) not at least one species from the sister branch of the tested branch in the alignment,

for example, for the N-branch the presence of A. striatum was demanded, (II) less than four species in the alignment, (III) remained only few columns (<60 or <66.67%) or N. furzeri codons (<60%) of the alignment after GBLOCKS filtering, (IV) disproportional dN/dS ratios (e.g., ≥100 in foreground branch, >1 in background branch, <0.85 in foreground branch) were calculated by CODEML or (V) had an unreliably high fraction of inferred positively selected sites (more than 20%). Finally, we inspected all candidates on the FKK- and PR-branch manually as well as roughly ten percent of those on the LCA-branch and removed ten additional candidates (<5%) in total.

Phylogenetic tree

The phylogenetic tree that is needed for the analysis with CODEML was calculated based on the concatenated alignment of CDS isoforms of those 4865 genes with aligned isoforms from all species (Table S6, Supporting information). The final tree was the consensus of 1046 different trees created by splitting the alignment in fragments of 15 knt and calculating a tree for each fragment with DNAML from the PHYLIP (Felsenstein, 2005) package. All PSGs that were predicted with this globally estimated tree were again tested for positive selection with the same methods described above but with a tree that was estimated on the alignment of the respective gene. 65 of 75, 79 of 88, and 100 of 106 candidates were confirmed by this approach in the N-, FKK-, and N-branch, respectively (candidates that could not be confirmed are marked in Tables S1–S3, Supporting information).

Hypothesis-driven GO enrichments

We determined potential enrichments for the GO categories mitochondrial ribosome (GO:0005761) and mitochondrial respiratory chain complex I (GO:0005747) with Fisher's exact test. The set of tested genes that could be converted to entrez IDs served as background, that is, 7523, 9416, and 8745 genes for the N-, FKK-, and PR-branch, respectively (Table S7, Supporting information). As this was an hypothesis-driven approach, the *P*-values were not corrected for multiple testing.

Mitochondrial biogenesis enrichment simulation

For each of three tested branches, we built two gene sets, a background gene set and a mitochondrial biogenesis gene set. The background gene sets consisted of the tested genes of the respective branch that could be converted to Entrez IDs (https://www.ncbi.nlm.nih.gov/Entrez), that is, 7523, 9416, and 8745 genes for the N-, FKK-, and PR-branch, respectively (Table S7, Supporting information). To avoid biases due to expression differences between gene sets, we reduced the background sets to those genes within the 5–95% expression quantile of the union of PSGs across the three branches. This resulted in 6803, 8336, and 7882 genes, respectively (Tables S8 and S11, Supporting information). For the mitochondrial biogenesis gene sets, a union was built from the genes enlisted in the following five mitochondrial-related GO terms (GO:0000959, 0032543, 0045333, 0033108, 0070584). This union encompassed 331 genes (Table S9, Supporting information). For each branch, the mitochondrial biogenesis gene set consisted of the genes from this union that were also present in the background of the respective branch, resulting in 221, 250, and 245 genes, respectively (Table S9, Supporting information). In each simulation, round drawings were done for each branch from the respective background set and as often as PSGs were identified in that branch and could be converted to Entrez IDs, that is, 65, 73, 89 times (Table S10, Supporting information), respectively. Our simulation was conservative in the way that we did not reduce the number of drawings in each branch to the 5–95% expression quantile of the PSGs, giving the simulation a higher chance to draw genes from mitochondrial biogenesis set than we had in reality. At the end of each simulation round, it was counted how many drawn genes were in the mitochondrial biogenesis gene set for each branch and, finally, the sum across the three branches was calculated. One million simulation rounds were done.

Author's contributions

AS and MB performed the analysis; MP and AC supervised the work; and AS, MP, and AC wrote the manuscript.

Funding

This work was supported by the Leibniz Association (SAW-2012-FLI) and the German Research Foundation (DFG: PL 173/8-1) and a grant from Scuola Normale Superiore (CELLSNS2015).

Conflict of interest

None declared.

References

Antonicka H, Shoubridge EA (2015) Mitochondrial RNA granules are centers for posttranscriptional RNA processing and ribosome biogenesis. *Cell Rep.* doi:10.1016/j.celrep.2015.01.030.

Baumgart M, Groth M, Priebe S, Savino A, Testa G, Dix A, Ripa R, Spallotta F, Gaetano C, Ori M, Terzibasi Tozzini E, Guthke R, Platzer M, Cellerino A (2014) RNA-seq of the aging brain in the short-lived fish *N. furzeri* – conserved pathways and novel genes associated with neurogenesis. *Aging Cell* **13**, 965–974.

Baumgart M, Di Cicco E, Rossi G, Cellerino A, Tozzini ET (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**, 63–69.

Baumgart M, Priebe S, Groth M, Hartmann N, Menzel U, Pandolfini L, Koch P, Felder M, Ristow M, Englert C, Guthke R, Platzer M, Cellerino A (2016) Longitudinal RNA-Seq analysis of vertebrate aging identifies mitochondrial complex I as a small-molecule-sensitive modifier of lifespan. *Cell Syst.* **2**, 122–132.

Bens M, Sahm A, Groth M, Jahn N, Morhart M, Holtze S, Hildebrandt TB, Platzer M, Szafranski K (2016) FRAMA: from RNA-seq data to annotated mRNA assemblies. *BMC Genom.* **17**, 54.

Broer L, Buchman AS, Deelen J, Evans DS, Faul JD, Lunetta KL, Sebastiani P, Smith JA, Smith AV, Tanaka T, Yu L, Arnold AM, Aspelund T, Benjamin EJ, De Jager PL, Eirkisdottir G, Evans DA, Garcia ME, Hofman A, Kaplan RC, Kardia SL, Kiel DP, Oostra BA, Orwoll ES, Parimi N, Psaty BM, Rivadeneira F, Rotter JI, Seshadri S, Singleton A, Tiemeier H, Uitterlinden AG, Zhao W, Bandinelli S, Bennett DA, Ferrucci L, Gudnason V, Harris TB, Karasik D, Launer LJ, Perls TT, Slagboom PE, Tranah GJ, Weir DR, Newman AB, van Duijn CM, Murabito JM (2015) GWAS of longevity in CHARGE consortium confirms APOE and FOXO3 candidacy. *J. Gerontol. A Biol. Sci. Med. Sci.* **70**, 110–118.

Brunet-Rossinni AK (2004) Reduced free-radical production and extreme longevity in the little brown bat (*Myotis lucifugus*) versus two non-flying mammals. *Mech. Ageing Dev.* **125**, 11–20.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421.

Cellerino A, Valenzano DR, Reichard M (2016) From the bush to the bench: the annual *Nothobranchius* fishes as a new model system in biology. *Biol. Rev. Camb. Philos. Soc.* **91**, 511–533.

de Magalhaes JP, Curado J, Church GM (2009) Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics* **25**, 875–881.

Davies KT, Bennett NC, Tsagkogeorga G, Rossiter SJ, Faulkes CG (2015) Family wide molecular adaptations to underground life in African mole-rats revealed by phylogenomic analysis. *Mol. Biol. Evol.* **32**, 3089–3107.

Dillin A, Hsu AL, Arantes-Oliveira N, Lehrer-Graiwer J, Hsin H, Fraser AG, Kamath RS, Ahringer J, Kenyon C (2002) Rates of behavior and aging specified by mitochondrial function during development. *Science* **298**, 2398–2401.

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

Duerr JM, Podrabsky JE (2010) Mitochondrial physiology of diapausing and developing embryos of the annual killifish Austrofundulus limnaeus: implications for extreme anoxia tolerance. *J. Comp. Physiol. B.* **180**, 991–1003.

Ewald CY, Landis JN, Porter Abate J, Murphy CT, Blackwell TK (2015) Dauer-independent insulin/IGF-1-signalling implicates collagen remodelling in longevity. *Nature* **519**, 97–101.

Fassone E, Duncan AJ, Taanman JW, Pagnamenta AT, Sadowski MI, Holand T, Qasim W, Rutland P, Calvo SE, Mootha VK, Bitner-Glindzicz M, Rahman S (2010) FOXRED1, encoding an FAD-dependent oxidoreductase complex-I-specific molecular chaperone, is mutated in infantile-onset mitochondrial encephalopathy. *Hum. Mol. Genet.* **19**, 4837–4847.

Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Fletcher W, Yang Z (2010) The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol. Biol. Evol.* **27**, 2257–2267.

Furness AI, Reznick DN, Springer MS, Meredith RW (2015) Convergent evolution of alternative developmental trajectories associated with diapause in African and South American killifish. *Proc. Biol. Sci.* **282**, 20142189.

Ghezzi D, Zeviani M (2012) Assembly factors of human mitochondrial respiratory chain complexes: physiology and pathophysiology. *Adv. Exp. Med. Biol.* **748**, 65–106.

Haque ME, Elmore KB, Tripathy A, Koc H, Koc EC, Spremulli LL (2010) Properties of the C-terminal tail of human mitochondrial inner membrane protein Oxa1L and its interactions with mammalian mitochondrial ribosomes. *J. Biol. Chem.* **285**, 28353–28362.

Helgadottir A, Gretarsdottir S, Thorleifsson G, Hjartarson E, Sigurdsson A, Magnusdottir A, Jonasdottir A, Kristjansson H, Sulem P, Oddsson A, Sveinbjornsson G, Steinthorsdottir V, Rafnar T, Masson G, Jonsdottir I, Olafsson I, Eyjolfsson GI, Sigurdardottir O, Daneshpour MS, Khalili D, Azizi F, Swinkels DW, Kiemeney L, Quyyumi AA, Levey AI, Patel RS, Hayek SS, Gudmundsdottir IJ, Thorgeirsson G, Thorsteinsdottir U, Gudbjartsson DF, Holm H, Stefansson K (2016) Variants with large effects on blood lipids and the role of cholesterol and triglycerides in coronary disease. *Nat. Genet.* **48**, 634–639.

Houtkooper RH, Mouchiroud L, Ryu D, Moullan N, Katsyuba E, Knott G, Williams RW, Auwerx J (2013) Mitonuclear protein imbalance as a conserved longevity mechanism. *Nature* **497**, 451–457.

Jemielity S, Chapuisat M, Parker JD, Keller L (2005) Long live the queen: studying aging in social insects. *Age* **27**, 241–248.

Joshi NA, Fass JN (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files.

Koc EC, Cimen H, Kumcuoglu B, Abu N, Akpinar G, Haque ME, Spremulli LL, Koc H (2013) Identification and characterization of CHCHD1, AURKAIP1, and CRIF1 as new members of the mammalian mitochondrial ribosome. *Front. Physiol.* **4**, 183.

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A (2008) Patterns of positive selection in six Mammalian genomes. *PLoS Genet.* **4**, e1000144.

Lambert AJ, Buckingham JA, Boysen HM, Brand MD (2010) Low complex I content explains the low hydrogen peroxide production rate of heart mitochondria from the long-lived pigeon, Columba livia. *Aging Cell* **9**, 78–91.

Lee YC, Zhou Q, Chen J, Yuan J (2016) RPA-binding protein ETAA1 is an ATR activator involved in DNA replication stress response. *Curr. Biol.* **26**, 3257–3268.

Li Y, de Magalhaes JP (2013) Accelerated protein evolution analysis reveals genes and pathways associated with the evolution of mammalian longevity. *Age* **35**, 301–314.

Litonin D, Sologub M, Shi Y, Savkina M, Anikin M, Falkenberg M, Gustafsson CM, Temiakov D (2010) Human mitochondrial transcription revisited: only TFAM and TFB2M are required for transcription of the mitochondrial genes in vitro. *J. Biol. Chem.* **285**, 18129–18133.

