Transcriptome analysis of the short-lived fish *Nothobranchius furzeri*

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 Andreas Petzold

geboren am 01.01.1982 in Merseburg/Saale

Die vorliegende Arbeit wurde in der Zeit vom Januar 2009 bis August 2013 am Leibniz-Institut für Altersforschung - Fritz-Lipmann-Institut (FLI) in Jena angefertigt.

Gutachter:

1. Dr. Matthias Platzer, Genomanalyse, Leibniz-Institut für Altersforschung – Fritz-Lipmann-Institut (FLI), Jena

2. Dr. Reinhard Guthke, Forschungsgruppe Systembiologie, Leibniz-Institut für Naturstoff-Forschung und Infektionsbiologie - Hans-Knöll-Institut (HKI), Jena

3. Dr. Andreas Dahl, CRTD / DFG-Forschungszentrum für Regenerative Therapien Dresden – Exzellenzcluster, Dresden

Datum der Verteidigung: 19.05.2014

Table of contents

Abbreviations

B

BLAST Basic Local Alignment Search Tool
bp base pair

C

cDNA complementary DNA
CDS coding sequence

D

DEG differentially expressed gene
DNA deoxyribonucleic acid
DR dietary restriction

E

EST expressed sequence tag
et al. et alii (and others)

F

FLI Fritz Lipmann Institute

G

Gb Gigabase
GO Gene Ontology
GRZ Gona Re Zhou

I

ID identifier
IIS insulin/insulin-like growth factor 1 signalling

K

kb kilobase

M

Mb Megabase
miRNA microRNA
mRNA messenger RNA
MZM Mozambique

N

NCBI National Center for Biotechnological Information
NGS next-generation sequencing
nt nucleotide

O

ORF open reading frame

P

PC principal component
PCA principal component analysis
PCR polymerase chain reaction

Q

qPCR quantitative PCR

R

RNA ribonucleic acid
RNA-seq RNA sequencing
ROS reactive oxygen species
RPKM reads per kilobase of exon model per million mapped reads
rRNA ribosomal RNA

S

Sir2 sirtuin-2

T

TOR Target Of Rapamycin
tRNA transfer RNA

U

UTR untranslated region

V

vs. versus

W

w week

List of Figures and Tables

Zusammenfassung

Altersforschung wird erschwert durch die lange Lebensspanne der verwendeten Modellorganismen und die damit verbundenen hohen Kosten beziehungsweise den großen Arbeitsaufwand. Der türkise Prachtgrundkärpfling *Nothobranchius furzeri* lebt in saisonal vorhandenen Gewässern in Südostafrika und hat die kürzeste Lebensspanne, die je für Wirbeltiere in Gefangenschaft ermittelt wurde. Die Tiere entwickeln sich sehr schnell und werden innerhalb weniger Wochen geschlechtsreif. Der Alterungsprozess des Fisches kann durch verschiedene Biomarker nachgewiesen und quantifiziert werden. Weiterhin zeigen *N. furzeri* Populationen aus Gebieten mit verschiedenen klimatischen Bedingungen große Unterschiede in der Lebensspanne. Diese Unterschiede lassen sich auch bei den daraus gezüchteten Laborstämmen zeigen. Nachkommen von Kreuzungen verschiedener Stämme zeigen eine intermediäre Lebensspanne. Es ist somit anzunehmen, dass die Lebensspanne von *N. furzeri* genetisch bestimmt wird. Aus diesen Gründen wurde *N. furzeri* über die letzten Jahre als ein neuer Modellorganismus für die Erforschung des Alterns und der Lebensdauer eingeführt.

Das erste Ziel dieser Arbeit war, einen umfassenden Katalog aller Transkripte der Protein-kodierenden Gene von *N. furzeri* zu erstellen. Dieser erlaubt die Planung von Experimenten und Studien und ist eine Grundvoraussetzung für einen Modellorganismus. Dafür wurden 13 cDNA-Banken aus verschiedenen Transkriptomproben hergestellt und mit den Sequenziertechnologien Sanger, 454/Roche und Solexa/Illumina sequenziert. Die Gesamtmenge an erzeugten Sequenzdaten betrug 47 Gigabasen. Für die optimale Assemblierung und Annotation dieser umfangreichen Daten wurden verschiedene bioinformatische Programme und Systeme eingesetzt beziehungsweise verbessert. Der so erstellte Transkriptkatalog enthält Sequenzen für 19.875 Proteinkodierende Gene. Dabei gibt es für 71% dieser Gene mindestens eine Transkriptsequenz, die den vollständigen Proteinkodierenden Bereich enthält. Außerdem wurde eine Internetplattform eingerichtet, um einen einfachen Zugang zu dem Transkriptkatalog zu ermöglichen.

Der zweite Teil dieser Arbeit beschreibt Untersuchungen zu Veränderungen der Genexpression in alternden *N. furzeri*. Mittels der neuen RNA-seq-Methode war es möglich, die Transkripthäufigkeiten in jungen und alten *N. furzeri* des Laborstammes GRZ und des Laborstammes MZM-0403, welcher fast doppelt so alt wie GRZ wird, zu bestimmen. Dabei wurden in Gehirn und Haut 86 Gene nachgewiesen, welche in alternden *N. furzeri* signifikante Veränderungen in der Expressionshöhe zeigten. Diese differentiell exprimierten Gene spielen eine Rolle in biologischen Prozessen mit bekanntem Alternsbezug, wie zum Beispiel im Zellzyklus, in der Zellteilung und im Zellwachstum, in Entzündungsprozessen und bei der Erhaltung von Geweben. Zur Bestätigung der Ergebnisse wurde ein zweites RNA-seq-Experiment im Zebrafisch durchgeführt. Für eine große Anzahl (41%) der in *N furzeri* differentiell exprimierten Gene konnten ähnliche Veränderungen auch im alternden Zebrafisch nachgewiesen werden, was die alterns-relevante Funktion dieser Gene bestätigte. Der

Vergleich der Veränderungen in den Genexpression zwischen den beiden Stämmen legt nahe, dass der Alterungsprozess in GRZ schneller verläuft als in MZM-0403.

Zusammengefasst habe ich in dieser Arbeit die Entwicklung eines umfassenden annotierten Transkriptkatalogs für *N. furzeri* beschrieben und erste Einblicke in die Veränderungen der Genexpression alternder *N. furzeri* gegeben.

Abstract

Age research is hindered by the long lifespan of the current vertebrate model organisms and the associated costs and efforts. The turquoise killifish *Nothobranchius furzeri* inhabits seasonal ponds in South-East Africa and has the shortest lifespan recorded for vertebrates in captivity. The fish grows very fast, reaches sexual maturity in a few weeks and shows the expression of typical ageing-related biomarkers. Furthermore, *N. furzeri* populations living in regions with different climatic conditions show large differences in lifespan, and derived strains maintain these differences even in captivity. The progeny of crosses of different strains show an intermediate lifespan, indicating that lifespan is genetically determined in *N. furzeri*. For these reasons, *N. furzeri* has been established as a new model organism for the studies of ageing and lifespan determination during the recent years.

The first aim of this thesis was to build a comprehensive transcript catalogue of protein-coding *N. furzeri* genes, which is one prerequisite for a model organism, for example to design experiments and studies. To this end, Sanger, 454/Roche and Solexa/Illumina sequencing was used to sequence 13 cDNA libraries from different transcriptomes, yielding 46 Gb sequence data. Efficient assembly and annotation of large datasets was ensured by applying and developing specifically designed bioinformatics tools and pipelines. The resulting transcript catalogue contains transcript contigs for 19,875 protein-coding genes. Of these, 71% are represented by at least one transcript contig with a complete coding sequence. Furthermore, a transcriptome browser was set up, to facilitate easy access to the transcript catalogue.

The second aim was to study gene expression changes in ageing *N. furzeri*. A RNA-seq experiment was conducted to study transcript levels of young and old fish of the strains GRZ and MZM-0403, which differ in lifespan by 100%. Eighty-six differentially expressed genes were detected in the analysed tissues brain and skin. These genes play a role in ageing-relevant processes like cell cycle, division and proliferation, inflammation and tissue maintenance. A similar RNA-seq experiment was conducted in zebrafish. A significant fraction (41%) of the *N. furzeri* genes was also found to be differentially regulated in ageing zebrafish, thus confirming their relevance in ageing. Finally, comparisons of fold changes of the two strains suggested that ageing is accelerated in the short-lived *N. furzeri* strain GRZ, compared to the longer-lived strain MZM-0403.

In summary, in this thesis, I describe the development of a comprehensive, annotated *N. furzeri* transcript catalogue and give first insights into transcriptome-wide changes during *N. furzeri* ageing.

.

1 Introduction

In the mid of the 19th century, the Industrial Revolution profoundly changed the social, economic and cultural conditions of human life in the Western civilisations. During this major turning point, scientific and technologic advances lead to large progress in food production, manufacturing, transportation and medical care resulting in much improved living conditions. At the beginning of the 20th century, newly established methods in hygiene, disease prevention and medical treatment considerably reduced the mortality rate, especially of infants, and extended the lives of the elderly. As a result, life expectancy increased significantly, and more people reached ages that had been previously considered as exceptional. Today, the increase in life expectancy is on-going but accompanied by a decline in fertility. As a result, the fraction of old people is steadily rising and the populations as such begin to age (Christensen et al. 2009). This development brings new social and economic challenges for the functioning of entire societies. Consequently, age research, the studies of ageing processes and lifespan determination, is of growing importance and will help in dealing with the ramifications of an ageing society.

1.1 Genes and pathways involved in ageing

The Handbook of the Biology of Aging describes ageing as a degenerative process affecting virtually all known organisms that is characterised by progressive deterioration of cellular components and deregulation of cellular processes, resulting in mortality (Masoro & Austad 2011, p.215). The ageing process itself is affected by both environmental and genetic factors. Environmental factors are very diverse, hard to discern and therefore difficult to study. Genetic factors, however, have been analysed and resulted in a number of candidate genes relevant for ageing and lifespan determination (Christensen et al. 2006). These genes have a role in genomic maintenance and repair, metabolism, inflammation and mitochondrial oxidation; below, a short overview of the most important findings is given.

1.1.1 Genomic maintenance and repair

Maintenance and renewal of cells, tissues and organs is ensured by the process of cell division. In this process, the genetic material of the parent cell is duplicated and the copies are distributed between the two daughter cells. Therefore, any errors introduced into the genetic material of the parent cell will be passed on to the daughter cells. With further cell divisions, more errors accumulate until the cells are not viable anymore. Thus, fail-safe replication as well as repair and maintenance of the genetic material are of paramount importance for any organism. Malfunctioning of these processes leads to damages and, as a consequence, to accelerated ageing. A prominent example which illustrates this relationship comes from the Werner syndrome. It is a rare autosomal recessive disorder that is caused by a loss-of-function mutation in the Werner gene, encoding a DNA helicase (Yu et al. 1996).

Mutations in this gene cause premature ageing. Patients show typical signs such as wrinkled skin, grey hair and frailty, and develop typical ageing-related diseases such as cataracts, diabetes and osteoporosis (Goto 1997).

1.1.2 Metabolism

In 1934, Mary F. Crowell and Clive M. McCay found that dietary restriction (DR, the restriction of nutrients without malnutrition) almost doubled the lifespan of rats (McCay & Crowell 1934). Later, DR has been shown to increase lifespan in a variety of species, including yeasts, worms, flies and rodents (Masoro 2003; Sinclair 2005). Accordingly, several pathways that play a role in metabolism have been found to be affected in longevity mutants. The most prominent is the insulin/insulin-like growth factor 1 signalling (IIS) pathway. It mediates cell growth, proliferation and cell death in response to environmental conditions and is regulated by both insulin and insulin-like growth factors. Its ageing-relevance was first found in nematodes, where worms with mutations in the gene *age-1* lived almost twice as long as normal (Friedman & Johnson 1988). Since then, IIS genes have been shown to affect lifespan in yeasts, worms, flies and rodents (Fabrizio et al. 2001; Henderson & Johnson 2001; Tatar et al. 2001; Blüher et al. 2003).

A second major pathway connecting metabolism and ageing is the Target Of Rapamycin (TOR) pathway. Its main regulator is the serine/threonine kinase TOR complex 1, which acts as a sensor for nutrients, energy levels, growth factors and various stress-induced conditions (Hay & Sonenberg 2004). Deletion of the TOR complex 1 in yeast resulted in a significant extension of lifespan (Kaeberlein et al. 2005). In worms, flies and rodents, inhibition of the TOR complex 1 also extended lifespan (Vellai et al. 2003; Kapahi et al. 2004; Harrison et al. 2009). Due to its role as nutrient and energy sensor, the TOR pathway is assumed to be involved in lifespan extension via DR, and experimental support came from yeasts under DR, which did not show significant extension of lifespan after deletion of the TOR complex 1 (Kaeberlein et al. 2005).

The third major candidate which may link metabolism and ageing is sirtuin-2 (Sir2). Sir2 is a NAD^+-dependent histone deacetylase which acts as a transcriptional repressor and is responsible for the maintenance of the genome stability (Imai et al. 2000). Deletion and overexpression of Sir2 shortened and extended lifespan in yeast, respectively (Kaeberlein et al. 1999). Similar findings were obtained for worms and flies (Tissenbaum & Guarente 2001; Rogina & Helfand 2004). The dependency of Sir2 on NAD^+, an important regulator of metabolism, suggests that it is also involved in the cellular response to DR. Accordingly, it was shown that Sir2 is required for DR in yeast (Lin et al. 2000). Moreover, Sir2 has been implicated to act on the tumour suppressor p53, which plays a central role in determining the cell's fate by regulating cell cycle, cell death and DNA repair (Vaziri et al. 2001).

1.1.3 Inflammation

Inflammation is considered a core process of ageing. With age, the risk of developing a chronic condition increases, and older people are more likely to have some form of low-grade, persistent inflammation (Krabbe et al. 2004). It was shown that blood levels of the pro-inflammatory proteins such as interleukin 6, tumour necrosis factor α and C-reactive protein are expressed constitutively and mediate the acute-phase response to inflammation events, tend to increase with age (Ferrucci et al. 2005). Permanently elevated levels of these proteins are known as a risk factor for cardiovascular problems in particular and therefore increase mortality in general (Danesh et al. 2008). However, the reasons for these elevated protein levels are unclear. Older people might simply show a higher vulnerability against diseases and thus the immune system is triggered more often. Additionally, inflammation also plays in the major degenerative diseases of the late life, for example in Alzheimer disease (Akiyama et al. 2000).

1.1.4 ROS and mitochondrial oxidation

In the course of the cellular metabolism, chemically reactive oxygen-containing by-products such as oxygen ions and hydrogen peroxide are produced; these molecules are called reactive oxygen species (ROS). Due to their high oxidising potential, they can cause significant oxidative damage to cell structures. Based on that observation, several theories linking oxidative stress to ageing have been introduced to the scientific community. The first of these theories, the *free radical theory of ageing*, was proposed by Harman in 1956 and states that ageing may be related to the deleterious effects of ROS on cell constituents and connective tissues (Harman 1956). An extension of this theory, the *mitochondrial theory of ageing*, suggests that the complex redox reactions in mitochondria can be considered as a major contributor of ROS and consequently determine the rate of ageing (Harman 1972). Another very intriguing theory is commonly referred to as the *mitochondrial "vicious cycle" theory of ageing*. In brief, the basic idea of this theory is that ROS-induced mutations in the mitochondrial genome lead to a compromised respiration, which, in turn, leads to an increased ROS production and, ultimately, to ageing (Miquel et al. 1980). However, in addition, there are a large number of other theories, and the effect of oxidative stress on ageing is still only rudimentary understood.

Consequently, this branch of age research is focused on mechanisms that, first, decrease ROS production and, second, provide defence against already generated ROS. The latter led to the discovery of antioxidants, which react with ROS and render them harmless. Antioxidants are for example vitamin E and C, coenzyme Q_{10}, and resveratrol. However, the effects of antioxidants on lifespan are unclear. For example in case of vitamin E, some studies report a significant increase in lifespan (Navarro et al. 2005), while others report only a decrease in oxidative damage (Lipman et al. 1998). Several antioxidant enzymes which catalyse the detoxification of ROS, such as superoxide dismutases, catalases, and glutathione peroxidase, were identified. Similar to the antioxidants, life-

prolonging effects of these enzymes were found only in a few studies (Schriner et al. 2005; Hu et al. 2007). Nonetheless, antioxidants and antioxidant enzymes are still regarded as key regulators of ageing.

1.2 *Nothobranchius furzeri* as model organism for age research

1.2.1 Established model organisms

Age research in human is limited due to practical (long life expectancy) and ethical (no experiments possible) reasons. Moreover, since it is difficult to quantitatively measure ageing, research is often limited to alternative traits that correlate with health- and lifespan such as metabolic parameters, histological/anatomic markers or stress resistance. However, these can serve only as an approximation and do not necessarily reflect health- or lifespan. Consequently, besides studying ageing in human cell cultures, scientists have turned to other invertebrate and vertebrate species which have been introduced as model organisms for age research.

Invertebrate species used as model organisms are the budding yeast *Saccharomyces cerevisiae*, the nematode *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster*. They are small, easy to maintain in a laboratory, produce a large number of offspring and have a short lifespan (yeast: 26 generations with a generation time of one hour; worm and fly: three weeks and one month, respectively). They are amenable to experimental manipulation, and many established resources and protocols exist, which allows researchers to conduct experiments relatively fast and cheap. Due to their small size and short lifespan, they can be analysed in large-scale genetic screens and functional genomics, to quickly identify the function of single genes. Furthermore, their genomes are relatively small (12, 100 and 130 Mb) and simple, compared to those of vertebrates, and genome sequences are available for all three species (Goffeau et al. 1996; The *C. elegans* Sequencing Consortium 1998; Adams et al. 2000).

Vertebrate species used as model organisms in age research are the zebrafish *Danio rerio* and the house mouse *Mus musculus*. They are larger than invertebrate models, live longer (in captivity, three years and 12 months year, respectively), produce less offspring, and require large expensive facilities and intensive care. Due to the higher organismal complexity of vertebrates, many of the invertebrate approaches for experimental manipulation are either more difficult or do not work at all in zebrafish and mouse. Furthermore, their genomes are large (2.7 and 1.4 Gb; Church et al. 2009; Howe et al. 2013) and complex, which complicates genetic analyses. Despite all these disadvantages, the main reason for using vertebrates as model organisms for biomedical age research is that they are evolutionary closer to and that their physiology is more alike humans. Thus, it is more likely that findings obtained in these species can be successfully transferred to human. Additionally, testing of

potentially life-extending drugs potentially suited for humans/medical applications is more realistic in vertebrate species.

Many of the ageing-related genes and pathways described in 1.1 were first discovered in an invertebrate model organism. While the general nature of ageing suggests that the involved genes and pathways are conserved between all organisms, until now, only inhibition of the TOR pathway was found to extend lifespan in each of the four species yeast, worm, fly and mouse (Kaeberlein & Kennedy 2009). Moreover, other ageing-related processes like inflammation act completely different between invertebrate and vertebrate species. Consequently, for age research/gerontology, findings obtained in vertebrate model organisms are obviously preferred to those obtained in invertebrates. However, such findings are complicated by the long lifespan of current vertebrate model organisms. This illustrates the need to have a vertebrate model organism that shares many of the advantages of the invertebrate model organisms, especially the short lifespan.

1.2.2 Nothobranchius furzeri

Nothobranchius furzeri (Figure 3) is a small freshwater fish that inhabits small ponds in South-East Africa. It was named after R.E. Furzer who, together with W. Warne, first collected the fish in Zimbabwe in 1968 (Jubb 1971). In 2005, *N. furzeri* was proposed as a new model for studies of ageing and longevity (Genade et al. 2005).

1.2.2.1 Life cycle

The habitat of *N. furzeri* is characterised by distinct, seasonal changes in the amount of precipitation (rainfall), that is, the year is divided in a wet and a dry season (Figure 1). *N. furzeri* have adapted to the regular drying by a specialised, annual life cycle with a complete reproductive cycle during the wet season (Figure 2). Female fish spawn up to 50 eggs per day on the ground of the pond (Haas 1976). The fertilised eggs then survive the dry season in the dried mud by entering into diapause in which they remain dormant and are protected against desiccation by a hard shell. As soon as the next wet season begins and precipitation starts, the fish hatches and reaches sexual maturity within a few weeks (Wildekamp 1993; Wourms 1972). Moreover, the eggs can remain in diapause for several years. This strategy ensures the survival of the species and allows *N. furzeri* to compensate an occasional absence of the wet season (Jubb 1971; Markofsky & Matias 1977).

Figure 1: The habitat of *N. furzeri*.
The left-hand picture shows a typical pond where *N. furzeri* can be found during wet season. The right-hand picture shows the same locality with the pond desiccated during dry season.

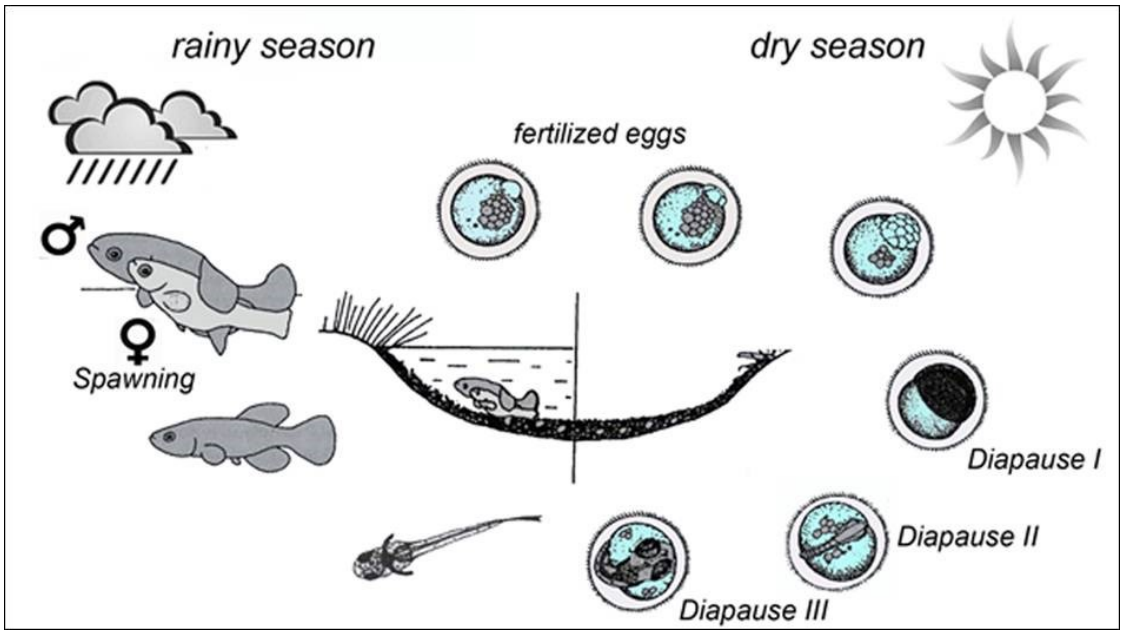

Figure 2: The life cycle of *N. furzeri*.
Female fish spawn fertilised eggs on the ground of the pond and die when the dry season starts. In contrast, the eggs are desiccation-resistant and survive in the dried mud in a diapause state. Shortly after beginning of the wet season, the embryo hatch, grow very fast and start to reproduce in a few weeks, thus completing the life cycle of *N. furzeri.* [adapted by E. Terzibasi from Seegers, L., 1997. *Aqualog*. *Killifishes of the world* ; Old World *killis* 2. 8 8, Mörfelden-Walldorf: Verl. A.C.S.]

1.2.2.2 Lifespan of N. furzeri populations

The length of the wet season of only a few months sets a natural upper limit to the lifespan of *N. furzeri*. However, in its natural habitat, lifespan data is difficult to obtain and is therefore mostly inferred from observations on when and where certain ponds exist. Nonetheless, lifespan has been monitored for fish in captivity, where over the years several strains have been established. The strain

GRZ, collected in the Gona Re Zhou (GRZ) Game Reserve in 1968, has a maximum lifespan of only 12-16 weeks in captivity, despite ideal conditions and unlimited water supply (Genade et al. 2005; Terzibasi et al. 2008; Hartmann et al. 2009). Thus, *N. furzeri* has the shortest maximum lifespan measured for vertebrates in captivity.

Several *N. furzeri* isolates/strains, derived from populations living under different climatic conditions, exhibit longer lifespans (Terzibasi et al. 2008; Terzibasi et al. 2013). In its geographic range in South-East Africa, a north-south decline in altitude causes an increase in yearly precipitation. The Gona Re Zhou Game Reserve, the original location of the strain GRZ, is located in the north and receives very little precipitation. Additional strains were sampled from the Limpopo River which is located 300 km south in Mozambique (MZM) and receives more precipitation. These MZM-strains live significantly longer in captivity. For example, strain MZM-0403 has a maximum lifespan of 29-32 weeks, which is twice as long as GRZ (Terzibasi et al. 2008). Furthermore, other Nothobranchius species from more humid areas live considerably longer, with a maximum lifespan of almost a year (Terzibasi et al. 2013).

Figure 3: *N. furzeri* strains MZM-0403 and GRZ.
The left-hand and the right-hand picture show a 25-weeks-old male and a 15-weeks-old female, respectively, of the strain MZM-0403. The middle picture shows a 11-weeks-old male GRZ.

1.2.2.3 N. furzeri *as model for age research*

Since the short-lived phenotype of *N. furzeri* is observed not only in the wild but also in captivity, it was assumed that the short lifespan is genetically determined. This assumption is supported by the emergence of strains with different lifespans as an adaption to the varying precipitation and by the finding that crosses between short- and long-lived strains show an intermediate lifespan (Kirschner et al. 2011). Several other characteristics additionally qualify *N. furzeri* as a model system for age research: captive maintenance and breeding, fast growth and maturity, a clearly visible and measurable ageing phenotype, the possibility of lifespan modulation, the existence of inbred strains and an evolutionary favourable relationship to other fish model systems (summarised in Genade et al. 2005); below, some of these characteristics are explained in more detail.

N. furzeri grows with a remarkable pace and shows symptoms of old age such as pale body colour, reduced body weight and fat, spine curvature and increasing frailty in general. At histological level, the fish develops degenerative lesions and neoplasia in different organs which may be the

primary cause of death but are not unusual for teleost fish (Di Cicco et al. 2011). At cellular level, ageing can be confirmed by a number of established biomarkers like lipofuscin, senescence-associated β-Galactosidase and Flouro-JadeB (Genade et al. 2005; Terzibasi et al. 2008; Di Cicco et al. 2011).

Lifespan and ageing can be manipulated in *N. furzeri*. The largest effects were achieved by administering the antioxidant resveratrol, resulting in an increase of up to 59% in both median and maximum lifespan of GRZ (Valenzano & Cellerino 2006; Valenzano et al. 2006b). Treated fish showed delayed physical and cognitive decay and were longer fertile. Similar effects were observed under dietary restriction, which prolonged median and maximum lifespan of GRZ by 13% and 33%, respectively, improved cognitive performance and retarded the expression of the ageing-related biomarkers lipofuscin and Flouro-JadeB (Terzibasi et al. 2009). Lifespan extension was also achieved by temperature reduction; a decrease from 25 to 20°C increased the maximum lifespan by 10% and delayed many of the ageing-related phenotypic symptoms described above (Valenzano et al. 2006a).

1.2.2.4 Genomic and genetic resources

One of the main requirements for model organisms is the availability of comprehensive genetic and genomic resources which serve as prerequisite for many up-to-date experiment and study designs. These may include genome or transcriptome sequences, genetic markers and maps, sequence variants, and more. However, since *N. furzeri* is under active research only for some years and the respective scientific community is still small, such resources are yet limited to a number of initial studies.

In 2009, our group provided an initial characterisation of the *N. furzeri* genome (Reichwald et al. 2009). The genome is diploid and contains 19 chromosomes (2n = 38). Furthermore, 5.4 Mb of the strain GRZ were sequenced at random and analysed. The genome size was estimated to be between 1.6 and 1.9 Gb. Compared to the four other fish species with a sequenced genome, *N. furzeri* has the biggest genome (medaka: 1 Gb, stickleback: 0.7, tetraodon: 0.4, zebrafish: 1.4; Kasahara et al. 2007; Jones et al. 2012; Jaillon et al. 2004; Howe et al. 2013). Additionally, *N. furzeri* showed the highest repeat content of all, with 45% of the genome (a second analysis increased this value to 64%; Koch 2010). The high repeat content included 21% tandem repeats, which is exceptional among fish as well as vertebrates in general. Importantly, the two most prominent tandem repeats localised preferentially to the centromeric regions where they contribute to large heterochromatic regions, also detected by chromosome staining. This pattern was only observed in *N. furzeri* and did not occur in other Nothobranchius species. Furthermore, the evolutionary relationship of *N. furzeri* to other known fish species was determined. Here, reconstruction of a phylogenetic tree based on protein-coding sequences revealed that medaka is the next relative with a sequenced genome. Finally, the genetic variation in the two different *N. fuzeri* strains was assessed. Genotyping of gene-associated single nucleotide variants and microsatellite markers showed that the long inbred GRZ strain is highly homozygous while the recently collected MZM-0403 strain still resembles the wild type.

Based on a cross between GRZ and MZM-0403, a first genome-wide genetic linkage map was constructed, which shows the positions of genetic markers relative to each other. In total, 413 *N. furzeri* F_2-individuals were genotyped for 152 microsatellites, and a linkage map with 25 linkage groups was constructed. Additionally recorded phenotypic data allowed identifying regions in the linkage map associated with sex determination and tail colour (Valenzano et al. 2009). In 2011, the cross was repeated in our group, and additional microsatellite marker plus gene-associated single nucleotide variants were included in the linkage map analysis. The resulting second-generation linkage map contained 22 linkage groups of which three likely represented fragments of the other linkage groups. The remaining 19 linkage groups agreed well with the number of chromosomes determined by karyotyping, and the total length of the map was similar to the estimated size of the genome (Kirschner et al. 2011).

End of 2009, our group also started a *N. furzeri* genome project, which aims at a high-quality reference sequence of the genome. Until 2011, approximately 190 Gb of genomic data was generated, and since 2010, the sequencing data is being assembled. Different genome assemblies were produced; the latest version, produced in 2012, contains 944 Mb in 7,675 scaffolds, which covers 59% of the genome, assuming a size of 1.6 Gb. Finishing of the assembly and subsequent comprehensive annotation represents a substantial effort and greatly benefits from independent experimental resources. One of those essential components of a genome project is the acquisition of transcript data. Furthermore, transcriptome sequencing represents a much simpler and less costly alternative to obtain valuable sequence information for species without a known genome sequence (Vera et al. 2008).

1.3 Transcriptome analysis

The transcriptome is defined as the set of all transcripts and their quantity in a cell. Species of transcript include the protein-coding mRNA as well as non-protein-coding RNA such as ribosomal RNA (rRNA), transfer RNA (tRNA), microRNA (miRNA), and so on. This thesis concentrates primarily on the protein-coding mRNA, and, therefore, the term transcriptome will henceforth refer to the set of all mRNA transcripts. It should be noted that the transcriptome is not fixed and that there is no general set of transcripts found in all cells at all times. Transcriptomes vary between tissues and different conditions. Moreover, the transcriptome within a single cell is also highly dynamic, because it underlies simultaneous RNA transcription and degradation. Consequently, transcriptome analysis always grasps only a snapshot of the transcriptome within a certain tissue at a certain time.

Understanding the transcriptome helps to identify the functional elements in the genome and to reveal the molecular constituents of cells and tissues. Therefore, transcriptome analysis aims at identifying all species of transcripts, their genic structure with exons, introns, 5' and 3' ends, and their splicing pattern as well as other posttranscriptional modifications. Moreover, transcriptome analysis also aims at quantifying the differences/changes of transcript levels between cells, tissues or organs

and under different conditions. For these purposes, various techniques and technologies have been developed deduce and measure transcriptomes.

1.3.1 Classical approaches for transcriptome analysis

Experimental detection of single transcripts was initially done with the northern blot technique that involves size separation by gel electrophoresis, followed by transfer to a membrane and detection by hybridisation with labelled probe (Alwine et al. 1977). However, the technique is sensitive to RNA degradation, and quantification is limited due to its small dynamic range. The current gold standard method is the quantitative PCR (qPCR; reviewed in Wong & Medrano 2005). At first, RNA is reverse-transcribed into complementary DNA (cDNA). The targeted cDNA molecule is then PCR-amplified using specific primers, and the accumulation is measured in each cycle with a fluorescent probe. The initial start amount of transcript can then be extrapolated from the increase per cycle. The qPCR method is highly accurate and covers several log orders of magnitude of transcript levels. However, qPCR as well as northern blotting can only detect known transcripts, that is, the detection of novel transcripts is not possible. Moreover, they are both very labour-intensive and not suited for larger numbers of transcripts to be quantified.

The development of automated sequencing methods enabled the direct read-out of the cDNA sequence, and a number of sequencing-based approaches were developed. Expressed sequence tag (EST) Sanger sequencing generates random, single-pass sequences of cDNA clones (National Center for Biotechnology 2004). These sequences allow the discovery of novel, unknown transcripts and greatly facilitate the discovery of new genes, for example in human (Adams et al. 1991). The main drawback is that the length of the EST is limited by the read length of the sequencing method, and many transcripts are only partially sequenced. Moreover, EST sequencing is not suited for transcript quantification because essentially only a single copy of a transcript is determined per sequencing reaction, and reaching statistically significant transcript counts is expensive and labour-intensive. Tag-based methods (reviewed in Harbers & Carninci 2005) were developed to overcome these limitations, including serial analysis of gene expression, cap analysis of gene expression and massively parallel signature sequencing. Briefly, all methods isolate only a small chunk from each cDNA clone, called tag, and concatenate these tags into larger chains. These chains are then sequenced and the tags counted. Thus, multiple transcripts can be profiled in parallel, providing ‘digital’ transcript levels. However, these methods are also based on expensive Sanger sequencing, and the source of a significant portion of the short sequence tags cannot always be uniquely identified.

In parallel, the development of DNA microarrays provided the means for relatively inexpensive large-scale quantifications of transcriptomes. A DNA microarray consists of thousands of DNA probes, which are specific for the targeted genes, attached on a glass surface (reviewed in Lockhart & Winzeler 2000). The sample, RNA or cDNA, is labelled with a fluorescent dye and hybridisation is detected by measuring the fluorescence signal. Effectively, DNA microarrays allow

the detection of thousands of transcripts in parallel. By co-hybridising a sample and a reference labelled with different fluorescent dyes and measuring the signal ratio, it is also possible to obtain an estimation of the transcript levels in the sample. Moreover, genomic tiling arrays with probes spanning exon junctions can be used to analyse spliced isoforms. However, DNA microarrays have also several limitations: reliance on sequence knowledge, cross-hybridisation between probes (Okoniewski & Miller 2006) and a limited dynamic range due to signal saturation (Dodd et al. 2004).

1.3.2 Next-generation sequencing technologies

Sanger sequencing is considered as a first-generation sequencing technology and has been widely used for over 30 years. However, in the past eight years, the technology has been gradually replaced by new high-throughput sequencing technologies, which are commonly summarised as next-generation sequencing (NGS) technologies and produce massive volumes of data at very low costs. NGS technologies are often divided into second-generation sequencing technologies, which employ PCR methods for signal amplification (454/Roche, Solexa/Illumina and ABI/SoLID), and third-generation sequencing technologies, which perform true single-molecule sequencing, thus avoiding the introduction of artefacts during the PCR amplification step (Pacific Biosciences, Oxford Nanopore). In this thesis, I used Sanger sequencing as well as the two NGS technologies 454/Roche and Solexa/Illumina; these three technologies are briefly summarised below.

1.3.2.1 Sanger

Sanger sequencing (Sanger et al. 1977, reviewed in Metzker 2005) involves the synthesis of a complementary DNA template by a DNA polymerase using both natural 2'-deoxynucleotides and modified 2',3'-dideoxynucleotides, which cannot be elongated and therefore terminate the synthesis reaction. Depending on the ratio of synthesis and termination, a number of fragments are produced which differ in the number of incorporated nucleotides and can be size-separated by gel electrophoresis. The primers or terminating 2',3'-dideoxynucleotides are radioactively or fluorescently labelled, which allows the identification the terminal nucleotide (A, C, G or T). By successively identifying these nucleotides of the length-sorted fragments, the DNA sequence of the template is determined. This sequence is commonly referred to as read and measured in base pairs (bp).

In modern automated Sanger sequencing, these processes are largely automated, for example by the use of nucleotide-specific fluorescent dyes (for example, green for A). The templates are kept in 96-well plates, which serve as reaction tubes for the sequencing. The average read length of Sanger sequencing is approximately 850 bp, and modern Sanger sequencer can process 96 templates in one run. However, for high-throughput, Sanger sequencing is quite expensive and labour-intensive, which is why it was only used to do an initial characterisation of the *N. furzeri* transcriptome.

1.3.2.2 454/Roche

The first commercially available NGS technology was developed by the company 454 Life Sciences, acquired by Roche in 2007. The major improvement compared to Sanger sequencing is that the sequencing reaction takes place in very small wells with a pico-litre volume (10^{-12} l), which allows running more than a million reactions in parallel on one plate (Margulies et al. 2005). Single DNA fragments are bound to beads that are surrounded by water droplets coated in an oil-based emulsion and serve as reactor for the clonal amplification of the DNA fragments (emulsion PCR). Each DNA bead is then deposited within one pico-litre well. During pyrosequencing, only one kind of nucleotides is added at a time, and a pyrophosphate is released every time the DNA polymerase incorporates a molecule. This pyrophosphate is converted by enzymes, resulting in a light signal. The number of incorporated nucleotides is then directly proportional to the number of released pyrophosphates and the signal intensity.

The first version of the 454/Roche sequencer, termed GS20, produced a total sequencing output of about 25 Mb per run, with an average read length of 100 bp (Margulies et al. 2005). Subsequent improvements in chemistry and signal detection increased average read length (FLX: 250 bp, Titanium: 400 bp) and sequencing output (FLX: 100 Mb, Titanium: 500 Mb); these versions of the 454/Roche sequencer are used in this thesis. The most recent version of the technology allows read lengths up to 1 kb (http://454.com/products/gs-flx-system/). Regarding sequence accuracy, 454/Roche sequencing errors occur predominantly in homopolymeric nucleotide stretches (Margulies et al. 2005). The exact number of incorporated nucleotides cannot be always reliably inferred from the signal intensity and is often over- or underestimated, resulting in insertion or deletion errors. Generally, 454/Roche sequencing is rather expensive, due to its low output compared to other NGS technologies.

1.3.2.3 Solexa/Illumina

The Solexa/Illumina technology uses reversible dye-terminator nucleotides for sequencing (Bentley et al. 2008). DNA fragments are hybridised to primers bound to the glass surface of so-called flow cells, which serve as reaction chamber for sequencing. The attached DNA fragments are then amplified using a 'bridging' PCR, in which the synthesised single-stranded DNA fragments bend over to form bridges with other primers, enabling the synthesis of the second strand. This step is repeated until cluster of clonal DNA fragments are formed. Sequencing is done by repeated cycles of polymerase-directed single nucleotide incorporation. Each nucleotide has a reversible modification that prevents further extension and is labelled with specific fluorescent dye. In each cycle, the identity of the last incorporated nucleotide is determined by fluorescent imaging. Subsequently, the modification is removed, thus allowing the incorporation of another nucleotide in the next cycle. Importantly, the number of incorporation cycles and thus the read length is fixed which is in contrast to Sanger and 454/Roche.

The first version of the Solexa/Illumina sequencer, termed GA1, produced 30-60 million 35-bp long reads, summarising to a total sequencing output of 1-2 Gb per run (Bentley et al. 2008). Improvements especially in the optics and the flow cell design rapidly lead to increased sequencing output. The GA2x, which is used in this thesis, produces reads with lengths up to 150 bp with a total sequencing output of 80 Gb. Solexa/Illumina also supports paired-end sequencing, that is, a longer DNA fragment is sequenced from both ends to obtain more sequence information. For example, 100-bp long reads are generated from both ends of a 500-bp long fragment. Current versions of the technology allow read length and sequence output up to 250 bp and 600 Gb, respectively. In contrast to 454/Roche, Solexa/Illumina errors are mostly substitutions, which preferentially occur at the 3' end of the reads (Dohm et al. 2008). During each sequencing cycle, exactly one nucleotide is incorporated, and all clonal DNA fragments of one cluster are supposed to give the same signal, indicating the incorporated nucleotide. However, it can happen that some DNA fragments are not elongated properly by exactly one nucleotide, that is, they are out of phase. As a result, the overall cluster signal suffers and the error rate increases. Moreover, AT- and GC-rich regions are underrepresented in the sequencing data, which is presumably due to a bias during the amplification step (Dohm et al. 2008).

1.3.3 RNA-seq as new approach for transcriptome analysis

The application of NGS technologies to sequence cDNA derived from cellular RNA enables to fully catalogue and quantify transcriptomes at low costs; this approach is termed RNA-seq (Nagalakshmi et al. 2008; Wilhelm et al. 2008). Main applications of RNA-seq are the development of transcript catalogues and the quantification of transcripts over wide ranges of transcript levels (reviewed in Wang et al. 2009). Furthermore, RNA-seq has been used to catalogue sense and antisense transcripts, to detect alternative splicing events and gene fusion transcripts, and to map transcription start sites (reviewed in Ozsolak & Milos 2011).

There are several protocols for RNA-seq, which differ in RNA extraction, cDNA library construction, fragmentation and sequencing strategy (reviewed in Wang et al. 2009; Wilhelm & Landry 2009). The general procedure starts with the extraction of the total RNA from the sample. The RNA is then reverse-transcribed into cDNA for sequencing; the result is a cDNA library. There are two methods to build a cDNA library from RNA that depend on the type of primer used to initiate reverse transcription. Oligo(dT) primer bind to the poly(A)-tails at the 3' end of the mRNA transcript. Random hexamer oligonucleotides bind to the entire mRNA transcript. When using random hexamer primer, cDNA library preparation is often preceded by a poly(A)-selection step for mRNA enrichment, which initially makes up only a small part of the total RNA. Furthermore, mRNA transcript levels vary considerably between genes, causing transcripts of highly-expressed genes to be overrepresented in the cDNA library and the sequencing results. Library normalisation attempts to equalise transcript levels in cases where the goal of the experiment is rather qualitative than quantitative. Alternatively, high sequencing depth can compensate for the differences in transcript levels. Prior to sequencing, the

cDNA is fragmented by sonication or DNase I treatment. However, several protocols involve fragmentation already before cDNA construction, by RNA hydrolysis or nebulisation. Finally, the cDNA fragments are sequenced with NGS. Which sequencing strategy is used depends on the goal of the RNA-seq experiment. For example, for the *de novo* development of a transcript catalogue without reference genome, longer reads, preferably from paired-end sequencing, are advantageous. For profiling of transcript levels, short reads are sufficient. Sequencing results in large RNA-seq datasets which provide the starting point for extensive computational analyses.

Analysis of RNA-seq data largely relies on two different bioinformatics approaches, that is, mapping and assembly, which depend on the availability of a reference. The first approach, mapping, is usually applied for organisms, for which high-quality genome sequences are available, and involves the alignment of millions of short reads, to identify their genomic locations and to assign them to transcripts and genes. The second approach, assembly, applies when reference is not available or is of low quality and aims at reconstructing the transcript sequences *de novo*, meaning without any already known sequence information. Note that these two approaches are not mutually exclusive. Successful reconstruction of transcript sequences often provides a reference for the mapping of RNA-seq reads. Conversely, for some applications, the prior assembly of the short reads into longer contigs often improves the mapping accuracy and reduces ambiguously mapped reads.

1.4 Thesis objectives

Aim 1: In the last few years, *Nothobranchius furzeri* has been established as a model organism for the studies of vertebrate ageing, which is mainly based on its exceptionally short lifespan and the presence of typical ageing-related characteristics (Genade et al. 2005). One prerequisite for a model organism is the availability of comprehensive genetic and genomic resources, for example to design experiments and studies. However, sequences for *N. furzeri* are still limited, that is, only a few genes have been sequenced so far as part of initial studies. Consequently, the first aim of this thesis is the development of a transcript catalogue, which provides a comprehensive sequence resource to the scientific community and thereby facilitates the acceptance of *N. furzeri* as an alternative model organism for vertebrate ageing.

Objectives: To this end, the *N. furzeri* transcriptome is sequenced using the Sanger technology as well the NGS technologies 454/Roche and Solexa/Illumina. Bioinformatics tools and approaches for the processing, assembly and annotation of transcriptome data are developed, which specifically account for the characteristic properties of NGS (high throughput, short read length). These are then applied to the *N. furzeri* datasets to develop a transcript catalogue that contains transcript sequences for the majority of protein-coding genes.

Aim 2: In addition, several *N. furzeri* strains, derived from populations living under different climatic conditions, exhibit large differences in lifespan (Terzibasi et al. 2008). For example, the short-lived strain GRZ has a maximum lifespan of 12-16 weeks, whereas the longer-lived strain MZM-0403 has a maximum lifespan of 29-32 weeks. Crosses between the two strains showed an intermediate lifespan (Kirschner et al. 2011). These observations suggest that the differences in lifespan between the *N. furzeri* strains are genetically determined. Therefore, the second aim of this thesis is to characterise transcript level changes in ageing *N. furzeri*.

Objectives: Transcriptomes of two tissues (brain, skin) and two time points (young, old) from two strains (GRZ, MZM-0403) are sequenced with Solexa/Illumina, and the transcript catalogue developed in (1) is employed to quantify transcript levels in these samples. The resulting transcriptome profiles are used to determine significant changes in transcript levels between young and old *N. furzeri* in general and to identify strain-specific differences between GRZ and MZM-0403.

2 Methods

The development of a transcript catalogue is a complex process which involves the generation and analysis of large amounts of data in a series of different analysis steps. Manually running these steps and managing their results is challenging or almost impractical. Thus, I decided to use an already available pipeline for the automated analysis of transcriptome data, EST2uni.

2.1 Analysis pipeline for transcriptome data

2.1.1 EST2uni

EST2uni (*EST analysis software TO create an annotated UNIgene database*) was published in 2008 and is an highly-configurable open-source pipeline for the pre-processing, assembly and annotation of expressed sequence tags Forment et al. 2008. The pipeline includes all major steps of EST pre-processing and assembly. Annotation of the resulting contigs relies largely on the sequence similarity searches with the Basic Local Alignment Search Tool (BLAST) against known databases with known sequences. The results are provided on a user-friendly website, which allows complex queries and data mining operations for sequence retrieval. EST2uni is written in the programming language Perl, and associated data storage is managed by a central MySQL database. The different analyses can be run in parallel on a multicore computer system or on a computing cluster using a batch-queuing system. EST2uni has been used in a number of EST projects Sunagawa et al. 2009; Lee et al. 2010; Yoshida et al. 2010 and serves as basis and framework for the analysis of the *N. furzeri* transcriptome data.

2.1.2 Installation and general modifications

I installed the pipeline as well as all additionally required programs and set up the MySQL database. The EST2uni website was installed on a local server of the Genome Analysis group at the FLI. Subsequently, to run the EST2uni pipeline successfully on the large transcriptome datasets generated for *N. furzeri*, a number of modifications were made to the pipeline as well as the database.

Most importantly, the tools for transcriptome assembly implemented in EST2uni work only with Sanger data but cannot process NGS data. Thus, I developed a separate pipeline to account for the different sequencing technologies (explained later). The resulting contigs were then fed into the EST2uni pipeline. Moreover, EST2uni annotation routines were extensively modified to speed up analysis of large transcript contig numbers. BLAST and HMMER were run in parallel on a computing cluster with up to 64 processes. For this purpose, I wrote the Perl script runSGE.pl, which splits a job into smaller sub-jobs that are then subsequently distributed across the cluster. Additionally, the BLAST parameters were optimised for speed with only slightly lowered sensitivity. Overall, these

modifications considerably reduced the time needed for a complete `EST2uni` analysis of the *N. furzeri* transcriptome data. Likewise, the database was modified to adequately deal with the large transcriptome datasets generated for *N. furzeri* (Supplementary File 1). Only Sanger and 454/Roche reads were stored in the database, and Solexa/Illumina reads were simply stored on the local file system. Several search indices were added to the database, which brought additional speed-up especially for large tables. Moreover, tables were modified to hold more detailed information about the results, and new tables were created for results of additional analyses run outside the `EST2uni` pipeline as part of the *N. furzeri* transcriptome annotation.

2.2 Transcriptome sequencing

2.2.1 Analysed fish species and strains

The two *N. furzeri* strains GRZ and MZM-0403 are being actively maintained at the FLI. Up to 12 fish per strain were kept in 40 litre tanks at 26°C under a light regime of 12:12 h light:dark and fed on red mosquito larvae (*Chironomidae*) *ad libitum* once a day. Water was filtered using air-driven foam filters; the water was changed once a week. More information, descriptions and laboratory protocols on care and breeding of the two strains have been published online (Genade et al. 2005; Genade 2007; Terzibasi et al. 2008). The zebrafish strain TüAB (AB/Tübingen) is a wild-type line and is also being actively maintained at the FLI. The fish were kept in groups in an open circulating standard zebrafish system (Aqua Schwarz).

2.2.2 RNA-isolation, cDNA library construction and sequencing

For *N. furzeri*, total RNA was isolated from skin and brain using the RNeasy Mini kit (Qiagen) and from whole body using the RNeasy Midi kit (Qiagen). Males and females were used as given in Supplementary Table 1. For zebrafish, total RNA was isolated from skin of male specimens aged 5 and 42 months, respectively, using Trizol (Invitrogen).

For Sanger sequencing (library #1), a normalised cDNA library was built by Evrogen and cloned into a pAL17.3 plasmid using the SMART cDNA Library Construction kit (Clontech). Recombinant plasmids were then amplified in *Escherichia coli*, purified, sequenced from both ends with the BigDye Terminator v.2.1 Cycle Sequencing Kit (ABI) and separated on ABI 3730xl capillary sequencers. The resulting trace files were processed by the in-house pipeline `Converge` to generate Sanger reads (http://genome.fli-leibniz.de/).

For 454/Roche sequencing, normalised and non-normalised cDNA libraries with ligated 454/Roche adaptors were prepared by Evrogen (library #2 and #3) and library #4 by Vertis. The resulting cDNA libraries were diluted and subjected to emulsion PCR according to the manufacturer's instructions (454/Roche Diagnostics). Sequencing was performed on 70×75 picotiter plates resulting

in two and a half runs using the older GS LR70 (FLX) sequencing kit and two runs with the new XLR70t (Titanium) kit, respectively.

For Solexa/Illumina sequencing, *N. furzeri* libraries #5-13 and zebrafish libraries were prepared using the mRNA-Seq sample prep kit 8 (Solexa/Illumina) according to the manufacturer's instructions. Clusters were generated using the Single Read Cluster Generation Kit v4 or the Paired End Cluster Generation Kit v4, respectively. Each library was loaded onto one lane of the flow cell at a concentration of seven Pico molar. Sequencing of *N. furzeri* and zebrafish libraries was performed on a Genome Analyzer IIx (Solexa/Illumina) for either 76, 101 or 150 cycles, using the 36 Cycle Sequencing Kit v4 following the manufacturer's protocol.

2.2.3 Read pre-processing and quality control

In 454/Roche reads, specific primer sequences ligated during cDNA library preparation were removed with the in-house `Perl` script `454-Primer-Trim.pl`. Subsequently, Sanger and 454/Roche reads were processed in a similar fashion. Low-quality reads were trimmed with `LUCY` (Chou & Holmes 2001). Vector, adaptor and poly(A)-stretches were removed using `SeqClean` (TIGR 2008). For this purpose, all vector sequences used for cDNA library preparation, normalisation and sequencing were compiled in one database. Additionally, standard vector sequences from the UniVec database (Kitts et al. 2012) of the National Center for Biotechnological Information (NCBI) were included. Low complexity regions were lowercase-masked by `SeqClean`. Additionally, `RepeatMasker` (Smit et al. 1996) was run with a library of Nothobranchius-specific repeats (Reichwald et al. 2009) to lowercase-mask complex repetitive elements. Finally, processed Sanger and 454/Roche reads were filtered for a minimum length of 80 bp and loaded into the `EST2uni` database. Solexa/Illumina reads were processed with the `String Graph Assembler` (`SGA`; Simpson & Durbin 2010). Reads with more than 20 low-quality bases were discarded with the `preprocess` command of `SGA`. Moreover, errors were corrected with the `correct` command, and duplicate reads removed with `rmdup` command.

2.3 Transcriptome assembly

2.3.1 Assembly of Sanger and 454/Roche reads

Sanger and 454/Roche reads were assembled with the classical overlap-layout-consensus strategy (Myers 1995). I tested three different programs specifically developed for the assembly of transcriptome data: *The Gene Indices Clustering tools* (`TGICL`; Pertea et al. 2003), *Newbler* (454 Life Sciences 2012) and *Program for Assembling and Viewing ESTs* (`PAVE`; Soderlund et al. 2009). `TGICL` was run with default parameters except for `-c 8 -O " -p 90"`. `Newbler` and `PAVE` were run with default parameters. Resulting contigs were filtered for a minimum length of 300 bp.

For each assembly, the number and total length of contigs as well as average, median and maximum contig length were recorded. Additionally, to estimate the success at assembling larger transcripts, the number and the fraction of contigs larger than 1 kb were computed. According to these metrics, PAVE produced the best assembly of Sanger and 454/Roche reads. Therefore, I chose this assembly as backbone for the integration of the Solexa/Illumina reads in the second step.

2.3.2 Integration of Solexa/Illumina data

Solexa/Illumina reads were assembled onto the PAVE contigs using the *CLC Assembly Cell* (CLCbio 2011), which is a *de Bruijn* graph-based assembly tool and was specifically developed for NGS reads. The Solexa/Illumina libraries were incrementally added to the assembly, that is, in each iteration, one single library was assembled onto the contigs of the recent assembly. Recent refers in the first iteration to the PAVE assembly and in all other iterations to the assembly of the previous iteration. The actual steps of the iterative assembly procedure were as follows (Figure 4): At first, Solexa/Illumina reads were mapped onto the contigs of the recent assembly using the *Burrows-Wheeler Alignment tool* (BWA; Li & Durbin 2009). Subsequently, unmapped reads served, together with the contigs, as input for the first assembly by the *CLC Assembly Cell*. Several assembly artefacts such as palindromic joins and N-stretches were fixed by custom Perl scripts. To further merge and extend the resulting contigs, they were re-assembled with TGICL, which can find and utilise weaker overlaps due to its classical overlap-layout-consensus approach. Since TGICL cannot use sequences larger than 2 kb as input, such contigs were split into 2 kb-fragments with a 300 bp overlap. Last, CD-HIT-EST (Li & Godzik 2006) was used to filter highly redundant contigs (99% alignment identity and 90% coverage of the smaller sequence), and contigs smaller than 300 bp were removed. The resulting contigs served as input for another assembly pass with relaxed parameters. After two assembly passes, the new assembly consisted of the previous one plus new sequence information introduced by the respective assembled Solexa/Illumina library. This assembly then served as input for the next iteration which integrated the reads of the next library.

After integrating the Solexa/Illumina libraries, the resulting assembly was improved by mapping GRZ reads onto the contigs and correcting the respective sequence. Therefore, GRZ libraries were mapped with BWA onto the contigs of the assembly. The resulting alignments were processed with SAMtools (Li et al. 2009). Alignments were filtered for a minimum quality of 10, and duplicate read alignments, suggestive for PCR duplicates, were removed. Filtered alignments then served as input for the BCFtools (Li et al. 2009) to determine GRZ-specific sequence differences, which were then used to edit the contig sequence accordingly. The edited transcript contigs constituted the final assembly of Sanger, 454/Roche and Solexa/Illumina data.

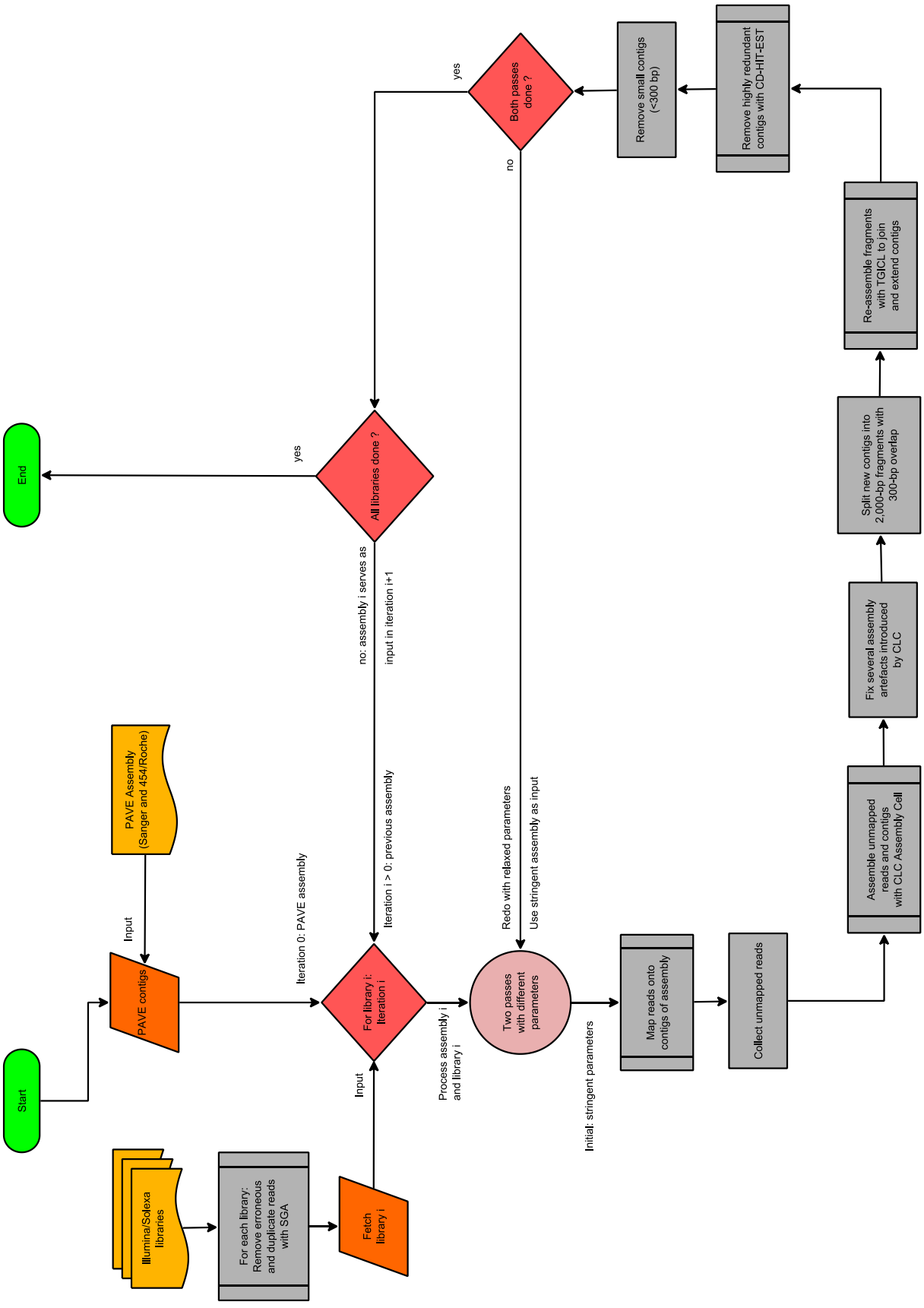

Figure 4: Flow chart of the iterative assembly procedure used for the Solexa/Illumina data.
Symbol legend: circle - labelled connector; parallelogram - input data; rectangle - processing step; rectangle with double-struck vertical edges: subroutine, external program; rectangle with wavy base - input file; rounded rectangle - start, end; rhombus - decision, iteration

2.3.3 Contamination analysis

Contamination analysis, that is, the identification of accidentally introduced foreign DNA, was done with BLAST. Ensembl transcript sequences from the fish species medaka, stickleback, tetraodon and zebrafish (Flicek et al. 2011) were pooled together with RefSeq transcript sequences from various non-fish species (Pruitt et al. 2005) to build a BLAST contamination database. *N. furzeri* transcript contigs were compared against this database using the program `tBLASTx` of the `WU-BLAST` package (Gish 1996). Only transcript contigs with BLAST hits (e-value $\leq 10^{-20}$) were subjected to further analysis. A transcript contig was considered as putative contamination if the best BLAST hit referred to a non-fish species and the e-value of the next fish-specific hit was at least 10 magnitudes higher. These criteria reduced the probability that BLAST mistakenly identifies a *N. furzeri* transcript contig as contamination. The identified transcript contigs were tagged as putative contamination and excluded from further analyses.

2.3.4 Post-processing and database import

For the import into the `EST2uni` database, the transcript contigs were renamed according to the naming scheme *Nofu_GRZ_cDNA_Y_XXXXXXX*, where a name starts with the prefix *Nofu_GRZ_cDNA* to indicate that this sequence is a *Nothobranchius furzeri* GRZ transcript contig, followed by a single-digit number for the assembly version and a unique seven-digit number for the contig. As an example, *Nofu_GRZ_cDNA_1_0000001*, *Nofu_GRZ_cDNA_2_0000001* and *Nofu_GRZ_cDNA_3_0000001* are contig names from three different transcriptome assemblies. Note that the first and the second name indicate earlier, deprecated assembly versions. The current assembly described here is the version three. Renamed transcript contigs were loaded into the `EST2uni` database.

2.4 Transcript annotation pipeline

2.4.1 BLAST searches

For BLAST searches, the `WU-BLAST` package (Gish 1996) was used, instead of the preinstalled `NCBI-BLAST` package (Altschul et al. 1990), and the `EST2uni` pipeline was modified accordingly. BLAST databases were built from several sequence resources (Table 1). Protein and protein-coding transcript sequences of the four fish species medaka, stickleback, tetraodon and zebrafish were downloaded from Ensembl and human protein sequences from RefSeq. Two large protein databases, the non-redundant protein database nr (National Center for Biotechnological Information 2011) and the UniProt database (Boeckmann et al. 2003) were downloaded as well. Fish-specific ESTs from the NCBI UniGene database (Pontius et al. 2003) complemented the set of BLAST databases. BLAST program options were defined for each database individually based on recommendations found in the

book *BLAST* (Korf et al. 2003). An overview of all databases including BLAST program options and hit criteria for hits is given in Supplementary Table 2.

Table 1: Protein and protein-coding transcript databases used for BLAST searches.

Database	Description	Maintained by	Entries	Program	Max. e-value
Ensembl fish proteins	Protein sequences from medaka, stickleback, tetraodon and zebrafish	Ensembl	116,833	`BLASTx`	10^{-07}
Ensembl fish transcripts	Transcript sequences from medaka, stickleback, tetraodon and zebrafish	Ensembl	123,997	`tBLASTx`	10^{-07}
UniProt	Large, partially curated, collection of protein sequences	EBI	17,035,495	`BLASTx`	10^{-07}
NCBI nr proteins	Collection of proteins from SwissProt, PIR, PDB; identical sequences merged	NCBI	15,270,974	`BLASTx`	10^{-07}
Refseq human proteins	Curated human protein sequences from the RefSeq database	NCBI	33,950	`BLASTx`	10^{-07}
NCBI UniGene transcripts	Clustered EST from catfish, cod, medaka, minnow, mummichog, pufferfish, salmon, stickleback, trout, zebrafish	NCBI	243,046	`BLASTn`	10^{-20}

BLAST searches were run in parallel on the available computing cluster using the `Sun Grid Engine` batch-queuing system and the Perl script `runSGE.pl`. Resulting BLAST reports were parsed with `BioPerl` (Stajich et al. 2002). The original `EST2uni` pipeline extracts only hit name and description, alignment coordinates, identity, similarity and e-value. Additional details on protein or transcript function, gene, species and cross references to other databases were retrieved through local installations of the Ensembl Core MySQL databases and the NCBI gene repository. Alternatively, this information was parsed from the description lines of the BLAST hits. Finally, the three best BLAST hits of each transcript contig were stored in the `EST2uni` database.

2.4.2 Transcript contig annotation

The BLAST results served as input for the annotation step of `EST2uni`. Basically, for each transcript contig, the description of its best BLAST hit constituted the new annotation for the transcript contig. For an optimal annotation, this process was done in three major steps: (1) filtering of BLAST hits, (2) annotation based on the BLAST hits, and (3) validation and correction of annotations.

Filter criteria for BLAST hits included maximum e-value, minimum overlap length and identity. These were defined for each database individually and are summarised in Supplementary Table 2. Furthermore, BLAST hits with descriptions that are inapplicable for annotation (for example *whole genome shotgun*) were excluded.

The filtered BLAST results against the protein and transcript databases were used for the annotation of the transcript contigs. For this purpose, I defined an annotation order, in which the BLAST results against the different databases are examined. Because the four fish species medaka,

stickleback, tetraodon and zebrafish were the closest relatives of *N. furzeri*, annotation was preferentially based on BLAST hits against the Ensembl protein and protein-coding transcript sequences of those species. In case annotation of the transcript contig was not successful, hits against human protein sequences from the well-curated and reliable RefSeq database were examined next. The next two databases in the annotation order were the two large protein sequence collections NCBI nr and UniProt, which were supposed to identify all remaining transcript contigs not yet annotated. Fish-specific ESTs were the last resort for annotation. BLAST hits against these sequences at least indicated that the transcript contig is indeed transcribed and not some sequencing artefact.

BLAST hits with uninformative descriptions were replaced, if possible. The corresponding BLAST hit was put back, and another hit with a similar score but a more informative description was searched as replacement. In case, no such BLAST hit was found, annotation reverted back to the original hit. Descriptions containing the following words were considered as uninformative: *novel*, *predicted*, *unknown*, *uncharacterized*, *unnamed* and *hypothetical*.

Assigned annotations were validated, and potential annotation errors were corrected, if possible. Potential annotation errors occurred primarily due to (i) obvious misannotations where other annotations with much higher scores exist and (ii) non-informative annotations which can be safely replaced by more informative annotations with comparable scores. These two issues could be relatively easily fixed by removing the problematic annotation followed by re-examining the remaining available BLAST hits. The (iii) cause for annotation errors were paralogous genes which originate from duplication events and are frequent in fish genomes. Since the copies can retain a relatively high sequence homology, BLAST annotation cannot always reliably distinguish between them. Comprehensive information about potential paralogous genes in different species is collected in the Ensembl Compara database (Vilella et al. 2009). For annotations obtained from Ensembl databases, this information was used to check for misannotations. The corresponding Ensembl gene ID of the annotation was used to fetch all paralogous genes. If other Ensembl BLAST hits predominantly pointed to a paralog, then this indicated a misannotation and, the annotation was corrected accordingly. Overall, these measures reduced annotation errors by including additional information.

2.4.3 Protein domain prediction

Protein domain prediction supplemented the BLAST-based annotation. Transcript contigs without annotation were searched for domain motifs which allow at least an initial annotation based on the function of the protein domain. To do so, transcript contigs were translated into putative proteins, and the proteins were compared against a database containing motifs of known protein domains.

Prediction of putative proteins was done with the external polypeptide prediction pipeline `prot4EST` (Wasmuth & Blaxter 2004), which utilises BLASTs against protein databases and `ESTScan` predictions (Iseli et al. 1999) to determine and translate the coding sequence (CDS).

Initially, the pipeline identified contigs derived from ribosomal transcripts, which are not translated, by running `BLASTn` searches against the SILVA ribosomal RNA gene database (Quast et al. 2013). The remaining transcript contigs were compared with `BLASTx` against Ensembl medaka proteins to identify nuclear proteins and against the GOBASE database (O'Brien et al. 2009), a collection of organelle sequences, to search for mitochondrial proteins. In order to use `ESTScan`, medaka protein sequences were reverse-translated following *N. furzeri*-specific codon usage frequencies, and an artificial *N. furzeri* transcriptome was built. Based on this artificial transcriptome, a *N. furzeri*-specific hidden Markov model was constructed that was used by `ESTScan` for predicting CDS. BLAST results and `ESTScan` predictions were fed into `prot4EST` to construct CDS. If neither BLAST results nor `ESTScan` predictions were available, the longest six-frame translation uninterrupted by a stop codon was used. Finally, predicted nuclear and mitochondrial CDS were translated into protein sequences using the standard and the vertebrate mitochondrial genetic code, respectively.