Loytynoja A, Goldman N (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**, 1632–1635.

Mahley RW (2016) Apolipoprotein E: from cardiovascular disease to neurodegenerative disorders. *J. Mol. Med.* **94**, 739–746.

Mallick S, Gnerre S, Muller P, Reich D (2009) The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res.* **19**, 922–933.

Markova-Raina P, Petrov D (2011) High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 *Drosophila* genomes. *Genome Res.* **21**, 863–874.

Mendes FK, Hahn MW (2016) Gene tree discordance causes apparent substitution rate variation. *Syst. Biol.* **65**, 711–721.

Miwa S, Jow H, Baty K, Johnson A, Czapiewski R, Saretzki G, Treumann A, von Zglinicki T (2014) Low abundance of the matrix arm of complex I in mitochondria predicts longevity in mice. *Nat. Commun.* **5**, 3837.

Munro D, Pichaud N, Paquin F, Kemeid V, Blier PU (2013) Low hydrogen peroxide production in mitochondria of the long-lived *Arctica islandica*: underlying mechanisms for slow aging. *Aging Cell* **12**, 584–592.

Reichwald K, Petzold A, Koch P, Downie BR, 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 SS, Sharma V, Kraus JM, Schmid F, Priebe S, Liehr T, Gorlach M, Than ME, Hiller M, Kestler HA, Volff JN, Schartl M, Cellerino A, Englert C, Platzer M (2015) Insights into sex chromosome evolution and aging from the genome of a short-lived fish. *Cell* **163**, 1527–1538.

Roux J, Privman E, Moretti S, Daub JT, Robinson-Rechavi M, Keller L (2014) Patterns of positive selection in seven ant genomes. *Mol. Biol. Evol.* **31**, 1661–1685.

Sacconi S, Salviati L, Trevisson E (2009) Mutation analysis of COX18 in 29 patients with isolated cytochrome c oxidase deficiency. *J. Hum. Genet.* **54**, 419–421.

Sahm A, Platzer M, Cellerino A (2016a) Outgroups and positive selection: the *Nothobranchius furzeri* case. *Trends Genet.* **32**, 523–525.

Sahm A, Bens M, Platzer M, Cellerino A (2016b) Convergent evolution of genes controlling mitonuclear balance in annual fishes. *bioRxiv*. doi: https://doi.org/10.1101/055780

Sanchez E, Lobo T, Fox JL, Zeviani M, Winge DR, Fernandez-Vizarra E (2013) LYRM7/MZM1L is a UQCRFS1 chaperone involved in the last steps of mitochondrial Complex III assembly in human cells. *Biochim. Biophys. Acta* **1827**, 285–293.

Sasarman F, Brunel-Guitton C, Antonicka H, Wai T, Shoubridge EA, Consortium L (2010) LRPPRC and SLIRP interact in a ribonucleoprotein complex that regulates posttranscriptional gene expression in mitochondria. *Mol. Biol. Cell* **21**, 1315–1323.

Seim I, Fang X, Xiong Z, Lobanov AV, Huang Z, Ma S, Feng Y, Turanov AA, Zhu Y, Lenz TL, Gerashchenko, MV, Hee Fan, D, Yim, S, Yao, X, Jordan, D, Xiong, Y, Ma, Y, Lyapunov, AN, Chen, G, Kulakova, OI, Sun, Y, Lee, SG, Bronson, RT, Moskalev, AA, Sunyaev, SR, Zhang, G, Krogh, A, Wang, J and Gladyshev, VN (2013) Genome analysis reveals insights into physiology and longevity of the Brandt's bat *Myotis brandtii*. *Nat. Commun.* **4**, 2212.

Shen YY, Liang L, Zhu ZH, Zhou WP, Irwin DM, Zhang YP (2010) Adaptive evolution of energy metabolism genes and the origin of flight in bats. *Proc. Natl Acad. Sci. USA* **107**, 8666–8671.

Stiburek L, Fornuskova D, Wenchich L, Pejznochova M, Hansikova H, Zeman J (2007) Knockdown of human Oxa1l impairs the biogenesis of F1Fo-ATP synthase and NADH:ubiquinone oxidoreductase. *J. Mol. Biol.* **374**, 506–516.

Tacutu R, Craig T, Budovsky A, Wuttke D, Lehmann G, Taranukha D, Costa J, Fraifeld VE, de Magalhaes JP (2013) Human ageing genomic resources: integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Res.* **41**, D1027–D1033.

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564–577.

Tozzini ET, Dorn A, Ng'oma E, Polacik M, Blazek R, Reichwald K, Petzold A, Watters B, Reichard M, Cellerino A (2013) Parallel evolution of senescence in annual fishes in response to extrinsic mortality. *BMC Evol. Biol.* **13**, 77.

Valenzano DR, Benayoun BA, Singh PP, Zhang E, Etter PD, Hu CK, Clement-Ziza M, Willemsen D, Cui R, Harel I, Machado BE, Yee MC, Sharp SC, Bustamante CD, Beyer A, Johnson EA, Brunet A (2015) The African turquoise killifish genome provides insights into evolution and genetic architecture of lifespan. *Cell* **163**, 1539–1554.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19**, 908–917.

Yang Z, Wong WS, Nielsen R (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* **22**, 1107–1118.

Yu C, Li Y, Holmes A, Szafranski K, Faulkes CG, Coen CW, Buffenstein R, Platzer M, de Magalhaes JP, Church GM (2011) RNA sequencing reveals differential

expression of mitochondrial and oxidation reduction genes in the long-lived naked mole-rat when compared to mice. *PLoS ONE* **6**, e26729.

Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22**, 2472–2479.

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article.

Table S1 Results of positive selection analysis in the N-branch, ranked by *P*-value.

Table S2 Results of positive selection analysis in the PR-branch, ranked by *P*-value.

Table S3 Results of positive selection analysis in the FKK-branch, ranked by *P*-value.

Table S4 Comparison of fold-changes in expression of mitonuclear and mitochondrial biogenesis gene sets.

Table S5 Overview of the tested genes.

Table S6 List of genes used to construct the tree in Fig. 1.

Table S7 List of background genes used for GO analysis.

Table S8 List of genes used as background for the simulation experiment (5–95% expression quantile of the PSGs).

Table S9 List of mitochondrial biogenesis genes within the background genes list (Table S7).

Table S10 Entrez orthologs of PSGs.

Table S11 Expression levels of all tested genes.

Table S12 Statistics of read data for each library.

Table S13 Complete list of *Nothobranchius* PSGs overlapping with PSGs in long-lived mammals.

4 Discussion

Model organisms are vital experimental systems to understand fundamental biological phenomena and particular features of human biology. Current large-scale molecular studies favour high-quality genome sequences to perform genome-wide biological and computational analyses. Not least because of this reason, the genomes of traditional mammalian laboratory animals (mouse (2002) [111], rat (2004) [112]) became sequenced soon after the human genome (2001) [113]. However, the introduction of 2GS technologies massively changed the way researchers can approach biological questions and enabled the generation of sequence resources for virtually any species of choice. This gave rise to extensive genome sequencing projects, such as the 10k Genome Project that aims to sequence at least one species of each vertebrate genus, unfolding research in non-model organisms covering a wide ecological diversity [114]. Yet, even after a decade of progress in 2GS technologies, the assembly of large and complex genomes to the level of continuous and low-error chromosome-scale sequences is still time consuming and costly [18,19]. Genome assemblies usually remain at a draft stage, consisting of thousands of fragments having high gap percentage and limiting downstream analyses [15]. Genome annotation is another complex task and depends on the quality (*i.e.* completeness, contiguity) of the underlying genome sequence [17,98].

The series of four studies that comprise this thesis follow a different route by investigating the protein-coding fraction of genomes in non-model organisms using RNA-seq. RNA-seq and subsequent computational reconstruction of mRNAs appeals to researchers as it avoids the complex process of genome reconstruction and annotation, and simultaneously enables gene expression profiling. Despite the dependency on mRNA sequences and increasing availability of RNA-seq data, no standardized workflow has been developed to obtain mRNA catalogues that are suitable for downstream analysis. The first publication (M1) aims to fill that gap by establishing a software pipeline (`FRAMA`) that delivers mRNA catalogues based on *de novo* assembly of RNA-seq data and additionally comprises custom-build tools to address certain issues in *de novo* assemblies. mRNA catalogues delivered by `FRAMA` provide a step towards the integration of non-model organisms into "daily routines" of wet laboratory (*e.g.* primer and small interfering RNA design) and *in silico* (*e.g.* gene expression analyses, comparative studies) research. The proof of use in *in silico* research was demonstrated by qualitative and quantitative investigation of promising non-model organisms in ageing research (Figure 1) as presented in M2 and M3 & M4.

4.1 Coping with *de novo* transcriptome assembly issues from RNA-seq data

The development of `FRAMA` in M1 was motivated by the lack of software pipelines that perform necessary steps to get from RNA-seq data of non-model organisms to mRNA

catalogues appropriate for downstream analysis. Software tools for key steps in such a pipeline were already available: (i) `Trinity` for *de novo* transcriptome assembly, (ii) `CD-HIT-EST` and `TGICL` to remove assembly redundancy, (iii) `BLAST` to detect sequence similarity for ortholog inference and (iv) `GENSCAN` for *ab initio* CDS prediction or `MAFFT` for CDS inference by multiple sequence alignment (referenced in M1). However, correct application (parameter choice), output processing (filtering, format conversion) and efficient connection (parallelisation) of these programs requires knowledge in bioinformatics and is a time-consuming process. Further, although such key steps deliver an mRNA catalogue eventually, shortcomings of *de novo* transcriptome assembly arising from the complexity of transcriptomes and short sequencing reads are not addressed. Besides efficiently implementing these key steps, `FRAMA` integrates custom-build tools to attenuate *de novo* assembly issues. In the following, I will focus on three specific issues addressed by `FRAMA` – fusion contigs, mRNA end clipping and fragmented contigs - and discuss possible alternative approaches.

The assessment of post-transcriptional mRNA regulation involves analysis of primary and secondary structure of UTRs and is an important aspect to assess, *e.g.* translation efficiency, localization and stability of mRNAs [99]. However, *de novo* assemblers experience difficulties in determining precise transcript ends arising from several biological factors, such as (i) imprecise molecular mechanism of transcription initiation and 3′ end cleavage, (ii) alternative promotors and polyadenylation events and (iii) antisense transcripts [25,28,115,116]. Adjacent gene loci can even lead to the reconstruction of artificially fused transcripts [40,117]. `FRAMA` addresses the fusion problem by identifying and splitting contigs showing consecutive ortholog alignments with minor overlaps. Despite from assembly (or reverse transcription [118]) artefacts, such fusion contigs can potentially arise from genuine chimeric RNAs that are translated into proteins [119] and generated, *e.g.* by read-through transcription of adjacent genes or by trans-splicing of distant gene loci [25,120]. However, assuming that genuine chimeras are contained in sequence databases and are conserved across species, genuinely fused contigs are still correctly annotated. Thus, split of contigs showing consecutive ortholog alignments is reasonable to increase the detectable number of genes. This approach reduces the need for strand-specific sequencing, which, however, would only allow to distinguish transcripts from (i) adjacent gene loci on opposite strands of the genome (*i.e.* head-to-head, tail-to-tail gene organization [117]) and (ii) antisense transcripts. However, my approach could be supplemented with fusion detection tools to exclude split of genuine chimeras. Performance of such tools was recently assessed, but most tools require a genomic reference and paired-end RNA-seq data [121]. `JAFFA` presents a suitable candidate by utilizing a transcript-centric approach and also works with single-end RNA-seq data [122].

To determine more reliable and evolutionary reasonable 3′ mRNA ends in split and all other contigs, `FRAMA` exploits the coverage drop biases in RNA-seq data near transcript ends, incorporates information from poly-(A) containing reads and uses sequence characteristics, including poly-(A) signals and conservation of 3′ UTR ends [116,123,124]. Yet, each feature does

not provide a precise cut position (*e.g.* presence of multiple poly-(A) signals, gradual decrease in read coverage). Therefore, each 3′ UTR position is assigned a score by fuzzy logic, connecting the weighted signals from each feature. The respective weights and a scoring threshold were empirically determined based on a set of gold standard mRNA sequences. However, there are different library preparation protocols for 2GS that allow to determine precise transcript ends experimentally rather than leveraging ortholog information and sequence characteristics. For instance, "cap analysis gene expression" is able to capture the cap found at the 5′ end of mature mRNAs and enables subsequent sequencing of 5′ ends [125]. Other approaches detect both ends simultaneously, such as "transcript isoform sequencing" and "gene identification signature" [126,127]. To my knowledge, data from such approaches have not yet been integrated directly into *de novo* assemblers. Although adding complexity to the sequencing process, appropriate integration should theoretically reduce assembly complexity by (i) delineating transcripts regardless of their origin and (ii) providing linkage information of start and ends of transcripts, thereby decreasing the number of possible assembly solutions.

The third issue are low expressed transcripts and low complexity regions (*e.g.* tandem repeats). Insufficient read coverage of low expressed transcripts leads to fragmented contigs. Similarly, low complexity regions present no unique assembly solution and respective low complexity reads might even be discarded during assembly [39]. To attenuate the fragmentation issue, `FRAMA` comprises a scaffolding step that combines fragments showing high sequence similarity to an ortholog into a contiguous transcript scaffold. Therefore, `FRAMA` approximates the set cover problem by a greedy algorithm and maximizes ortholog coverage using as few fragments as possible. An alternative approach could adapt parameters of the assembly program or combine outputs of different assemblers. Especially k-mer length has an impact on the reconstruction of low expressed genes and incorporating assemblers using different k-mer lengths might be beneficial [41]. Also, recent developments to assess the quality of *de novo* transcriptome assemblies solely based on assembled contigs and RNA-seq data, provide opportunities to dynamically optimize assembly parameter in `FRAMA` [128]. On the downside, performing multiple assemblies based on different parameters and/or assemblers will increase the runtime dramatically. However, parameter estimation based on a subsample of the input RNA-seq data might prove sufficient.

The application of `FRAMA` was exemplified by assembling the transcriptome of the NMR using RNA-seq data obtained from multiple tissues. Human protein-coding transcripts were chosen as reference for annotation as these showed higher sequence similarity to the NMR than mouse transcripts. This could be caused by the long generation time in NMRs, potentially leading to a slower rate of molecular evolution compared to mice. The delivered mRNA catalogue comprises transcripts corresponding to 16,887 protein-coding genes and `FRAMA` was able to increase the length in the case of 3,488 genes by scaffolding and corrected transcript ends in respect to 4,774 genes. We assessed the quality of the mRNA catalogue based on available NMR draft genome sequences and respective publicly available gene annotations. Not

only did we demonstrate FRAMA's competitiveness with publicly available catalogues, we also reported striking differences in the quality of these resources, depending on the genome sequence used for reconstruction. Further, mRNA sequences delivered by FRAMA are potentially able to improve the hetgla2 genome sequence by spanning 1,695 gaps and adding potentially 408,293 bp novel sequence. Taken together, M1 (i) comprises the development of a software pipeline automating transcriptome assemblies and post-assembly tasks, (ii) reveals striking differences in publicly available mRNA catalogues for the NMR and (iii) provides an independent resource of NMR mRNA sequences as the basis for further studies.

4.2 Gene expression profiling in naked mole-rats

NMRs show a remarkable life history, characterized by an exceptional long lifespan and extreme resistance to age-associated deterioration. This enables to identify gene candidates involved in a long and healthy lifespan and, due to its close phylogenetic proximity, potential transfer of insights to humans. M2 focuses on the lifelong fertility of NMRs, which is a remarkable trait when considering that a single female functions as the colony's "breeding unit" throughout her life. In addition, despite the associated huge metabolic burden of pregnancy and lactation, breeding animals presumably live at least as long as non-breeders, contradicting the "disposable soma theory" of ageing [95]. To investigate this interesting trait, I analysed gene expression profiles of two experimental groups: (i) reproductively supressed non-breeders kept in their natal colonies and (ii) reproductively active breeders. For the breeding group, non-breeding siblings of (i) were removed from their natal colonies and paired with the opposite sex, thereby turning them into breeders. After the lactation period of the second litter, tissues were collected in both groups. Guinea pigs (GPs) (AnAge*: Cavia porcellus*, 12 years) were subjected to a similar experimental design (non-breeder: same sex housing, breeders: opposite sex housing) as a baseline for reproductive changes. Although not as well-established as mouse, GPs are a suitable reference in this study due to their (i) close phylogenetic relationship to mole-rats [67], (ii) similar reproductive biology to NMR [129,130] and (iii) similar endocrine profiles during pregnancy to humans [131]. Gene expression profiles in both species were investigated based on *de novo* assembled mRNAs by FRAMA (NMR as part of M1). To avoid loss of highly tissue-specific genes, RNA-seq data from the same set of tissues were used in assembly and gene expression profiling. Transcriptomes of both species were annotated using human protein-coding genes, which provides a comprehensive resource for functional annotations, essential for subsequent interpretation of identified DEGs.

To avoid any possibility of confusion during sample collection, exchange and sequencing of all 480 biological samples, I ensured the identity of tissue and species shortly after sequencing. Therefore, I aligned small subsets of each RNA-seq data set to highly tissue-specific marker genes, previously collected from VeryGene [132]. This enabled prompt reaction to wrong labelling prior to in-depth analyses on completion of sequencing. The detection of DEGs in such an experiment is influenced by several factors, including (i) number of biological

replicates (*i.e.* independent biological samples of an experimental group), (ii) biological variability between biological replicates, (iii) effect size between experimental groups, (iv) sequencing depth and (v) choice of DEG-calling tool [36–38,59]. Without preparatory studies, biological variability and effect sizes are usually unknown and to increase the likelihood of detecting potentially small effects within variable experimental groups, we chose a comparably high number of six biological replicates. Further, we aimed to identify DEGs rather than differences in isoform expression and accordingly chose a moderate sequencing depth of ~20 million single-end reads per RNA-seq sample [35–37]. A variety of different tools for DEG-calling are available and my choice for a specific tool was motivated by theoretical reflection and results obtained from permutation tests [48,49,59]. For instance, t-test was excluded, because random sampling by RNA-seq is approximable by Poisson rather than normal distribution [133], and non-parametric tests were excluded, because these require large numbers of biological replicates [48]. Appropriate candidates tools (`edgeR`, `DESeq`, `DESeq2` [134–136]), based on generalizations of the Poisson distribution, were subjected to permutation tests. These tests assessed sensitivity and specificity of each tool, under the assumption that detected DEGs in incorrectly assigned experimental groups are false positives. In contrast to `edgeR` and `DEseq`, which generally identified low numbers of DEGs, `DESeq2` showed the highest sensitivity in correctly grouped data sets while producing neglectable numbers of false positives in incorrect assignments. Additional gene filtering approaches (*e.g.* based on expression quantiles, low inner-group variance, high signal-to-noise ratio) across different thresholds prior to DEG-calling showed minor changes in `DESeq2`, indicating that its default filter mechanisms are robust and further filtering is not necessary [137].

I used `DEseq2` to determine differential expression (DE) tissue-wise between (i) sexes and (ii) reproductive statuses within each species and (iii) general differences across species. While DE within each species was assessed using respective mRNA catalogues, we additionally determined highly similar regions in NMR and GP orthologs for inter-species comparison. As `DESeq2` does not account for gene lengths, the latter approach avoids artefacts arising from different levels of gene completeness and additionally prevents comparing different transcript isoforms. After adjusting p-values for multiple testing (false discovery rate [138]), resulting DEGs were tested for enrichment in Gene Ontology [54] and Digital Ageing Atlas [57] using `GoMiner` [53] and Fisher's exact test, respectively.

Across all tissues in NMR non-breeders, we found only scattered instances of DEGs between sexes, in contrast to GP non-breeders and described DE in adult animals of other rodents and humans. Interestingly, pronounced sex-related DE manifested in NMR breeders, associated with massive changes to sex steroid metabolism. This supports previous studies indicating that sexual maturation in NMR non-breeder is delayed until transition into breeder. Further, these minor sex-related differences in non-breeding NMR fit their outwardly identical morphology and identical behaviour. Although status-related DE in females of both species is substantial, both species showed only minor overlaps in DEGs, indicating that physiological

adaptations to breeding proceed differently in both species. Potentially, this is caused by unique and targeted adaptations in female NMR breeder to their lifelong role as the colony's „breeding unit", accompanied, *e.g.* by severe pregancy-induced spine-elongation [139]. Interestingly, only NMRs showed significant enrichment of status-related DEGs in age-associated genes, indicating also differences in the impact of reproduction on lifespan in both species. Taken together, M2 (i) presents the transcriptional characterization of changes induced by reproduction in NMRs and GPs and (ii) provides a unique resource of extensive transcriptome data obtained from multiple tissues in breeders and non-breeders of NMRs and GPs.

4.3 Investigating positively selected genes in annual fishes

Initially and as part of the large collaborative study M3, the application of *de novo* assembled transcriptomes to identify PSGs was driven by the interest in phylogenetic approaches to detect genetic determinants of the short lifespan in *N. furzeri*. Besides positive selection analyses, M3 presents a milestone in establishing the *N. furzeri* as a model organism by providing a chromosome-scale draft genome sequence, including annotation of protein-coding genes and several classes of ncRNAs. Further ageing-related investigations included the identification of DEGs between young and old animals in multiple tissues. Interestingly, revealed DEGs are non-randomly distributed in the *N. furzeri* genome and DEGs in certain hotspots show significant functional association, suggesting co-regulation of functionally-associated temporally-regulated genes.

The short lifespan of *N. furzeri* evolved naturally from a longer-lived ancestor and presumably in adaptation to its transient habitat. In such an environment, gene variants promoting, *e.g.* fast growth and early sexual maturation could provide a selective advantage and become fixated by the process of positive selection. In consequence, positive selection analysis should provide useful pointers to genes and site-specific variants that have modulating effects towards a rapid life cycle and a short lifespan. To follow the idea, I used `FRAMA` to assembled and annotated transcriptomes of five annual fishes and the non-annual sister taxon (*Aphyosemion striatum)* based on RNA-seq data from brain tissues. The thorough *N. furzeri* annotation of protein-coding genes was used as reference for transcript annotation in all fishes. Unfortunately, mRNA catalogs were initially compromised by several misassembled contigs, mainly arising from misassemblies of highly similar paralogous genes. Presumably, this was caused by the teleost-specific genome duplication event [140]. Compared to previous tests in mammalian species, this additional genome duplication potentially challenged *de novo* assemblies by an increased transcriptome complexity caused by the high similarity of protein-coding paralogs and processed pseudogenes. However, catalogues were cleaned by demanding higher sequence similarity in the *N. furzeri* alignment step. This enabled PSG identification using the branch-site test to investigate terminal phylogenetic branches leading to (i) *N. furzeri*, as the shortest-lived annual fish, and (ii) *N. pieenaari*, faced with comparably short availability

of its habitat and showing a similar short lifespan. Despite the small number of identified PSGs (*N. furzeri* 7, *N. pienaari* 1), five PSGs in *N. furzeri* were DE between yong and old animals (two down and three up with age), suggesting functional adaptations to temporally-regulated genes. Further, both short-lived fish evolved different sequence adaptations to the same gene (inhibitor of DNA binding 3, PSG in *N. furzeri*), which is otherwise conserved in respective but longer-lived sister taxons. This indicates that parallel evolution of short lifespan in *N. furzeri* and *N. pieenari* is accompanied by changes to similar genes.