Conserved protein domains were identified with `HMMER` (Eddy 2012) by comparing the predicted protein sequences against the Pfam database (Punta et al. 2011), which contains conserved protein families with shared domains. Only Pfam-A entries, which represent high-quality, manually curated protein families, were downloaded from the Pfam website at the Sanger Wellcome Trust Institute (http://pfam.sanger.ac.uk/). `HMMER` was run in parallel on the computing cluster using the batch-queuing system. The results were parsed with `BioPerl`, and the hits were filtered for a maximum e-value of 10^{-20}. Last, the predicted protein domains were loaded into the `EST2uni` database to be used for annotation.

2.4.4 Gene ontology

Gene Ontology (GO; Ashburner et al. 2000) terms were assigned to *N. furzeri* transcript contigs based on BLAST hit to proteins which are already associated with GO terms. GO ontology and definition files (OBO v1.2) as well as GO term associations for the UniProt protein database were downloaded from the Gene Ontology web site (http://www.geneontology.org/) and stored in a MySQL database for faster access. Subsequently, GO terms were retrieved for transcript contigs with UniProt annotations based on the UniProt - GO term associations. For transcript contigs with annotations from Ensembl protein databases, the corresponding GO terms were directly obtained from the Ensembl MySQL server. For a broader overview, GO terms were mapped to GO Slim terms, which represent only high-level functional annotations, using the `map2slim` script (`go-perl` package, see Gene Ontology web site). Associated GO and GO Slim terms were stored in the `EST2uni` database.

2.4.5 Gene family analysis

Gene families were identified by taking advantage of the Ensembl Compara database, which contains gene family information for the four fish species medaka, stickleback, tetraodon and zebrafish. The longest *N. furzeri* transcript contig per gene was compared against Ensembl proteins of these four fish

species using BLASTx (e-value ≤ 10^{-07}). The Ensembl gene ID of the best BLAST hit was then used to retrieve the associated gene family. Thereby, *N. furzeri* transcript contigs, and genes, could be grouped into gene families.

A similar approach was followed to identify genes which are duplicated in *N. furzeri* as result of the teleost-specific genome duplication. The longest transcript contig per gene was compared using BLASTx (e-value ≤ 10^{-07}) against Ensembl protein sequences of medaka, stickleback, tetraodon and zebrafish. Subsequently, the corresponding Ensembl gene ID was used to fetch information about paralogous genes from the Ensembl Compara gene trees. In case, two different transcript contigs pointed to orthologous genes of two different species, the Ensembl gene ID of the first species was replaced with the corresponding ID of the orthologous gene in the second species.

To identify genes which are exclusively duplicated in *N. furzeri*, all transcript contigs belonging to the same gene were retrieved and translated into putative protein sequences. Only protein sequences with a minimum length of 100 aa were considered for further analysis. The protein sequence of the longest (primary) transcript contig was then aligned to each of the other (secondary) transcript contigs using the local alignment tool water from the EMBOSS package (Rice et al. 2000). The generated pairwise protein alignments were filtered for a minimum length of 100 aa, and protein sequence identities were calculated. The transcript contigs were then clustered based on their protein sequence divergence to the longest transcript contig, with a minimum distance of 10% between each cluster. Two or more clusters indicated genes that are duplicated exclusively in *N. furzeri*.

2.5 Transcriptome browser

2.5.1 Changes in structure and design

The EST2uni system includes a website which provides easy access to the generated results. To accommodate the website to the large amount of *N. furzeri* transcriptome data, a number of modifications were made. The original EST2uni website contained for each contig a graphical representation of the assembly which showed the individual reads that were used to build the contig. Because the large read numbers of the Solexa/Illumina datasets severely slowed down the website, this feature and several others which also involved the analysis of individual reads were removed from the website. Moreover, the website was adapted to the modified structure of the EST2uni database. For these reason, large sections of the PHP code were re-written. The changes allowed using the website for the *N. furzeri* transcript catalogue within good performance.

Also, the design of the website was changed to improve the presentation of the *N. furzeri* transcript catalogue. The website is now presented in a green-coloured theme, which follows the colours of the FLI's corporate design and distinguishes it from other transcriptome browsers available in the internet. The navigation bar was placed at the top of the web interface for a better overview, and

the individual result sections were re-arranged to separate them more clearly. Finally, a number of layout problems, such as inconsistent font usage or formatting errors, were fixed.

2.5.2 New features

Overall, the transcriptome browser now provides many more details regarding the transcript contigs and their annotation. Results of additional novel analyses were also integrated into the transcriptome browser. Several new panels provide the additional information gained from these analyses which include for example orthologous and paralogous genes, predicted miRNA locations and gene expression values from RNA-seq experiments. Furthermore, the transcriptome browser is now connected to the *N. furzeri* genome browser, which maintains the current build of the *N. furzeri* genome. The transcriptome browser reports for each transcript contig its putative genomic location, and a special hyperlink allows quickly changing to the genomic region in the genome browser.

Moreover, the query interface for searching transcript contigs was extended by adding several new search options. These include several annotation-related criteria such as gene symbol, CDS coverage, protein domains or associated GO terms. Importantly, the user can restrict the query to the best transcript contig per protein or gene. This way, the user obtains only the best transcript contig but does not have to deal with smaller, unwanted transcript contigs. These features facilitate a faster and more efficient search for transcript contigs of interest.

2.5.3 Security issues

The original `EST2uni` website contained a number of serious security issues which had to be fixed before making the website publicly accessible. Most importantly, user input is now checked for malicious code fragments as well as important Unix commands and control characters. Such input is rejected from further analysis. Permissions were restricted as much as possible, to prevent unauthorised access to server and database. The database structure was changed so that only reading from the database is needed and writing to the database is deactivated, which effectively protects the database. Furthermore, the list of programs that can be run on the server is now explicitly limited to the few programs that are actually needed. Finally, minor changes were made to stop web robots, small automated web applications which systematically browse the web, typically to build web indices. In case of `EST2uni`, web robots constantly download all analysis results and cause massive data traffic for the `EST2uni` server. This behaviour was restricted by the use of the file `robots.txt`, which denies the access to the website for web robots. However, since these instructions are purely advisory and web robots can choose to ignore them, an additional JavaScript routine was implemented which identifies web robots and denies them.

2.6 Analysis of differential gene expression

2.6.1 Mapping and quantification of transcript levels

The libraries #6-13, sequenced by Solexa/Illumina, were used to quantify the transcript levels in *N. furzeri*. However, in contrast to their use during transcriptome assembly, the Solexa/Illumina reads were not subjected further error correction and duplicate removal since that would introduce a bias into transcript levels. Prior to read mapping, a reference sequence was constructed from the longest transcript contig per gene, to avoid multiple mappings to transcript fragments and isoforms. Read mapping was done with `BWA`. For this purpose, a *Burrows-Wheeler Transform* index was constructed from the *N. furzeri* reference sequence. Subsequently, `BWA` was run with default parameters to map the Solexa/Illumina reads against the indexed reference. Only unique mappings were accepted. The resulting alignments were filtered for a minimum alignment quality of 10 and then sorted by coordinate to serve as input for the following quantification.

Read alignments were parsed and transcript levels determined by a custom `Perl` script. The transcript level of a gene was inferred from the number of reads mapping the corresponding transcript contig. Total read numbers were then normalised as *reads per kilobase of exon model per million mapped reads* (RPKM). All counts and RPKM values together with information about the corresponding samples were stored in the `EST2uni` database.

2.6.2 Statistical analysis and identification of DEGs

Statistical analysis was done within `R`, a software environment for statistical computing and graphics, and data plotting with the `R` package `ggplot2` (Wickham 2009), if not stated otherwise. Counts and RPKM levels for transcripts were directly fetched from the `EST2uni` database using the `RMySQL` package (Horner 2012) and imported into `R`. Correlations between the samples were calculated with the standard function `cor`, Spearman method, and RPKM values as input. The graphical representation of these correlation values was produced with the `heatmap.2` function of the `gplots` package (Warnes 2011). The principal component analysis (PCA, was done with the `R` function `prcomp` and the RPKM values as input.

Identification of differentially expressed genes (DEGs) was done with the `DESeq` package (Anders & Huber 2010). Raw read counts were used as input. Samples of the two different strains were treated as replicates, and transcript levels were compared between young and old *N. furzeri*. DEGs were called following the instructions in the `DESeq` package vignette and filtered for a maximal p-value (adjusted for multiple testing with the Benjamini-Hochberg procedure) of 0.01.

DEGs were submitted to the *Database for Annotation, Visualisation and Integrated Discovery Analysis* (Dennis et al. 2003; Huang et al. 2009) to identify enriched biological functions. Input gene

lists were constructed from the respective gene symbols of the identified *N. furzeri* DEGs. Gene symbols which were not recognised were exchanged for synonymous symbols, listed in the NCBI Gene repository. Clusters of enriched biological functions were filtered for a minimum enrichment score ≥ 1.

2.6.3 Confirmation by qPCR

DEGs were confirmed by qPCR in MZM-0403. RNA was extracted from the same skin and brain samples analysed in the RNA-seq experiment. Subsequent cDNA synthesis was performed in a 20 µl volume using 500 ng total RNA, 10 pmol oligo(dT) primer and 200 U SuperScript II reverse transcriptase (Invitrogen). Real-time PCR was performed with the SYBR GreenER qPCR SuperMix (Invitrogen) using the iCycler iQ5 detection system (Bio-Rad). Primer sequences are given in Supplementary Table 5. Cycle threshold values were normalised to the gene *INSR* (insulin receptor) which already showed to be very stably expressed at different ages. Fold changes were determined with the relative expression software tool (Pfaffl et al. 2002).

2.6.4 Confirmation of DEGs in zebrafish

Sequenced skin samples of young and old zebrafish were used as confirmation of the ageing-related DEGs identified in *N. furzeri*. The reads were mapped with `Bowtie` (Langmead et al. 2009) against the zebrafish genome (*danRer7*) and a precompiled database of exon junction sequences. Zebrafish DEGs were called from raw read counts following the same procedure and maximal adjusted p-value applied for *N. furzeri*.

Orthologous genes between *N. furzeri* and zebrafish were identified from best bidirectional hits between *N. furzeri* putative protein sequences (see 2.4.3) and zebrafish protein sequences from Ensembl. Only the longest protein sequence per gene was used for the comparison. *N. furzeri* protein sequences were compared with `BLASTx` against zebrafish protein sequences, and vice versa. A *N. furzeri* gene and a zebrafish gene were considered orthologous, if the corresponding protein sequences were identified as the best hits in the two reciprocal BLAST searches.

2.7 Data access

The *N. furzeri* transcriptome sequencing data was made publicly available in the following public databases: The Sanger reads were submitted to the NCBI dbEST database (Boguski et al. 1993) under accessions JZ200028-JZ330399. Since dbEST only accepts processed high-quality sequences but no raw data, not all Sanger reads could successfully deposited at dbEST. 454/Roche and Solexa/Illumina reads were submitted to the NCBI sequence read archive (SRA; Kodama et al. 2011) under submission accession SRA050046. Zebrafish Solexa/Illumina reads were submitted to the NCBI SRA under submission accession SRA054207.

The *N. furzeri* transcriptome assembly was submitted to the NCBI transcriptome shotgun assembly archive (TSA; Benson et al. 2012) under the project accession GAIB00000000. The accession for this assembly version is GAIB01000000. Future versions of the assembly will have different accessions, for example GAIB02000000, GAIB03000000 and so on. The individual transcript contigs can be retrieved with the accessions GAIB01000001-GAIB01210031. To comply with TSA requirements, contigs were modified as follows: Transcript contigs with stretches of 15 or more Ns into separate contigs, and the parts were numbered accordingly. Only contigs > 200 bp were submitted.

The NCBI BioProject constitutes a collection of biological data associated with a biological project or imitative (Barrett et al. 2011). For the *N. furzeri* transcriptome project, a BioProject was set up under the accession PRJNA85613. The project website provides all relevant information about the *N. furzeri* transcriptome including reads, assemblies, publications and related projects (http://www.ncbi.nlm.nih.gov/bioproject/PRJNA85613).

3 Results

The results of my thesis are grouped in four main parts. The first part, described mainly in the methods section, covers the installation of a transcriptome analysis pipeline and of a transcriptome browser as well as all necessary modifications. The second part (results section, 3.1) describes the development of a comprehensive catalogue of transcripts, which greatly improves the availability of genetic resources and provides sequence information for most genes in *N. furzeri*. In the third part (3.2), the transcript catalogue is compared against the protein sequences of four model fish species to obtain more insights into general relation to other fish species. In the fourth part (3.3), I quantitatively analyse the transcript level in young and old *N. furzeri* from the short-lived strains GRZ and the longer-lived strain MZM-0403 to characterise the age-dependent changes in these fish.

3.1 Development of a *N. furzeri* transcript catalogue

3.1.1 Transcriptome sequencing

3.1.1.1 Sample preparation and cDNA library construction

Initially, the short-lived *N. furzeri* strain GRZ was studied to construct a transcript catalogue. In total, 28 male and female GRZ individuals of ages between one and 14 weeks were used for tissue collection. RNA was isolated from the whole body of the fish and, if the amount of collected material permitted, from the two tissues brain and skin. RNA samples then served as templates for reverse transcription into cDNA. Depending on library type and applied sequencing protocol, cDNA libraries were constructed by external service providers as well as in-house approaches. The companies Evrogen and Vertis prepared the libraries #1-4 for Sanger and 454/Roche sequencing. Except for library #2, these libraries were normalised, equalising the abundances of cDNA molecules that reflect the RNA proportion in the cell. The remaining four libraries (#5-9) were prepared in-house for Solexa/Illumina sequencing. Thus a total of nine GRZ cDNA libraries were subjected to sequencing analysis (Supplementary Table 1).

For additional transcriptome sequencing in longer-lived *N. furzeri*, the MZM-0403 strain was selected. Of 17 individuals of ages between five and 31 weeks, brain and skin tissues were collected for RNA extraction. These RNA samples were used to construct four MZM-0403 cDNA libraries (#10-13) following the protocols for Solexa/Illumina cDNA library construction (Supplementary Table 1).

3.1.1.2 Sequencing data

First, the classical Sanger technology was applied for sequencing. For library #1, cDNA clones with lengths between 500-3,000 bp were picked and sequenced from both ends. After processing of the raw

sequencing signals, the resulting read dataset included 131,808 reads, representing 97 Mb of transcriptome data. The mean read length of 734 bp was typical for Sanger sequencing (Figure 5).

The low throughput and the high price/bp of the Sanger sequencing technology were prohibitive to capture transcripts with low abundances. Therefore, the NGS technology 454/Roche, introduced in the Genome Analysis group at the FLI in 2008, was applied to obtain higher read numbers. The libraries #2 and #3 were sequenced with the FLX chemistry of the 454/Roche system and the library #4 with the Titanium chemistry. The Titanium chemistry represents a technological enhancement of the FLX, and facilitates longer read lengths and higher sequencing output. Overall, 454/Roche yielded 1,131 Mb, thereof 262 Mb by FLX and 869 Mb by Titanium. Besides a higher sequencing output, the Titanium chemistry also produced longer reads (mean read length Titanium: 369 bp vs. FLX: 224 bp). Figure 5 also shows the read length distributions of the libraries #2-4, compared to the Sanger reads of library #1.

In 2010, *N. furzeri* transcriptome sequencing continued with the Solexa/Illumina technology, which produces several magnitudes more data. The hereby applied strategies included normal and paired-end sequencing, with differing read lengths (76, 101 and 150 bp). Thus, sequencing of the remaining libraries #5-13 resulted in 467 million reads with a total of 45,180 Mb, which represents by far largest portion of the primary *N. furzeri* transcriptome data.

In summary, 46 Gb transcriptome data was obtained from two different *N. furzeri* strains using three sequencing technologies with varying read lengths and data output. All sequenced cDNA libraries together as well as their raw sequencing outputs are listed in Table 2. The generated transcriptome data served as basis for both the assembly of a transcript catalogue as well as subsequent analyses of differential expression.

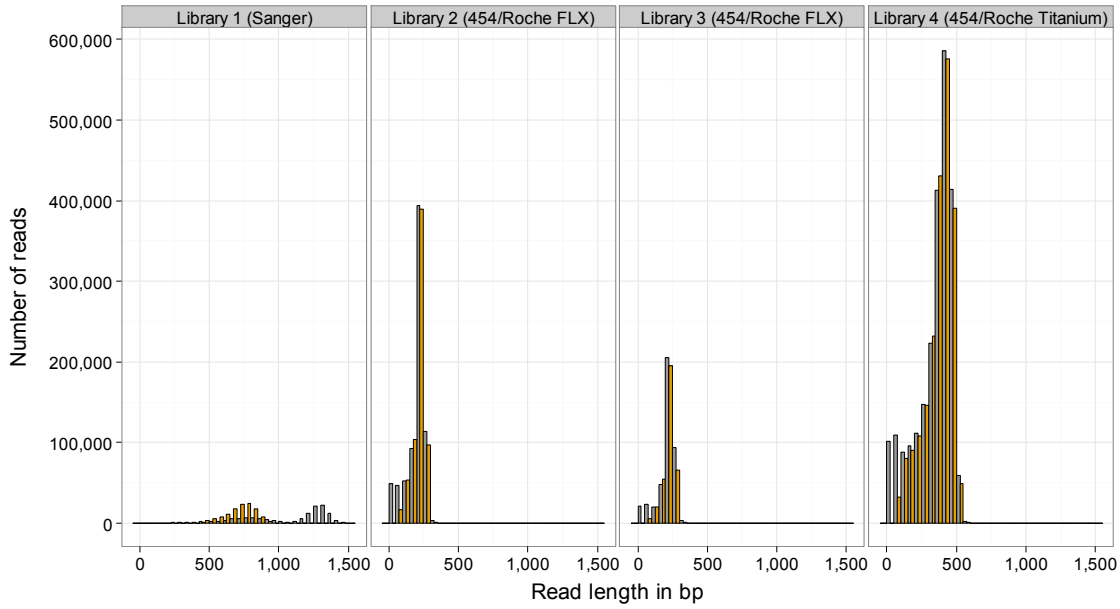

Figure 5: Read length distribution of Sanger and 454/Roche data.
Sanger and 454/Roche reads from libraries #1-4 are sorted according to their length in bins of 50 bp (x-axis). Grey and ochre bars show raw and processed reads, respectively.

3.1.1.3 Pre-processing of the generated datasets

Work in wet laboratories is never unbiased, that is, it is always subjected to a number of confounding factors. Therefore, in raw sequencing data, a number of quality problems can occur which include bad sequence quality, as well as sequencing artefacts and contamination. These issues have to be resolved before any further analysis. Thus, the transcriptome data were first quality-curated and processed accordingly. However, the available datasets derived from three sequencing technologies that show considerable differences in read length and output, especially when comparing reads from Sanger and 454/Roche with reads from Illumina/Solexa. Therefore, Sanger and 454/Roche datasets were processed similarly, but Solexa/Illumina datasets were processed differently.

Because Sanger and 454/Roche datasets contained small to medium-sized read numbers, they could be processed by tools designed for the output of traditional sequencing projects. Read processing included clipping of low-quality regions with `LUCY` as well as removing of vector sequence as well as of poly(A)-tails, and masking of low-complexity/repeat regions with `SeqClean`. Low-quality regions in the Sanger data comprised 6.1 Mb data. Vector sequence made up 0.1 and 2.9 Mb of the Sanger and 454/Roche data, respectively, and poly(A)-tails made up another 1.5 and 8.1 Mb. All these sequences were removed. In addition, reads shorter than 80 bp were discarded. Furthermore, repeat analysis with `RepeatMasker` identified 25 Mb of low-complexity/dispersed repeats. These sequences can create problems in similarity searches and, therefore, they were masked from further analysis. In summary, 89 and 1,001 Mb (92 and 89%) remained as Sanger and 454/Roche assembly-ready data, respectively.

In contrast, the Illumina/Solexa datasets contained shorter reads but in very large numbers, which prohibited read processing steps like those applied for Sanger and 454/Roche data. Instead, the Solexa/Illumina datasets were processed with `SGA`. Reads were quality-filtered, discarding 11,906 Mb low-quality sequence, and putative-sequencing errors were corrected. To reduce the amount of data used for the assembly of the transcript catalogue, exact-match read duplicates were removed. Applying these three processing steps to the Solexa/Illumina datasets resulted in an additional reduction of 13,996 Mb. In summary, 19,278 Mb (43%) high-quality Solexa/Illumina data remained after processing.

As a result of the applied processing steps, 26 Gb transcriptome data were discarded during pre-processing. A comprehensive summary of the processed data is given in Table 2. The remaining 20 Gb high-quality transcriptome data served as input for the subsequent transcriptome analyses.

Table 2: Transcriptome sequencing data.

Library	Specimen	Sequencing technology / layout	Raw data		Processed data		Sequence in Mb clipped due to:			Reads discarded due to:	
			Number of Reads	Total length in Mb (Mean length in bp)	Number of Reads	Total length in Mb (Mean length in bp)	Low quality	Vector/ Adaptor	Poly(A)	Length filter	Contam-ination
1	GRZ, 9w, whole body	Sanger	131,808	97 (734)	129,784	89 (684)	6.1	0.1	1.5	1,668	356
2	GRZ, 10w, whole body	454/Roche GS-FLX	751,885	168 (223)	663,286	141 (212)	-	0.8	2.3	85,506	3,093
3	"	454/Roche GS-FLX	415,426	94 (225)	342,940	75 (218)	-	0.3	1.8	42,898	29,588
4	GRZ, 8w, whole body	454/Roche GS-Titanium	2,351,236	869 (369)	2,138,135	785 (367)	-	1.8	4.0	202,652	10,449
			Raw data		**Processed data**		**Reads (Sequence in Mb) discarded due to:**				
			Number of Reads	Total length in Mb	Number of Reads	Total length in Mb	Low quality filter			Duplicate read filter	
5	GRZ, 1w, whole body	Solexa/Illumina, 2 x 150 bp	89,730,910	13,549	54,826,045	8,279	33,267,342 (5,023)			1,637,523 (247)	
6a	GRZ, 5w, skin	Solexa/Illumina, 2 x 101 bp	55,410,042	5,596	8,511,138	859	17,002,812 (1,717)			29,896,092 (3,020)	
6b	"	Solexa/Illumina, 76 bp	33,947,044	2,579	12,361,342	939	2,249,443 (170)			19,336,259 (1,470)	
7a	GRZ, 5w, brain	Solexa/Illumina, 2 x 101 bp	60,997,820	6,161	12,040,350	1,216	25,334,894 (2,559)			23,622,576 (2,386)	
7b	"	Solexa/Illumina, 76 bp	36,117,265	2,744	19,807,913	1,504	4,481,760 (341)			11,827,592 (899)	
8	GRZ, 14w, skin	"	35,963,921	2,733	14,594,109	1,109	5,139,078 (391)			16,230,734 (1,233)	
9	GRZ, 14w, brain	"	27,369,310	2,080	16,295,440	1,238	4,113,140 (313)			6,960,730 (529)	
10	MZM-0403, 5w, skin	"	31,727,935	2,411	7,207,454	548	11,966,406 (909)			12,554,075 (954)	
11	MZM-0403, 5w, brain	"	34,720,289	2,639	17,963,204	1,365	2,467,553 (188)			14,289,532 (1,086)	
12	MZM-0403, 31w, skin	"	30,268,649	2,300	11,848,006	900	1,185,720 (90)			17,234,923 (1,310)	
13	MZM-0403, 31w, brain	"	31,423,327	2,388	17,387,383	1,321	2,688,578 (205)			11,347,366 (862)	
	Total			**46,408 Mb**		**20,368 Mb**					

The table summarises the *N. furzeri* transcriptome data obtained by sequencing 13 cDNA libraries with the sequencing technologies Sanger, 454/Roche and Solexa/Illumina. For each library, the first six columns provide library number, the respective sequencing technology / layout and the data output, directly after sequencing ("raw") and after pre-processing ("processed"). For Sanger and 454/Roche, the table additionally specifies the average read length, while, for Solexa/Illumina, read length is determined by the sequencing layout. The remaining columns give an overview of the amount of raw data that was discarded during pre-processing. Note that, to account for the larger data output compared to Sanger and 454/Roche, Solexa/Illumina libraries were treated differently.

3.1.2 Assembly of a transcript catalogue

N. furzeri transcriptome sequencing produced large datasets from three different sequencing technologies, that is, Sanger, 454/Roche and Solexa/Illumina, which exhibited largely varying read lengths and output. To accommodate for these differences, I decided to assemble the transcriptome data in two steps. First, the Sanger and 454/Roche data was assembled with tools based on the traditional overlap-layout-consensus approach (Myers 1995) since these are well-suited for longer reads and small to medium-sized datasets. In the second step, the Solexa/Illumina libraries were successively integrated into the contigs generated from the Sanger and 454/Roche data. To deal with the large Solexa/Illumina datasets, an alternative assembly approach based on *de Bruijn* graphs (Pevzner et al. 2001) was used, which is particularly suited for NGS data. Ultimately, applying these two different approaches lead to an assembly which optimally combines the different sequencing technologies.

During assembly, it was necessary to evaluate the different assembly results and to quantify their improvements. Usually, assembly success is described by number of contigs, total contig length, mean and maximum contig length. These metrics have been originally defined for genome assembly where the aim is to build a complete genome sequence, that is, few large contigs which cover the entirety of the genome. In transcriptome assembly, however, this is different. Here, one aims at reconstructing as many transcripts as possible. Moreover, these transcripts have differing lengths, from some hundred to several ten-thousand base pairs. Thus, the metrics mentioned above alone are insufficient to describe the success of an assembly. Therefore, I additionally recorded the number of large contigs, that is, the number of contigs with a minimum length of one kb. Since most transcripts usually are longer than one kb, this metric provides a good proxy to estimate the number of complete transcripts and to assess the quality of the assemblies.

3.1.2.1 Assembly of the Sanger and 454/Roche data

Since Sanger and 454/Roche datasets contained long reads but in small to medium-sized numbers, I decided to assemble these reads with the overlap-layout-consensus approach, which makes optimal use of the available sequence data. Three different tools which employ this approach were tested for the assembly of transcriptome data: `TGICL`, `Newbler` and `PAVE` (see methods, 2.3.1, for parameters and references). The metrics of the three assemblies are summarised in Table 3.

Table 3: Transcriptome assembly metrics.

	First assembly*			Second assembly†
	TGICL	**Newbler**	**PAVE**	**Iterative assembly procedure**
Datasets	Sanger, 454/Roche	Sanger, 454/Roche	Sanger, 454/Roche	PAVE contigs, Solexa/Illumina
Number of contigs (≥ 300 bp)	134,225	141,973	118,795	213,621 (+80%)
Total contig length	85.1 Mb	78.1 Mb	86.9 Mb	252.9 Mb (+192%)
Median length	474 bp	434 bp	495 bp	
Mean length	634 bp	549 bp	731 bp	1,183 bp (+62%)
Maximum length	9,221 bp	7,841 bp	9,241 bp	64,116 bp (+594%)
Number of large contigs (≥ 1 kb)	15,516	11,756	23,534	79,035 (+235%)
Total length/Fraction of large contigs	24.9 Mb / 29%	19.1 Mb / 24%	38.3 Mb / 44%	183.0 Mb / 73% (+378%)

*Results of the three assembly programs that were tested for the Sanger and 454/Roche data ('First assembly'). †Subsequent assembly of the Solexa/Illumina data onto the PAVE contigs using the iterative assembly procedure outlined in methods, 2.3.2 ('Second assembly').

Evaluation of the three assemblies gave the following results: The Newbler assembly was last in all metrics, and therefore it was considered inferior to the other two assemblies generated by TGICL and PAVE. The TGICL assembly was slightly better, especially in maximum length and number of large contigs. The best assembly was produced by PAVE. This program generated the fewest contigs with the highest total contig length, and the largest mean, median and maximum contig length. In particular, the number and fraction of large contigs (≥1 kb) was much higher when compared with the Newbler and TGICL assemblies, and these contigs already made up a considerable fraction of the assembly. Furthermore, when comparing the contig length distributions of all three assemblies, PAVE provided the most contigs in almost each length bin; this was especially evident for larger contig lengths (Figure 6). To conclude, PAVE produced the best assembly of the available Sanger and Roche/454 data and, therefore, I chose this assembly as backbone for the second step, the integration of the Solexa/Illumina data.

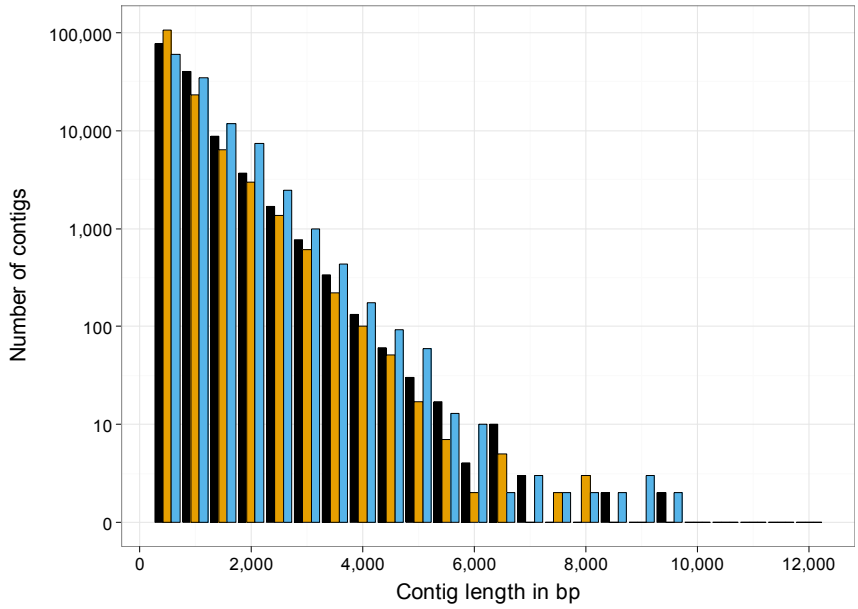

Figure 6: Contig length distribution of the Sanger and 454/Roche assemblies. Contigs are sorted according to their length in bins of 500 bp. Three different tools were used for assembly: `TGICL` (black), `Newbler` (ochre) and `PAVE` (blue).

3.1.2.2 Assembly of the Illumina/Solexa data

The Solexa/Illumina sequencing output was magnitudes larger than that of Sanger and 454/Roche, and, assembly required a different strategy to efficiently handle these large datasets. First and foremost, switching to the *CLC Assembly Cell*, a tool based the new *de Bruijn* graph approach, enabled an effective assembly within reasonable run times. Moreover, rather than processing all data in one assembly, the individual datasets were iteratively assembled onto the contigs of the most recent assembly; starting with the `PAVE` assembly of the Sanger and 454/Roche data (the iterative assembly procedure is outlined in the methods, 2.3.2). Thus, the successive integration of the Solexa/Illumina libraries extended existing contigs, and added new contigs to the assembly. Overall, these measures allowed an efficient integration of the available Solexa/Illumina data.

The iterative assembly procedure started with the Sanger and Roche/454 assembly generated by `PAVE`, which contained 118,795 contigs with a total length of 87 Mb. During integration of the individual Solexa/Illumina libraries, mean and maximum contig length as well as the number of large contigs improved continuously and considerably (Figure 8). After integration of all available Solexa/Illumina libraries, both number of contigs and total length were extensively increased, and the final assembly contained 213,621 contigs (+80%) with a total length of 253 Mb (+192%). The mean and maximum length of these contigs accounted for 1,183 and 64,116 bp, respectively, which represented a substantial improvement compared to the initial `PAVE` assembly.

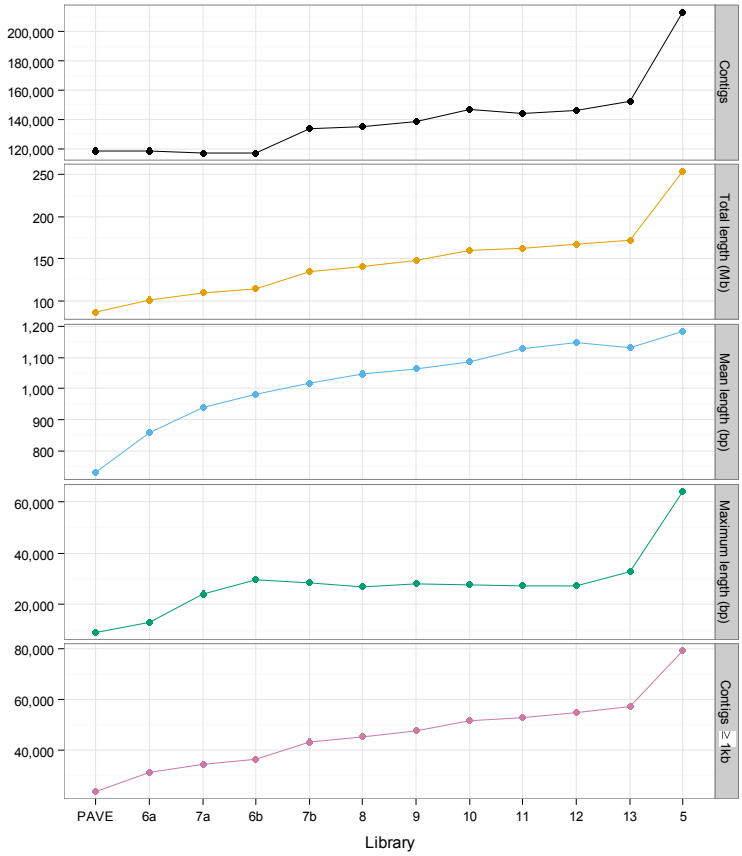

Figure 7: Progress of the iterative assembly of Solexa/Illumina libraries.
Assembly metrics were recorded for the starting PAVE assembly and all Solexa/Illumina library iterations. Mean and maximum contig length, and the number of large contigs (≥ 1kb) improve continuously and considerably.

It should be noted that these metrics are already close to those recorded for transcriptomes of other fish species. For example, the mean and maximum length of the medaka transcripts available in Ensembl (Flicek et al. 2011) account for 1,550 and 78,426 bp, respectively. Generally, compared to the PAVE assembly, the new assembly contained more very large contigs, that is, with lengths from 10 up to 64 kb (Figure 8).

Furthermore, the assembly also contained a much higher number of contigs larger than 1 kb, which made up almost 72% (183 Mb) of the total assembly. Again, when relating to the mean transcript length of 1,550 bp in medaka, the high fraction of these contigs suggested that many *N. furzeri* transcripts were almost half or even fully represented by a contig. Thus, in summary, the Solexa/Illumina libraries largely contributed to the generation of the final *N. furzeri* transcriptome assembly (Table 3).