Notably, coincidentally with M3, Valenzano et al. [110] also investigated positive selection in *N. furzeri*, but reported substantially different results comprising almost 500 PSGs. However, these differences could be attributed to the choice of species in both studies [141]. In M3, we used previously uncharacterized annual fishes as close relatives to the *N. furzeri,* while Valenzano et al. used publicly available data and included the platyfish (*Xiphophorus maculatus*) as the closest relative. Thus, while we investigated PSGs mainly within the ~8 million years old *Nothobranchius* clade [142], Valenzano et al. investigated PSGs across species that separated ~70-50 million years ago [141]. This indicates that the majority of detected PSGs in Valenzano et al. likely reflect adaptations that predate the evolution of annualism and can also be found in longer-lived fish that separated along the phylogenetic path to annualism.

To provide a more extensive investigation of the evolution of annual life history and follow up the search for parallel evolution, we extended the set of non-annual species to increase the depth of the phylogeny and analysed deeper branches in the phylogenetic tree in M4. Again, we used the branch-site test that allows testing of ancestral sequences and accounts for variable selective pressures among codons and branches, enabling sensitive PSG detection [64]. In particular, we tested (i) the last common ancestor of annual fishes and two *Nothobranchius* clades, each leading to either short-lived fish (ii) *N. furzeri* or (iii) *N. pienaari*. While the first test should reveal adaptations that coincide with the manifestation of annual life history, the other two investigate possible parallel evolution in both *Nothobranchius* lineages. Functional enrichment analysis of detected PSGs revealed that each set of PSGs is significantly enriched for categories related to mitochondrial biogenesis. Interestingly, mitochondrial dysfunction is a hallmark of ageing and further overlaps in PSG by temporally-regulated DEGs from M3 as well as by PSGs in long-lived mammals, support our result that adaptations to mitochondrial biogenesis plays a major role in the evolution of lifespans and specifically in short lifespan of annual fishes. Further, the enrichment of mitochondrial biogenesis in PSGs of all three lineages suggests recurring positive selection events in this pathway, starting with the manifestation of annual life history in last common ancestor of annual fishes (4 genes) and continuing in both *Nothobranchius* lineages (other 9 genes in both branches). Remarkably, both *Nothobranchius* lineages show adapations in a common subset of PSGs also related to mitochondrial biogenesis and even encoding for different subunits of the same protein complex (mitochondrial RNA polymerase). This supports previous indications in M3 that parallel evolution of short lifespans in annual fishes is accompanied by changes in similar genes.

Notably, this analysis also revealed that positive selection of short lifespans influenced whole pathways and multiple subunits of protein complexes rather than unrelated genes across the transcriptome.

Although M3 and M4 demonstrated the power of positive selection analyses by shedding light on the evolution of short lifespans in annual fishes, the functional role of specific animo acid changes is usually unknown and could cause either mild or severe effects as well as increased or decreased protein functionality. Consequently, no conclusive link can be drawn from sequence adaptations to impacts on protein function and lifespan. Nevertheless, both studies present starting points for follow-up experiments, either based on extant or reconstructable ancestral gene variants. Such follow-up studies could comprise transgenic experiments in *N. furzeri* using CRISPR/Cas9 [143] to (i) functionally characterize PSGs with unknown function or (ii) assess the impact of site-specific adaptation on the whole organism and in respect to lifespan. For instance, the latter could be investigated by editing a PSG in *N. furzeri* to reflect an ortholog sequence of a longer-lived species. Under the assumption that the *N. furzeri* gene variant contributes to its short lifespan, such an approach potentially elongates its lifespan and thereby verifies the impact of site-specific changes on lifespan. Another approach could use crystallographic methods to compare the three dimensional shape of proteins that are encoded by different gene variants. This potentially elucidates the impact of amino acid changes on protein stability or on interactions with ligands and other proteins.

Both studies demonstrated the successful application of positive selection analyses. Nevertheless, there are alternative approaches to investigate protein-coding gene sequences apart from the comparably strict positive selection criteria. For instance, the search for accelerated protein evolution in long-lived mammalian lineages revealed longevity-associated gene candidates [23]. Interestingly, despite utilizing a different approach to identify selective pressure on genes, these candidates overlapped with positive selection studies [23]. Another study investigated predictions of functionally significant amino acid changes in one particular gene (hyaluron synthase 2) based on multiple sequence alignment of orthologs from different long- and short-lived mammals [144]. Notably, although pointing to potentially relevant amino acid changes in the NMR and other long-lived mole-rats, signals from the multiple species alignment indicated that this genes underlies negative rather than positive selection [144]. This demonstrates the power of functionally- compared to evolutionary-centred approaches, at least at single gene level. Finally, another study focused on the selection of gene expression rather than gene sequence in mammals and revealed, among other things, orthologs whose expression variation is associated with lifespan differences [145]. Nevertheless, researchers must keep in mind that adaptations are not necessarily linked to the trait under investigation and might reflect other adaptive processes. Regardless of the approach, RNA-seq of previously uncharacterized species and subsequent assembly and annotation by `FRAMA` offers an uncomplicated procedure that provides the basis for comparative studies.

4.4 Alternative technologies and future applications for transcriptome investigations

Microarrays (MAs) and 3rd-generation sequencing (3GS) technologies [146] provide alternative strategies to quantify and/or sequence transcripts in non-model organisms. However, in the following, I will argue that both technologies are not (yet) appropriate to investigate whole-transcriptomes, especially in non-model organisms, and discuss possible niche application.

For more than a decade, researchers relied on MAs and developed a wide range of matured workflows for data analysis [147]. MAs quantify the abundance of fluorecently labeled transcripts by hybridisation to complementary sequences (probes). Because of that, the design of MAs requires prior knowledge of transcript sequences and confines its application to model organisms with available MAs. Considering constantly improving transcript annotations in model organisms (Ensembl is updated every 3-6 months [98]), this approach as such is prone to errors caused by outdated information. In principle, the investigation of non-model organisms is possible by in-house construction of anonymous MAs [148]. Yet, this involves spotting complementary DNA libraries constructed from biological samples on to MAs, subsequent identification of candidate probes (*e.g.* by differential expression analyses) and, finally, their sequencing. Such „fishing-expeditions" are costly in terms of labour and time as well as limited to a comparably small number of short probes. Not only does RNA-seq offer near-complete snapshots of whole-transcriptomes in a single organism, it even offers to investigate a mixture of transcriptomes originating from multiple unknown species simultaneously. Such metatranscriptomes, particularly those obtained from gut microbiota, are becoming increasingly relevant in ageing research [149] and have recently been studied in *N. furzeri* [150].

The construction of MAs based on sequence knowledge obtained from RNA-seq data seems like the next obvious choice to investigate gene expression profiles. Yet, also in this respect RNA-seq outperforms MAs. For instance, although RNA-seq data might be dominated by a few highly expressed genes, consequently limiting the detection of lowly expressed genes, relative expression level among detected genes can be accurately quantified [34]. However, MAs suffer from technical problems in lowly (cross-hybridisation) as well as highly (signal saturation) expressed genes [151]. In addition, comparison of gene expression across a single MA is compromised by sequence-dependent binding affinities that influence signals [151]. Another important topic, also in respect to ageing and age-associated diseases [152], is the investigation of alternative splicing. Although appropiate MA design enables detection of alternative splicing, RNA-seq benefits from its single-base resolution and is significantly more sensitive [153]. Thus, while both technologies show low technical variability [133,154], RNA-seq is more reliable and more sensitive across a wider range of expression levels and transcripts. In this thesis, single-based resolution of RNA-seq has been the driving force, and apart from presented applications, offers the detection of allele-specific expression, single-nucleotide

polymorphisms, chimeras [121] and post-transcriptional regulation [155], which cannot be efficiently studied using MAs.

Nevertheless, RNA-seq suffers from its short read sequences, which hamper correct reconstruction of full-length transcripts even in genome-based approaches [156]. Meanwhile, 3GS technologies are gaining ground and generate full-length sequences of complete RNA molecules without the need for reconstruction [146]. Currently, low throughputs and high error rates of 3GS technologies hamper comprehensive and accurate transcript identification at low cost. Although, hybrid approaches successfully reduce sequencing costs by utilizing accurate but short 2GS reads to correct long but error prone 3GS [157–159], this approach depends on the availability of both sequencing generations. In principle, 3GS presents a perfectly suitable alternative to 2GS, but currently at a higher financial burden and wet laboratory workload. However, in the long run, 3GS technologies likely render 2GS technologies and computation sequence reconstruction obsolete.

Beyond research, measuring transcript expression is relevant for clinical application and could provide insights into patients' individual biology. Pathological patterns of gene expression and nucleotide variations have already been characterized for various diseases, including age-associated diseases [9,10,160]. Also, chronological age (elapsed time since birth) is an important factor to select appropiate therapeutics, but since individuals show large variations in the rate of ageing and different organs of a single individual age differently, the chronological age provides limited information about the patients' general health status [10]. Once mechanisms of ageing are well-characterized in humans, knowledge of genome sequence together with gene expression, which can both be obtained from easy accessible biological samples (*e.g.* blood, skin, saliva) or via biopsies from different organs, could provide valuable information of biological age and general health status. Besides enhancing selection of therapeutics, this potentially enables to predict the pace of age-associated deterioration and could suggest beneficial lifestyle interventions. Clinical application of all three technologies offer advantages. MAs are already used in clinics [161] and might still be appropriate for future application as long as required information is confined to a comparably small set of marker genes. However, unbiased RNA analysis by 2GS and 3GS in combination with whole-genome sequencing offers the possibility, *e.g.* to identify malignant private mutations, allele-specific expression or easily investigate more complex datasets (microbiota compositions) [161]. An unprecedented and practical development within the 3GS sector are portable devices. These can be read by laptops and offer in-field application, *e.g.* in developing countries or after catastrophes [162]. Eventually, clinical use also depends considerably on (i) associated costs, including personnel (*e.g.* laboratory assistant, bioinformatician, biostatistician) as well as acquisition and maintenance of platforms, (ii) fast and automatable workflows and (iii) comprehensive *in silico* analyses to predict phenotypic consequences from molecular variations, give reliable prognosis and individualise pharmacological treatments and/or lifestyle.

4.5 Conclusion and Outlook

As part of this thesis, I developed `FRAMA`, a software framework that automatically assembles and annotates transcriptomes from data obtained by RNA-seq. This approach is particularly useful and effective in non-model organisms without availability of a sufficiently well-assembled and well-annotated genome reference. I incorporated custom-build steps in `FRAMA` to attenuate common issues in *de novo* transcriptome assembly by leveraging ortholog information and sequence characteristics of assembled contigs. Thereby, `FRAMA` offers non-experts easy access into the transcriptional landscape of non-model organisms and enables downstream analyses in their field of expertise. Application and quality assessment of `FRAMA` was exemplified using the NMR, a promising non-model organism in ageing research. The convenience of delivered mRNA catalogues for downstream analyses was demonstrated by two different but commonly used *in silico* analyses. These analyses were performed in promising non-model organisms in ageing research. mRNA catalogues and molecular characterizations presented in this thesis provide valuable resources for further research and steps toward the integration of these non-model organisms into molecular research, regardless whether *in silico*, *in vitro* or *in vivo*.

Gene expression analyses in NMRs and GPs characterized the molecular patterns associated with reproductive status in both species. Although insights into ageing were limited by the availability of one age cohort, changes in reproductive status are associated with significant enrichment of age-associated genes only in NMRs. This indicates differences in the impact of reproduction on ageing in long-lived NMRs and shorter-lived GPs. The extensive accumulation and analyses of transcriptomes from a variety of different tissues and two eusocial castes, together with the provided mRNA catalogues for both species offer valuable resources for future investigations.

Analyses of positive selection in annual fishes shed light on the evolution of short lifespans by revealing that positive selection recurrently shaped genes encoding for components of mitochondrial biogenesis. Adaptations in these components were dated back to the last common ancestor of annual fishes and continued in parallel in two analysed *Nothobranchius* lineages. These naturally evolved adaptations provide useful pointers for further wet laboratory experiments that could investigate potential lifespan modulating effects of extant and extinct gene variants. For this purpose, *N. furzeri* with its extremely short lifespan, typical signs of ageing and established genetic manipulation by CRISPR/Cas9 provides a reasonable model system.

Presented *in silico* analyses using non-model organisms provided new insights into the evolution of short lifespans, reproductive biology and ageing. Although final conclusions from such *in silico* studies are limited, the presented results indicate the adequateness of selected non-model organisms for ageing research and the revealed candidate genes and pathways provide useful pointers for further follow-up studies. For instance, candidate genes could

narrow down genome-wide association studies in humans. Such studies investigate the molecular basis of ageing by searching for genetic variants that are associated with long lifespans and healthy ageing, *e.g.* by comparing frequencies of genetic variants of a long-lived and healthy population (*e.g.* centenarians) to the general population. However, such large-scale studies are statistically challenging (*e.g.* enormous amount of variants, confounding effects, multiple testing problem) and appropiate candidate genes could aid in restricting the search space. Gene candidates could also be experimentally examined for potential lifespan modulating effects in model organisms, *e.g.* by manipulating gene sequences using CRISPR/Cas9 or by manipulating gene expression using RNA interference. Moreover, the annotation of CDSs by `FRAMA` provides a reference for protein expression measurements by mass-spectrometry or for *in silico* protein structure predictions. Finally, the refinement of mRNA ends offers a reliable basis to identify potential micoRNA targets or to investigate DNA methylation patterns in these regions. Taken together, regardless of the concrete approach, `FRAMA` delivers valuable sequence resources for previously uncharacterized non-model organisms and simplifies the process of transcriptome characterization by uncomplicated and cost-effective RNA-seq. This offers especially non-experts easier and faster access to a wide range of downstream analyses.

Although `FRAMA` provides a simplification of the mRNA reconstruction process, its execution via command line and required minor modifications to a configuration file might act as a deterrent to users. The usability could be enhanced by integrating `FRAMA` into scientific workflow platforms such as `Galaxy` [163]. The provided graphical user interface would offer biologists without computational background easier access to and configuration of `FRAMA`. Additionally, this would enable straight integration into up- and downstream workflows offered by other software packages.

5 References

[1] Jones, O. R. *et al.* Diversity of ageing across the tree of life. *Nature* **505,** 169–173 (2013).

[2] Kirkwood, T. B. & Austad, S. N. Why do we age? *Nature* **408,** 233–8 (2000).

[3] Oeppen, J. Demography: Enhanced: Broken Limits to Life Expectancy. *Science* **296,** 1029–1031 (2002).

[4] Tuljapurkar, S., Li, N. & Boe, C. A universal pattern of mortality decline in the G7 countries. *Nature* **405,** 789–792 (2000).

[5] Vaupel, J. W. Biodemography of human ageing. *Nature* **464,** 536–542 (2010).

[6] Hayflick, L. Biological aging is no longer an unsolved problem. *Ann. N. Y. Acad. Sci.* **1100,** 1–13 (2007).

[7] Höhn, A. *et al.* Happily (n)ever after: Aging in the context of oxidative stress, proteostasis loss and cellular senescence. *Redox Biol.* **11,** 482–501 (2017).

[8] Kaeberlein, M. Longevity and aging. *F1000Prime Rep.* **5,** R132 (2013).

[9] López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The Hallmarks of Aging. *Cell* **153,** 1194–1217 (2013).

[10] Karasik, D., Demissie, S., Cupples, L. A. & Kiel, D. P. Disentangling the genetic determinants of human aging: biological age as an alternative to the use of survival measures. *J. Gerontol. A. Biol. Sci. Med. Sci.* **60,** 574–87 (2005).

[11] Tacutu, R. *et al.* Human Ageing Genomic Resources: Integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Res.* **41,** D1027–D1033 (2013).

[12] Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17,** 333–51 (2016).

[13] Hu, J., Adar, S., Selby, C. P., Lieb, J. D. & Sancar, A. Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. *Genes Dev.* **29,** 948–960 (2015).

[14] Hodges, E. *et al.* Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* **39,** 1522–1527 (2007).

[15] Zimin, A. V. *et al.* Mis-assembled 'segmental duplications' in two versions of the Bos taurus genome. *PLoS One* **7,** (2012).

[16] Fierst, J. L. Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. *Front. Genet.* **6,** 220 (2015).

[17] Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13,** 329–342 (2012).

[18] Alkan, C., Sajjadian, S. & Eichler, E. E. Limitations of next-generation genome sequence assembly. *Nat. Methods* **8,** 61–65 (2011).

[19] Chain, P. S. G. *et al.* Genome Project Standards in a New Era of Sequencing. *Science* **326,** 236–237 (2009).

[20] Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).

[21] Baumgart, M. *et al.* Longitudinal RNA-Seq Analysis of Vertebrate Aging Identifies Mitochondrial Complex I as a Small-Molecule-Sensitive Modifier of Lifespan. *Cell Syst.* **2,** 122–32 (2016).

[22] Longo, V. D. & Kennedy, B. K. Sirtuins in Aging and Age-Related Disease. *Cell* **126,** 257–268 (2006).

[23] Li, Y. & de Magalhães, J. P. Accelerated protein evolution analysis reveals genes and pathways associated with the evolution of mammalian longevity. *Age (Omaha).* **35,** 301–314 (2013).

[24] Roux, J. *et al.* Patterns of Positive Selection in Seven Ant Genomes. *Mol. Biol. Evol.* **31,** 1661–1685 (2014).

[25] ENCODE Project Consortium *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).

[26] Smith-Vikos, T. & Slack, F. J. MicroRNAs and their roles in aging. *J. Cell Sci.* **125,** 7 LP-17 (2012).

[27] Pang, K. & Lenhard, B. Antisense Transcription in the Mammalian Transcriptome. *Science* **309,** 1564–1566 (2005).

[28] Carninci, P. *et al.* The Transcriptional Landscape of the Mammalian Genome. *Science* **309,** 1559–1563 (2005).

[29] Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467–70 (1995).

[30] Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–80 (2005).

[31] Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26,** 1135–1145 (2008).

[32] Sultan, M. *et al.* A simple strand-specific RNA-Seq library preparation protocol combining the Illumina TruSeq RNA and the dUTP methods. *Biochem. Biophys. Res. Commun.* **422,** 643–646 (2012).

[33] Oshlack, A. & Wakefield, M. J. Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct* **4,** 14 (2009).

[34] Marinov, G. K. *et al.* From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Res.* **24,** 496–510 (2014).

[35] Francis, W. R. *et al.* A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. *BMC Genomics* **14,** 167 (2013).

[36] Liu, Y., Zhou, J. & White, K. P. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics* **30,** 301–304 (2014).

[37] Vijay, N., Poelstra, J. W., Künstner, A. & Wolf, J. B. W. Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Mol. Ecol.* **22,** 620–634

(2013).

[38] Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: a matter of depth. *Genome Res.* **21,** 2213–23 (2011).

[39] Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29,** 644–652 (2011).

[40] Martin, J. a & Wang, Z. Next-generation transcriptome assembly. *Nat. Rev. Genet.* **12,** 671–682 (2011).

[41] Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28,** 1086–1092 (2012).

[42] Fitch, W. M. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19,** 99–113 (1970).

[43] Conrad, B. & Antonarakis, S. E. Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annu. Rev. Genomics Hum. Genet.* **8,** 17–35 (2007).

[44] Tatusov, R. L., Koonin, E. V & Lipman, D. J. A Genomic Perspective on Protein Families. *Science* **278,** 631 LP-637 (1997).

[45] Moreno-Hagelsieb, G. & Latimer, K. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics* **24,** 319–324 (2008).

[46] Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131,** 281–285 (2012).

[47] Mortazavi, A., Williams, B. a, McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5,** 621–628 (2008).

[48] Seyednasrollah, F., Laiho, A. & Elo, L. L. Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief. Bioinform.* **16,** 1–12 (2013).

[49] Rapaport, F. *et al.* Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* **14,** R95 (2013).

[50] Fang, X. *et al.* Genome-wide adaptive complexes to underground stresses in blind mole rats Spalax. *Nat. Commun.* **5,** 3966 (2014).

[51] Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11,** R14 (2010).

[52] Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14,** 7 (2013).

[53] Zeeberg, B. R. *et al.* GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* **4,** R28 (2003).

[54] Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25,** 25–29 (2000).

[55] Ogata, H. *et al.* KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **27,** 29–34 (1999).

[56] Fabregat, A. *et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Res.* **44,** D481–

D487 (2016).

[57] Craig, T. *et al.* The Digital Ageing Atlas: integrating the diversity of age-related changes into a unified resource. *Nucleic Acids Res.* **43,** D873--D878 (2014).

[58] Hühne, R., Thalheim, T. & Sühnel, J. AgeFactDB - The JenAge Ageing Factor Database - Towards data integration in ageing research. *Nucleic Acids Res.* **42,** 892–896 (2014).

[59] Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17,** 13 (2016).

[60] Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D. & Woolf, P. J. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* **10,** 161 (2009).

[61] Yang, Z. & Bielawski, J. P. R. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15,** 496–503 (2000).

[62] Sahm, A., Bens, M., Platzer, M. & Szafranski, K. PosiGene: automated and easy-to-use pipeline for genome-wide detection of positively selected genes. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkx179

[63] Zhang, J. Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. *Mol. Biol. Evol.* **22,** 2472–2479 (2005).

[64] Yang, Z. & dos Reis, M. Statistical properties of the branch-site test of positive selection. in *Molecular biology and evolution* **28,** 1217–28 (2011).

[65] Keane, M. *et al.* Insights into the Evolution of Longevity from the Bowhead Whale Genome. *Cell Rep.* **10,** 112–122 (2015).

[66] Morgan, C. C. *et al.* Molecular adaptation of telomere associated genes in mammals. *BMC Evol. Biol.* **13,** 251 (2013).

[67] Fang, X. *et al.* Adaptations to a Subterranean Environment and Longevity Revealed by the Analysis of Mole Rat Genomes. *Cell Rep.* **8,** 1354–1364 (2014).

[68] Magalhaes, J. P. d., Costa, J. & Church, G. M. An Analysis of the Relationship Between Metabolism, Developmental Schedules, and Longevity Using Phylogenetic Independent Contrasts. *Journals Gerontol. Ser. A Biol. Sci. Med. Sci.* **62,** 149–160 (2007).

[69] Buffenstein, R. The naked mole-rat: a new long-living model for human aging research. *J. Gerontol. A. Biol. Sci. Med. Sci.* **60,** 1369–77 (2005).

[70] Fontana, L., Partridge, L. & Longo, V. D. Extending healthy life span--from yeast to humans. *Science* **328,** 321–6 (2010).