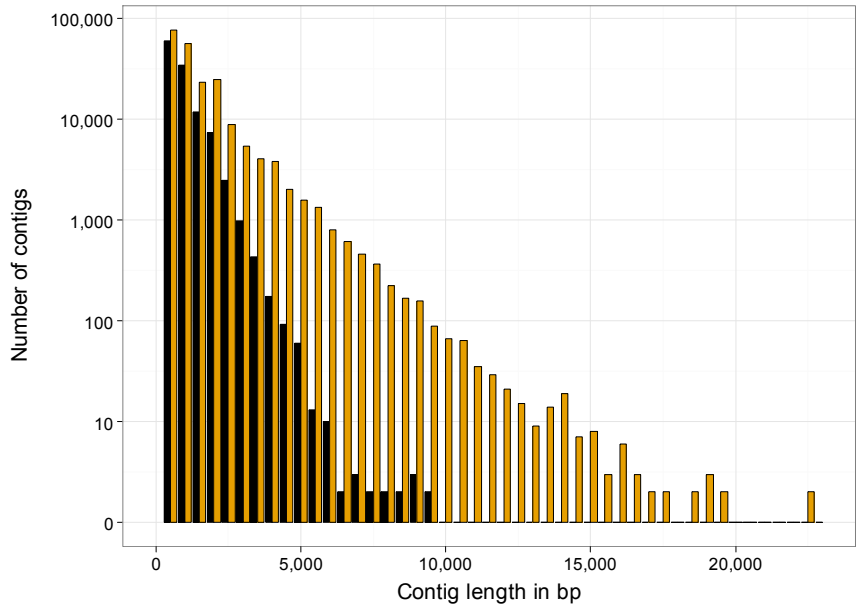

Figure 8: Contig length distribution before and after integration of the Solexa/Illumina data. Contigs are sorted according to their length in bins of 200 bp. Black bars show the Sanger and 454/Roche assembly generated by PAVE. Ochre bars show the final transcriptome assembly after successive integration of the Solexa/Illumina data.

3.1.2.3 Validation of the assembly

The *N. furzeri* transcriptome assembly provides valuable sequence information for further analyses and experiments. However, this means that the assembled transcript contigs will have a major impact on future analyses. Therefore, it is essential to evaluate the quality of the assembly to avoid the introduction of potential biases into present and/or future results. In the case of the *N. furzeri* transcriptome assembly, a first concrete example for a potential bias was observed in the contig length distribution which showed several unexpected peaks at different lengths. In the following, this artefact is discussed in more detail. Moreover, I analysed the contribution of each sequencing technology to the assembly and evaluated the contig quality by comparing them against a reference sequence. Last, contamination from different external sources, a common problem, is assessed. These analyses should give more insight into the quality of the *N. furzeri* transcriptome assembly.

3.1.2.3.1 Analysis of an assembly artefact

Usually, transcriptome assembly produces many short and medium-sized contigs, and relatively few long contigs. For this reason, the contig length distribution of the assembly typically resembles a bell-shaped curve that is skewed to the right side of the figure. This was also true for the *N. furzeri* transcriptome assembly, however, a closer examination of the rather smooth contig length distribution revealed an unexpected excess of contigs with lengths around 2 kb. Additionally, at least three other peaks could be observed at 3.7, 5.4 and 7.1 kb (Figure 9). These peaks indicated a potential assembly artefact which had to be examined in more detail before continuing with the annotation of the transcript contigs.

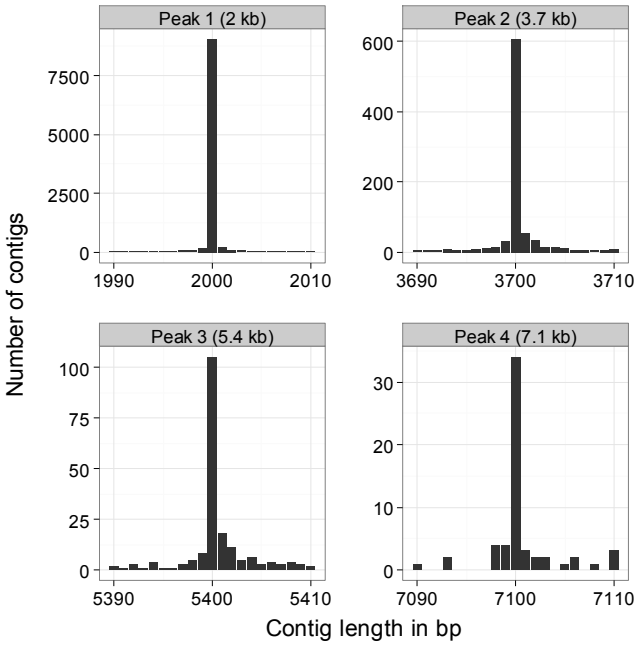

Figure 9: Unexpected peaks in the contig length distribution.
The contig length distribution reveals four unexpected peaks at 2, 3.7, 5.4 and 7.1 kb, which indicate putative assembly artefacts.

Concerning the origin of these unexpected peaks, I noticed that the first peak at 2 kb coincided with the fragmentation step used in the iterative assembly procedure. In this step, contigs are split into 2 kb-pieces with a 300 bp-overlap and then re-assembled with relaxed parameters to further join and extend contigs (see methods, 2.3.2). Contig pieces which were not re-assembled might have accumulated and thus contributed to the 2 kb peak. Respectively, the distance between the four peaks accounted for 1.7 kb, which equals to fragment length minus overlap. This observation suggested that the other three peaks also originate from the fragmentation procedure. The 3.7 kb-peak, for example, might contain contig fragments built from two 2 kb-pieces with a 300 bp-overlap. Thus, I hypothesised that these peaks represent a decent fraction of the artificial 2 kb-pieces which could not be extended by the assembler due to repetitive motives at one (resulting in the 3.7, 5.4 and 7.1 kb peaks) or both ends (2 kb peak).

To further investigate this question, all 2 kb-contigs (9,033; 7%) were compared against the remaining transcript contigs. If these contigs indeed originated from the fragmentation process, they would show at least partial overlaps to any of the other contigs. Almost all 2 kb-contigs (9,026; 99%) showed significant hits (`BLASTn`, e-value $\leq 10^{-60}$). The average overlap length and sequence identity of 379 bp and 97%, respectively, supported the assumption that these contigs originated from the fragmentation procedure in which a 300 bp-overlap was used. Moreover, of the contigs with hits, the large majority (8,793; 97%) also showed multiple hits, with an average of 18 hits per contig. These findings supported the assumption that the unusual peaks represent artificial 2 kb-pieces which could not or only partially be extended.

A number of possible explanations which might explain the failed extension of the split contigs can be thought of. Repeat-rich sequences, for example, are typically difficult to assemble and often result in contig fragmentation. However, the 2 kb-contigs did not show an elevated repeat content, compared to the overall transcriptome assembly (12% vs. 15%). Another explanation could be that these contigs originated from abundant transcripts. Higher read coverage, especially from Solexa/Illumina sequencing, causes more sequencing errors, which, in turn, can cause contig breaks during assembly. Indeed, when evaluating the Solexa/Illumina coverage (see below), the 2 kb-contigs were, on average, covered by more than twice as many reads, compared to the remaining transcript contigs (243 vs. 117 reads). Apparently, the majority of 2 kb-contigs represented transcripts with a high coverage, which might be problematic for transcriptome assembly.

3.1.2.3.2 Contribution of the different sequencing technologies to the assembly

The *N. furzeri* transcriptome assembly was built from a number of read datasets, which, depending on the applied sequencing technology, differed widely in read length and throughput and therefore, contributed to the assembly to different extents. However, these differences might lead to an assembly that is biased towards a single sequencing technology. For example, the longer reads of Sanger and 454/Roche sequencing might result in more full-length transcripts, compared to transcripts assembled from short Solexa/Illumina reads. This may have consequences for further down-stream analyses. For example, during transcript quantification using RNA-seq, longer transcripts are covered by more reads, which, in turn, result in higher transcript levels. Consequently, transcript levels would be biased by the sequencing technology used for assembly. To quantify the impact of the three different sequencing technologies in more detail, I assessed the contribution of each technology to the total contig consensus of the assembly.

For a thorough analysis of the sequencing coverage, transcriptome reads of all three technologies were mapped (Sanger and 454/Roche with `Newbler`; Solexa/Illumina with `BWA` (Li & Durbin 2009)) onto the contigs of the final transcriptome assembly. As a result, the vast majority of reads were mapped successfully (Sanger: 99%, 454/Roche: 92%, Solexa/Illumina: 97%), which suggested that the large majority of reads were actually used during assembly. Sanger, 454/Roche and Solexa/Illumina reads covered about 20, 56 and 237 Mb, respectively, of the 253 Mb total contig length. However, only 1 and 7 Mb was covered exclusively by Sanger and 454/Roche reads, respectively. In contrast, over 178 Mb of the total transcriptome assembly was covered solely by Solexa/Illumina reads, which shows the large contribution of the Solexa/Illumina technology to the assembly.

3.1.2.3.3 Evaluation using a set of medaka proteins as reference

While a number of metrics exist which evaluate the quality of a genome assembly, only a few have been proposed for transcriptome assembly, and only recently (Martin & Wang 2011). Moreover, these metrics rely on the availability of a reference, usually an already sequenced genome or a set of

transcripts, against which the assembled transcript contigs can be compared and evaluated. In the case of the *N. furzeri* transcriptome assembly, genome information was not available for most of the project duration, and only a few transcripts had already been fully sequenced. Therefore, medaka protein sequences served as substitute reference for an evaluation of the *N. furzeri* transcriptome assembly. Medaka is the closest relative with a sequenced genome, and protein sequence conservation between medaka and *N. furzeri* is on average high enough to allow for a meaningful comparison.

Medaka protein sequences from Ensembl served as reference for a basic evaluation of the *N. furzeri* transcriptome assembly. To avoid misinterpretation due to annotation errors in medaka, the set of protein sequences was limited to 1,750 entries with experimental support (Ensembl category *known*). The remaining entries, which are derived from BLAST projection and computational gene prediction, were excluded from the analysis. Altogether, 13,451 *N. furzeri* transcript contigs showed significant overlaps (`BLASTx`, e-value $\leq 10^{-20}$) against 1,290 medaka protein sequences (74%). This result implied that the transcriptome assembly already contained transcript contigs for the large majority of the protein sequences. More importantly, 74% of the protein sequences with hits were nearly completely covered by transcript contigs, that is, more than 90% of their length was hit. Moreover, for 67%, a single transcript contig was sufficient to achieve this coverage. Thus, it is reasonable to assume that many *N. furzeri* transcripts are already fully recovered by the transcriptome assembly.

The large number of transcript contigs indicated that a certain fraction of transcript contigs are actually fragments. The medaka protein sequences allowed assessing the degree of fragmentation of the *N. furzeri* transcriptome assembly in more detail. Fragmentation occurred at different regions of the medaka protein. In only 13% of all cases, the *N. furzeri* transcript contigs reached the N-terminus of the protein (within a tolerance of three amino acids). A similar fraction was found for the C-terminus (16%). For most of the proteins, the respective transcript contigs did not reach the N- or the C-terminus at all (64%). In only 7%, a transcript contig spanned the complete protein and additionally included untranslated regions (UTR).

Alternatively to fragmentation, transcript contigs can be redundant, meaning two or more contigs partially align to the same protein sequence. This was observed for the majority of the matched medaka protein entries. Over 92% of the total concatenated sequence was hit by at least two overlapping *N. furzeri* transcript contigs (nine contigs on average). Thus, it was reasoned that the *N. furzeri* transcriptome assembly contained multiple transcript contigs for the majority of genes.

Finally, the medaka proteins also allowed assessing the rate of putative chimeric transcript contigs in the *N. furzeri* transcriptome assembly. Chimeric transcript contigs derive from two different transcripts and usually indicate potential assembly errors. A chimeric *N. furzeri* transcript contig was required to have two non-overlapping hits to different medaka proteins which meet the following

criteria: e-value ≤ 10^{-20}, coverage of both protein hits ≥ 75% and identity ≥ 50%. According to this definition, 43 out of 13,451 (0.3%) transcript contigs were considered chimeric.

3.1.2.3.4 Analysis of contamination

Contamination can be inherent to the sequenced organism, such as undigested food or parasites, or is accidentally introduced during sample preparation in the laboratory. Unfortunately, during annotation, contigs derived from contamination may be mistaken for transcripts of the original sample and have to be excluded in advance. Therefore, `tBLASTx` (e-value ≤ 10^{-20}) compared all transcript contigs against a special database which contained both fish-specific protein-coding transcripts from Ensembl as well as non-fish transcripts from RefSeq (Pruitt et al. 2005). A contig was considered as contamination if and only if the best `tBLASTx` hit identified a non-fish species and the e-value of the next fish-specific hit was at least 10 orders of magnitude higher. This definition ensured that the transcript contig indeed originated from contamination instead from *N. furzeri*.

BLAST analysis marked 3, 961 of the 213,621 transcript contigs as putative contamination; these made up 2% (4 Mb) of the total transcriptome assembly. The contigs were rather small; the median length was 531 bp. The comparable median overlap length of 394 bp indicated that most contamination overlaps covered almost the complete transcript contig. However, mean overlap identity was only 64%. This might be explained by the limited representation of potential contamination species in the database which, naturally, cannot contain all contaminations common for *N. furzeri*. Consequently, the BLAST hits may not have identified the actual contaminating species but rather another, more distantly related species.

BLAST analysis identified most contigs as mammalian origin (1,576 contigs), especially human and chimp (*Pan troglodytes*), which hinted at a certain degree of contamination in the laboratory. Another group of 1,493 contigs contained invertebrate species. These included the purple sea urchin (*Strongylocentrotus purpuratus*) and the fresh water polyp (*Hydra magnipapillata*), which are both common aquatic animals. The third largest group of contaminating contigs related to plant species (582 contigs). Here, the common grape vine (*Vitis vinifera*) and the castor oil plant (*Ricinus communis*) were among the most frequently hit species.

To summarise, a specifically designed BLAST procedure tagged 3,961 contigs as putative contamination. These contigs most likely represented sequences from species which were introduced as by-products during sample and library preparation, and thus they were excluded from further analysis. The remaining 209,660 contigs (249 Mb) constituted the final *N. furzeri* transcript catalogue available for subsequent annotation.

3.1.3 Annotation of the transcript catalogue

3.1.3.1 BLAST similarity searches

Annotation of the *N. furzeri* transcript catalogue started with a series of BLAST similarity searches against known protein and nucleotide sequences collected from several databases (methods, Table 1 in 2.4.1). Most importantly, these included (i) Ensembl protein and (ii) protein-coding transcript sequences of medaka, stickleback, tetraodon and zebrafish, which are the closest relatives of *N. furzeri* with a sequenced genome. Additional annotation databases contained (iii) human protein sequences from RefSeq and from two large multi-species protein collections, (iv) UniProt (Boeckmann et al. 2003) and the (v) NCBI non-redundant protein database nr (National Center for Biotechnological Information 2011). Finally, (vi) fish-specific ESTs from the NCBI UniGene database (Pontius et al. 2003) were also included, to complement annotation. Depending on the respective database, different BLAST searches with individual parameters and hit thresholds were conducted (Supplementary Table 2). The resulting BLAST hits are summarised in Table 4.

Table 4: BLAST results used for the annotation of the *N. furzeri* transcript contigs.

Database	Contigs hit	Database entries hit	Mean identity	Database entries used	Annotated contigs
Ensembl fish proteins	98,941	38,355	73%	37,739 (81%)	92,857 (85%)
Ensembl fish transcripts	98,099	39,301	77%	1,369 (3%)	3,690 (3%)
Refseq human proteins	80,671	18,071	61%	305 (1%)	497 (1%)
UniProt	98,053	42,361	71%	2,643 (6%)	3,425 (4%)
NCBI nr proteins	96,548	41,500	71%	956 (2%)	1,450 (1%)
NCBI UniGene transcripts	107,901	36,593	77%	3,206 (7%)	7,113 (6%)
Total	122,177			46,218 (100%)	109,032 (100%)

In total, BLAST identified database hits for 122,177 (58%) of the 209,660 contigs in the *N. furzeri* transcript catalogue. These transcript contigs had a considerable higher mean length (1,604 bp), compared to contigs without hits (600 bp). Most hits were found for databases with fish-specific sequences, that is, the NCBI UniGene database, and the Ensembl fish protein and protein-coding transcript databases. The two large protein collections UniProt and NCBI nr did have slightly less hits, which confirmed the comprehensive fish-specific sequence resources of the NCBI UniGene and Ensembl databases. Expectedly, human protein sequences from RefSeq obtained the fewest hits. Generally, hit numbers were comparable between the databases. A set of 75,091 (36%) transcript contigs showed hits to all six databases. In contrast, only 16,256 (8%) transcript contigs had hits to just one single database. Overall, these findings indicated that the obtained BLAST results may serve for a reliable and comprehensive annotation of the *N. furzeri* transcript catalogue.

3.1.3.2 Transcript contig annotation based on BLAST results

Prior to annotation, BLAST hits were filtered. Subsequently, for each *N. furzeri* transcript contig, the best BLAST hit constituted its annotation. That way, 109,032 (52%) of the 209,660 transcript contigs obtained protein-coding gene annotations from 46,218 different database entries of all six BLAST databases (Table 4). Fish-specific protein and protein-coding transcript sequences from Ensembl and NCBI UniGene, which presumably cover almost all known fish genes, contributed over 90% of all annotations. This was corroborated by the small percentage of transcripts which were additionally annotated by the large protein collections UniProt and NCBI nr as well as by the curated human protein sequences from RefSeq. To conclude, the *N. furzeri* transcript catalogue contained contigs annotated by 46,218 protein and protein-coding transcript entries (Figure 10).

The difference between the annotated transcript contigs (109,032) and the database entries used for annotation (46,218) indicated that, in several cases, a single database entry annotated multiple contigs. Indeed, this was observed for about half of all database entries (21,728) which were hit by multiple transcript contigs (in total 87,304), with an average of four transcript contigs per database entry. These database entries presumably represented *N. furzeri* transcripts which, due to assembly problems, were represented by multiple contigs. As shown in the comparison against the set of medaka protein sequences (3.1.2.3.3), the likely cause was the high redundancy of the assembly. However, transcript fragmentation also contributed a decent fraction to the large number of multiple transcript contig per database entry.

The BLAST approach assigns features to contigs by searching for sequence similarities to annotated sequence databases. The obtained hits, which are used for annotation, represent mostly protein or protein-coding transcripts entries, depending on the type of database. However, for the sake of convenience, protein and protein-coding transcript entries identified in the transcript catalogue are from now on both generally referred to as (protein-coding) transcripts.

As already demonstrated, the *N. furzeri* transcript catalogue contained contigs representing 46,218 protein-coding transcripts. Since there are approximately 20,000 protein-coding genes in common fish genomes, and presumably also in *N. furzeri*, it was reasonable to assume that a number of transcript contigs derived from the same gene. Thus, gene symbols were assigned to transcript contigs based on the available gene information of the database entry used for annotation. For example, when using an Ensembl protein sequence for the annotation of a transcript contig, the Ensembl database also provides a gene symbol, which can then be assigned to the transcript contig. Assigned gene symbols grouped the 46,218 transcripts of the *N. furzeri* transcript catalogue into 19,875 protein-coding genes.

Since the *N. furzeri* transcript catalogue covered 46,218 transcripts for 19,875 protein-coding genes, it can be assumed that several genes were represented by transcript isoforms. In total, for about

half of all *N. furzeri* genes (10,392/52%), transcript isoforms were found, with three isoforms per gene on average. Interestingly, this would suggest that almost 80% (46,218-(19,875+10,392)) of the transcripts found in the *N. furzeri* transcript catalogue are actually isoforms.

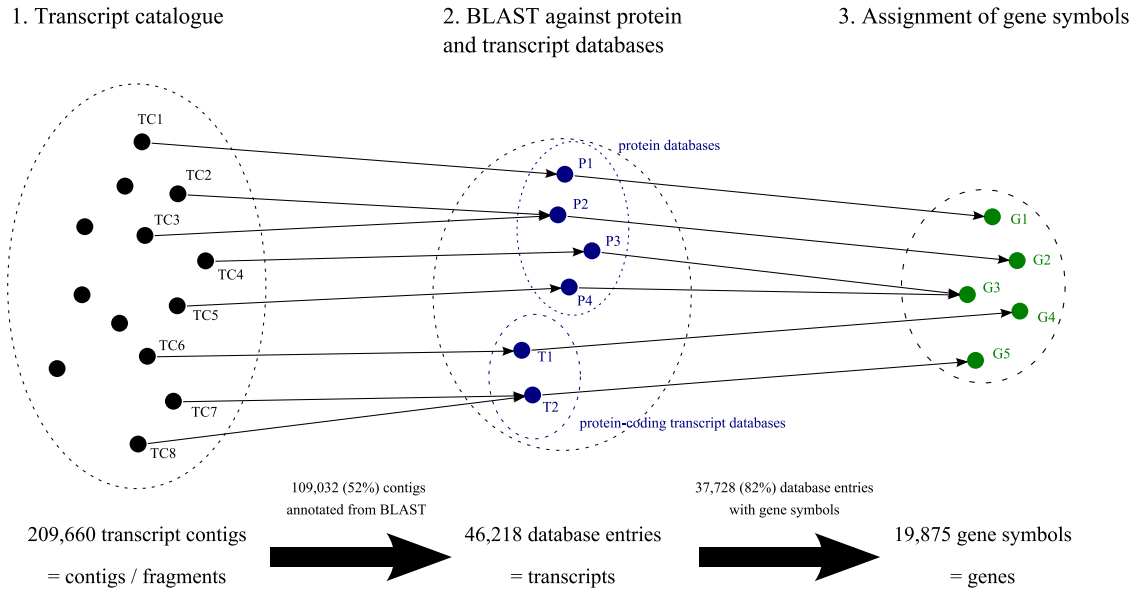

Figure 10: Schematic representation of the annotation process.
The annotation process of *N. furzeri* transcript contigs (TC) relied on BLAST hits to databases with protein (P) or protein-coding transcript sequences (T), followed by the assignment of gene symbols (G). In this example, TC1-TC5 and TC6-TC8 show BLAST hits to the database entries P1-P4 and T1-T2, respectively. Note that TC2 and TC3 share the same protein sequence (P2), and, therefore, they are derived from the same *N. furzeri* transcript. This also applies for TC7 and TC8. Subsequently, gene symbols are assigned to transcript contigs based on available gene information. Here, TC4 and TC5 share the same gene symbol (G3), and, therefore, they represent alternative transcripts of the same *N. furzeri* gene. Eventually, of the 209,660 contigs in the transcript catalogue, 109,032 contain sequence information for 46,218 *N. furzeri* transcripts coding for 19,875 protein-coding *N. furzeri* genes.

3.1.3.3 Transcript contig annotation based on protein domain prediction

So far, annotation relied completely on sequence similarity searches with BLAST against databases of protein and protein-coding transcript sequences known from other, related species. Of course, this strategy requires that these databases contain as many sequences as possible, and that they are also present in *N. furzeri*, having a sufficiently high conservation. However, transcripts that are very divergent or even exclusive to *N. furzeri* will be presumably missed by the BLAST annotation. BLAST based annotation identified only 109,032 (52%) of the 209,660 assembled transcript contigs. Thus, prediction of conserved protein domains was done to extend the annotation. Therefore, it was necessary to translate the transcript contigs into putative protein sequences.

The identification of the particular protein sequence encoded by the transcript contig is not straightforward since the contig contains, besides the CDS, also UTR at its ends. Thus, the encoded protein does not necessarily start at the beginning of the transcript contig. Moreover, transcript contigs can contain assembly errors, for example small insertion or deletions, which may disrupt the reading frame or result in a premature stop codon. Therefore, the correct CDS has to be identified beforehand,

which is commonly done by searching for the open reading frame (ORF) which contains the encoded protein.

Transcript contigs were translated with `prot4EST` (Wasmuth & Blaxter 2004), which integrated BLAST searches against medaka proteins and predictions with `ESTScan` (Iseli et al. 1999) to find the correct ORF and CDS. For 94,814 (45%) of the 209,660 analysed transcript contigs, a BLAST hit against a medaka protein identified the likely CDS. Subsequently, `ESTScan` predicted probable CDS for 69,118 (33%) of the remaining contigs without BLAST hits. Last, the pipeline simply searched for the longest transcript sequence uninterrupted by a stop codon, which suggested potential CDS for additional 45,617 (22%) transcript contigs. Translation into amino acids yielded in total 209,549 putative protein sequences, with an average length of 210 aa. Over 65% of these had a minimum length of at 100 aa. Interestingly, of the 109,032 transcript contigs without annotation, the large majority (98,252; 90%) had predictions for putative protein sequences. However, there was a considerable difference in the length of the predicted protein sequence between annotated and unannotated transcript contigs (307 vs. 119 aa).

The putative protein sequences served as input for the identification of conserved protein domain motifs with the `HMMER` package (Eddy 2012). In total, 37,380 (18%) protein sequences showed significant (e-value $\leq 10^{-20}$) matches to 3,869 protein domains maintained in the Pfam database (Punta et al. 2011), which is a large collection of protein families with shared domains. Common domain motifs included protein kinase (*Pkinase*, 1,529 hits), tyrosine kinase (*Pkinase*_Tyr, 1,378 hits) and rhodopsin-like receptors (*7tm*_1, 591 hits). Protein domain predictions were then used to extend the annotation of the transcript catalogue. However, only five additional transcript contigs were annotated by this approach. Based on the corresponding protein domains, they were identified as an alanine racemase (*Ala_racemase_N*), an alcohol dehydrogenase transcription factor (*MADF_DNA_bdg*), an ubiquitin-activating enzyme (*UBA_e1_C*), a gap junction protein (*Neuromodulin_N*) and a transmembrane ion channel protein (*Ion_trans*).

3.1.3.4 Completeness of the transcript catalogue

Naturally, assembly of a transcript catalogue aims at identifying and reconstructing transcripts from as many genes as possible. Ideally, this results in transcript sequences for every gene that is expressed at a specific level under defined conditions. Furthermore, the transcripts should be assembled as complete as possible, that is, for example they encode not only the complete protein but also contain both UTR. Depending on the success in these two criteria, it might be necessary to produce more sequencing data, preferably from different individuals or tissues, to identify additional and extend existing transcripts. For *N. furzeri*, the identification of transcripts and genes during annotation allowed for a thorough estimation of the completeness of the transcript catalogue.

Annotation identified 19,875 protein-coding genes represented in the *N. furzeri* transcript catalogue. Gene numbers in related fish species with a sequenced genome lie between 19,602 for tetraodon, the fish species with the smallest genome (Jaillon et al. 2004), and 26,206 genes for zebrafish, the fish species with the largest genome (Howe et al. 2013). Moreover, as will be extensively described in 3.2, comparisons against protein sequences of medaka, stickleback, tetraodon and zebrafish found homologous *N. furzeri* transcript contigs for the majority of sequences of the respective species. This suggested that the *N. furzeri* transcript catalogue contains the majority of genes commonly found in fish species.

However, because the comparison of gene numbers relied on the sequences from the four model fish species, it could be biased towards the quality of the respective genome annotation. For example, several genes are only inferred from gene model prediction without any real sequence confirmation. Therefore, I additionally evaluated the number of gene families present in the transcript catalogue. Gene families are groups of genes which are evolved from common evolutionary ancestor by duplication and speciation (see also 3.2.2). The number of gene families per species does not change much between species and therefore it is better suited to estimate completeness. Gene family information was obtained from the Ensembl Compara database (Vilella et al. 2009), which describes such families and their evolution as phylogenetic gene trees. Here, zebrafish had the highest number of gene families (8,241), followed by stickleback (7,663), medaka (7,501) and tetraodon (7,261). Based on the `BLASTx` results against these species (see 3.2.1), I identified 7,489 gene families hit by at least one *N. furzeri* transcript contig. Moreover, for 85% (6,428 families), at least one of the four model fish species shared the same number of genes per family. These results further emphasised a completeness of the provided transcript catalogue comparable to that of model fish species for which an annotated genome is available.

Besides gene representation, the completeness of the *N. furzeri* transcript catalogue was also evaluated by analysing the amount of CDS that is translated into the respective protein. However, in the case of the transcript catalogue, genes were represented by multiple redundant and fragmented transcript contigs, which shared parts of the CDS. This redundancy made estimating the CDS for a gene rather difficult. Therefore, only the longest transcript contig per gene was considered for the calculation of the CDS. Of the 19,875 identified *N. furzeri* genes, 14,164 (71%) were represented by a transcript contig with an almost complete CDS (Figure 11), that is, the contig covered >90% of the database entry used for its annotation. Moreover, when setting a minimum CDS coverage of 50%, the gene number increased to 16,961 (85%). The findings suggested that the transcript catalogue contained complete CDSs for the large majority of *N. furzeri* genes.

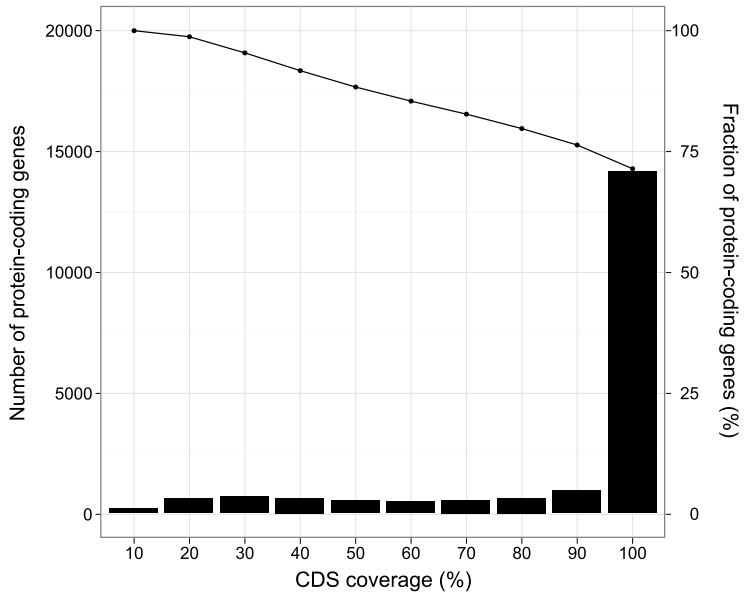

Figure 11: Fractions of putative CDS represented in the longest transcript contig per *N. furzeri* gene.
For each gene, the CDS of the longest transcript contig was determined, and predicted CDS fractions were binned into deciles. The histogram bars show the number of respective transcript contigs (genes) per decile. The line shows the cumulative number. More than 70% of the protein-coding genes are represented by a transcript contig with a complete (> 90%) CDS.

CDS lengths generally differ largely between genes. Consequently, the previous CDS estimation may be biased towards genes with smaller CDS because they are more likely to be complete. In other words, the 71% calculated above may be simply reflecting a high fraction of small *N. furzeri* genes. Therefore, I conducted a second analysis based on the total CDS length summed over all genes. As above, only the longest transcript contig was analysed for each gene. Altogether, the overall CDS of these contigs covered 29 Mb (83%) of the 35 Mb estimated from the database entries used for annotation. Furthermore, compared to the four model fish species, the *N. furzeri* transcript catalogue represents 74% of zebrafish and 99% of medaka annotated CDS. In summary, besides identifying the large majority of protein-coding genes and transcripts, the *N. furzeri* transcript catalogue also reconstructed most of the CDS in the transcripts.

3.1.3.5 Functional annotation with Gene Ontology

While transcriptome assembly and annotation identified the individual transcript contigs with associated gene symbols, GO terms gave a more general overview of the different biological functions encoded in the *N. furzeri* transcript catalogue. Since protein product descriptions may differ between databases and thereby inhibit effective grouping of similar biological functions, GO attempts to develop controlled vocabularies, also called ontologies, to describe gene products consistently across species. Thereby, GO characterises a gene by assigning GO terms in three different domains: the *biological process* (BP), the *molecular function* (MF) and the *cellular component* (CC). The resulting GO classification of the *N. furzeri* transcript catalogue allowed analysing complete groups of genes with shared biological functions instead of single selected genes.

GO terms were copied from the database entries used for transcript contig annotation, that is, from Ensembl and UniProt entries. The current ontology (OBO v1.2, 27/07/10) contained 32,901 different GO terms in all three domains. Of 109,032 annotated transcript contigs, 76,013 (72%) were

linked to 340,452 GO counts, that is, counting all occurrences of GO terms. MF represented the largest class of assignments (166,476 GO counts/70,305 contigs), followed by BP (108,336/49,972) and CC (65,640/40,551). After removing redundant counts, 6,796 unique GO terms were identified, which constituted the set of biological functions in the transcript catalogue.

The identified GO terms provided descriptions of biological functions down to the smallest detail. To obtain a broader overview, observed GO terms were mapped to 131 GO-Slim terms which are a subset of GO terms and provide high-level functional annotation without the detailed specific fine grained terms. Furthermore, to avoid redundant counts due to multiple transcript contigs per genes, only the longest contig per gene was considered. In total, 137,683 GO-Slim hits were assigned to 15,040 (76%) transcript contigs representing 124 unique GO-Slim terms. Most frequent terms among the three GO domains were *metabolic process* (6,108 contigs, BP), *binding* (10,567 contigs, MF) and *cell* (7,668 contigs, CC). Figure 12 shows the distribution of the 56 GO-Slim terms present at the second level of the GO-Slim annotation tree; these provide the most general descriptions next to the three first-level GO-Slim terms *biological process*, *molecular function* and *cellular component*.