[71] Pearl, R. The rate of living. (1928).

[72] Keller, L. & Genoud, M. Extraordinary lifespans in ants: a test of evolutionary theories of ageing. *Nature* **389,** 958–960 (1997).

[73] Promislow, D. E. L. & Harvey, P. H. Living fast and dying young: A comparative analysis of life-history variation among mammals. *J. Zool.* **220,** 417–437 (1990).

[74] Fischer, K. E. & Austad, S. N. The development of small primate models for aging research. *ILAR J.* **52,** 78–88 (2011).

[75] Looso, M., Borchardt, T., Krüger, M. & Braun, T. Advanced Identification of Proteins in

Uncharacterized Proteomes by Pulsed in Vivo Stable Isotope Labeling-based Mass Spectrometry. *Mol. Cell. Proteomics* **9,** 1157–1166 (2010).

[76] Novina, C. D. & Sharp, P. A. The RNAi revolution. *Nature* **430,** 161–4 (2004).

[77] Cong, L. *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* **339,** 819–823 (2013).

[78] Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126,** 663–676 (2017).

[79] Austad, S. N. Comparative Biology of Aging. *Journals Gerontol. Ser. A Biol. Sci. Med. Sci.* **64A,** 199–201 (2009).

[80] Szigeti, B. *et al.* OpenWorm: an open-science approach to modeling Caenorhabditis elegans. *Frontiers in Computational Neuroscience* **8,** 137 (2014).

[81] Drasdo, D. *et al.* The virtual liver: state of the art and future perspectives. *Arch. Toxicol.* **88,** 2071–2075 (2014).

[82] Jacob, F. Evolution and tinkering. *Science* **196,** 1161–6 (1977).

[83] Seok, J. *et al.* Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 3507–12 (2013).

[84] Austad, S. N. Diverse aging rates in metazoans: Targets for functional genomics. *Mech. Ageing Dev.* **126,** 43–49 (2005).

[85] Gorbunova, V., Bozzella, M. J. & Seluanov, A. Rodents for comparative aging studies: from mice to beavers. *Age (Dordr).* **30,** 111–9 (2008).

[86] Abele, D., Brey, T. & Philipp, E. Bivalve models of aging and the determination of molluscan lifespans. *Exp. Gerontol.* **44,** 307–15 (2009).

[87] Han, B. A., Schmidt, J. P., Bowden, S. E. & Drake, J. M. Rodent reservoirs of future zoonotic diseases. *Proc. Natl. Acad. Sci.* **112,** 7039–7044 (2015).

[88] Lucas-Sánchez, A., Almaida-Pagán, P. F., Mendiola, P. & de Costa, J. Nothobranchius as a model for aging studies. A review. *Aging Dis.* **5,** 281–291 (2014).

[89] Buffenstein, R. Negligible senescence in the longest living rodent, the naked mole-rat: Insights from a successfully aging species. *J. Comp. Physiol. B Biochem. Syst. Environ. Physiol.* **178,** 439–445 (2008).

[90] Seluanov, A. *et al.* Hypersensitivity to contact inhibition provides a clue to cancer resistance of naked mole-rat. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 19352–7 (2009).

[91] Edrey, Y. H., Hanes, M., Pinto, M., Mele, J. & Buffenstein, R. Successful aging and sustained good health in the naked mole rat: a long-lived mammalian model for biogerontology and biomedical research. *ILAR J.* **52,** 41–53 (2011).

[92] Jarvis, J. U. Eusociality in a mammal: cooperative breeding in naked mole-rat colonies. *Science* **212,** 571–3 (1981).

[93] Hochberg, M. E., Noble, R. J. & Braude, S. A Hypothesis to Explain Cancers in Confined Colonies of Naked Mole Rats. *bioRxiv* (2016).

[94] Dammann, P., Šumbera, R., Maßmann, C., Scherag, A. & Burda, H. Extended longevity

of reproductives appears to be common in Fukomys mole-rats (Rodentia, Bathyergidae). *PLoS One* **6,** 2–8 (2011).

[95] Kirkwood, T. B. Evolution of ageing. *Nature* **270,** 301–304 (1977).

[96] Kim, E. B. *et al.* Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature* **479,** 223–227 (2011).

[97] Keane, M. *et al.* The Naked Mole Rat Genome Resource: facilitating analyses of cancer and longevity-related adaptations. *Bioinformatics* **30,** 3558–3560 (2014).

[98] Aken, B. L. *et al.* The Ensembl gene annotation system. *Database* **2016,** baw093 (2016).

[99] Mignone, F., Gissi, C., Liuni, S. & Pesole, G. Untranslated regions of mRNAs. *Genome Biol.* **3,** reviews0004.1-0004.10 (2002).

[100] Tian, X. *et al.* High-molecular-mass hyaluronan mediates the cancer resistance of the naked mole rat. *Nature* **499,** 346–349 (2013).

[101] Yang, Z., Zhang, Y. & Chen, L. Investigation of anti-cancer mechanisms by comparative analysis of naked mole rat and rat. *BMC Syst. Biol.* **7,** S5 (2013).

[102] Thieme, R. *et al.* Analysis of Alpha-2 Macroglobulin from the Long-Lived and Cancer-Resistant Naked Mole-Rat and Human Plasma. *PLoS One* **10,** e0130470 (2015).

[103] Yu, C. *et al.* RNA sequencing reveals differential expression of mitochondrial and oxidation reduction genes in the long-lived naked mole-rat when compared to mice. *PLoS One* **6,** e26729 (2011).

[104] Genade, T. *et al.* Annual fishes of the genus Nothobranchius as a model system for aging research. *Aging Cell* **4,** 223–33 (2005).

[105] Valdesalici, S. & Cellerino, A. Extremely short lifespan in the annual fish Nothobranchius furzeri. *Proc. R. Soc. B Biol. Sci.* **270,** S189–S191 (2003).

[106] Platzer, M. & Englert, C. Nothobranchius furzeri: A Model for Aging Research and More. *Trends in Genetics* **32,** 543–552 (2016).

[107] Terzibasi, E., Valenzano, D. R. & Cellerino, A. The short-lived fish Nothobranchius furzeri as a new model system for aging studies. *Exp. Gerontol.* **42,** 81–89 (2007).

[108] Terzibasi, E. *et al.* Effects of dietary restriction on mortality and age-related phenotypes in the short-lived fish Nothobranchius furzeri. *Aging Cell* **8,** 88–99 (2009).

[109] Reichwald, K. *et al.* Insights into Sex Chromosome Evolution and Aging from the Genome of a Short-Lived Fish. *Cell* **163,** 1527–38 (2015).

[110] Valenzano, D. R. *et al.* The African Turquoise Killifish Genome Provides Insights into Evolution and Genetic Architecture of Lifespan. *Cell* **163,** 1539–54 (2015).

[111] Mouse Genome Sequencing Consortium *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–62 (2002).

[112] Gibbs, R. A. *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521 (2004).

[113] Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

[114] Koepfli, K.-P., Paten, B. & O'Brien, S. J. The Genome 10K Project: A Way Forward. *Annu. Rev. Anim. Biosci.* **3,** 57–111 (2015).

[115] Yelin, R. *et al.* Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol.* **21,** 379–86 (2003).

[116] Tian, B., Hu, J., Zhang, H. & Lutz, C. S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* **33,** 201–212 (2005).

[117] Veeramachaneni, V., Makałowski, W., Galdzicki, M., Sood, R. & Makałowska, I. Mammalian overlapping genes: The comparative perspective. *Genome Res.* **14,** 280–286 (2004).

[118] Houseley, J. & Tollervey, D. Apparent Non-Canonical Trans-Splicing Is Generated by Reverse Transcriptase In Vitro. *PLoS One* **5,** e12271 (2010).

[119] Frenkel-Morgenstern, M. *et al.* Chimeras taking shape: Potential functions of proteins encoded by chimeric RNA transcripts. *Genome Res.* **22,** 1231–1242 (2012).

[120] Herai, R. H. & Yamagishi, M. E. B. Detection of human interchromosomal trans-splicing in sequence databanks. *Brief. Bioinform.* **11,** 198–209 (2010).

[121] Kumar, S., Vo, A. D., Qin, F. & Li, H. Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. *Sci. Rep.* **6,** 21597 (2016).

[122] Davidson, N. M., Majewski, I. J. & Oshlack, A. JAFFA: High sensitivity transcriptome-focused fusion gene detection. *Genome Med.* **7,** 43 (2015).

[123] Morgan, M., Iaconcig, A. & Muro, A. F. Identification of 3′ gene ends using transcriptional and genomic conservation across vertebrates. *BMC Genomics* **13,** 708 (2012).

[124] Griebel, T. *et al.* Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res.* **40,** 10073–10083 (2012).

[125] Kodzius, R. *et al.* CAGE: cap analysis of gene expression. *Nat. Methods* **3,** 211–222 (2006).

[126] Ng, P. *et al.* Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat. Methods* **2,** 105–11 (2005).

[127] Pelechano, V., Wei, W. & Steinmetz, L. M. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497,** 127–131 (2013).

[128] Li, B. *et al.* Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biol.* **15,** 553 (2014).

[129] Roellig, K., Drews, B., Goeritz, F. & Hildebrandt, T. B. The long gestation of the small naked mole-rat (Heterocephalus glaber Rüppell, 1842) studied with ultrasound biomicroscopy and 3D-ultrasonography. *PLoS One* **6,** e17744 (2011).

[130] Peaker, M. & Taylor, E. Sex ratio and litter size in the guinea-pig. *Reproduction* **108,** 63–67 (1996).

[131] Dyson, R. M., Palliser, H. K., Kelleher, M. A., Hirst, J. J. & Wright, I. M. R. The guinea pig as an animal model for studying perinatal changes in microvascular function. *Pediatr. Res.* **71,** 20–4 (2012).

[132] Yang, X. *et al.* VeryGene: linking tissue-specific genes to diseases, drugs, and beyond for

knowledge discovery. *Physiol. Genomics* **43,** 457–460 (2011).

[133] Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18,** 1509–1517 (2008).

[134] Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2009).

[135] Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

[136] Love, M. I. *et al.* Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).

[137] Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 9546–9551 (2010).

[138] Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57,** 289–300 (1995).

[139] Dengler-Crish, C. M. & Catania, K. C. Cessation of reproduction-related spine elongation after multiple breeding cycles in female naked mole-rats. *Anat. Rec.* **292,** 131–137 (2009).

[140] Jaillon, O. *et al.* Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* **431,** 946–957 (2004).

[141] Sahm, A., Platzer, M. & Cellerino, A. Outgroups and Positive Selection: The Nothobranchius furzeri Case. *Trends Genet.* **32,** 523–525 (2016).

[142] Dorn, A., Musilová, Z., Platzer, M., Reichwald, K. & Cellerino, A. The strange case of East African annual fishes: aridification correlates with diversification for a savannah aquatic group? *BMC Evol. Biol.* **14,** 210 (2014).

[143] Harel, I., Valenzano, D. R. & Brunet, A. Efficient genome engineering approaches for the short-lived African turquoise killifish. *Nat. Protoc.* **11,** 2010–2028 (2016).

[144] Faulkes, C. G., Davies, K. T. J., Rossiter, S. J. & Bennett, N. C. Molecular evolution of the hyaluronan synthase 2 gene in mammals: implications for adaptations to the subterranean niche and cancer resistance. *Biol. Lett.* **11,** (2015).

[145] Fushan, A. A. *et al.* Gene expression defines natural changes in mammalian lifespan. *Aging Cell* **14,** 352–365 (2015).

[146] Oikonomopoulos, S., Wang, Y. C., Djambazian, H., Badescu, D. & Ragoussis, J. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Sci. Rep.* **6,** 31602 (2016).