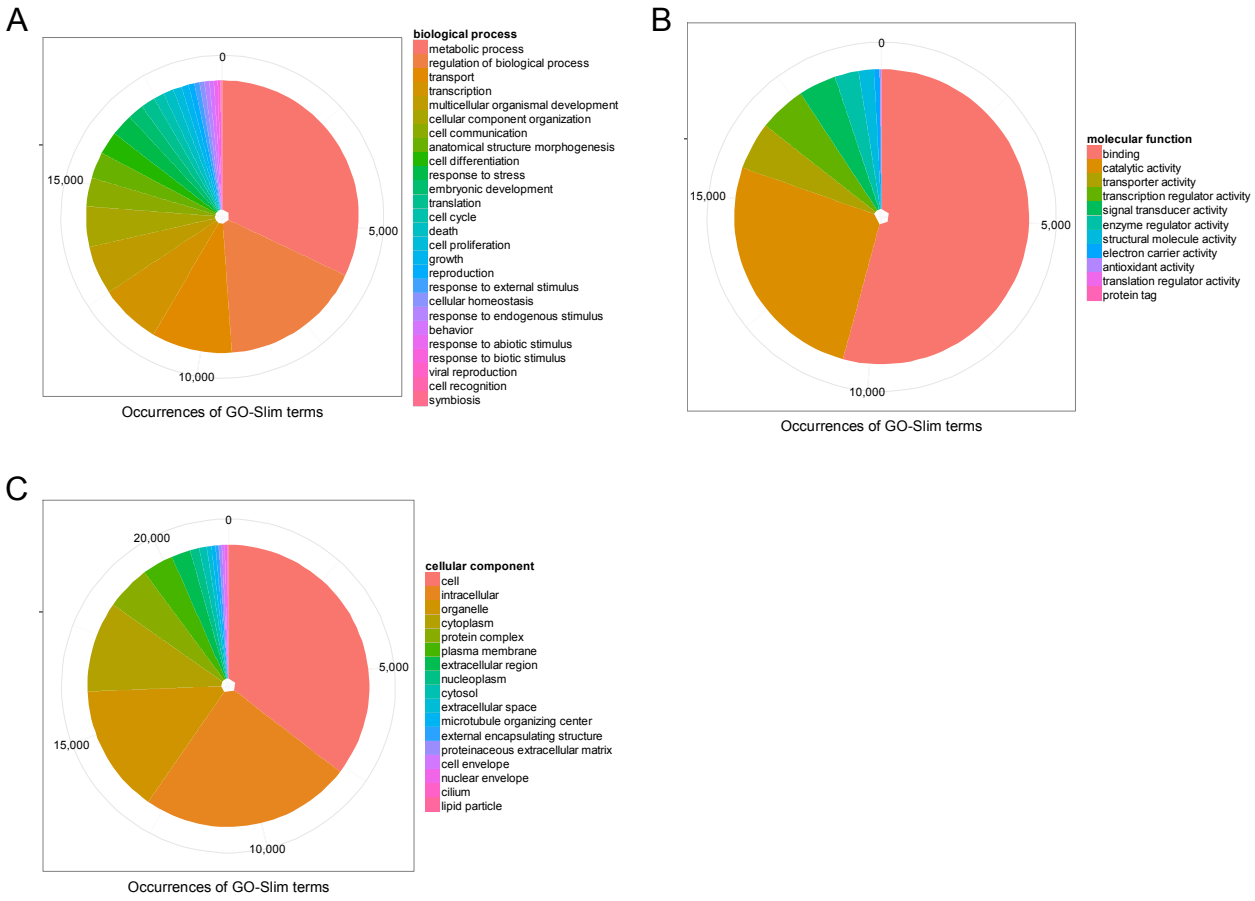

Figure 12: Functional annotation of the longest *N. furzeri* transcript contigs per gene based on second-level GO Slim terms.
A) Biological process, **B)** Molecular function and **C)** Cellular component. In total, for 15,040 (76%) *N. furzeri* genes, 60,201 second-level GO-Slim annotations were counted. These cover 54 of the 56 unique second-level GO Slim terms. The two missing second-level GO Slim terms are *thylakoid* and *nutrient reservoir activity*.

Taken together, GO-Slim terms provided a condensed overview of the Gene Ontology annotations assigned to transcript contigs. Moreover, almost all GO-Slim terms were covered by at least one *N. furzeri* transcript contig thereby showing that the transcript catalogue contained genes for almost all currently annotated biological functions.

3.1.3.6 Analysis of the remaining transcript contigs without annotation

Annotation could be assigned to only 52% of all assembled *N. furzeri* transcript contigs but another 100,628 (63 Mb) contigs remained unannotated; of these, 12,890 (13%) were larger than 1 kb. Although library preparation and sequencing can produce sequence artefacts, which might be assembled as contigs, they cannot explain the large number of transcript contigs without annotation. Several possible explanations might provide deeper insights into the nature of these transcript contigs.

First, a number of transcript contigs probably consist mostly of UTR. However, without a genome, this is difficult to validate. UTR are less conserved between species and therefore sequence comparison against distantly related species often fails. Also, there are no special UTR characteristics that can be searched for. Nevertheless, some indirect evidence could be found that supported this assumption. In a large number of unannotated transcript contigs (70,357; 70%), predicted CDS fragments were located near the ends of the contig, and the remaining sequence presumably represented either 5' or 3' UTR. Generally, the average CDS length of the unannotated transcript contigs was also much smaller, compared with the average CDS length of the annotated transcript contigs (356 vs. 883 bp). As a result, the contigs did not contain enough CDS to find BLAST hits in other sequence databases and remained unannotated. Additionally, 1,275 (1%) contigs had poly(A)-stretches at one end, which are commonly added to the poly(A)-site of the 3'UTR. Moreover, for 448 of these contigs, a polyadenylation signal (*AAUAAA* or one of ten related variants, Ulitsky et al. 2012) was found within a 30 bp region upstream, suggesting that that these contigs might have derived from UTR.

Second, several transcript contigs might have derived from dispersed repeats which can contain cryptic ORFs that are still being transcribed. A `RepeatMasker` (Smit et al. 1996) search with a library of *N. furzeri*-specific repeats identified in genomic sequence (Reichwald et al. 2009) found repetitive elements in 53,107 (53%) unannotated transcript contigs. Many hits were only partial, but in 21,341 (21%) contigs, repetitive elements made up at least half of the entire sequence. In 5,895 (6%) transcript contigs, repetitive elements covered the complete sequence (over 90%). Common repeat elements in transcript contigs included DNA transposons (6,542 contigs), short and long interspersed nuclear elements (3,302 and 4,791) and long terminal repeat retrotransposons (1,028). However, for the majority of transcript contigs (47,419), the origin of the repetitive sequence could not be determined.

Third, some transcript contigs might have derived from other RNA species such as rRNA, tRNA or other non-protein-coding RNAs. Although library preparation involved a poly(A)-selection step, which enriches for mRNA, this step may not have completely removed other RNA species. However, `BLASTn` similarity searches against databases with representative rRNA and tRNA sequences (Quast et al. 2013; Chan & Lowe 2009) showed significant hits (e-value $\leq 10^{-10}$) for only 7 and 24 transcript contigs, respectively. Additionally, some transcript contigs might represent precursor transcripts of miRNA. Similarly, a `BLASTn` search (e-value $\leq 10^{-10}$) against the database miRBase (Kozomara & Griffiths-Jones 2011) identified 984 transcript contigs as putative precursor sequences of annotated miRNA.

Ultimately, definite identification of the unannotated transcript contigs requires a genome sequence. In 2012, a high-quality draft assembly of the *N. furzeri* genome was built by the Genome Analysis group at the FLI, which covers approximately 60% of the total sequence. Using `BLASTn` (e-value $\leq 10^{-20}$), 95% of the 100,628 unannotated transcript contigs were successfully aligned to a genomic location. The very high average sequence identity of 98% indicates that these hits are true positives. Furthermore, over 54% of the contigs with hits were fully aligned, that is, the alignment covered at least 90% of the contig length.

In summary, the large majority of the unannotated transcript contigs truly originate from transcribed regions of the *N. furzeri* genome. However, identifying the exact nature of these contigs is currently difficult and mostly speculative. Once the *N. furzeri* genome is fully assembled and comprehensively annotated, a more detailed analysis of these unannotated transcript contigs will be possible.

3.1.4 Representation of the transcriptome data in a browser

The *N. furzeri* transcript catalogue contains over 200,000 transcript contigs, for which, annotation produced large amounts of analysis data. Although the results are organised in a database, accessing them is almost impossible without programming skills. The `EST2uni` pipeline used for the annotation also provides a web interface which allows accessing the results within a normal internet browser. For the *N. furzeri* transcript catalogue, I installed and considerably improved the web interface, as described in the corresponding methods part. The resulting website, namely the *Nothobranchius furzeri Information Network transcriptome browser* (NFINtb, https://gen100.imb-jena.de/EST2UNI/nfintb/), is publicly available and provides fast and easy access to the transcript catalogue (Figure 13).

Figure 13: The *Nothobranchius furzeri* Information Network transcriptome browser.

3.2 Comparison to gene and transcript data of other fish species

3.2.1 BLASTx against the protein sequences of other fish species

Initially, I compared the *N. furzeri* transcript catalogue against the protein sequences annotated for the four fish species using `BLASTx` (Table 5). Of the 209,660 transcript contigs, 46% had significant BLAST hits (e-value $\leq 10^{-07}$) to at least one fish protein sequence. Interestingly, this number coincides well with the number of annotated transcript contigs in the transcript catalogue. Analysed by species, 42% showed hits to zebrafish and medaka, 40% to stickleback and 38% to tetraodon. Regarding protein sequence conservation, average amino acid identities of BLAST hits ranged between 68% for zebrafish and 75% for stickleback, confirming values measured in previous species comparisons done for *N. furzeri* (Reichwald et al. 2009).

Conversely, the fish protein sequences were generally well represented, that is, most were hit by at least one *N. furzeri* transcript contig. Only the results against zebrafish deviated slightly from this pattern since only half of the protein sequences were hit. Similarly, when counting the Ensembl gene

entries that are associated with the protein sequences, over 80% of all medaka, stickleback and tetraodon genes were also found in *N. furzeri*. Again, for zebrafish, the number was slightly lower (71%). Nevertheless, these numbers suggested that *N. furzeri* shares the large majority of proteins and genes with the other four fish species. Moreover, they also supported the estimate of approximately 20,000 genes present in *N. furzeri* transcript catalogue.

A set of 73,238 (35%) transcript contigs showed hits to all four fish species (Figure 14), corresponding to approximately 16,000 genes. Most presumably, these genes are generally conserved between all four fish species and *N. furzeri*. Accordingly, average amino acid identities measured for this subset higher compared to overall values (medaka: 76% vs. 74%, stickleback: 77% vs. 75%, tetraodon: 75% vs. 74%, zebrafish: 70% vs. 67%). On the other hand, 8,175 transcript contigs (4%), corresponding to 2,902 genes, had hits exclusively in one species (Figure 14, Table 5). Average amino acid identities measured for these hits were lower compared to overall values (medaka: 56% vs. 74%, stickleback: 57% vs. 75%, tetraodon: 62% vs. 74%, zebrafish: 50% vs. 67%). These values suggested that the corresponding genes likely represent potential misannotations in this species.

Table 5: `BLASTx` comparison of *N. furzeri* transcript contigs to the protein/gene annotations of four other fish genomes.

	Medaka	Stickleback	Tetraodon	Zebrafish
Number of protein/gene entries annotated	24,661 / 19,686	27,576 / 20,787	23,118 / 19,602	41,478 / 26,095
Contigs with hits	87,378	84,761	79,594	87,219
Protein/gene entries hit	19,272 / 17,173 78% / 87%	20,534 / 17,795 74% / 86%	18,613 / 16,744 81% / 85%	22,669 / 18,620 55% / 71%
Average amino acid identity	73.5%	74.6%	73.7%	67.5%
Contigs with hits exclusively in this species	3,597	1,256	531	2,791
Protein/gene entries hit	919 / 914	702 / 690	358 / 356	984 / 942
Average amino acid identity	55.7%	57.0%	62.0%	49.5%

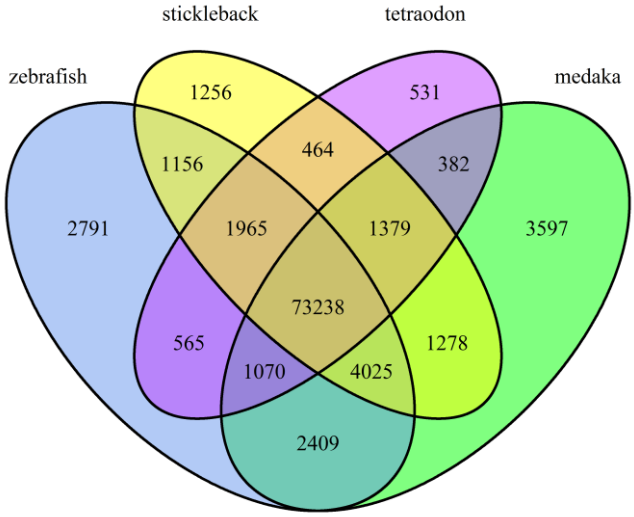

Figure 14: `BLASTx` comparison of *N. furzeri* transcript contigs to the protein/gene annotations of four other fish genomes.
Venn diagram showing 96,106 (46%) *N. furzeri* transcript contigs with BLASTx hits to Ensembl protein annotations of the four fish genomes. A set of 73,238 transcript contigs (35%) had hits in all four fish species, whereas 8,175 transcript contigs (4%) had hits exclusively in one species.

3.2.2 Duplicated *N. furzeri* genes

Gene duplication is considered as a major mechanism of evolution, and genes that originate from a duplication event are called paralogs. Genome duplication is a major source of gene duplication, and vertebrate genomes are believed to have undergone at least two rounds of whole genome duplication (Ohno 1970). Studies in teleost fish species showed that their genomes underwent an additional round of genome duplication (Amores et al. 1998; Taylor et al. 2003). Most of the original fish paralogs are lost during evolution. However, several genes which are found only once in vertebrates occur twice in fish species, that is, they are paralogs. For *N. furzeri*, the transcript catalogue provided the opportunity to assess the fraction of paralogous genes in more detail. This was done in two steps: Initially, I concentrated on genes which are also duplicated in other fish species, and then searched for genes which are duplicated exclusively in *N. furzeri*.

First, identification of duplicated *N. furzeri* genes relied on sequence similarity searches against known paralogous genes of the four other model fish species medaka, stickleback, tetraodon and zebrafish. All relevant information was retrieved from the Ensembl Compara Database, which contains phylogenetic trees of genes across species (Figure 15A). To avoid redundant counts due to multiple transcript contigs per gene, only the longest transcript contig was selected and then compared against the Ensembl protein sequences of the four fish species. Of the 19,875 transcript contigs/genes, 17,261 showed significant `BLASTx` hits (e-value $\leq 10^{-07}$) to a fish protein sequence. Subsequently, the Ensembl gene ID of the best hit was used to query the corresponding Ensembl Compara gene tree for paralogs of this gene. Thereby, analysis was limited to those genes which were duplicated as result of the teleost-specific genome duplication, that is, paralogs of the earlier genome duplications in vertebrates were not analysed. If any of the paralogs were hit by another *N. furzeri* gene, then the two

N. furzeri genes were considered as paralogs. Using this approach, I analysed 17,261 *N. furzeri* genes with `BLASTx` hits and identified 4,498 paralogous genes grouped in 2,148 paralog families (on average two paralogs per family, Supplementary File 2). Consequently, these families can be seen as representatives of the ancestral vertebrate genes, which remained duplicated since the teleost genome duplication.

For further validation of the results, I selected the transcript contigs representing the putative *N. furzeri* paralogs annotated as *SOX4A* and *SOX4B* and constructed a phylogenetic tree based on protein sequence data from seven fish species; human, mouse and the Western clawed frog protein sequences were used as outgroup (Figure 15B). Indeed, the phylogeny resembles that of the corresponding Ensembl Compara gene tree of *SOX4* (Figure 15A), and the fish-specific duplication node with its paralog sub trees is clearly visible. Thus, *SOX4* is most likely also duplicated in *N. furzeri*. Moreover, this demonstrates the usability of the first approach to identify paralogous genes in the *N. furzeri* transcriptome.

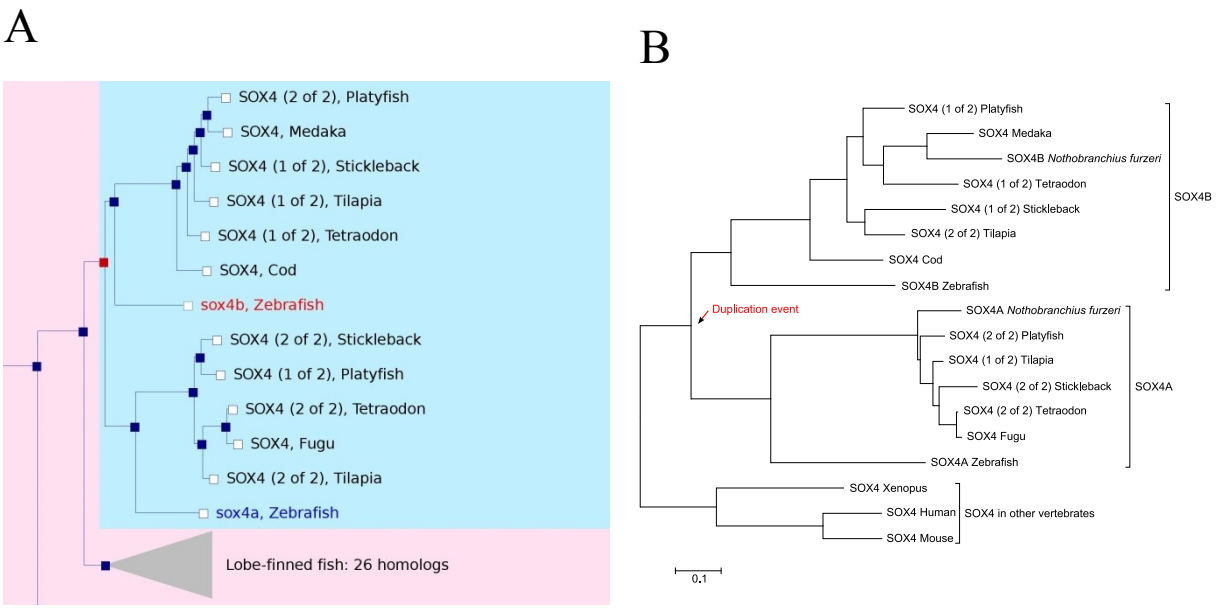

Figure 15: Analysis of putative *N. furzeri* paralogs using the Ensembl Compara gene trees.
(A) shows the Ensembl Compara gene tree for the gene *SOX4* in the currently sequenced fish species (light-blue shaded area). The tree contains a duplication node (in red) with the two sub trees for each copy of *SOX4*. **(B)** The *SOX4* gene tree was independently reconstructed including the two paralogs found in *N. furzeri*. Again, the gene duplication node is clearly visible and each of the two sub trees contains one *N. furzeri SOX4* paralog. Note that the notation for gene symbols of paralogs (for example *1 of 2*) is officially used by Ensembl.

Because the first approach relied on data about genes known to be duplicated in other fish species, it cannot find genes which are exclusively duplicated in *N. furzeri*. Moreover, sequence similarity searches like BLAST cannot discern between paralogs and will assign them a single gene symbol. Thus, some *N. furzeri* transcript contigs which, according to annotation, share the same gene symbol might represent in fact paralogous genes. Consequently, in the second approach, I searched those remaining non-duplicated genes which are represented by multiple transcript contigs for

indications of *N. furzeri*-specific paralogs. Therefore, I developed a clustering approach based on the protein sequence identities of the transcript contigs. For each gene, all transcript contigs were translated into putative protein sequences (see 3.1.3.3) and pairwise aligned. The protein sequence identities of these alignments were then clustered based on a minimum distance of 10% identity between each cluster. Different clusters then hinted at potential paralogs. For 1,249 genes, transcript contigs formed more than one cluster, corresponding to a total of 2,619 (on average two clusters per gene, Supplementary File 2). These clusters might represent paralogous genes which collapsed during BLAST annotation. They are exclusive to *N. furzeri*, at least when considering the four other fish species.

Taken together, the results indicated that the current annotated *N. furzeri* transcript catalogue consists of 17,525 (19,875-4,498+2,148; see first approach) protein-coding genes when the paralogs are counted only once. This number most presumably reflects the set of protein-coding genes commonly found in all vertebrates. Consequently, of those genes, 19% (2,148+1,249, see first and second approach) showed evidences of multiplication, that is, they have fish-specific paralogs in *N. furzeri*. In other words, of the complete set of genes that resulted from the teleost genome duplication, about 19% can be still found in the *N. furzeri* transcript catalogue, whereas over 80% are either not transcribed anymore or even lost during evolution. Indeed, approximately, 90% (4,134+2,274) of the 7,117 (4,498+2,619) paralogs occurred as pairs. The remaining 709 (364+345) paralogs were grouped in 193 paralog families with 3 and more paralogs (average: 4 paralogs per family, maximum: 23). The higher paralog numbers in these families cannot be attributed to the teleost whole genome duplication but are presumably caused by subsequent segmental duplications or by assembly/annotation errors.

3.3 Transcriptome changes in ageing *N. furzeri*

3.3.1 Mapping of RNA-seq datasets

RNA-seq experiments analyse transcript levels of genes by mapping transcriptome reads against a reference sequence, either an annotated genome or a reference set of transcript sequences. In the first case, the transcript level of a gene is then inferred from the number of reads mapping to the gene's locus. Since for *N. furzeri* an annotated genome sequence was not available, the transcript catalogue served as reference for mapping. Consequently, the transcript levels were directly inferred from the reads mapping to the assembled transcript contigs of the individual genes. However, due to assembly fragmentation and redundancy, the transcript catalogue often contained multiple transcript contigs per gene. In such case, reads map equally well to multiple reference sequences, which complicates subsequent quantification. To avoid such problems, only the longest transcript contig per gene was considered as the representative reference sequence. Consequently, the transcriptome reference built for RNA-seq mapping comprised 19,875 transcript contigs with a total length of 53 Mb.

Subsequently, RNA-seq datasets (#6–13, Table 2) were used to obtain transcriptome-wide insights into transcript levels changes during ageing in *N. furzeri*. Therefore, analysis included two tissues, two time points as well as two strains - the short-lived strain GRZ vs. the longer-lived strain MZM-0403. More precisely, transcript levels were measured in skin and brain of young and old GRZ (5 and 14 weeks, respectively) and MZM-0403 (5 and 31 weeks, respectively). In contrast to assembly, the respective datasets were not quality-filtered because read mapping is less sensitive to sequencing errors. Moreover, the duplicate read removal done during quality processing was also not done since it would distort the transcript levels. Consequently, the raw datasets contained ~20 Gb of transcriptome data, or roughly 30 million reads per library (Table 6).

Reads were mapped with `BWA` against the *N. furzeri* transcriptome reference. Between 41% and 65% reads per library were initially mapped. However, a number of read mapped ambiguously to multiple transcript contigs and therefore had to be excluded. Thus, mapping rates dropped, that is, between 36% and 53% uniquely mapped reads per library were available for quantification (Table 6).

One possible explanation for this relatively low mapping success provided the design of the reference sequence, which contained only the longest transcript contig per *N. furzeri* gene. To validate whether inclusion of further transcript contigs would improve read mappability, two other references were compiled, one containing all transcript contigs with gene annotations (85,431 contigs, 155 Mb) and one containing the complete *N. furzeri* transcript catalogue (209,660 contigs, 249 Mb). Fractions of mapped reads increased up to 86% and 91%, respectively. Hence, a considerable fraction of the reads mapped to the additional, secondary transcript contigs which were not included in the first reference sequence. However, the corresponding fractions of reads with unique mappings dropped to 21% and 19%, respectively. To avoid the aforementioned problems with the ambiguous mappings, I decided to use the transcript levels derived from the first reference sequence. Finally, to allow comparing different samples from different sequencing runs, count values were normalised to RPKM (Mortazavi et al. 2008).

Table 6: Summary of the RNA-seq analysis.

Strain	GRZ				MZM-0403			
Library	6	7	8	9	10	11	12	13
Age (weeks)	5		14		5		31	
Organ	Skin	Brain	Skin	Brain	Skin	Brain	Skin	Brain
Reads	33,947,044	36,117,265	35,963,921	27,369,310	31,727,935	34,720,289	30,268,649	31,423,327
Mapped / Unique (%)	58 / 46	42 / 37	53 / 43	41 / 36	65 / 52	54 / 48	64 / 53	52 / 46
Transcribed genes	17,350	17,638	17,456	17,535	17,170	17,590	17,375	17,644
Median transcript level (RPKM)	2.3	4.7	2.6	5.7	2.5	5.4	3.5	5.8
Range of transcript levels (RPKM)	0.005 – 13,043	0.006 – 40,801	0.005 – 39,533	0.01 – 5,587	0.006 – 10,409	0.003 – 5,288	0.005 – 3,878	0.009 – 3,946

Nevertheless, although at most 50% of the reads were used for quantification, transcripts could be detected at a wide range of transcription intensity. In almost all samples, extreme transcript levels below a RPKM of 0.01 or above 1,000 RPKM (in some cases even over 10,000 RPKM) were detected. Interestingly, the number of transcripts (genes) with non-zero read counts was always roughly around 17,000, irrespective of tissue. When applying a minimum threshold of ten reads, the number of profiled transcripts dropped to approximately 15,000-16,500 per analysed sample. Notably, filtered transcript profiles still contained transcripts with very low expression levels; per analysed sample, between 1,500 and 3,000 profiled transcripts had RPKMs smaller than 1, respectively. These numbers suggested that, although effectively only half of the reads were used for quantification, they were still sufficient to determine transcript levels across a wide dynamic range.

3.3.2 Correlation and principal component analyses

Statistical analysis first concentrated on the comparison of the different samples to find similar transcript level signatures. Therefore, I first calculated the pairwise correlations between the samples based on the RPKM values of the genes. According to Spearman correlation (Figure 16), samples from the same tissue showed the highest correlation, ranging from 0.93 to 0.97. In contrast, samples from the same strain or age correlated much less, and the calculated values differed widely (0.62 to 0.97). Interestingly, when considering only the tissue skin, sample of the same age clustered better than those of the same strain, whereas in brain, samples of the same strain seemed to cluster better than age.

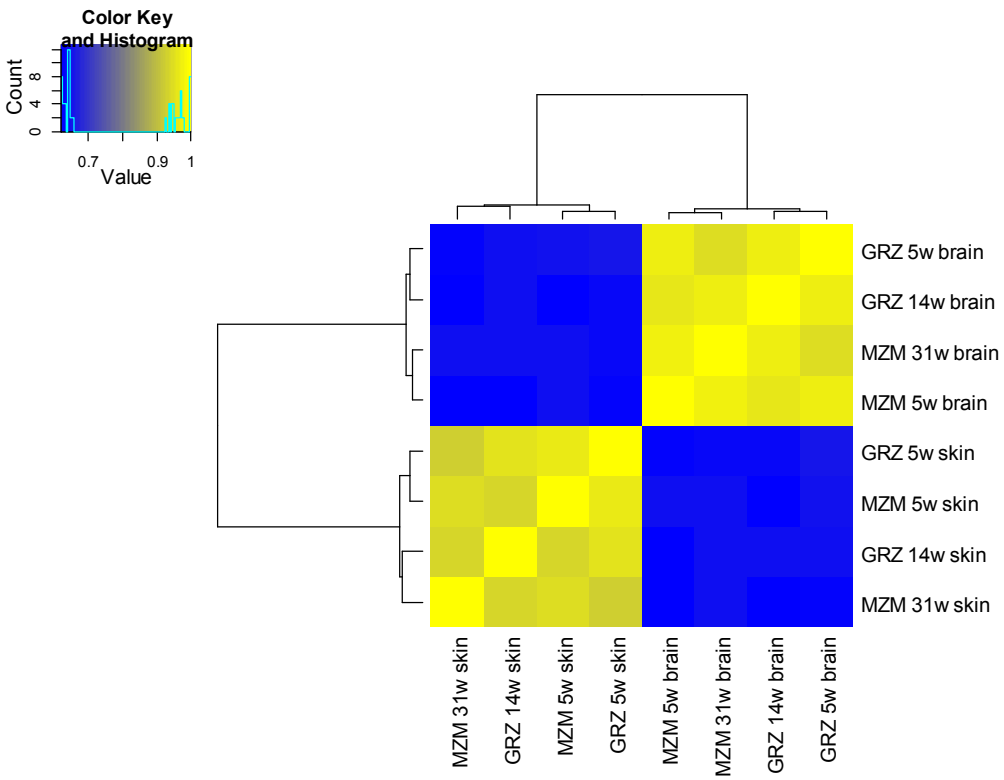

Figure 16: Cluster heat map of ***N. furzeri*** **RNA-Seq data**.
Correlation analysis of datasets, which were derived from skin and brain of young and old specimens of the short-lived strain GRZ and the longer-lived strain MZM-0403 (MZM). Young age represents 5 weeks (w) in both strains, whereas old age is reached at 14 w in GRZ and 31 w in MZM-0403. Analysis is based on Spearman correlation coefficients of log-transformed RPKM transcript levels.

To analyse the underlying effects more closely, I conducted a principal component analysis (PCA). This type of analysis aims to reduce a highly complex dataset to a system of principal components (PC), which ideally explain the most variance found in the data. For *N. furzeri*, PCA on the RPKM values of the eight samples identified eight principal components. Only the top three components indicated a strong influence, that is, they explained over 96% of the total variation in all samples (PC1: 56%; PC2 22%; PC3: 18%). To get a better impression of the individual PC's nature, they were visualised in so-called biplots, in which two of the components are plotted against each other. In the first biplot (PC1 vs. PC2; Figure 17A), samples were arranged along PC1 by their age, and a similar pattern could be also observed for the second biplot (PC1 vs. PC3; Figure 17B). Concerning the other two components, samples were arranged along PC2 and PC3 according to strain and tissue. Taken together, the PCA provided more detailed insights into the principal structure of the *N. furzeri* transcriptome data.

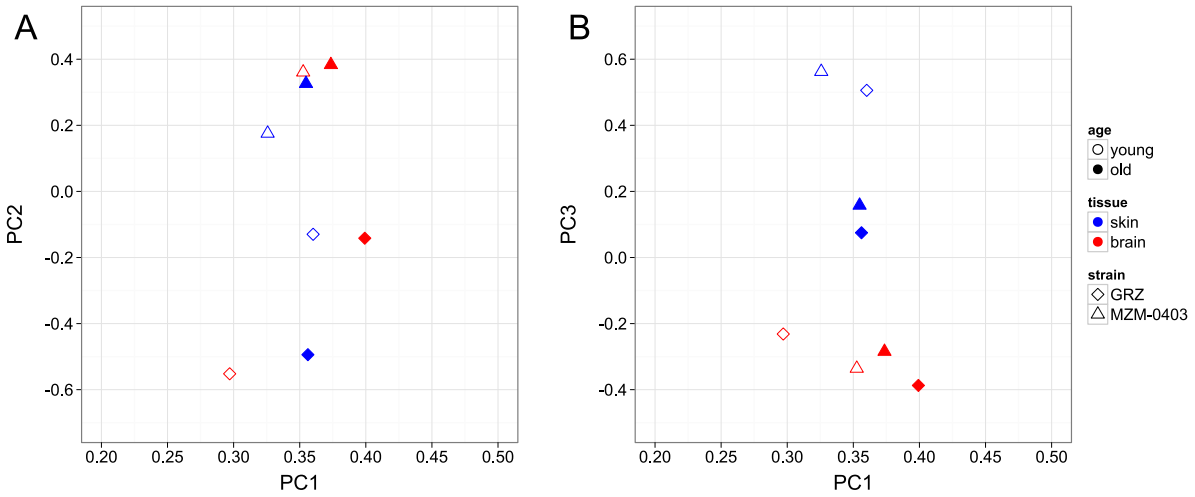

Figure 17: Principal component analysis of the ***N. furzeri*** RNA-Seq data.
Biplots of the **(A)** first vs. the second and **(B)** first vs. third component.

3.3.3 Identification of differentially expressed genes

The main goal was to identify differentially expressed genes (DEGs) during ageing of *N. furzeri*. Therefore, I compared compare the transcript levels between young and old individuals and searched for genes with significant differences. Statistical tests for such tasks usually rely on the availability of replicates to discern real (biological) differences from statistical noise and technical errors. Unfortunately, the available RNA-seq datasets lacked biological replicates because the primary goal was the completion of the transcript catalogue. Instead, multiple individuals were pooled together, which might have been otherwise used as replicates, to obtain enough usable RNA material. To overcome this limitation, I decided to ignore strain-specific differences and to use GRZ and MZM-0403 samples of the same tissue and age as replicates. Consequently, I compared young and old *N. furzeri* in two tissues, with two samples per data point.

Subsequently, DEGs were called with the `R` package `DESeq` (Anders & Huber 2010). The tool identified 43 genes which showed differential expression with age in each of the two tissues, respectively ($p \leq 0.01$; skin: Supplementary Table 3; brain: Supplementary Table 4; additional information: Supplementary File 3). In skin, the majority of genes were downregulated with age (11 upregulated vs. 32 downregulated). Interestingly, in brain, the opposite was observed, that is, more genes were up- than downregulated (24 vs. 19). Furthermore, changes in transcript levels were slightly more pronounced in skin, compared to brain, indicated by the stronger fold changes. Finally, the intersection of both DEG sets contained only one gene, namely *VDRB* (*vitamin D receptor b*), which was upregulated in both tissues.

Next, to identify biological functions which may be affected by the detected transcript level changes, DEGs were submitted to the *Database for Annotation, Visualization and Integrated*

Discovery (Dennis et al. 2003; Huang et al. 2009), which is an online resource for the functional interpretation of large gene lists. DEGs were analysed for shared GO terms, pathways from the database and protein domains. Only gene groups with a minimum enrichment score ≥ 1 were considered as significant enriched annotation cluster. The 34 upregulated genes were grouped into the following three annotation clusters: *activation of immune response*, *cell adhesion at the plasma membrane* and *positive regulation of apoptosis*. The first cluster, *activation of immune response*, contained only genes which were exclusively upregulated in brain, whereas the other clusters contained genes from both tissues. The 51 downregulated genes seemed to affect mostly *collagens/proteoglycans* and were located in the *extracellular region*; these annotation clusters contained skin DEGs only. Additional annotation clusters of downregulated genes from both tissues described processes like *cell cycle*, *cell division* and *cell proliferation* as well as *DNA replication* (Supplementary File 4).

To verify the DEGs identified by RNA-seq, three up- and downregulated genes were selected from each tissue and analysed in MZM-0403 using qPCR. The same skin and brain samples of young and old MZM-0403 analysed by RNA-seq were reverse-transcribed into cDNA, and the amount of cDNA was measured by qPCR to obtain transcript levels. Except for *VDRB*, qualitative changes in transcript levels were confirmed, with quantitative values being in good agreement (Figure 18).

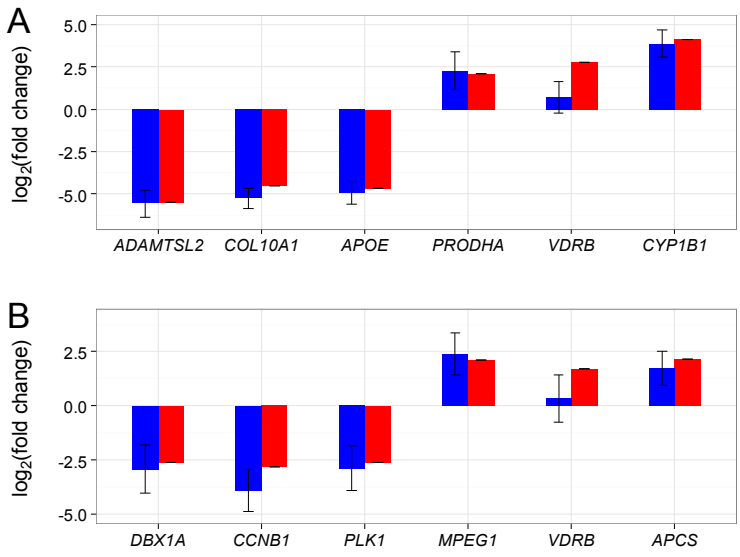

Figure 18: Validation of selected DEGs in MZM-0403 using qPCR.
Three down- and upregulated genes detected by RNA-seq (blue) were selected from skin **(A)** and brain **(B)**, respectively, and qPCR validation (red) was performed in MZM-0403 on the same RNA samples. Positive values indicate overexpression in old age. Note that, for RNA-seq, error bars are meaningless, since only one measurement per gene was available.

3.3.4 Confirmation of DEGs in zebrafish

Next, I analysed whether the observed ageing-related DEGs are specific for *N. furzeri* and repeated this type of analysis in another fish species. I chose zebrafish, because it is well characterised and widely used as model organism for studying many aspects of biology, for example development and ageing. Because zebrafish brain samples were not available, identification of DEGs concentrated on the comparison of skin samples from young and old individuals. Since enough material was available

for the RNA-seq experiment, it was possible to consider biological replicates from the beginning. Consequently, RNA-seq data was generated from skin of 5 and 42 month old individual zebrafish (10 and 5 specimens, respectively). In total, 1,191 zebrafish DEGs were found ($p \leq 0.01$). This number is much higher than in *N. furzeri*, which is likely due to the presence of a more appropriate selection of biological replicates. Interestingly, like in *N. furzeri*, the majority of genes were downregulated with age (547 upregulated vs. 644 downregulated), although the difference was not that pronounced.

Of the 43 DEGs identified in *N. furzeri* skin, 19 (43%) were also present in the zebrafish DEG set and also showed the same direction in fold change (Supplementary File 5). This overlap was significantly different from chance ($p = 2.4x10^{-14}$, binomial test). Of the other 24 genes, additional 16 (33%) showed the same direction, but did not reach significance in zebrafish. Of the remaining eight genes, only one (*cytokine-like 1*, *CYTL1*) was significant in both species but show a different direction in fold change (*N. furzeri*: downregulated, zebrafish: upregulated). In summary, a significant number of DEGs was similarly regulated in ageing of *N. furzeri* and zebrafish. Therefore, these DEGs are presumably involved in general processes of ageing.

3.3.5 Analysis of strain-specific differences during ageing

Before, I used samples of the same tissue and age from both *N. furzeri* strains as replicates for the identification of DEGs. However, this approach disregards any strain-specific differences in ageing that exist between the two strains. MZM-0403 live significantly longer than GRZ, and this difference in lifespan may be also reflected by the changes in transcript levels of the ageing-related DEGs.

Consequently, I analysed the 43 DEGs identified in skin and brain, respectively, for significant strain-specific differences in transcript level fold changes. Initially, only a slight trend was observed which suggested that, in both tissues, fold changes are more extreme in the longer-lived strain MZM-0403 (Figure 19A). However, these differences were not statistically significant (brain: $p = 0.529$, skin: $p = 0.239$; two-sample, paired Wilcoxon test). One reason for these results might be that the calculated fold changes occur over different periods of time. Baseline ("young") for determining transcript levels was for both strains at an age of five weeks, that is, when they reach sexual maturity. However, because GRZ reaches an old age already at 14 weeks while MZM-0403 does so not before 31 weeks, the second time point ("old") differed between the two strains by 17 weeks. To account for this difference, I simplistically assumed that transcript levels change linearly over time and normalised the MZM-0403 fold changes accordingly. Thus, the fold changes originally determined at 31 weeks were scaled down to 14 weeks (Figure 19B). The normalised fold changes were significantly different between GRZ and MZM 0403 in skin ($p = 0.039$) and brain ($p = 0.027$). Moreover, when up- and downregulated genes were analysed separately, the difference was highly significant for brain ($p = 9.1x10^{-05}$ and $3.8x10^{-06}$, respectively) and at least partly significant for skin ($p = 0.067$ and $2.3x10^{-10}$). These results indicate that fold changes over chronological time were more pronounced in GRZ than in MZM 0403, which suggested that ageing is accelerated in GRZ.

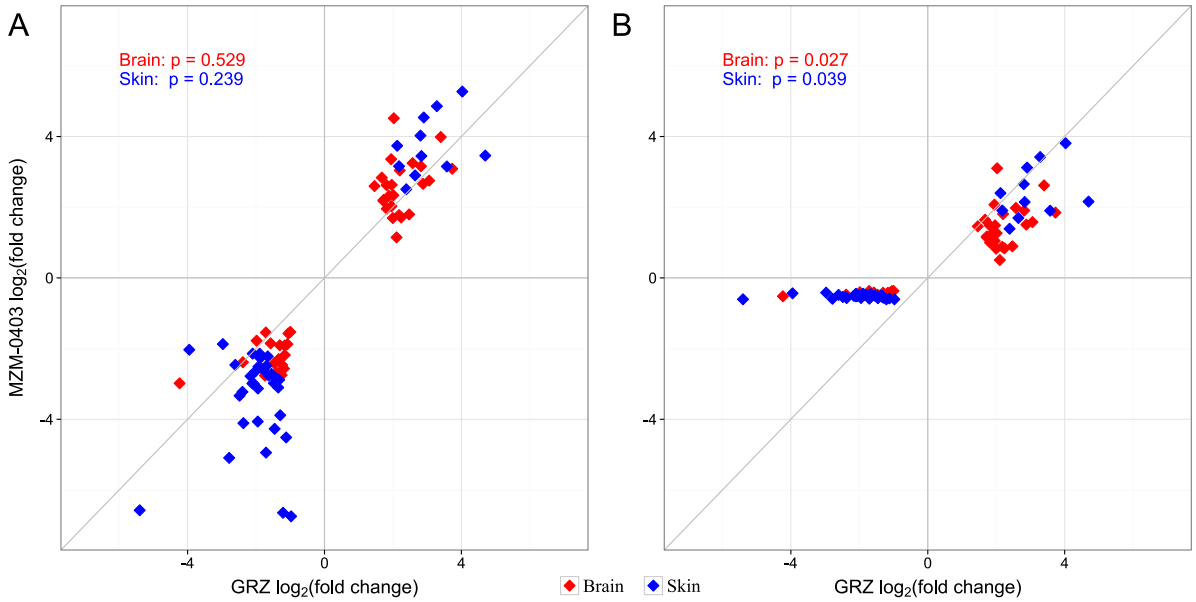

Figure 19: Differences in DEG fold changes between GRZ and MZM-0403.
(A) Fold changes are shown as determined over the time span of 9 and 26 weeks for GRZ and MZM-0403, respectively. **(B)** MZM-0403 fold changes were normalised to the GRZ time span of 9 weeks. P-values are calculated to test for differences in fold changes between GRZ and MZM-0403 using two-sample, paired Wilcoxon test.

4 Discussion

In this thesis, I have described the development of a comprehensive transcript catalogue for the non-model species *N. furzeri* (Petzold et al. 2013), the vertebrate species with the shortest lifespan recorded in captivity (Valdesalici & Cellerino 2003). Classical Sanger as well as NGS technologies were used to sequence the *N. furzeri* transcriptome in great depth across several tissues. To make optimal use of the produced datasets, I developed several new algorithmic approaches which outline general recommendations for the joined processing, assembly and annotation of EST and NGS transcriptome data. The resulting transcript catalogue contains sequences representing the vast majority of all *N. furzeri* protein-coding genes, of which many represent full-length transcripts. Furthermore, I have used the transcript catalogue as reference for the identification of differentially expressed genes during *N. furzeri* brain and skin ageing (Petzold et al. 2013).

4.1 Strategies for the development of a transcript catalogue

Until a few years ago, experimental determination of a transcript catalogue relied on high-throughput Sanger sequencing of random cDNA clones (Gerhard et al. 2004). The non-uniform distribution of transcripts and the number of sequences required made such an undertaking expensive and labour-intensive. With the advent of the NGS technologies, it has been possible to sequence complete transcriptomes in short time and at reasonable costs (Wang et al. 2009). This approach is not limited to the availability of a reference sequence and allows analysing transcriptomes even in non-model species with complex genomes (Vera et al. 2008). However, NGS technologies introduce new challenges for transcriptome analysis, and several aspects concerning library preparation, sequencing, assembly and annotation have to be made, which are discussed below.

Already beginning with RNA extraction, library preparation and sequencing, several issues that critically influence the success of the transcriptome assembly have to be considered. In case of multi-cellular organisms, RNA should be extracted from tissues instead of whole body. Animals are largely composed of muscle, which is one of the least complex transcriptomes; for example, in mammalian muscle tissues, the ten most highly expressed genes make roughly 20–40% thereof (Ramsköld et al. 2009). Consequently, transcripts of these genes are over-sequenced, thereby wasting sequencing power. Ideally, to maximise the number of identified transcripts, different tissues are analysed separately or in pools. For cDNA library synthesis, random hexamer primers should be considered, because oligo(dT) priming can bias the sequence coverage towards the 3' end of the transcripts (Nam et al. 2002). The disadvantage of random hexamer primers is that the priming success is influenced by the nucleotide composition, which leads to an uneven sequence coverage (Hansen et al. 2010). This effect can be mitigated by fragmenting the RNA prior to reverse transcription (Mortazavi et al. 2008). Furthermore, cDNA library normalisation can be done to equalise transcript

levels and to identify more rare transcripts, but this step is not essential since the depth of most NGS technologies is often high enough. Ultimately, many of the described steps are also depending on the applied sequencing technology. For an initial characterisation of the transcriptome or for the analysis of a selected set of transcripts, 454/Roche with its current read lengths of up 1,000 bp may be a good choice. However, 454/Roche lacks the necessary sequencing coverage to identify the rarer transcripts, and other technologies like Solexa/Illumina are better suited. Generally, the applied sequencing strategy should include long reads (≥ 150 bp), which allow spanning multiple exon/intron boundaries. For the same reason, pair-end sequencing should also be considered, because it effectively increases for example Solexa/Illumina read length to approximately 300-500 bp.

The raw sequence data should be pre-processed and quality-checked. Low-quality sequence can cause contig breaks during assembly, leading to transcript fragmentation. Unclipped primer or vector sequence can have even more detrimental effects, because transcript contigs sharing the same primer sequence may be assembled into one chimeric transcript contig; this kind of assembly error is much less obvious and hard to detect. Depending on the sequencing technology, different programs should be used for pre-processing. Sanger and 454/Roche are best processed with classical programs already developed for Sanger data (Bonfield et al. 1995; Chou & Holmes 2001; TIGR 2008). These programs employ sensitive search algorithms and are well-suited for that task. Other NGS technologies have to be processed with new, dedicated programs (several are listed in Patel & Jain 2012). For example, Solexa/Illumina reads show a considerable drop in sequence quality towards the 3' end, which increases with read length (Dohm et al. 2008). Furthermore, reads containing primer are another (underestimated) problem, which occurs when sequencing over the 3' end of the template (Kircher et al. 2011). Errors within reads can be corrected with methods based on k-mers (reviewed in Yang et al. 2013); however, such methods may also remove rare transcripts (Martin & Wang 2011). Finally, removing exact match duplicates should be definitely considered for *de novo* transcriptome assembly. Such duplicates can either originate from highly expressed genes or arise due to biases in the PCR amplification step (Dohm et al. 2008), and they form large read stacks, which are difficult to assemble. Removing duplicate reads equalises the coverage across the transcripts and also reduces the amount of data to be assembled, without losing transcript sequence information. It must be noted that, while duplicate removal is beneficial for *de novo* transcriptome assembly, it should not be applied when determining transcript levels. Ultimately, large amounts of sequence are usually discarded during pre-processing. This loss may seem a waste of resources, but it rather prevents problems caused by the large read numbers produced by NGS.

The assembly of transcriptome data produced by NGS is still a challenging task. The classical programs for assembly were designed for only up to several ten thousand reads and cannot process the massive numbers of short reads due to high computational demands. Though several new programs have been introduced to solve this task (for example, Velvet; Zerbino & Birney 2008), they were

primarily designed for genomic data and, more importantly, are based on assumptions which do not hold for transcriptomes. The goal of genome assembly is to generate few large contiguous genome sequences, whereas transcriptome assembly aims at generating rather short sequence contigs for each expressed gene. Additionally, due to alternative splicing, multiple transcript sequences may be possible for an individual gene in variable tissues and under different conditions (Xu et al. 2002). Furthermore, genome assembly assumes that the read coverage of the sequences to be assembled is relatively homogenous. In contrast, transcript abundances are highly variable, thus leading to a very heterogeneous coverage (Birol et al. 2009). This situation causes two main problems. Reads derived from rare transcripts may be removed by error correction algorithms, which regard low coverage as an indication for putative sequencing errors (Martin & Wang 2011). Moreover, the high read coverage of very abundant transcripts may be misinterpreted by genome assembly programs as an indication for repeat structures (Butler et al. 2008; Zerbino et al. 2009); common strategies for resolving repeats may lead to the fragmentation of these abundant transcripts. Finally, additional problems arise when multiple sequencing technologies with different properties (read length, accuracy and more) have been applied. To summarise, genome and transcriptome assembly face different challenges which require algorithms specifically designed for the respective tasks. However, dedicated assembly programs for transcriptome data were introduced only recently and are still under development (the first program, `Trans-ABySS`, is described in Robertson et al. 2010).

I have addressed these problems by developing a two-step strategy, which accounts for the differences between the sequencing technologies. In the first step, Sanger and 454/Roche reads were assembled using `PAVE` (Soderlund et al. 2009), a classical overlap-layout-consensus program which is ideally suited for long reads in small- to medium-sized numbers (Myers 1995). It compensates well for sequencing errors, is sensitive enough to find less perfect overlaps and generates long contigs within a reasonable runtime. Moreover, it was shown to be superior to already existing programs for transcriptome assembly (Soderlund et al. 2009). In the second step, the Solexa/Illumina datasets were iteratively assembled onto a backbone of `PAVE` contigs using an assembly program based on the *de Bruijn* graph approach (Pevzner et al. 2001). This approach is far better suited for the high read numbers, because it basically breaks reads into smaller words of a fixed length, which effectively reduces the computational complexity of the assembly (for instance, the number of possible overlaps decreases with sequence length). Three other measures additionally improve the assembly. The first is handling of individual Solexa/Illumina libraries in separate assembly iterations. This reduces the impact of eventual library biases introduced during library preparation. Second, the input complexity of a Solexa/Illumina library can be greatly reduced by mapping the reads onto the existing contigs and discarding those which align to a contig and hence do not contribute further to the assembly. Third, as *de Bruijn* graph-based assembly methods are susceptible to sequencing errors and tend to breaks contigs at such position, a subsequent assembly with overlap-layout-consensus programs can merge these fragmented contigs.

The massive sequencing output of the NGS technologies also indirectly affects the subsequent annotation process. Accordingly, transcriptome assembly often results in a very large number of fragmented and redundant transcript contigs (Robertson et al. 2010). However, existing pipelines for transcript annotation like `EST2uni` or `Blast2GO` (Conesa et al. 2005) were originally designed for the small to medium-sized sequence sets obtained by Sanger sequencing and need to be modified accordingly. Furthermore, many sequence analyses, especially similarity searches, are extremely time-consuming with large contig numbers and require a powerful IT infrastructure, for example, large multicore systems or computing clusters. Furthermore, the increased rate of fragmented and redundant transcript contigs creates additional problems for the transcript annotation. Determining whether transcript contigs of the same gene are derived from one or from multiple transcript isoforms is computationally difficult. Furthermore, the need to consider multiple transcript contigs for a gene (for example, as mapping reference) complicates subsequent down-stream analyses.

4.2 Assessment of the assembly quality

Assembly algorithms rely on certain assumptions which try to reduce the complexity of sequence datasets into simple sets of rules that can be efficiently computed. However, because sequencing data is experimentally derived, it is error-prone and underlies stochastic fluctuations. Therefore, these assumptions are not always given. To control the impact of the assembled contigs on the down-stream analyses, a quality control of the obtained assembly is needed.

While a number of criteria and protocols have been established for genome assemblies (Salzberg & Yorke 2005), there are only a few comparable criteria for assessing the quality of transcriptome assemblies (Martin et al. 2010; Martin & Wang 2011). Moreover, these criteria require a genomic reference with annotated gene structures, which is used to evaluate/score the transcript contigs; although, for a minimal quality control, a set of transcript sequences might also serve. Nevertheless, for many non-model species, neither genome nor transcript sequences are available, which makes the quality assessment of *de novo* transcriptome assemblies much more difficult. This was also the case for the *N. furzeri* transcriptome assembly.

Therefore, I used protein sequences from medaka (the closest relative of *N. furzeri* with a sequenced genome, Reichwald et al. 2009) as reference sequences against which the transcript contigs were compared with `BLASTx`. I used only protein sequences with experimental support (1,750), to avoid problems due to misannotations in medaka. The majority of the medaka protein sequences (74%) were hit by at least one *N. furzeri* transcript contig. Of these, 74% were completely covered by multiple transcript contigs, and, notably, 67% were completely covered by a single transcript contig. The latter results indicate that the corresponding transcripts are also present in *N. furzeri* and that their CDSs were fully sequenced. Note that the sequence divergence between the two species might

interfere with the results. Thus, the results represent a rather conservative estimate, and it is likely that several other *N. furzeri* transcripts are also successfully reconstructed, at least to a certain degree.

Hits to the remaining medaka protein sequences (26%) might be missed for several reasons. First of all, the protein sequence identity between *N. furzeri* and medaka may be too low, so that the protein sequences were not identified by the sequence similarity search. Alternatively, the expression of the respective genes might be restricted to tissues not analysed within the present effort. It should be mentioned that transcriptome sequencing included the whole body, which is essentially a mixture of tissues. Thus, besides brain and skin, other tissues may also have been sequenced, at least in smaller amounts. Furthermore, analysed individuals were aged between one week and 14 weeks, and some of the respective genes might be exclusively expressed in the earlier developmental stages. Moreover, genes might be also missed because they were expressed at very low levels.

The protein sequences also allowed assessing the fragmentation and redundancy of the assembly. Transcript fragmentation means that the transcript sequence is assembled into multiple transcript contigs. The rate of fragmentation was rather low and occurred mainly close to the ends of the protein sequences. The main reasons for contig fragmentation are presumably low-quality sequence and shallow sequencing depth. However, alternative splicing events can also cause contig breaks. Many assembly programs assume that there is always only one, unambiguous, possibility to extend a contig. If they encounter a situation where there are two possibilities for a contig extension, they most likely break the contig at this position.

In contrast, the transcriptome assembly exhibited a high degree of redundancy. This was observed for the vast majority of the medaka protein sequences. Over 92% of the total sequence was hit by at least two transcript contigs (nine contigs on average). One conclusion could be that, in *N. furzeri*, almost all corresponding genes are represented at least two transcript isoforms, which would indicate exhaustive alternative splicing and is very unlikely. Another more reasonable explanation is that, although the redundant transcript contigs differ only slightly (for example, in a few positions), assembly failed to merge them into a single contig. Assembly based on *de Bruijn* graphs is known to be susceptible to nucleotide variants and sequence errors (Pevzner et al. 2001). With increasing sequencing depth, errors accumulate, which, in turn, lead to more redundant contigs. This assumption is supported by the observation that the number of *N. furzeri* transcript contigs increases with every assembled Solexa/Illumina library; without errors, this number ought to reach a plateau. Nevertheless, this does not exclude that many redundant transcript contigs actually represent transcript isoforms.

Chimeric transcript contigs may originate from gene or transcript fusions, but they can also indicate severe misassemblies which involve merging contigs of two unrelated transcripts. This type of error is normally hard to detect, but the medaka protein sequences provided the opportunity to at least

estimate the rate of chimeric transcript contigs in the *N. furzeri* transcriptome assembly. Only 0.3% of the transcript contigs with hits showed evidences for chimerism. This relatively low rate can be regarded as an indication for a negligible misassembly rate which does not require further computational or experimental efforts.

4.3 Completeness of the transcript catalogue

Ideally, a transcript catalogue contains transcript sequences for all protein-coding genes, and these transcript sequences are completely reconstructed, that is, they consist of the full CDS plus the UTR. However, this goal is unrealistic, mostly due to spatiotemporal expression (within specific tissues at specific times), library preparation biases, lack of sequencing depth, and problems during assembly and annotation. For these reasons, it is necessary to determine the completeness of a transcript catalogue. The results provide an estimate of the representativeness of the current transcript catalogue and serves as basis of decision-making for further transcriptome sequencing.

The presented *N. furzeri* transcript catalogue contains transcript contigs for almost 20,000 protein-coding genes. This number is typically observed for vertebrate genomes (Volff 2006), and also true for the currently known fish genomes. Nevertheless, without a high-quality *N. furzeri* genome sequence, the overall number of protein-coding genes is not known which complicates estimating the number of genes still missing from the transcript catalogue. Therefore, I first related the gene number to the respective overall numbers observed for medaka, stickleback, tetraodon and zebrafish. BLAST hits to annotated genes in the other fish species indicated that between 71% and 87% of the *N. furzeri* protein-coding genes are present in the transcript catalogue. However, this approach heavily depends on the quality of the gene annotations in the respective fish species. Gene families comprise all genes derived from a common ancestor and are much more stable among fish species. Consequently, gene families are much better suited as a proxy for completeness of transcript catalogues. Identification of gene families in the *N. furzeri* transcript catalogue showed that their number is comparable to those found in the other fish species (according to gene family assignments in the Ensembl Compara database (Vilella et al. 2009)). This result is supported by the observation that, for most gene families, the number of genes determined in *N. furzeri* does not deviate very much from those obtained in the other species. To summarise, the presented *N. furzeri* transcript catalogue is in terms of completeness comparable to gene catalogues of fish species for which genomes have been sequenced.

A much more interesting question is which and how many genes are missing in the *N. furzeri* transcript catalogue. Transcriptome sequencing did not include embryonic tissue and the analysed specimens were at least one week old. Consequently, genes that are expressed in the early development of *N. furzeri* are most likely missing. For example, in zebrafish, between 1,400 and 2,400 genes have been estimated to be involved in embryonic/early larval development (Haffter et al. 1996; Amsterdam et al. 2004). Generally, recent gene annotation in zebrafish identified roughly 26,000

protein-coding genes (Howe et al. 2013), which is far more comprehensive than in the other three fish species medaka, stickleback and tetraodon (all roughly 20,000 genes; Kasahara et al. 2007; Jones et al. 2012; Jaillon et al. 2004). But the conclusion that some 6,000 genes are still missing in the *N. furzeri* transcript catalogue is misleading for several reasons. First of all, a large fraction of the additional genes presumably originates from the additional, teleost-specific, whole genome duplication. However, the fraction of retained gene copies (that are not lost shortly after the duplication event) varies between fish species (Brunet et al. 2006). Thus, it might be that the additional zebrafish genes do not have orthologs in *N. furzeri* or that the respective *N. furzeri* orthologs were turned into pseudogenes without measurable expression. On the other hand, the Ensembl zebrafish gene annotation may still contain errors, and the number of protein-coding genes may actually be an overestimation. The zebrafish genome was recently improved by incorporating RNA-seq expression data (Collins et al. 2012). About 46% of the original gene models obtained by traditional evidence (cDNA sequences and prediction of orthologs) were refined, suggesting that the zebrafish annotation accuracy can be still improved.

Besides overall gene number, percentage of assembled CDS is the second important criterion for the completeness of transcript catalogue. More than 85% of the *N. furzeri* protein-coding genes were represented by a transcript contig which contained at least half of the CDS. More importantly, for more than 70% of the protein-coding genes, the transcript contig contained the complete (> 90%) CDS. This high degree of CDS coverage was not exclusively restricted to small- or normal-sized genes but also applied to those genes which code for the large structural proteins. For example, one *N. furzeri* transcript contig of the gene *TTN* was found to encode the complete ~30,000 aa large *Titin* protein. Furthermore, when I compared the sum of the non-redundant CDS identified in the transcript catalogue to the total CDS of all reference entries used for annotation, the completeness was estimated to be 83%. That means that, simply put, over 80% of the *N. furzeri* overall protein sequence is known.

4.4 Transcript contigs without annotation

Assembly of the *N. furzeri* transcriptome data resulted in 209,660 transcript contigs, of which, however, only 109,032 (52%) were annotated based on sequence similarity to known protein and protein-coding transcript sequences (Petzold et al. 2013). Thus, 100,628 (58%) transcript contigs remained for which annotation could not be obtained and whose source is currently unclear. Theoretically, these contigs might represent artefacts from library preparation or sequencing, or undetected contamination. Fortunately, the draft assembly of the *N. furzeri* genome allowed verifying the source of the unannotated transcript contigs. The large majority of contigs (95%) was successfully mapped to the draft assembly, thus showing that they indeed originate from transcribed regions of the *N. furzeri* genome. However, the draft assembly is not yet annotated, and, thus, the nature of the unannotated transcript contigs remains unclear. Several explanations for these unannotated contigs can be thought of.

Some transcript contigs might remain unannotated due to problems during the annotation process. The annotation process relied on sequence similarity searches against databases with known protein and protein-coding transcript sequences, mainly of four fish species with a sequenced genome - medaka, stickleback, tetraodon and zebrafish. Hence, successful annotation of a *N. furzeri* transcript contig depends on the availability of homologous sequences. This assumption might be problematic for three main reasons. First, the respective *N. furzeri* gene was not assembled or not annotated in the genome reference of the other fish species. Three of the four fish species' genome assemblies available at Ensembl are older, and annotation of these genomes was largely based on similarity searches and gene prediction (Flicek et al. 2011). Only the zebrafish genome was recently comprehensively re-assembled and annotated with RNA-seq data (Collins et al. 2012; Howe et al. 2013). Second, the *N. furzeri* transcript contigs and their orthologs were too divergent, and, thus, sequence similarity searches failed. This happens for genes which are not under negative evolutionary pressure and therefore may diverge very quickly. Third, the respective gene is species-specific, that is, it only exists in *N. furzeri*. However, all three discussed points presumably account only for a small part of the unannotated transcript contigs. To address these two issues, the annotation pipeline additionally included two large protein sequence collections NCBI nr and UniProt as well ESTs from several fish species, which should have compensated for the missing sequence information or should have provided sequences of other, more closely related fish species. And concerning the third point, although CDS were predicted for the majority of the unannotated transcript contigs, they were on average shorter by two thirds, compared to those of annotated transcript contigs (356 vs. 883 bp). Moreover, subsequent search for conserved protein family domains by HMMER found hits for only four additional contigs. Given that HMMER is based on profile hidden Markov models, which are more sensitive for remote sequence matches (Johnson et al. 2010), it is likely that there are only a few novel protein-coding transcripts hidden among the unannotated transcript contigs.

A considerable part of the contigs might therefore derive from UTR fragments of already annotated protein-coding transcripts. UTR refers to the untranslated regions 5' and 3' to the CDS and can make up a considerable fraction of the transcript (Grillo et al. 2010). Generally, sequencing and assembly of UTR is problematic because UTR sequence coverage is strongly affected by the methods used for library construction and template fragmentation. For example, when using oligo(dT) primer for cDNA synthesis, internal priming can cause a high frequency of truncated cDNAs (Nam et al. 2002), which, in turn, leads to a sequence coverage biased towards the 3' end of the transcript (Hoque et al. 2013). Furthermore, methods for cDNA fragmentation such as DNase I treatment or sonication are also biased towards the 3' ends of the transcript (Mortazavi et al. 2008). These methods were applied to some extent during library preparation and may have resulted in low or fluctuating sequencing coverage at the UTRs. However, many assemblies programs try to reconstruct contigs that show a balanced read coverage and will introduce contig breaks at positions with aberrant coverage. Some weak evidence for this assumption comes from the observation that the majority of unannotated

transcript contigs contained CDS predictions at one of its ends. However, these predicted CDS had no sequence similarity to any known protein or transcript sequence. Alternatively, they might encode for upstream ORFs, which can be found before the main start codon and play a role in the regulation of the gene expression (Mignone et al. 2002). Finally, the prediction methods applied for the unannotated transcript contigs were rather crude (for example, identifying the longest six-frame translation uninterrupted by a stop codon), and, therefore, many CDS predictions are presumably false positives.

Part of the unannotated transcript contigs might originate from dispersed repeats. One class of dispersed repeats, the retrotransposons, employs reverse transcription of RNA intermediates for transposition. Likely, these RNA intermediates can be detected by the high depth of transcriptome sequencing with NGS technologies. Indeed, in 2009, a study analysed the contribution of repetitive elements to the transcriptomes of mouse and human and revealed an intensive transcriptional activity of retrotransposons (Faulkner et al. 2009). In *N. furzeri*, almost 41% of the genome consist of dispersed repeats (Koch 2010). Unfortunately, because many of these repeats are still unclassified, the percentage of retrotransposons is not known yet. Nevertheless, it is likely that a number of retrotransposons is still transcribed. Dispersed repeat elements were identified in 53% of the unannotated *N. furzeri* transcript contigs. Whether and to which part these contigs really originate from active retrotransposons is difficult to say. Retrotransposons also occur in the 3' UTR of protein-coding transcripts (Yulug et al. 1995). Consequently, the transcript contigs could have been derived from transcription of non-repetitive sequences or retrotransposons. Interestingly, retrotransposon insertions in 3'UTRs have been demonstrated to repress transcription (Chen et al. 2008). More importantly, retrotransposons in 3'UTR have also been proposed to truncate full-length transcripts by providing an alternative terminator or to facilitate transcript degradation by forming double-stranded RNA with other retrotransposon transcripts (Faulkner et al. 2009).

Though functionally most important, protein-coding mRNA transcripts make up only a small part of the transcriptome (2-4%; Lindberg & Lundeberg 2010), and the larger part can be attributed to a number of non-protein-coding RNA species. Thereof, rRNA and tRNA transcripts make up the largest part. However, only few hits were found for the unannotated *N. furzeri* transcript contigs, which presumably is due to an efficient poly(A)-selection during library preparation. Small non-protein-coding RNA species such as miRNAs or small nuclear RNAs may be theoretically sequenced as by-products or as part of their precursor transcripts. Long non-protein-coding RNA (≥ 200 nt), often referred to as the dark matter of the genome, have been proposed to play an important role as regulators of gene activity (Nagano & Fraser 2011). Unfortunately, so far, little is known about non-protein-coding RNA species in *N. furzeri*.

Several studies in mammalian genomes demonstrated that the vast majority of the genome is pervasively transcribed, including large parts of the intergenic regions (Carninci et al. 2005; Birney et al. 2007). One recent study reported that almost 74% of the human genome contains primary

transcripts and implied that the number of genes in human is actually twice as large as originally assumed (Djebali et al. 2012). Whether the additionally detected transcribed units represent functional entities/genes or are the product of technical/biological noise is currently subject of a hot debate (Graur et al. 2013; Eddy 2013). Nevertheless, these reports shed some light on the large number of unannotated *N. furzeri* transcript contigs. The finding that over 95% can be aligned to the draft assembly of the *N. furzeri* genome suggests that they are derived from pervasive transcription. However, similar to the situation in the mammalian genomes, the exact reasons for this phenomenon remain unclear.

4.5 Duplicated genes

Gene duplication is considered as an important mechanism of evolution, which facilitates the creation of new genes with novel functions (Ohno 1970). One major causative event for duplicated genes is the whole genome duplication, which is caused by improper chromosome pairing during meiosis and results in additional copies of the entire genome transferred to the offspring. Vertebrate genomes are believed to have undergone at least two rounds of whole genome duplication (Spring 1997); this hypothesis is best known as 2R hypothesis. Later, it was shown that the genomes of teleost fish species underwent a 3rd round of genome duplication (Amores et al. 1998; Taylor et al. 2003); this event is referred to as teleost-specific whole genome duplication or as 3R duplication.

About 19% of the protein-coding *N. furzeri* genes show evidences for teleost-specific paralogs. Conversely, this indicates that, after whole genome duplication, >80% of the gene paralogs were lost from the *N. furzeri* genome. These numbers are roughly comparable to those calculated for other fish species (zebrafish: 22%, medaka: 23%, stickleback: 23% and tetraodon: 24%; Kassahn et al. 2009, synteny results). The vast majority of the identified *N. furzeri* paralogs occur in pairs, concordant with the teleost-specific whole genome duplication. A number of genes are represented by more than two paralogs; however, it is likely that many were erroneously identified as result of methodological problems, for example misannotations, false-positive BLAST hits or errors in the gene phylogeny. Moreover, it has to be considered that not all of these duplications may date back to the teleost-specific genome duplication. Still, some of these multiple copies might represent novel species-specific paralogs, which may originate for example from a subsequent gene duplication event. Finally, it should be noted that the transcriptome approach could only detect paralogs whose expression was sufficiently high in the analysed samples.