[147] Reimers, M. Making Informed Choices about Microarray Data Analysis. *PLoS Comput. Biol.* **6,** e1000786 (2010).

[148] Hegarty, M. J. *et al.* Development of anonymous cDNA microarrays to study changes to the Senecio floral transcriptome during hybrid speciation. *Mol. Ecol.* **14,** 2493–510 (2005).

[149] O'Toole, P. W. & Jeffery, I. B. Gut microbiota and aging. *Science* **350,** 1214 LP-1215 (2015).

[150] Smith, P. *et al.* Regulation of Life Span by the Gut Microbiota in The Short-Lived African

Turquoise Killifish. *bioRxiv* (2017).

[151] Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLoS One* **9,** e78644 (2014).

[152] Heintz, C. *et al.* Splicing factor 1 modulates dietary restriction and TORC1 pathway longevity in C. elegans. *Nature* **541,** 1–21 (2016).

[153] Sultan, M. *et al.* A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science* **321,** 956–960 (2008).

[154] Bryant, P. A., Smyth, G. K., Robins-Browne, R. & Curtis, N. Technical Variability Is Greater than Biological Variability in a Microarray Experiment but Both Are Outweighed by Changes Induced by Stimulation. *PLoS One* **6,** 1–8 (2011).

[155] Gaidatzis, D., Burger, L., Florescu, M. & Stadler, M. B. Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. *Nat Biotech* **33,** 722–729 (2015).

[156] Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10,** 1177–1184 (2013).

[157] Au, K. F., Underwood, J. G., Lee, L. & Wong, W. H. Improving PacBio long read accuracy by short read alignment. *PLoS One* **7,** e46679 (2012).

[158] Koren, S. *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30,** 693–700 (2012).

[159] Goodwin, S. *et al.* Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res.* **25,** 1750–1756 (2015).

[160] Cooper-Knock, J. *et al.* Gene expression profiling in human neurodegenerative disease. *Nat. Rev. Neurol.* **8,** 518–30 (2012).

[161] Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D. & Craig, D. W. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat Rev Genet* **17,** 257–271 (2016).

[162] Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530,** 228–232 (2016).

[163] Goecks, J., Nekrutenko, A. & Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11,** R86 (2010).

6 Abbreviation

2GS
: 2nd-generation sequencing.

3GS
: 3rd-generation sequencing.

AnAge
: The Animal Ageing and Longevity Database.

CDS
: protein-coding sequence.

CRISPR/Cas9
: clustered regularly interspaced short palindromic repeats/CRISPR asscociated protein 9.

DE
: differential expression.

DEG
: differentially expressed gene.

dN
: rate of substitutions at non-silent sites.

DNA
: deoxyribonucleic acid.

dS
: rate of substitutions at silent sites.

F. mechowii
: *Fukomys mechowii.*

FRAMA
: From RNA-seq to annotated mRNA assemblies.

GP
: guinea pig.

mRNA
: messenger RNA.

N. furzeri
: *Nothobranchius furzeri.*

N. pieenari
: *Nothobranchius pieenari.*

ncRNA
: non-coding RNA.

NMR
: naked mole-rat.

PSG
: positively selected gene.

RNA
: ribonucleic acid.

RNA-seq
: RNA sequencing.

UTR
: untranslated region.

7 Appendix

The following sections correspond to the manuscripts presented in section 3. Each subsection describes material deposited on the enclosed CD-ROM.

7.1 Manuscript 1 (M1)

7.1.1 **M1_MainDocument.pdf**

Digital version of manuscript M1.

7.1.2 **M1_SupplementTables.xlsx**

Additional File 1

Table S1. List of external software.

Table S2. NMR transcript data set TCUR, and orthologous transcripts from human, mouse and guinea pig. Multi-species mRNA alignments were constructed independently from those described in the main text, using the sequence database entries as listed.

Table S3. Naked mole-rat samples for strand- specific RNA-seq, and produced RNA-seq data.

Table S4. Pairwise transcript sequence identities between NMR and related mammals. The analysis is based on 142 multiple sequence alignments of the CDSs of NMR, guinea pig, human and mouse (as listed in Additional file 1: Table S2). Identity values were computed based on gap-masked alignments.

Table S5. Statistics of the transcriptome data produced by Trinity (column "transcript assembly") and subsequently processed using FRAMA (column "transcript catalog").

Table S6. CEGMA results on transcriptome datasets. As defined by CEGMA, 'complete proteins' are recovered with >70 % in comparison to CEGMA's core proteins. 'Partial proteins' additionally include proteins, which exceed a certain alignment score threshold. CEGMAs software components were used as suggested: geneid (v1.4), genewise (wise2.2.3-rc7), hmmer (HMMER 3.0), NCBI BLAST+ (2.2.25).

Table S7. Source of transcript sequence sets and underlying input data. Table S8: Transcript-genome alignment statistics of curated dataset (TCUR) in hetgla1. The alignments comprise 1473 well-aligned blocks and 81 unaligned or mismatching blocks. Transcripts show 99.9 % average identity within well-aligned blocks.

Table S9. Transcript-genome alignment of curated dataset (TCUR) in hetgla2. The alignments comprise 1525 well-aligned blocks and 16 unaligned or mismatching blocks. Transcripts show 99.9 % average identity within well-aligned blocks.

Table S10. Correspondence of gene symbols between transcript sets. The evaluation considered gene loci overlapping in the hetgla2 genome sequence, where all transcript-genome alignments of a gene were considered to define the gene locus. Only genes with ascertained function (non-LOC gene symbol) were compared.

Table S11. Accession numbers of sequences that are shown in the genome-based transcript map (hetgla2, scaffold JH602043; Fig. 4). Accession numbers for each sequence are listed in the same order as shown in Fig. 4 (from top to bottom).

7.1.3 **M1_SupplementNotes.docx**

Additional File 2

Figure S1. Multiple sequence alignments of CALM1, CALM2 and CALM3 in human and NMR. (A) protein coding sequence (B) protein sequence. All protein coding sequences encode for the same protein sequence. The nucleotide identity between human and NMR orthologs is higher (97 % CALM1, 98 % CALM2, 95 % CALM3) than the intra-species paralog identity (e.g., human CALM1/CALM2 highest identity with 85 %).

Figure S2. Recovery of transcripts is predicted by the expression level in the reference organism - (A) human liver, (B) human kidney. Public human Illumina RNA-seq data were obtained from the Short Read Archive at the EBI (accessions ERR030895 and ERR030893, respectively). Box plots show the human expression levels in log-scale FPKM; zero FPKM values were initially transformed to 0.80 times the lowest finite value. Human genes are displayed in three groups: all genes (“all”), genes recovered as orthologous NMR transcripts (“recovered”), and genes missing in the NMR transcript catalog (“missing”). Boxes enclose the data ranges of the central two-third quantiles, and central bars indicate the data medians. Note that the group-wise medians are significantly influenced by the fraction of zero-expression genes; these are 12 % in the liver-recovered group, 56 % in the liver-missing group, 7% in the kidney-recovered group, and 49 % in the kidney-missing group.

Figure S3. Results of structural agreement between transcript sets. The evaluation considered gene loci overlapping in the hetgla2 genome. Each transcript set was compared to TCUR.

Figure S4. Classification of exons into four categories (exact, overlapping, missing and wrong) based on the reference transcript model. Exact exons share the same boundaries. Overlapping exons share base pairs, but not necessarily any boundary. Exons only present in the predicted transcript model are classified as wrong. Exons only present in the reference transcript model are classified.

7.2 Manuscript 2 (M2)

7.2.1 **M2_MainDocument.pdf**

Digital version of manuscript M2.

7.2.2 **M2_Supplementary_Tables_S1-28.xlsx**

S1. Age at death for NMRs and GPs.

S2. Number of analyzed biological replicates per group.

S3. Reason for exclusion of samples from further analysis.

S4. Number of uniquelly aligned RNA-seq reads.

S5. Mean pairwise Pearson correlation coeffiencents between biological replicates in (A) naked mole-rat and (B) guinea pig.

S6. Number of differentially expressed genes (FDR < 0.01) between NMR and GP (A) without logFC threshold and (B) with logFC threshold (|logFC| > 2)

S7. DEGs in cross-species comparison that show an overlap with DAA.

S8. Gene set enrichment analysis for cross-species DEGs (FDR < 0.01, |logFC| > 2) that are age-related (DAA).

S9. Number of sex-related (female vs. male) differentially expressed genes in (A) non-breeder and (B) breeder (FDR < 0.01).

S10. Functional enrichment analysis of 790 DEGs shared between GP-B-FvM and GP-N-FvM.

S11. Functional gene set enrichment analysis of DEGs in NMR-B-FvM.

S12. Gene description of sex-related DEGs in NMR-N-FvsM.

S13. Number of status-related (breeder vs.non-breeder) differentially expressed genes in (A) females and (B) males (FDR < 0.01).

S14. Functional gene set enrichment analysis of DEGs in Ovr of NMR-F-BvsN.

S15. Functional gene set enrichment analysis of DEGs in Adr of NMR-F-BvsN.

S16. Functional gene set enrichment analysis of DEGs in Cer of NMR-F-BvsN.

S17. DEGs in intersection between NMR-F-BvsN and NMR-M-BvsN.

S18. Functional gene set enrichment analysis of DEGs in Tes of NMR-M-BvsN.

S19. Functional enrichment analysis in cellular components of DEGs in Tes NMR-M-BvsN.

S20. Functional gene set enrichment analysis of DEGs in Skn of NMR-M-BvsN.

S21. Proportion of mitochondrial transcriptonal output.

S22. Differentially expressed nuclear genes (FDR<0.05) associated with mitochondrial respiratory chain complexes in Tes and Skn of NMR-M-BvsN.

S23. Antioxidant enzymes differentially expressed in Tes and Skn of NMR-M-BvsN (FDR<0.05)

S24. Enrichment of age-related genes (Digital Ageing Atlas) in status-related DEGs. Only tissues having at least 50 DEGs were tested for enrichment.

S25. Shared status-related genes between NMR and GP.

S26. Functional enrichment analysis of the non-redundant set of ageing-related 20%-quantiles that show the greatest interspecies difference in males (Skn, Tes).

S27. Functional enrichment analysis of the non-redundant set of ageing-related 20%-quantiles that show the greatest interspecies difference in females (Hrt, Pit, Ovr).

7.2.3 **M2_Supplementary_Figures_S1-14.pdf**

S1. Collected tissues exemplifed in schematic fgure of NMR

S2. Hierarchical clustering of gene expression profles. The clustering tree shows a clear separation of tissues in both species, but less pronounced differences between sex and breeding status.

S3. Top 15 highest ranked GO sets based on enrichment analysis of cross-species DEGs between NMR and GP. GO sets are ranked by number of summarized GO terms (x-axes) and number of DEGs (alongside bar)

S4. Correlation analyses. log2(fold change) of shared DEGs (FDR<0.05) in cross-species comparison (NMR vs. GP, y-axes) and status change (breeder vs. non-breeder, x-axes) are shown for each species across all tissues. DEGs in NMR (left) show a significant positive correlation (DEGs:3,695; spearman:0.17, p=1.7 × 10-24), while DEGs in GP show a negative correlation (DEGs:1,380; spearman:-0.1; p=1.6 × 10-4)

S5. Top 15 highest ranked GO sets based on enrichment analysis of sex-related DEGs that are shared between GP breeder and non-breeder. GO sets are ranked by number of summarized GO terms (x-axes) and number of DEGs (alongside bar).

S6. Top 15 highest ranked GO sets based on enrichment analysis of sex-related DEGs in NMR breeder. GO sets are ranked by number of summarized GO terms (x-axes) and

number of DEGs (alongside bar, together with number of up- and downregulated genes).

S7. Top 15 highest ranked GO sets based on enrichment analysis of status-related DEGs in ovary of NMR females. GO sets are ranked by number of summarized GO terms (x-axes) and number of DEGs (alongside bar, together with number of up- and downregulated genes).

S8. Top 15 highest ranked GO sets based on enrichment analysis of status-related DEGs in adrenal gland of NMR females. GO sets are ranked by number of summarized GO terms (x-axes) and number of DEGs (alongside bar, together with number of up- and downregulated genes).

S9. Top 15 highest ranked GO sets based on enrichment analysis of status-related DEGs in cerebellum of NMR females. GO sets are ranked by number of summarized GO terms (x-axes) and number of DEGs (alongside bar, together with number of up- and downregulated genes).

S10. Top 15 highest ranked GO sets based on enrichment analysis of status-related DEGs in testis of NMR males. GO sets are ranked by number of summarized GO terms (x-axes) and number of DEGs (alongside bar, together with number of up- and downregulated genes).

S11. Top 15 highest ranked GO sets based on enrichment analysis of status-related DEGs in skin of NMR males. GO sets are ranked by number of summarized GO terms (x-axes) and number of DEGs (alongside bar, together with number of up- and downregulated genes).

S12. Mitonuclear ratios in non-breeders and breeders per sex, tissue and species. Boxplots shows median, 2nd/3rd quartiles, whiskers extend to 1.5 the interquartile range and dots values outside this range. P-values were calculated using a two-tailed t-test; *: $p<0.05$, **: $p<0.01$.

S13. Top 15 highest ranked GO sets based on enrichment analysis of the non-redundant set of ageing-related 20%-quantiles that show the greatest interspecies difference in males (Tes, Skn). GO sets are ranked by number of summarized GO terms (x-axes) and number of DEGs (alongside bar).

S14. Top 15 highest ranked GO sets based on enrichment analysis of the non-redundant set of ageing-related 20%-quantiles that show the greatest interspecies difference in females (Hrt, Pit, Ovr). GO sets are ranked by number of summarized GO terms (x-axes) and number of DEGs (alongside bar).

7.2.4 **M2_Supplementary_Text_S1.pdf**

Selection of significantly up-regulated in naked mole-rat breeders that are involved in steroid metabolism and endocrine system.

7.2.5 **M2_Supplementary_Data_S1.zip**

DEseq2 results of interspecies comparison of naked mole-rat vs. guinea pig.

7.2.6 **M2_Supplementary_Data_S3.zip**

DESeq2 results for the comparison of sexes within each species and group.

7.2.7 **M2_Supplementary_Data_S3.zip**

DESeq2 results for the comparison of statuses within each species and sex.

7.3 Manuscript 3 (M3)

7.3.1 **M3_MainDocument.pdf**

Digital version of manuscript M3.

7.3.2 **M3_Supplement_Document_S1.pdf**

Document S1. Supplemental Experimental Procedures

7.3.3 **M3_Supplement_Data_S2.xlsx**

Data S1. Assembly and Annotation. Sequence data are listed that were used for the genome assembly and its quality assessment, for gene and repeat annotation, as well as for assembling the sex-determining region of the Y chromosome.

S1A. Sequence Data Used for Assembling the Reference Sequence

S1B. Sequence Data Obtained by Roche Sequencing

S1C. Sequence Data Obtained by Sanger Sequencing

S1D. Mapping of BAC and Fosmid End Sequences

S1E. Integration of Cytogenetic and Sequencing Data

S1F. Repeat Annotation

S1G. BACs Sequenced Using Illumina MiSeq Technology and Corresponding Contig Assemblies

S1H. WGS Sequencing Data Obtained by PacBio Technology

S1I. Assemblies of BACs Sequenced Using PacBio Technology

S1J. RNA-Seq Data Used for Annotation of Protein-Coding Genes

S1K. In Silico Gene Predictions

S1L. De Novo Transcript Assemblies with Trinity

S1M. Reference-Based Transcript Assemblies with STAR and Cufflinks

S1N. Annotated Protein-Coding Genes

S1O. Teleost-Specific Duplicated Genes in the N. furzeri Reference Sequence

S1P. Assessment of Completeness of the Reference Sequence

7.3.4 **M3_Supplement_Data_S2.xlsx**

Data S2. XY Sex Chromosome Evolution.

WGS sequence data used for genome-wide variation analyses, PCR-based analyses of selected variations, as well as identification of regions of suppressed recombination.

S2A. PCR- and Sanger Sequencing-Based Validation of Sex-Linkage of Selected SNVs

S2B. WGS Data Used for Variation Analyses in Four N. furzeri Strains

S2C. Regions of Suppressed Recombination on sgr05

S2D. Estimation of the Age of the Secondary Recombination Suppression

7.3.5 **M3_Supplement_Data_S3.xlsx**

Data S3. Newly Identified Sex-Determining Gene: gdf6Y

Local variation and positive selection analyses in the sex-determining region as well as expression analyses and transcript assembly of gdf6Y.

S3A. Variations at the gdf6/gdf6Y Locus Obtained by GATK and Data in Data S2b

S3B. Variations at the gdf6/gdf6Y Locus Analyzed Using PCR

S3C. Local Positive Selection within the GRZ Male-Specific Region of the Y Chromosome

S3D. RNA-Seq Data Used for Expression Analysis and Assembly of the gdf6Y Transcript

7.3.6 **M3_Supplement_Data_S4.xlsx**

Data S4. Genomic Positional Enrichment, Positive Selection, Diapause, and Aging.

Aging-related differentially expressed genes and analyses of positional gene enrichment, analyses of positive selection, as well as the overlap in expression between diapause and aging.

S4A. RNA-Seq Data Used for Identification of Aging-Related Differentially Expressed Genes

S4B. Differentially Expressed Genes in MZM-0410 Brain Used for Positional Enrichment Analysis

S4C. Differentially Expressed Genes in MZM-0410 Liver Used for Positional Enrichment Analysis

S4D. Differentially Expressed Genes in MZM-0410 Skin Used for Positional Enrichment Analysis

S4E. Differentially Expressed Genes in Brain Used for Positional Enrichment Analysis

S4F. Differentially Expressed Genes in Liver Used for Positional Enrichment Analysis

S4G. Differentially Expressed Genes in Skin Used for Positional Enrichment Analysis

S4H. Regions of Positional Gene Enrichment

S4I. RNA-Seq Data Used for Phylogenetic and Positive Selection Analyses

S4J. Positive Selection Analyses

S4K. Differentially Expressed Genes in MZM-0410 Brain Used for Positive Selection Analysis

S4L. Differentially Expressed Genes in MZM-0410 Liver Used for Positive Selection Analysis

S4M. Differentially Expressed Genes in MZM-0410 Skin Used for Positive Selection Analysis

S4N. RNA-Seq Data Used for Diapause-Related Differentially Expressed Genes

S4O. Differentially Expressed Genes in Diapause versus Non-Diapause Embryos

S4P. Differentially Expressed Genes in Brain Aging of MZM-0410, Showing Monotonic Up- or Down-Regulation with Age and Significant Difference Between 5 Weeks and 39 Weeks

S4Q. Differentially Expressed Genes in Liver Aging of MZM-0410, Showing Monotonic Up- or Down-Regulation with Age and Significant Difference Between 5 Weeks and 39 Weeks

S4R. Differentially Expressed Genes in Skin Aging of MZM-0410, Showing Monotonic Up- or Down-Regulation with Age and Significant Difference Between 5 Weeks and 39 Weeks

S4S. Overlap of Differentially Expressed Genes in Diapause Embryos and Brain Aging

S4T. Overlap of Differentially Expressed Genes in Diapause Embryos and Liver Aging

S4U. Overlap of Differentially Expressed Genes in Diapause Embryos and Skin Aging

S4V. Overlap of Differentially Expressed Genes in C. elegans Dauer Larvae and N. furzeri Diapause Embryos

7.4 Manuscript 4 (M4)

7.4.1 **M4_MainDocument.pdf**

Digital version of manuscript M4.

7.4.2 **M4_Supplement.xlsx**

Supplement Information

Table S1. Results of positive selection analysis in the N-branch, ranked by P-value.

Table S2. Results of positive selection analysis in the PR-branch, ranked by P-value.

Table S3. Results of positive selection analysis in the FKK-branch, ranked by P-value.

Table S4. Comparison of fold-changes in expression of mitonuclear and mitochondrial biogenesis gene sets.

Table S5. Overview of the tested genes.

Table S6. List of genes used to construct the tree in Fig. 1.

Table S7. List of background genes used for GO analysis.

Table S8. List of genes used as background for the simulation experiment (5–95% expression quantile of the PSGs).

Table S9. List of mitochondrial biogenesis genes within the background genes list (Table S7).

Table S10. Entrez orthologs of PSGs.

Table S11. Expression levels of all tested genes.

Table S12. Statistics of read data for each library.

Table S13 Complete list of Nothobranchius PSGs overlapping with PSGs in long-lived mammals.

8 Acknowledgements

First, I would like to express my gratitude to Matthias Platzer for giving me the opportunity to explore the world of science as a doctoral student in his group, for his scientific guidance and for his calmness during hard times. The same gratitude goes to Karol Szafranski for his supervision, active support and fruitful as well as distracting discussions. I also want to thank all current and former members of the Genome Analyses group for their nuggets of wisdom about science and life as well as for the great working atmosphere in all workplaces, starting in the construction trailer in the institute's yard.

I also want to thank my collaboration partners in the Research Group General Zoology at the University of Duisburg-Essen (Yoshiyuki Henning, Christiane Vole, Philip Dammann) and the Leibniz Institute for Zoo and Wildlife Research in Berlin (Michaela Wetzker, Susanne Holtze, Thomas B. Hildebrandt) for giving me the opportunity to experience mole-rats at first hand and for sharing their knowledge about these fascinating animals.

Further, I want to express my gratitude to all people across the Beutenberg campus that regularly joined the lunch club (Bianca Hoffmann, Thomas Fabisch, Nora Adam, Thomas Brockmöller, Sven Boese) for interesting thoughts and discussions about biology, technology, socio-economics and the regular struggles of PhD students.

Finally, I want to thank those friends (Janine Freitag) and family members that supported me during these years by giving comfort and mental stimuli outside of work.

9 Statement of Authorship

English

I confirm that I am familiar with the relevant course of examination for doctoral candidates in the Faculty of Biology and Pharmacy at Friedrich-Schiller-University in Jena. I have composed and written the dissertation by myself and I have acknowledged all additional assistance, personal communication, and sources within the work.

I have not enlisted the assistance of a doctoral consultant and no third parties have received either direct or indirect monetary benefits from me for work connected to the submitted dissertation. I have not submitted the dissertation in an exact or modified version for a state or other scientific examination.

Deutsch

Ich erkläre, dass mir die geltende Promotionsordnung der Biologisch-Pharmazeutischen Fakultät der Friedrich-Schiller-Universität in Jena bekannt ist. Ich versichere, dass ich die vorliegende Dissertation selbst angefertigt, keine Textabschnitte eines Dritten oder eigener Prüfungsarbeiten ohne Kennzeichnung übernommen und alle verwendeten Hilfsmittel, persönlichen Mitteilungen und Quellen in meiner Arbeit angegeben habe.

Ich bestätigte, dass ich die Hilfe eines Promotionsberaters nicht in Anspruch genommen habe und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die in Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Ich versichere, dass ich die Dissertation weder in gleicher noch in ähnlicher Form zuvor als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe.

……..……...…….……...

Martin Bens