4.6 Differentially expressed genes in ageing *N. furzeri*

After completion of the transcript catalogue, I employed selected RNA-seq datasets to conduct an initial characterisation of ageing-related changes in gene expression in *N. furzeri*. The crucial requirement for analysing transcript levels with RNA-seq is the availability of a reference for read

mapping, which is commonly a sequenced genome. In case of *N. furzeri*, the transcript catalogue served as reference. This type of reference raises several important issues for read mapping and counting. In the simplest case, a gene is represented in the reference by exactly one transcript contig. The RNA-seq reads derived from that gene/transcripts can then be unambiguously mapped to the corresponding transcript contig, and the read count can be seen as a measure for the transcript level of the gene. Genes, however, are often represented by multiple transcript isoforms; between 17 and 43% of the genes in teleost fish species undergo alternative splicing (Lu et al. 2010). Consequently, the RNA-seq reads of a gene are essentially a mixture of its transcript isoforms. This situation complicates both read mapping and counting. First of all, many reads map to multiple transcript contigs and cannot be unambiguously assigned. Second, for genes with multiple transcript isoforms, integrating the read counts from the different transcripts is not straightforward. Several solutions to deal with such multiple mappings were described previously. The simplest approach discards them and uses only the uniquely mapping reads (Marioni et al. 2008). A first strategy to 'rescue' multiple mapping reads proposed to distribute fractions of them to genes in proportion to the coverage of the uniquely mapped reads (Morin et al. 2008). More sophisticated methods try to model the RNA-seq experiment using a maximum likelihood function which fits the expected transcript isoform abundances to the observed read coverage (Trapnell et al. 2010; Katz et al. 2010). However, these programs were designed for genomic mapping references and cannot be applied for transcript references. Moreover, even it were possible, such programs would presumably introduce additional bias which would have to be controlled for.

To reduce the impact of these problems on the subsequent analyses of differential expression, I constructed a reference from the longest transcript contig of each gene (19,875 contigs; 53 Mb). Only 53% of the Solexa/Illumina reads were mapped uniquely. This relatively low mapping efficiency could have been caused either by the incompleteness of the transcript catalogue or by the design of the reference. Therefore, I decided to include more transcript contigs and constructed another reference from all transcript contigs with gene annotations (85,431 contigs; 144 Mb). Accordingly, mapping efficiency increased to 91%, which indicated that the initial mapping results were not due to missing gene information but rather reflected the large amount of valuable sequence information removed as result of the design of the reference. However, the second reference caused a high fraction of multiple mappings (~80%), presumably as result of the additionally introduced redundant transcript contigs. These might represent transcript isoforms, but they might also indicate unrecognised gene paralogs or assembly problems due to sequencing errors or repetitive motifs. Interestingly, I observed a similar situation for the zebrafish data, for which a genome is available. Using the genome, mapping efficiency is usually between 75 and 80% (Marco Groth, personal communication). However, when I used transcript instead of genome sequences as reference, mapping efficiency dropped below 50% (data not shown). This result confirms that the low mapping rate is the result of the design of the

respective reference. Generally, it suggests that using a genome as reference is superior to using all known transcripts.

In all samples, over 17,000 genes (Table 6) were profiled. At the first glance, this was surprising, because transcriptomes are highly specific for the analysed tissue or organ (Whitehead & Crawford 2005). However, many genes were covered only by a few reads, and their detection may indicate technical noise or very low expression levels. When a minimum threshold of ten reads was applied, the number of profiled genes dropped to roughly 15,000-16,500. However, of these, several (between 1,500 and 3,000 per respective sample) still showed very low transcript levels, that is, below 1 RPKM. Considering that values of 1-4 RPKM roughly correspond to 1 copy per cell (Mortazavi et al. 2008), it can be assumed that approximately 12,000-15,000 genes are expressed in all samples. On the other hand, the relatively large number of genes with very low transcript levels still needs further clarifying. It could be that the high depth of RNA-seq has picked up some kind of spurious/pervasive transcription whose biological implications cannot be assessed by the available data.

Transcriptome profiles of the same tissue showed the highest correlation, which was expected and demonstrates the tissue-specific gene expression in *N. furzeri*. Interestingly, other weaker trends were observed for the individual tissues. In skin, samples of the same age showed the second-best correlations, whereas in brain, this was the case for samples of the same strain. These weak but observable differences suggest that the ageing-related transcript level changes are specific for the respective tissues, which, in turn, would mean that tissues age by different ways. Furthermore, a PCA conducted on all samples indicated that the *N. furzeri* transcriptome data is composed of at three major and five minor components, which map to factors of gene expression in the respective analysed samples. The observation that the three main components were associated with tissues, ages and strains agrees well with the results of the correlation analysis.

In total, I have identified 43 ageing-related DEGs in *N. furzeri* brain and skin, respectively. Interestingly, the two tissues differed in the predominant direction of the transcript level change. In skin, the majority of DEGs was downregulated with age, whereas in brain, the opposite was found. This observation agrees well with the results found by correlation analysis and PCA and further supports the general picture of an ageing process that differs between tissues. The biological functions of the DEGs reflected the generally expected changes in an ageing phenotype (Magalhães et al. 2009). DEGs involved in apoptosis were upregulated, whereas DEGs involved in cell cycle control, cell division and proliferation were downregulated. The pattern was more prominent in brain than in skin, indicating more dynamic changes of transcript levels over lifetime. Furthermore, transcript levels also showed an increase in inflammation in brain, which previously has been reported as a marker for ageing-related decline in brain (Schumacher et al. 2008; Park et al. 2009) and for general ageing (Lee et al. 1999). Notably, downregulated skin DEGs were mainly involved in the maintenance of

collagenous tissue and cartilage, indicative for a decline in tissue homeostasis typically for ageing (Fisher et al. 2002).

The major drawback of the RNA-seq experiment was the lack of biological replicates. Initially, the *N. furzeri* RNA-seq was aimed at the completion of the transcript catalogue, and due to technical reasons (material availability, stage of technological development) multiple specimens were pooled instead of being analysed separately. On the other hand, one can assume that the pooling strategy efficiently reduces the relative abundance of biological noise derived from single individuals. Hence, those pooled datasets can be considered reliable. Nevertheless, to obtain an initial set of statistically significant DEGs, I decided to ignore strain-specific differences and treated samples of the same tissue and age as replicates. The validation of selected DEGs by qPCR showed that the rate of false positives was low (2 of 12). However, it is very likely that this approach missed DEGs which showed considerable strain-specific variation or were exclusive to one strain. A higher number of strain-specific replicates would presumably increase the sensitivity of the statistical tests, thus resulting in more candidate DEGs.

It should be noted that, theoretically, the obtained DEGs might simply reflect transcript level changes that were caused by the laboratory handling of the fish; for example, they might be a response to certain changes in feeding or care conditions. However, I could confirm many DEGs by an independent RNA-seq experiment in zebrafish, that is, 43% of the *N. furzeri* DEGs were significantly regulated in the same manner in zebrafish and another 33% showed at least the same direction of fold change. This suggested that transcript level dynamics over lifetime are similar in both *N. furzeri* and zebrafish. Thus, *N. furzeri*, in spite of its exceptionally short lifespan of only few months, exhibits general characteristics of vertebrate ageing.

4.7 Accelerated ageing in the *N. furzeri* strain GRZ

The initial approach for identifying DEGs ignored potential strain-specific differences, that is, I essentially treated GRZ and MZM-0403 individuals as one *N. furzeri* strain. However, the two strains significantly differ in lifespan, and previous studies of ageing-related biomarkers suggested that ageing is accelerated in GRZ, compared to MZM-0403 (Terzibasi et al. 2008; Di Cicco et al. 2011). This difference might be also reflected by the transcript level changes of the identified ageing-related DEGs. However, a simple comparison of the fold changes may be misleading because it ignores the different periods of times in which the changes occurred. To obtain comparable fold changes, I assumed that gene expression changes linearly over time and normalised the fold changes to the same period of time. Indeed, time-normalised fold changes were stronger in GRZ, compared to MZM-0403, which supports the hypothesis that ageing is accelerated in the short-lived GRZ strain. Of course, the assumption of a linear change in gene expression is most likely in many cases an oversimplification. Nevertheless, this two-point analysis represents only a first approximation. On-going and future

studies include/will have to include additional time points to better dissect the chronological course of gene expression during ageing in the *N. furzeri* strains.

4.8 Outlook

Future work is aimed at completing the *N. furzeri* transcript catalogue, for example, by identification of genes which are primarily expressed during embryonic development. For this purpose, RNA was isolated from MZM-0403 embryos, 24 hours post fertilisation and during the somite stage, and sequenced with Solexa/Illumina (2 x 150 bp). To take optimal advantage of this additional dataset, I performed a complete new *de novo* assembly including the `PAVE` contigs and all available Solexa/Illumina datasets as input. The resulting transcriptome assembly is currently being annotated by the `EST2uni` pipeline. Preliminary results indicate that the transcript catalogue is complemented by roughly 500 new genes, which presumably represent those exclusively expressed during embryonic development. Moreover, the number of genes with complete transcript contigs seems to be moderately larger. Subsequent to successful annotation, this new version of the transcript catalogue will be made available in the transcriptome browser.

In parallel, the *N. furzeri* genome is being assembled by the Genome Analysis group at the FLI and will be released in the very near future. After completion of this effort, the transcript catalogue will serve as a valuable resource for the genome annotation. The transcript contigs will be spliced-aligned against the genome, which allows determining exonic and intronic regions as well as associated CDS and UTR. The resulting gene structures will be supported by and complemented with gene models obtained from gene prediction programs to reconstruct the complete set of genes in *N. furzeri*, including protein- and non-protein-coding genes. Moreover, the transcript contigs will also facilitate the identification and quantification of transcript isoforms in the genome. Conversely, a *N. furzeri* genome will help to improve the transcript catalogue. The transcript contigs can be safely assigned to genes, based on their genomic mapping position, which allows to merge fragmented transcript contigs and to distinguish multiple transcript isoforms. Furthermore, an annotated genome will also allow clarifying the source and potential function of many of the remaining unannotated transcript contigs.

In future, further RNA-seq experiments have to be done to better understand ageing in *N. furzeri*. First of all, these should generally utilise biological replicates to identify more subtle ageing-related changes in gene expression. The design of the experiments should include additional time points, to monitor the changes over the complete lifespan of *N. furzeri*, and additional longer-lived strains, to determine which gene expression changes are involved in the lifespan variation. Finally, transcriptome analysis should also be extended to other Nothobranchius species, other teleost fish and finally other vertebrate species. The goal of this interspecies analysis would be to identify a general shared signature of ageing in vertebrates.

References

454 Life Sciences, 2012. *GS Data Analysis Software package*, Available: http://454.com/products/analysis-software/index.asp.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., George, R.A., Lewis, S.E., Richards, S., Ashburner, M., Henderson, S.N., Sutton, G.G., Wortman, J.R., Yandell, M.D., Zhang, Q., Chen, L.X., et al., 2000. The Genome Sequence of *Drosophila melanogaster*. *Science*, 287(5461), pp.2185–2195.

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F. & Et, A., 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252(5013), pp.1651–1656.

Akiyama, H., Barger, S., Barnum, S., Bradt, B., Bauer, J., Cole, G.M., Cooper, N.R., Eikelenboom, P., Emmerling, M., Fiebich, B.L., Finch, C.E., Frautschy, S., Griffin, W.S., Hampel, H., Hull, M., Landreth, G., Lue, L., Mrak, R., Mackenzie, I.R., McGeer, P.L., et al., 2000. Inflammation and Alzheimer's disease. *Neurobiology of aging*, 21(3), pp.383–421.

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

Alwine, J.C., Kemp, D.J. & Stark, G.R., 1977. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp.5350–5354.

Amores, A., Force, A., Yan, Y.L., Joly, L., Amemiya, C., Fritz, A., Ho, R.K., Langeland, J., Prince, V., Wang, Y.L., Westerfield, M., Ekker, M. & Postlethwait, J.H., 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science (New York, N.Y.)*, 282(5394), pp.1711–1714.

Amsterdam, A., Nissen, R.M., Sun, Z., Swindell, E.C., Farrington, S. & Hopkins, N., 2004. Identification of 315 genes essential for early zebrafish development. *Proceedings of the National Academy of Sciences of the United States of America*, 101(35), pp.12792–12797.

Anders, S. & Huber, W., 2010. Differential expression analysis for sequence count data. *Genome Biology*, 11(10), p.R106.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. & Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1), pp.25–29.

Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., Kimelman, M., Pruitt, K.D., Resenchuk, S., Tatusova, T., Yaschenko, E. & Ostell, J., 2011. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Research*, 40(D1), pp.D57–D63.

Benson, D.A., Karsch-Mizrachi, I., Clark, K., Lipman, D.J., Ostell, J. & Sayers, E.W., 2012. GenBank. *Nucleic Acids Research*, 40(D1), pp.D48–D53.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira Cheetham, R., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., et al., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218), pp.53–59.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., Kuehn, M.S., Taylor, C.M., Neph, S., Koch, C.M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J.A., Andrews, R.M.,

Flicek, P., et al., 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146), pp.799–816.

Birol, I., Jackman, S.D., Nielsen, C.B., Qian, J.Q., Varhol, R., Stazyk, G., Morin, R.D., Zhao, Y., Hirst, M., Schein, J.E., Horsman, D.E., Connors, J.M., Gascoyne, R.D., Marra, M.A. & Jones, S.J.M., 2009. De novo transcriptome assembly with ABySS. *Bioinformatics*, 25(21), pp.2872–2877.

Blüher, M., Kahn, B.B. & Kahn, C.R., 2003. Extended longevity in mice lacking the insulin receptor in adipose tissue. *Science (New York, N.Y.)*, 299(5606), pp.572–574.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. & Schneider, M., 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, 31(1), pp.365–370.

Boguski, M.S., Lowe, T.M. & Tolstoshev, C.M., 1993. dbEST--database for 'expressed sequence tags'. *Nature genetics*, 4(4), pp.332–333.

Bonfield, J.K., Smith, K.F. & Staden, R., 1995. A new DNA sequence assembly program. *Nucleic Acids Research*, 23(24), pp.4992–4999.

Brunet, F.G., Roest Crollius, H., Paris, M., Aury, J.-M., Gibert, P., Jaillon, O., Laudet, V. & Robinson-Rechavi, M., 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Molecular Biology and Evolution*, 23(9), pp.1808–1816.

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C. & Jaffe, D.B., 2008. ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. *Genome research*, 18(5), pp.810–820.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V.B., Brenner, S.E., Batalov, S., Forrest, A.R.R., Zavolan, M., Davis, M.J., Wilming, L.G., Aidinis, V., et al., 2005. The Transcriptional Landscape of the Mammalian Genome. *Science*, 309(5740), pp.1559–1563.

Chan, P.P. & Lowe, T.M., 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic acids research*, 37(Database issue), pp.D93–97.

Chen, L.-L., DeCerbo, J.N. & Carmichael, G.G., 2008. Alu element-mediated gene silencing. *The EMBO journal*, 27(12), pp.1694–1705.

Chou, H.H. & Holmes, M.H., 2001. DNA sequence quality trimming and vector removal. *Bioinformatics (Oxford, England)*, 17(12), pp.1093–1104.

Christensen, K., Doblhammer, G., Rau, R. & Vaupel, J.W., 2009. Ageing populations: the challenges ahead. *The Lancet*, 374(9696), pp.1196–1208.

Christensen, K., Johnson, T.E. & Vaupel, J.W., 2006. The quest for genetic determinants of human longevity: challenges and insights. *Nature reviews. Genetics*, 7(6), pp.436–448.

Church, D.M., Goodstadt, L., Hillier, L.W., Zody, M.C., Goldstein, S., She, X., Bult, C.J., Agarwala, R., Cherry, J.L., DiCuccio, M., Hlavina, W., Kapustin, Y., Meric, P., Maglott, D., Birtle, Z., Marques, A.C., Graves, T., Zhou, S., Teague, B., Potamousis, K., et al., 2009. Lineage-Specific Biology Revealed by a Finished Genome Assembly of the Mouse. *PLoS Biology*, 7(5). Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2680341/ [Accessed August 28, 2013].

Di Cicco, E., Tozzini, E.T., Rossi, G. & Cellerino, A., 2011. The short-lived annual fish Nothobranchius furzeri shows a typical teleost aging process reinforced by high incidence of age-dependent neoplasias. *Experimental gerontology*, 46(4), pp.249–256.

CLCbio, 2011. *CLC Assembly Cell*, Available: http://www.clcbio.com/.

Collins, J.E., White, S., Searle, S.M.J. & Stemple, D.L., 2012. Incorporating RNA-seq data into the zebrafish Ensembl genebuild. *Genome Research*, 22(10), pp.2067–2078.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. & Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics (Oxford, England)*, 21(18), pp.3674–3676.

Danesh, J., Kaptoge, S., Mann, A.G., Sarwar, N., Wood, A., Angleman, S.B., Wensley, F., Higgins, J.P.T., Lennon, L., Eiriksdottir, G., Rumley, A., Whincup, P.H., Lowe, G.D.O. & Gudnason, V., 2008. Long-term interleukin-6 levels and subsequent risk of coronary heart disease: two new prospective studies and a systematic review. *PLoS medicine*, 5(4), p.e78.

Dennis, G., Jr, Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C. & Lempicki, R.A., 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*, 4(5), p.P3.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G.K., Khatun, J., Williams, B.A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., et al., 2012. Landscape of transcription in human cells. *Nature*, 489(7414), pp.101–108.

Dodd, L.E., Korn, E.L., McShane, L.M., Chandramouli, G.V.R. & Chuang, E.Y., 2004. Correcting log ratios for signal saturation in cDNA microarrays. *Bioinformatics (Oxford, England)*, 20(16), pp.2685–2693.

Dohm, J.C., Lottaz, C., Borodina, T. & Himmelbauer, H., 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic acids research*, 36(16), p.e105.

Eddy, S.R., 2012. *HMMER v3.0*, Available: http://www.hmmer.org/.

Eddy, S.R., 2013. The ENCODE project: Missteps overshadowing a success. *Current Biology*, 23(7), pp.R259–R261.

Fabrizio, P., Pozza, F., Pletcher, S.D., Gendron, C.M. & Longo, V.D., 2001. Regulation of longevity and stress resistance by Sch9 in yeast. *Science (New York, N.Y.)*, 292(5515), pp.288–290.

Faulkner, G.J., Kimura, Y., Daub, C.O., Wani, S., Plessy, C., Irvine, K.M., Schroder, K., Cloonan, N., Steptoe, A.L., Lassmann, T., Waki, K., Hornig, N., Arakawa, T., Takahashi, H., Kawai, J., Forrest, A.R.R., Suzuki, H., Hayashizaki, Y., Hume, D.A., Orlando, V., et al., 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nature genetics*, 41(5), pp.563–571.

Ferrucci, L., Corsi, A., Lauretani, F., Bandinelli, S., Bartali, B., Taub, D.D., Guralnik, J.M. & Longo, D.L., 2005. The origins of age-related proinflammatory state. *Blood*, 105(6), pp.2294–2299.

Fisher, G.J., Kang, S., Varani, J., Bata-Csorgo, Z., Wan, Y., Datta, S. & Voorhees, J.J., 2002. Mechanisms of photoaging and chronological skin aging. *Archives of dermatology*, 138(11), pp.1462–1470.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kahari, A.K., Keefe, D., Keenan, S., Kinsella, R., Komorowska, M., et al., 2011. Ensembl 2012. *Nucleic Acids Research*, 40(D1), pp.D84–D90.

Forment, J., Gilabert, F., Robles, A., Conejero, V., Nuez, F. & Blanca, J., 2008. EST2uni: an open, parallel tool for automated EST analysis and database creation, with a data mining web interface and microarray expression data integration. *BMC Bioinformatics*, 9(1), p.5.

Friedman, D.B. & Johnson, T.E., 1988. A Mutation in the Age-1 Gene in Caenorhabditis Elegans Lengthens Life and Reduces Hermaphrodite Fertility. *Genetics*, 118(1), pp.75–86.

Genade, T., 2007. Laboratory manual for culturing *N. furzeri*. Available: http://www.nothobranchius.info/pdfs/lab_protocols_1.pdf.

Genade, T., Benedetti, M., Terzibasi, E., Roncaglia, P., Valenzano, D.R., Cattaneo, A. & Cellerino, A., 2005. Annual fishes of the genus Nothobranchius as a model system for aging research. *Aging Cell*, 4(5), pp.223–233.

Gerhard, D.S., Wagner, L., Feingold, E.A., Shenmen, C.M., Grouse, L.H., Schuler, G., Klein, S.L., Old, S., Rasooly, R., Good, P., Guyer, M., Peck, A.M., Derge, J.G., Lipman, D., Collins, F.S., Jang, W., Sherry, S., Feolo, M., Misquitta, L., Lee, E., et al., 2004. The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome research*, 14(10B), pp.2121–2127.

Gish, W., 1996. *WU-BLAST*, Available: http://blast.wustl.edu/.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H. & Oliver, S.G., 1996. Life with 6000 Genes. *Science*, 274(5287), pp.546–567.

Goto, M., 1997. Hierarchical deterioration of body systems in Werner's syndrome: Implications for normal ageing. *Mechanisms of Ageing and Development*, 98(3), pp.239–254.

Graur, D., Zheng, Y., Price, N., Azevedo, R.B.R., Zufall, R.A. & Elhaik, E., 2013. On the immortality of television sets: 'function' in the human genome according to the evolution-free gospel of ENCODE. *Genome Biology and Evolution*.

Grillo, G., Turi, A., Licciulli, F., Mignone, F., Liuni, S., Banfi, S., Gennarino, V.A., Horner, D.S., Pavesi, G., Picardi, E. & Pesole, G., 2010. UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Research*, 38(Database issue), pp.D75–D80.

Haas, R., 1976. Behavioral Biology of the Annual Killifish, Nothobranchius guentheri. *Copeia*, 1976(1), pp.80–91.

Haffter, P., Granato, M., Brand, M., Mullins, M.C., Hammerschmidt, M., Kane, D.A., Odenthal, J., van Eeden, F.J., Jiang, Y.J., Heisenberg, C.P., Kelsh, R.N., Furutani-Seiki, M., Vogelsang, E., Beuchle, D., Schach, U., Fabian, C. & Nüsslein-Volhard, C., 1996. The identification of genes with unique and essential functions in the development of the zebrafish, Danio rerio. *Development (Cambridge, England)*, 123, pp.1–36.

Hansen, K.D., Brenner, S.E. & Dudoit, S., 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, 38(12), pp.e131–e131.

Harbers, M. & Carninci, P., 2005. Tag-based approaches for transcriptome research and genome annotation. *Nature Methods*, 2(7), pp.495–502.

Harman, B., 1956. Aging: a theory based on free radical and radiation chemistry. *Journal of gerontology*, 11(3), pp.298–300.

Harman, D., 1972. The biologic clock: the mitochondria? *Journal of the American Geriatrics Society*, 20(4), pp.145–147.

Harrison, D.E., Strong, R., Sharp, Z.D., Nelson, J.F., Astle, C.M., Flurkey, K., Nadon, N.L., Wilkinson, J.E., Frenkel, K., Carter, C.S., Pahor, M., Javors, M.A., Fernandez, E. & Miller, R.A., 2009. Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. *Nature*, 460(7253), pp.392–395.

Hartmann, N., Reichwald, K., Lechel, A., Graf, M., Kirschner, J., Dorn, A., Terzibasi, E., Wellner, J., Platzer, M., Rudolph, K.L., Cellerino, A. & Englert, C., 2009. Telomeres shorten while Tert expression increases during ageing of the short-lived fish *Nothobranchius furzeri*. *Mechanisms of Ageing and Development*, 130(5), pp.290–296.

Hay, N. & Sonenberg, N., 2004. Upstream and downstream of mTOR. *Genes & Development*, 18(16), pp.1926–1945.

Henderson, S.T. & Johnson, T.E., 2001. daf-16 integrates developmental and environmental inputs to mediate aging in the nematode *Caenorhabditis elegans*. *Current biology: CB*, 11(24), pp.1975–1980.

Hoque, M., Ji, Z., Zheng, D., Luo, W., Li, W., You, B., Park, J.Y., Yehia, G. & Tian, B., 2013. Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nature methods*, 10(2), pp.133–139.

Horner, J., 2012. *RMySQL*, Available: http://biostat.mc.vanderbilt.edu/wiki/Main/RMySQL [Accessed May 10, 2013].

Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L., McLaren, S., Sealy, I., Caccamo, M., Churcher, C., Scott, C., Barrett, J.C., Koch, R., Rauch, G.-J., White, S., Chow, W., et al., 2013. The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, 496(7446), pp.498–503.

Hu, D., Cao, P., Thiels, E., Chu, C.T., Wu, G., Oury, T.D. & Klann, E., 2007. Hippocampal Long-term Potentiation, Memory, and Longevity in Mice that Overexpress Mitochondrial Superoxide Dismutase. *Neurobiology of learning and memory*, 87(3), pp.372–384.

Huang, D.W., Sherman, B.T. & Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1), pp.44–57.

Imai, S., Armstrong, C.M., Kaeberlein, M. & Guarente, L., 2000. Transcriptional silencing and longevity protein Sir2 is an NAD-dependent histone deacetylase. *Nature*, 403(6771), pp.795–800.

Iseli, C., Jongeneel, C.V. & Bucher, P., 1999. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, pp.138–148.

Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D., Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M., Levy, M., Boudet, N., et al., 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*, 431(7011), pp.946–957.

Johnson, L.S., Eddy, S.R. & Portugaly, E., 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, 11(1), p.431.

Jones, F.C., Grabherr, M.G., Chan, Y.F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M.C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M.C., Myers, R.M., Miller, C.T., Summers, B.R., Knecht, A.K., Brady, S.D., et al., 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, 484(7392), pp.55–61.

Jubb, R.A., 1971. A new Nothobranchius (Pisces, Cyprinodontidae) from Southeastern Rhodesia. *Journal of the American Killifish Association*, 8(12), pp.314–321.

Kaeberlein, M. & Kennedy, B.K., 2009. Ageing: A midlife longevity drug? *Nature*, 460(7253), pp.331–332.

Kaeberlein, M., McVey, M. & Guarente, L., 1999. The SIR2/3/4 complex and SIR2 alone promote longevity in *Saccharomyces cerevisiae* by two different mechanisms. *Genes & development*, 13(19), pp.2570–2580.

Kaeberlein, M., Powers, R.W., Steffen, K.K., Westman, E.A., Hu, D., Dang, N., Kerr, E.O., Kirkland, K.T., Fields, S. & Kennedy, B.K., 2005. Regulation of Yeast Replicative Life Span by TOR and Sch9 in Response to Nutrients. *Science*, 310(5751), pp.1193–1196.

Kapahi, P., Zid, B.M., Harper, T., Koslover, D., Sapin, V. & Benzer, S., 2004. Regulation of Lifespan in Drosophila by Modulation of Genes in the TOR Signaling Pathway. *Current biology : CB*, 14(10), pp.885–890.

Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y., Jindo, T., Kobayashi, D., Shimada, A., Toyoda, A., Kuroki, Y., Fujiyama,

A., Sasaki, T., Shimizu, A., Asakawa, S., Shimizu, N., et al., 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature*, 447(7145), pp.714–719.

Kassahn, K.S., Dang, V.T., Wilkins, S.J., Perkins, A.C. & Ragan, M.A., 2009. Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. *Genome Research*, 19(8), pp.1404–1418.

Katz, Y., Wang, E.T., Airoldi, E.M. & Burge, C.B., 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature methods*, 7(12), pp.1009–1015.

Kircher, M., Heyn, P. & Kelso, J., 2011. Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genomics*, 12(1), p.382.

Kirschner, J., Weber, D., Neuschl, C., Franke, A., Böttger, M., Zielke, L., Powalsky, E., Groth, M., Shagin, D., Petzold, A., Hartmann, N., Englert, C., Brockmann, G.A., Platzer, M., Cellerino, A. & Reichwald, K., 2011. Mapping of quantitative trait loci controlling lifespan in the short-lived fish *Nothobranchius furzeri*- a new vertebrate model for age research. *Aging Cell*. Available: http://www.ncbi.nlm.nih.gov/pubmed/22221414 [Accessed January 24, 2012].

Kitts, P.A., Madden, T.L., Sicotte, H., Black, L. & Ostell, J.A., 2012. UniVec Database.

Koch, P., 2010. *Repetitive Elemente im Genom von Nothobranchius furzeri – Charakterisierung und Implikationen für die Genomsequenzierung*. Diploma thesis. Jena: Friedrich Schiller University.

Kodama, Y., Shumway, M., Leinonen, R. & on behalf of the International Nucleotide Sequence Database Collaboration, 2011. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Research*, 40(D1), pp.D54–D56.

Korf, I., Yandell, M. & Bedell, J., 2003. *BLAST*, Sebastopol, CA, USA: O'Reilly & Associates, Inc.

Kozomara, A. & Griffiths-Jones, S., 2011. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, 39(suppl 1), pp.D152–D157.

Krabbe, K.S., Pedersen, M. & Bruunsgaard, H., 2004. Inflammatory mediators in the elderly. *Experimental gerontology*, 39(5), pp.687–699.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3), p.R25.

Lee, B.-Y., Howe, A.E., Conte, M.A., D'Cotta, H., Pepey, E., Baroiller, J.-F., di Palma, F., Carleton, K.L. & Kocher, T.D., 2010. An EST resource for tilapia based on 17 normalized libraries and assembly of 116,899 sequence tags. *BMC Genomics*, 11, p.278.

Lee, C.-K., Klopp, R.G., Weindruch, R. & Prolla, T.A., 1999. Gene Expression Profile of Aging and Its Retardation by Caloric Restriction. *Science*, 285(5432), pp.1390–1393.

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14), pp.1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), pp.2078–2079.

Li, W. & Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, 22(13), pp.1658–1659.

Lin, S.J., Defossez, P.A. & Guarente, L., 2000. Requirement of NAD and SIR2 for life-span extension by calorie restriction in *Saccharomyces cerevisiae*. *Science (New York, N.Y.)*, 289(5487), pp.2126–2128.

Lindberg, J. & Lundeberg, J., 2010. The plasticity of the mammalian transcriptome. *Genomics*, 95(1), pp.1–6.

Lipman, R.D., Bronson, R.T., Wu, D., Smith, D.E., Prior, R., Cao, G., Han, S.N., Martin, K.R., Meydani, S.N. & Meydani, M., 1998. Disease incidence and longevity are unaltered by dietary

antioxidant supplementation initiated during middle age in C57BL/6 mice. *Mechanisms of ageing and development*, 103(3), pp.269–284.

Lockhart, D.J. & Winzeler, E.A., 2000. Genomics, gene expression and DNA arrays. *Nature*, 405(6788), pp.827–836.

Lu, J., Peatman, E., Wang, W., Yang, Q., Abernathy, J., Wang, S., Kucuktas, H. & Liu, Z., 2010. Alternative splicing in teleost fish genomes: same-species and cross-species analysis and comparisons. *Molecular genetics and genomics: MGG*, 283(6), pp.531–539.

Magalhães, J.P. de, Curado, J. & Church, G.M., 2009. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics*, 25(7), pp.875–881.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., et al., 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), pp.376–380.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. & Gilad, Y., 2008. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9), pp.1509–1517.

Markofsky, J. & Matias, J.R., 1977. The effects of temperature and season of collection on the onset and duration of diapause in embryos of the annual fish *Nothobranchius guentheri*. *The Journal of experimental zoology*, 202(1), pp.49–56.

Martin, J., Bruno, V.M., Fang, Z., Meng, X., Blow, M., Zhang, T., Sherlock, G., Snyder, M. & Wang, Z., 2010. Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics*, 11(1), p.663.

Martin, J.A. & Wang, Z., 2011. Next-generation transcriptome assembly. *Nature Reviews Genetics*, 12(10), pp.671–682.

Masoro, E.J., 2003. Subfield History: Caloric Restriction, Slowing Aging, and Extending Life. *Science of Aging Knowledge Environment*, 2003(8), p.re2.

Masoro, E.J. & Austad, S.N., 2011. *Handbook of the biology of aging*, Available: http://www.sciencedirect.com/science/book/9780123786388 [Accessed December 4, 2012].

McCay, C.M. & Crowell, M.F., 1934. Prolonging the Life Span. *The Scientific Monthly*, 39(5), pp.405–414.

Metzker, M.L., 2005. Emerging technologies in DNA sequencing. *Genome Research*, 15(12), pp.1767–1776.

Mignone, F., Gissi, C., Liuni, S. & Pesole, G., 2002. Untranslated regions of mRNAs. *Genome Biology*, 3(3), p.reviews0004.

Miquel, J., Economos, A.C., Fleming, J. & Johnson, J.E., Jr, 1980. Mitochondrial role in cell aging. *Experimental gerontology*, 15(6), pp.575–591.

Morin, R., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T., McDonald, H., Varhol, R., Jones, S. & Marra, M., 2008. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques*, 45(1), pp.81–94.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth*, 5(7), pp.621–628.

Myers, E.W., 1995. Toward simplifying and accurately formulating fragment assembly. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 2(2), pp.275–290.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. & Snyder, M., 2008. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, 320(5881), pp.1344–1349.

Nagano, T. & Fraser, P., 2011. No-Nonsense Functions for Long Noncoding RNAs. *Cell*, 145(2), pp.178–181.

Nam, D.K., Lee, S., Zhou, G., Cao, X., Wang, C., Clark, T., Chen, J., Rowley, J.D. & Wang, S.M., 2002. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9), pp.6152–6156.

National Center for Biotechnological Information, 2011. The non-redundant protein sequence database nr. Available: ftp://ftp.ncbi.nlm.nih.gov/blast/db/.

National Center for Biotechnology, 2004. ESTs Factsheet. Available: http://www.ncbi.nlm.nih.gov/About/primer/est.html.

Navarro, A., Gómez, C., Sánchez-Pino, M.-J., González, H., Bández, M.J., Boveris, A.D. & Boveris, A., 2005. Vitamin E at high doses improves survival, neurological performance, and brain mitochondrial function in aging male mice. *American journal of physiology. Regulatory, integrative and comparative physiology*, 289(5), pp.R1392–1399.

O'Brien, E.A., Zhang, Y., Wang, E., Marie, V., Badejoko, W., Lang, B.F. & Burger, G., 2009. GOBASE: an organelle genome database. *Nucleic acids research*, 37(Database issue), pp.D946–950.

Ohno, S., 1970. *Evolution by gene duplication.*, London; New York: Allen & Unwin; Springer-Verlag.

Okoniewski, M.J. & Miller, C.J., 2006. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics*, 7(1), p.276.

Ozsolak, F. & Milos, P.M., 2011. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, 12(2), pp.87–98.

Park, S.-K., Kim, K., Page, G.P., Allison, D.B., Weindruch, R. & Prolla, T.A., 2009. Gene expression profiling of aging in multiple mouse strains: identification of aging biomarkers and impact of dietary antioxidants. *Aging cell*, 8(4), pp.484–495.

Patel, R.K. & Jain, M., 2012. NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLoS ONE*, 7(2), p.e30619.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J. & Quackenbush, J., 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics (Oxford, England)*, 19(5), pp.651–652.

Petzold, A., Reichwald, K., Groth, M., Taudien, S., Hartmann, N., Priebe, S., Shagin, D., Englert, C. & 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(1), p.185.

Pevzner, P.A., Tang, H. & Waterman, M.S., 2001. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, 98(17), pp.9748–9753.

Pfaffl, M.W., Horgan, G.W. & Dempfle, L., 2002. Relative expression software tool (REST) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. *Nucleic Acids Research*, 30(9), p.e36.

Pontius, J., Wagner, L. & Schuler, G., 2003. UniGene: a unified view of the transcriptome. In *The NCBI handbook*. Bethesda (MD): National Center for Biotechnology Information. Available: http://www.ncbi.nlm.nih.gov/books/NBK21101/.

Pruitt, K.D., Tatusova, T. & Maglott, D.R., 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 33(Database issue), pp.D501–504.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., Bateman, A. &

Finn, R.D., 2011. The Pfam protein families database. *Nucleic Acids Research*, 40(D1), pp.D290–D301.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. & Glöckner, F.O., 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research*, 41(Database issue), pp.D590–596.

Ramsköld, D., Wang, E.T., Burge, C.B. & Sandberg, R., 2009. An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data. *PLoS Comput Biol*, 5(12), p.e1000598.

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

Rice, P., Longden, I. & Bleasby, A., 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics: TIG*, 16(6), pp.276–277.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D., Mungall, K., Lee, S., Okada, H.M., Qian, J.Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y.S., Newsome, R., Chan, S.K., She, R., Varhol, R., Kamoh, B., et al., 2010. De novo assembly and analysis of RNA-seq data. *Nat Meth*, 7(11), pp.909–912.

Rogina, B. & Helfand, S.L., 2004. Sir2 mediates longevity in the fly through a pathway related to calorie restriction. *Proceedings of the National Academy of Sciences of the United States of America*, 101(45), pp.15998–16003.

Salzberg, S.L. & Yorke, J.A., 2005. Beware of mis-assembled genomes. *Bioinformatics*, 21(24), pp.4320–4321.

Sanger, F., Nicklen, S. & Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp.5463–5467.

Schriner, S.E., Linford, N.J., Martin, G.M., Treuting, P., Ogburn, C.E., Emond, M., Coskun, P.E., Ladiges, W., Wolf, N., Van Remmen, H., Wallace, D.C. & Rabinovitch, P.S., 2005. Extension of murine life span by overexpression of catalase targeted to mitochondria. *Science (New York, N.Y.)*, 308(5730), pp.1909–1911.

Schumacher, B., van der Pluijm, I., Moorhouse, M.J., Kosteas, T., Robinson, A.R., Suh, Y., Breit, T.M., van Steeg, H., Niedernhofer, L.J., van Ijcken, W., Bartke, A., Spindler, S.R., Hoeijmakers, J.H.J., van der Horst, G.T.J. & Garinis, G.A., 2008. Delayed and accelerated aging share common longevity assurance mechanisms. *PLoS genetics*, 4(8), p.e1000161.

Simpson, J.T. & Durbin, R., 2010. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics (Oxford, England)*, 26(12), pp.i367–373.

Sinclair, D.A., 2005. Toward a unified theory of caloric restriction and longevity regulation. *Mechanisms of ageing and development*, 126(9), pp.987–1002.

Smit, A., Hubley, R. & Green, P., 1996. *RepeatMasker Open-3.0*, Available: http://www.repeatmasker.org/.

Soderlund, C., Johnson, E., Bomhoff, M. & Descour, A., 2009. PAVE: program for assembling and viewing ESTs. *BMC Genomics*, 10, p.400.

Spring, J., 1997. Vertebrate evolution by interspecific hybridisation – are we polyploid? *FEBS Letters*, 400(1), pp.2–8.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G.R., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C.J., Osborne, B.I., Pocock,

M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E., Wilkinson, M.D., et al., 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome research*, 12(10), pp.1611–1618.

Sunagawa, S., Wilson, E.C., Thaler, M., Smith, M.L., Caruso, C., Pringle, J.R., Weis, V.M., Medina, M. & Schwarz, J.A., 2009. Generation and analysis of transcriptomic resources for a model system on the rise: the sea anemone *Aiptasia pallida* and its dinoflagellate endosymbiont. *BMC Genomics*, 10, p.258.

Tatar, M., Kopelman, A., Epstein, D., Tu, M.P., Yin, C.M. & Garofalo, R.S., 2001. A mutant Drosophila insulin receptor homolog that extends life-span and impairs neuroendocrine function. *Science (New York, N.Y.)*, 292(5514), pp.107–110.

Taylor, J.S., Braasch, I., Frickey, T., Meyer, A. & Van de Peer, Y., 2003. Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Research*, 13(3), pp.382–390.

Terzibasi, E., Dorn, A., Ng Oma, E., Pola Ik, M., Bla Ek, 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 evolutionary biology*, 13(1), p.77.

Terzibasi, E., Lefrançois, C., Domenici, P., Hartmann, N., Graf, M. & Cellerino, A., 2009. Effects of dietary restriction on mortality and age-related phenotypes in the short-lived fish *Nothobranchius furzeri*. *Aging Cell*, 8(2), pp.88–99.

Terzibasi, E., Valenzano, D.R., Benedetti, M., Roncaglia, P., Cattaneo, A., Domenici, L. & Cellerino, A., 2008. Large differences in aging phenotype between strains of the short-lived annual fish *Nothobranchius furzeri*. *PloS One*, 3(12), p.e3866.

The C. elegans Sequencing Consortium, 1998. Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology. *Science*, 282(5396), pp.2012–2018.

TIGR, 2008. *SeqClean*, Available: http://compbio.dfci.harvard.edu/tgi/software/.

Tissenbaum, H.A. & Guarente, L., 2001. Increased dosage of a sir-2 gene extends lifespan in *Caenorhabditis elegans*. *Nature*, 410(6825), pp.227–230.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., Baren, M.J. van, Salzberg, S.L., Wold, B.J. & Pachter, L., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5), pp.511–515.

Ulitsky, I., Shkumatava, A., Jan, C.H., Subtelny, A.O., Koppstein, D., Bell, G.W., Sive, H. & Bartel, D.P., 2012. Extensive alternative polyadenylation during zebrafish development. *Genome Research*, 22(10), pp.2054–2066.

Valdesalici, S. & Cellerino, A., 2003. Extremely short lifespan in the annual fish *Nothobranchius furzeri*. *Proceedings. Biological Sciences / The Royal Society*, 270 Suppl 2, pp.S189–191.

Valenzano, D. & Cellerino, A., 2006. Resveratrol and the pharmacology of aging: a new vertebrate model to validate an old molecule. *Cell cycle (Georgetown, Tex.)*, 5(10), pp.1027–1032.

Valenzano, D., Kirschner, J., Kamber, R.A., Zhang, E., Weber, D., Cellerino, A., Englert, C., Platzer, M., Reichwald, K. & Brunet, A., 2009. Mapping loci associated with tail color and sex determination in the short-lived fish *Nothobranchius furzeri*. *Genetics*, 183(4), pp.1385–1395.

Valenzano, D., Terzibasi, E., Cattaneo, A., Domenici, L. & Cellerino, A., 2006a. Temperature affects longevity and age-related locomotor and cognitive decay in the short-lived fish *Nothobranchius furzeri*. *Aging Cell*, 5(3), pp.275–278.

Valenzano, D., Terzibasi, E., Genade, T., Cattaneo, A., Domenici, L. & Cellerino, A., 2006b. Resveratrol Prolongs Lifespan and Retards the Onset of Age-Related Markers in a Short-Lived Vertebrate. *Current Biology*, 16(3), pp.296–300.

Vaziri, H., Dessain, S.K., Eaton, E.N., Imai, S.-I., Frye, R.A., Pandita, T.K., Guarente, L. & Weinberg, R.A., 2001. hSIR2SIRT1 Functions as an NAD-Dependent p53 Deacetylase. *Cell*, 107(2), pp.149–159.

Vellai, T., Takacs-Vellai, K., Zhang, Y., Kovacs, A.L., Orosz, L. & Müller, F., 2003. Genetics: Influence of TOR kinase on lifespan in *C. elegans*. *Nature*, 426(6967), pp.620–620.

Vera, J.C., Wheat, C.W., Fescemyer, H.W., Frilander, M.J., Crawford, D.L., Hanski, I. & Marden, J.H., 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, 17(7), pp.1636–47.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R. & Birney, E., 2009. EnsemblCompara GeneTrees: Complete, Duplication-Aware Phylogenetic Trees in Vertebrates. *Genome Research*, 19(2), pp.327–335.

Volff, J.-N., 2006. *Vertebrate genomes*, Basel; New York: Karger.

Wang, Z., Gerstein, M. & Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1), pp.57–63.

Warnes, G., 2011. *gplots package*, Available: http://cran.r-project.org/web/packages/gplots/index.html.

Wasmuth, J.D. & Blaxter, M.L., 2004. prot4EST: translating expressed sequence tags from neglected genomes. *BMC Bioinformatics*, 5, p.187.

Whitehead, A. & Crawford, D.L., 2005. Variation in tissue-specific gene expression among natural populations. *Genome biology*, 6(2), p.R13.

Wickham, H., 2009. *ggplot2: elegant graphics for data analysis*, Springer New York. Available: http://had.co.nz/ggplot2/book.

Wildekamp, R.H., 1993. *The genera Adamas, Adinia, Aphanius, Aphyoplatys and Aphyosemion*,

Wilhelm, B.T. & Landry, J.-R., 2009. RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*, 48(3), pp.249–257.

Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J. & Bähler, J., 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453(7199), pp.1239–1243.

Wong, M.L. & Medrano, J.F., 2005. Real-time PCR for mRNA quantitation. *BioTechniques*, 39(1), pp.75–85.

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

Xu, Q., Modrek, B. & Lee, C., 2002. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Research*, 30(17), pp.3754–3766.

Yang, X., Chockalingam, S.P. & Aluru, S., 2013. A survey of error-correction methods for next-generation sequencing. *Briefings in Bioinformatics*, 14(1), pp.56–66.

Yoshida, S., Ishida, J.K., Kamal, N.M., Ali, A.M., Namba, S. & Shirasu, K., 2010. A full-length enriched cDNA library and expressed sequence tag analysis of the parasitic weed, *Striga hermonthica*. *BMC Plant Biology*, 10(1), p.55.

Yu, C.E., Oshima, J., Fu, Y.H., Wijsman, E.M., Hisama, F., Alisch, R., Matthews, S., Nakura, J., Miki, T., Ouais, S., Martin, G.M., Mulligan, J. & Schellenberg, G.D., 1996. Positional cloning of the Werner's syndrome gene. *Science (New York, N.Y.)*, 272(5259), pp.258–262.

Yulug, I.G., Yulug, A. & Fisher, E.M., 1995. The frequency and position of Alu repeats in cDNAs, as determined by database searching. *Genomics*, 27(3), pp.544–548.

Zerbino, D.R. & Birney, E., 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome research*, 18(5), pp.821–829.

Zerbino, D.R., McEwen, G.K., Margulies, E.H. & Birney, E., 2009. Pebble and Rock Band: Heuristic Resolution of Repeats and Scaffolding in the Velvet Short-Read *de Novo* Assembler. *PLoS ONE*, 4(12), p.e8407.

Supplementary Tables

Supplementary Table 1: *N. furzeri* cDNA libraries prepared for transcriptome sequencing.

Library	Specimens used for RNA-isolation				Library preparation		
	Strain	Age (weeks)	Tissue	Fish / Sex	cDNA construction method	Normalisation	Prepared by
1	GRZ	9	Whole body	One male individual	Oligo(dT) primer	Yes	Evrogen,
2	GRZ	10	Whole body	One male individual	Oligo(dT) primer	Yes	Evrogen
3	GRZ	10	Whole body	One male individual	Oligo(dT) primer	No	Evrogen
4	GRZ	8	Whole body	Six male and female individuals	Random hexamer primers	Yes	Vertis
5	GRZ	1	Whole body	Three individuals of unknown sex	"	No	FLI
6	GRZ	5	Skin	Five male individuals	"	"	"
7	GRZ	5	Brain	Five male individuals	"	"	"
8	GRZ	14	Skin	Three male individuals	"	"	"
9	MZM-0403	14	Brain	Three male individuals	"	"	"
10	MZM-0403	5	Skin	Four male individuals	"	"	"
11	MZM-0403	5	Brain	Five male individuals	"	"	"
12	MZM-0403	31	Skin	Five male individuals	"	"	"
13	MZM-0403	31	Brain	Three male individuals	"	"	"

Supplementary Table 2: Programs, parameter and filter criteria used for BLAST annotation.

Database	Program	Parameter	Max. e-value	Min. overlap	Min. identity (%)
Ensembl fish proteins	BLASTx	wordmask=seg lcmask W=4 T=20	10^{-07}	30 aa	40
Ensembl fish transcripts	tBLASTx	wordmask=seg lcmask W=4 T=20	10^{-07}	100 bp	40 (on amino acid level)
UniProt	BLASTx	wordmask=seg lcmask W=4 T=20 hitdist=40	10^{-07}	30 aa	40
NCBI nr proteins	BLASTx	wordmask=seg lcmask W=4 T=20 hitdist=40	10^{-07}	30 aa	40
Refseq human proteins	BLASTx	wordmask=seg lcmask W=4 T=20	10^{-07}	30 aa	40
NCBI UniGene transcripts	BLASTn	M=1 N=-1 Q=3 R=2 W=9 wordmask=seg lcmask	10^{-20}	100 bp	75

BLAST options are based on recommendations from Korf et al. 2003.

Supplementary Table 3: *N. furzeri* ageing-related genes in skin.

Gene	Annotation	Fold change with age	Adj. p-value
AND2	Actinodin2	0.012247584	0.002029267
ADAMTSL2	Adamts-Like 2	0.08382646	3.76E-05
COL10A1	Collagen, Type X, Alpha 1	0.102937036	6.23E-07
HSP90AA1	Heat Shock Protein 90Kda Alpha (Cytosolic), Class A Member 1	0.110707245	2.21E-05
PYROXD2	Pyridine Nucleotide-Disulphide Oxidoreductase Domain 2	0.112218693	0.001136298
SGMS2	Sphingomyelin Synthase 2	0.117150579	4.60E-05
IGF2BP3	Insulin-Like Growth Factor 2 Mrna Binding Protein 3	0.134382444	0.00094006
TMEM119A	Transmembrane Protein 119A	0.140915716	0.00023659
AND1	Actinodin 1	0.143427451	0.008443258
FKBP9	Fk506 Binding Protein 9, 63 Kda	0.147235125	7.04E-05
GPX7	Glutathione Peroxidase 7	0.148329924	0.000185381
ENPEP	Glutamyl Aminopeptidase	0.161094687	0.002029267
PHGDH	Phosphoglycerate Dehydrogenase	0.164231298	0.001017207
FMOD	Fibromodulin	0.165056018	0.008465704
ASPN	Asporin	0.165087169	0.002029267
APOEA	Apolipoprotein E	0.175146154	0.005061209
SMPD3	Sphingomyelin Phosphodiesterase 3, Neutral Membrane (Neutral Sphingomyelinase Ii)	0.176841241	0.001486907
CYTL1	Cytokine-Like 1	0.182018695	0.00094006
CKAP4	Cytoskeleton-Associated Protein 4	0.182353971	0.00094006
IFITM5	Interferon Induced Transmembrane Protein 5	0.186727994	0.003879688
RRBP1	Ribosome Binding Protein 1 Homolog 180Kda	0.191539958	0.001486907
CREB3L1	Camp Responsive Element Binding Protein 3-Like 1	0.195713314	0.008153735
CX43	Gap Junction Protein, Alpha 1, 43Kda	0.199205798	0.001486907
TPO	Thyroid Peroxidase	0.1999015	0.004117939
BAMBIA	Bmp And Activin Membrane-Bound Inhibitor Homolog	0.200413763	0.001915442
LOXL4	Lysyl Oxidase-Like 4	0.203815982	0.004053163

COL11A1A	Collagen, Type Xi, Alpha 1	0.205054701	0.002511042
RCN3	Reticulocalbin 3, Ef-Hand Calcium Binding Domain	0.208484999	0.002029267
SP7	Sp7 Transcription Factor	0.209215809	0.003755911
LEPREL4	Leprecan-Like 4	0.211038397	0.006475011
PECR	Peroxisomal Trans-2-Enoyl-Coa Reductase	0.219936613	0.006475011
PLEKHH1	Pleckstrin Homology Domain Containing, Family H (With Myth4 Domain) Member 1	0.219998218	0.006382469
PRODHA	Proline Dehydrogenase (Oxidase) 1	4.457928504	0.007135947
VDRB	Vitamin D (1,25- Dihydroxyvitamin D3) Receptor	4.865498711	0.006475011
PCMTD1	Protein-L-Isoaspartate (D-Aspartate) O-Methyltransferase Domain Containing 1	5.274725969	0.004178069
C19H1ORF51	Chromosome 1 Open Reading Frame 51	5.593602335	0.008621025
MKNK2A	Map Kinase Interacting Serine/Threonine Kinase 2	7.160644756	5.38E-05
CABZ01039845	Uncharacterized Protein	8.030231074	2.95E-05
BM1_33690	Cell Wall Protein Dan4, Putative	8.233943189	0.005061209
CYP1B1	Cytochrome P450, Family 1, Subfamily B, Polypeptide 1	9.949835083	0.00094006
MFAP4	Microfibrillar-Associated Protein 4	13.49906985	2.96E-07
V1G57470	Uncharacterized Protein	13.69815852	0.006475011
C5H9ORF25	Chromosome 9 Open Reading Frame 25	21.5846481	0.004981189

Gene expression levels of young (GRZ, MZM-0403: 5 weeks) and old N. furzeri (GRZ: 14 weeks, MZM-0403: 31 weeks) in skin were tested for significant differences (p-value<=0.01). Fold changes below and above 1 indicate genes that are downregulated (coloured red) and upregulated (blue), respectively, with age. P-values are adjusted for multiple testing (Benjamini-Hochberg).

Supplementary Table 4: *N. furzeri* ageing-related genes in brain.

Gene	Annotation	Fold change with age	Adj. p-value
CYP1A	Cytochrome P450 1A	0.07630841	8.77E-05
SLC25A22	Solute Carrier Family 25 (Mitochondrial Carrier, Glutamate), Member 22	0.176896407	4.26E-08
DBX1A	Developing Brain Homeobox 1	0.210969982	0.009501616
CCNB1	Cyclin B1	0.213035249	0.005688501
ANLN	Anillin, Actin Binding Protein	0.251717751	0.00053191
CKAP2L	Cytoskeleton Associated Protein 2-Like	0.251962853	0.003703754
RACGAP1	Rac Gtpase Activating Protein 1	0.252429271	0.004424137
MCM5	Mcm5 Minichromosome Maintenance Deficient 5	0.255938925	0.004770879
MCM4	Minichromosome Maintenance Complex Component 4	0.260244037	0.000665594
CDCA7	Cell Division Cycle Associated 7	0.260801976	0.000676268
RRM2	Ribonucleotide Reductase M2	0.281686743	4.73E-05
PLK1	Polo-Like Kinase 1	0.285462218	0.002462939
ISYNA1	Inositol-3-Phosphate Synthase 1	0.296648058	0.002157374
TUBB5	Tubulin, Beta 5	0.307853482	2.85E-05
DLGAP5	Discs, Large Homolog-Associated Protein 5	0.312772559	0.005688501
MCM6	Minichromosome Maintenance Complex Component 6	0.319305429	0.004770879
BA1	Ba1 Globin	0.335343403	0.000512919
H2AFV	H2A Histone Family, Member V	0.378519563	0.004684264
H3F3B	H3 Histone, Family 3B (H3,3B)	0.380353373	0.004979366
DEDD2	Death Effector Domain Containing 2	2.800589334	0.004770879
SLC16A9A	Solute Carrier Family 16, Member 9 (Monocarboxylic Acid Transporter 9)	3.270483407	0.000676268

PTPRC	Protein Tyrosine Phosphatase, Receptor Type, C	3.352443552	0.004979366
CEBPB	Ccaat/Enhancer Binding Protein (C/Ebp), Beta	3.395174241	0.002836483
MPEG1	Macrophage Expressed 1	3.465945292	0.002229189
APCS	Amyloid P Component, Serum	3.524989277	0.001553384
VDRB	Vitamin D (1,25- Dihydroxyvitamin D3) Receptor	3.621157708	0.00021162
ARRDC3	Arrestin Domain Containing 3	3.654158099	0.000717741
AHNAK	Ahnak Nucleoprotein	3.665358687	0.000619388
RNF213	Ring Finger Protein 213	3.904100317	0.002157374
BX470254	Uncharacterized Protein	4.035880514	0.00010831
C21H5ORF41	Chromosome 5 Open Reading Frame 41	4.039721157	3.49E-05
ITGAM	Integrin, Alpha M	4.143194462	0.00096585
ZNFX1	Zinc Finger, Nfx1-Type Containing 1	4.328657989	0.004979366
ARRDC2	Arrestin Domain Containing 2	4.466725228	1.29E-06
IRF1	Interferon Regulatory Factor 1	5.229031373	3.83E-06
Q8UUL6_ORYLA	Mhc Class I A	5.498792629	2.39E-11
CABZ01039845	Uncharacterized Protein	6.261570956	1.47E-11
CTSS	Cathepsin S	6.471474368	4.33E-09
NCF1	Neutrophil Cytosolic Factor 1	6.640774427	0.000443124
FKBP5	Fk506 Binding Protein 5	6.931212649	1.67E-07
GPNMB	Glycoprotein (Transmembrane) Nmb	6.964355619	0.004158807
PRL	Prolactin	9.405446397	1.27E-13
GPR84	G Protein-Coupled Receptor 84	11.27671459	4.01E-06

Gene expression levels of young (GRZ, MZM-0403: 5 weeks) and old *N. furzeri* (GRZ: 14 weeks, MZM-0403: 31 weeks) in brain were tested for significant differences (p-value<=0.01). Fold changes below and above 1 indicate genes that are downregulated (coloured red) and upregulated (blue), respectively, with age. P-values are adjusted for multiple testing (Benjamini-Hochberg).

Supplementary Table 5: Primers used for the qPCR validation of selected ageing-related *N. furzeri* genes.

Gene	Annotation	Forward primer	Reverse primer
INSR	Insulin receptor	TGCCTCTTCAAACCCTGAGT	AGGATGGCGATCTTATCACG
ADAMTSL2	ADAMTS-like 2	GCAGGCCTTGCTGTAGTACC	AAACCGGTGTCCAAACAGAC
APOE	Apolipoprotein E	GCATAAGGACACCCAGGAGA	GGAGCAGGTCATTCAGGGTA
DBX1A	Developing brain homeobox 1a	CATCAGCAAGCCAGACAGAA	GACATCCACCGGATGACAG
CCNB1	Cyclin B1	CCGTCACATAGGCAAAGTCC	CTGCTGCAGGAGACCATGTA
PLK1	Polo-like kinase 1	TGTGTTTGTGGTCCTGGAGA	TTGCCCAGTTTCAGGTCTCT
PRODHA	Proline dehydrogenase (oxidase) 1a	GTGGATGCAGAGCAGACGTA	CCAAAATACCAGCCTTCTCG
VDRB	Vitamin D receptor b	CATGCAGACTCAAACGCTGT	CGTGCTTCCTTTTCTGCTTC
CYP1B1	Cytochrome P450, family 1, subfamily B, polypeptide 1	CGGACATATTTGGAGCCAGT	AGCTGTTGCTGGTCTTCGAT
MPEG1	Macrophage expressed gene 1	CAGAAAAGCACCACAGCTCA	GGCCTTCGCTGTGTACATAAA
APCS	Amyloid P component, serum	GAAGCTGGTCTGGTCCATGT	TCTGGAGCAAACATCACAGG
COL10A1	Collagen, type X, alpha 1	CCACTGGAAAGGGGTATGTG	GGCAGACCAATTCCATTCTC

Supplementary Files

Supplementary File 1 - Diagram of the `EST2uni` MySQL database.
Schematic of the `EST2uni` MySQL database with modifications highlighted in red

Supplementary File 2 - Genes duplicated in *N. furzeri*
List of duplicated genes identified in the *N. furzeri* transcript catalogue

Supplementary File 3 - *N. furzeri* ageing-related genes in skin and brain
Lists of differentially expressed genes in skin and brain of *N. furzeri* detected by RNA-seq

Supplementary File 4 - Enriched functional annotation terms obtained for *N. furzeri* ageing-related genes
Enriched functional annotation terms obtained by DAVID for up- and downregulated genes as well as for skin- and brain-specific genes that were differentially expressed in *N. furzeri* with age

Supplementary File 5 – *N. furzeri* ageing-related genes in skin and their confirmation in zebrafish
Lists of *N. furzeri* differentially expressed genes in skin and their confirmation skin of old vs. young zebrafish

The Supplementary Files as well as electronic versions of this thesis in Word and PDF format are available on the enclosed CD.

Curriculum Vitae

Personal Data

Name:	Andreas Petzold
Date of birth:	01.01.1982
Place of birth:	Merseburg
Nationality:	German
Address:	Sorbenweg 8, 06217 Merseburg
Email:	andpet@gmx.de

Education and Qualification

since 2009	PhD student, Genome Analysis Group, Leibniz Institute for Age Research - Fritz Lipmann Institute, Jena, Germany Topic: "Transcriptome analysis of the short-lived fish *Nothobranchius furzeri*"
09/2007 - 12/2008	Bioinformatician, Leibniz Institute for Age Research - Fritz Lipmann Institute, Jena, Germany: "Management and analysis of next-generation sequencing data"
02/2007 - 08/2007	Research assistant, Leibniz Institute for Age Research - Fritz Lipmann Institute, Jena, Germany
10/2001 - 01/2007	Study of bioinformatics, Friedrich Schiller University, Jena, Germany Title of the diploma thesis: "MOGA: Mutationsorientierte Genomanalyse am Beispiel von *E. coli* LW1655F+"
07/2000	Abitur (University entrance qualification), Johann - Gottfried Herder Gymnasium, Merseburg, Germany

Publications

Mayer KFX, Taudien S, Martis M, Simková H, Suchánková P, Gundlach H, Wicker T, Petzold A, Felder M, Steuernagel B, Scholz U, Graner A, Platzer M, Dolezel J, Stein N: **Gene content and virtual gene order of barley chromosome 1H**. *Plant Physiol* 2009, **151**:496–505.

Steuernagel B, Taudien S, Gundlach H, Seidel M, Ariyadasa R, Schulte D, Petzold A, Felder M, Graner A, Scholz U, Mayer KFX, Platzer M, Stein N: **De novo 454 sequencing of barcoded BAC**

pools for comprehensive gene survey and genome analysis in the complex genome of barley. *BMC Genomics* 2009, **10**:547.

Kenanov D, Kaleta C, Petzold A, Hoischen C, Diekmann S, Siddiqui RA, Schuster S: **Theoretical study of lipid biosynthesis in wild-type *Escherichia coli* and in a protoplast-type L-form using elementary flux mode analysis**. *FEBS J* 2010, **277**:1023–1034.

Krügel H, Licht A, Biedermann G, Petzold A, Lassak J, Hupfer Y, Schlott B, Hertweck C, Platzer M, Brantl S, Saluz H-P: **Cervimycin C resistance in *Bacillus subtilis* is due to a promoter up-mutation and increased mRNA stability of the constitutive ABC-transporter gene bmrA**. *FEMS Microbiol Lett* 2010, **313**:155–163.

Taudien S, Groth M, Huse K, Petzold A, Szafranski K, Hampe J, Rosenstiel P, Schreiber S, Platzer M: **Haplotyping and copy number estimation of the highly polymorphic human beta-defensin locus on 8p23 by 454 amplicon sequencing**. *BMC Genomics* 2010, **11**:252.

Burmester A, Shelest E, Glöckner G, Heddergott C, Schindler S, Staib P, Heidel A, Felder M, Petzold A, Szafranski K, Feuermann M, Pedruzzi I, Priebe S, Groth M, Winkler R, Li W, Kniemeyer O, Schroeckh V, Hertweck C, Hube B, White TC, Platzer M, Guthke R, Heitman J, Wöstemeyer J, Zipfel PF, Monod M, Brakhage AA: **Comparative and functional genomics provide insights into the pathogenicity of dermatophytic fungi**. *Genome Biol* 2011, **12**:R7.

Kirschner J, Weber D, Neuschl C, Franke A, Böttger M, Zielke L, Powalsky E, Groth M, Shagin D, Petzold A, Hartmann N, Englert C, Brockmann GA, Platzer M, Cellerino A, Reichwald K: **Mapping of quantitative trait loci controlling lifespan in the short-lived fish *Nothobranchius furzeri*- a new vertebrate model for age research**. *Aging Cell* 2011.

Taudien S, Steuernagel B, Ariyadasa R, Schulte D, Schmutzer T, Groth M, Felder M, Petzold A, Scholz U, Mayer KF, Stein N, Platzer M: **Sequencing of BAC pools by different next generation sequencing platforms and strategies**. *BMC Res Notes* 2011, **4**:411.

Taudien S, Szafranski K, Felder M, Groth M, Huse K, Raffaelli F, Petzold A, Zhang X, Rosenstiel P, Hampe J, Schreiber S, Platzer M: **Comprehensive assessment of sequence variation within the copy number variable defensin cluster on 8p23 by target enriched in-depth 454 sequencing**. *BMC Genomics* 2011, **12**:243.

Felder M, Romualdi A, Petzold A, Platzer M, Sühnel J, Glöckner G: **GenColors-based comparative genome databases for small eukaryotic genomes**. *Nucleic Acids Res* 2012.

Fluch S, Kopecky D, Burg K, Šimková H, Taudien S, Petzold A, Kubaláková M, Platzer M, Berenyi M, Krainer S, Doležel J, Lelley T: **Sequence composition and gene content of the short arm of rye (*Secale cereale*) chromosome 1**. *PLoS ONE* 2012, **7**:e30784.

Kirschner J, Weber D, Neuschl C, Franke A, Böttger M, Zielke L, Powalsky E, Groth M, Shagin D, Petzold A, Hartmann N, Englert C, Brockmann GA, Platzer M, Cellerino A, Reichwald K: **Mapping of quantitative trait loci controlling lifespan in the short-lived fish *Nothobranchius furzeri*--a new vertebrate model for age research**. *Aging Cell* 2012, **11**:252–261.

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

Terzibasi E, Dorn A, Ng Oma E, Pola Ik M, Bla Ek R, Reichwald K, Petzold A, Watters B, Reichard M, Cellerino A: **Parallel evolution of senescence in annual fishes in response to extrinsic mortality**. *BMC Evol Biol* 2013, **13**:77.

Scientific Talks

- "Transcriptome analysis of the short-lived *Nothobranchius furzeri* using next-generation sequencing technologies",
  Jena Centre for Bioinformatics Workshop, Jena, Germany, November 2009

Posters

- „Initial de novo transcriptome and genome assemblies of *Nothobranchius furzeri* – a new model for ageing research",
  Genome Informatics, Hinxton, UK, September 2010

- "*De novo* assembly and annotation of the *Nothobranchius furzeri* transcriptome – a new model for ageing research",
  Jena Centre for Bioinformatics, Jena, Germany, March 2011

- "The transcript catalogue of the short-lived fish *Nothobranchius furzeri* provides insights into age-dependent changes of mRNA levels",
  Meeting on Advances and Challenges of RNA-Seq Analysis, Halle, Germany, June 2012
  Conference on Systems Biology of Mammalian Cells, Leipzig, Germany, July 2012
  German Conference on Bioinformatics, Jena, Germany, September 2012

Acknowledgements

I am grateful to PD Dr. Matthias Platzer, for giving me the opportunity to work on this interesting project and for his continuous support and advice. Especially, I would like to thank Dr. Kathrin Reichwald, for her excellent mentoring, guidance and support. In addition, I would like to thank both for their help during the formation process and the critical reading of my dissertation.

I would like to thank all former and current colleagues of the Genome Analysis Group (also known as the Platzer Lab) for the very nice and constructive working environment, especially: Eileen Powalszky, Kathleen Seitz, Ivonne Heinze, Ivonne Görlich, Klaus Huse, Marco Groth, Stefan Taudien, Patricia Möckel, Beate Szafranski, Silke Förste, Marcel Kramer, Karol Szafranski, Marius Felder, Andrew Heidel, Roman Siddiqui, Niels Jahn, Bernd Senf, Gernot Glöckner, Rileen Sinha, Andrew Heidel, Bryan Downie, Philipp Koch, Martin Benz, Cornelia Luge, Nadine Zeise and Ulrike Gausmann.

Moreover, I would like to thank my thesis committee members: Matthias Platzer, Jürgen Sühnel and Reinhard Guthke from the Hans-Knöll Institut in Jena.

Furthermore, I would like to thank all the member of the Nothobranchius Consortium on the FLI for their on-going efforts regarding all kinds of *N. furzeri*-related topics, in particular: Nils Hartmann, Alesandro Cellerino, Mario Baumgart, Sabine Matz, Enoch N´goma, Christin Hahn, and many more.

Lastly, I should thank many individuals, friends and colleagues who have not been mentioned here personally in making this educational process a success. Maybe I could not have made it without your supports.

Finally, I want to thank my parents, my brother and my grandparents. You have both encouraged and supported me to finish this thesis.

Selbstständigkeitserklärung

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.

Insbesondere waren folgende Personen in genannter Art und Weise direkt an der Entstehung der vorliegenden Arbeit beteiligt:

- Dr. Nils Hartmann (FLI Jena) isolierte die RNA aus den Proben und führte die qPCR Experimente durch.
- Dr. Dmitry Shagin erstellte die Banken für die Sanger- und 454/Roche-Sequenzierung.
- Dr. Kathrin Reichwald, Dr. Marco Groth und Dr. Stefan Taudien (FLI Jena) führten die Sanger, 454/Roche und Solexa/Illumina Sequenzierungen durch.
- Dr. Steffen Priebe (HKI Jena) führte die initiale Analyse der Zebrafischdaten durch.

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.

Jena, 30.08.2013 ……..……...

Andreas Petzold