Molecular analysis of the *Oryzias latipes* (Medaka) transcriptome

Dissertation zur Erlangung des akademischen Grades des

Doktors der Naturwissenschaften (Dr. rer. nat.)

eingereicht im Fachbereich Biologie, Chemie, Pharmazie

der Freien Universität Berlin

vorgelegt von

Anja Berger

aus Schwedt/Oder

2010

Dissertation am MPI für Molekulare Genetik in Berlin,

Abteilung Prof. Dr. Lehrach, Arbeitsgruppe PD Dr. Himmelbauer,

23.04.2001-30.05.2005

1. Gutachter: PD Dr. H. Himmelbauer

2. Gutachter: Prof. Dr. R. Mutzel

Disputation am 24.09.2010

Contents

Abstract

Based on oligonucleotide fingerprinting (OFP) analysis and subsequent EST production a non-redundant set of 10,016 medaka cDNA clones was established from three different embryonic stages (gastrula, neurula and organogenesis) and one adult tissue (ovary) as a resource of high value for further research on the medaka transcriptome.

In a first round 26,880 medaka gastrula clones were subjected to OFP cluster analysis and representatives of each cluster or clones left as singletons were chosen for producing ESTs. In total 7680 cDNA clones were sequenced and 6909 high-quality 5'reads were obtained. The advantage of OFP lies not only in the normalisation but it is also possible to get insight into differential expression by subjecting cDNA libraries of different developmental stages or tissues to fingerprinting analysis. Therefore in a second round in addition to the gastrula clones, cDNA inserts from libraries of the ovary tissue and neurula and organogenesis stages were included. From this approach another 11,468 high-quality 5'ESTs were produced. All EST sequence data was published in GenBank EST database with the accession numbers from AM137442 to AM156757. The 18,377 high-quality sequences obtained were, by EST clustering, grouped into 3268 clusters and 7274 singletons providing us with 10542 unique sequences. Further clustering reduced this set to 10,016 unique sequences. High-quality EST clusters and singletons were annotated. To 8155 of these sequences functions were assigned, with many sequences showing similarity to proteins with important functions, e.g. in development. EST data which showed no similarity to any other known proteins includes by a large amount valuable and high-quality sequence information and must therefore be seen as new Medaka sequence data, either protein-coding or non-coding.

Zusammenfassung

Mit Hilfe der Oligonukleotid Fingerprintinganalyse (OFP) und anschließender EST Sequenzierung wurde ein nicht redundanter Satz von 10.016 cDNA Klonen von drei verschiedenen embryonalen Stadien (Gastrula, Neurula und Organogenese) und einem adulten Gewebe (Ovar) erstellt. Dieser Datensatz stellt eine wertvolle Resource für die weitere Arbeit am Medaka Transkriptom dar. Während erster Experimente wurden 26.880 Medaka Gastrula Klone einer OFP Clusteranalyse unterzogen und Repräsentanten resultierender OFP Cluster und Singletons wurden für die Produktion von ESTs ausgewählt. Insgesamt wurden 7.680 cDNA Klone sequenziert und daraus entstanden 6.909 qualitativ hochwertige 5' Sequenzen. Der Vorteil der OFP Analyse liegt aber nicht nur in der Normalisierung von cDNA Bibliotheken, sondern diese Methode bietet auch die Möglichkeit einen Einblick in differentielle Expression zu erhalten, wenn cDNA Bibliotheken verschiedener Entwicklungsstadien oder Gewebe verwendet werden. Deshalb wurden in weiteren Experimenten zusätzlich zu Gastrulaklonen auch cDNA Klone von drei anderen Bibliotheken, Ovar, Neurula und Organogenese, in die Analyse einbezogen. Davon wurden dann weitere 11.468 5' ESTs produziert. Die EST-Sequenzen wurden in der GenBank EST Datenbank unter den Accession Numbers AM137442 bis AM156757 publiziert. Die 18.377 EST Sequenzen hoher Qualität wurden durch Clusteranalyse in 3.268 Cluster und 7.274 Singletons gruppiert, die 10.542 verschiedene Sequenzen darstellen. Durch weitere Clusteranalyse wurde dieser Datensatz auf 10.016 Sequenzen reduziert. Diese Sequenzen wurden annotiert und für 8.155 dieser Sequenzen konnte eine Funktion zugeordnet werden, wobei viele Sequenzen Ähnlichkeit zu wichtigen Proteinen aufwiesen, z.B. zu Proteinen mit Funktion in der Embryonalentwicklung. EST Daten, denen keine Funktion zugewiesen werden konnte, bestehen zu einem großen Teil aus wertvoller, hoch-qualitativer Sequenzinformation und können somit als neue, proteinkodierende oder nicht-kodierende, Medaka Sequenzinformation gesehen werden.

Chapter 1

Introduction

1.1 Fish World - Introducing teleosts as model organisms

Among vertebrates fishes comprise the most versatile group as around 28,900 fish species are described [Froese and Pauly, 2005], which have evolved in just 450 [Kumar and Hedges, 1998] million years. Fishes exhibit the highest speciation rates known so far in vertebrates, and show therefore amazing differences in their morphology and ecological and behavioural adaptations [Venkatesh, 2003]. To get insight into evolutionary aspects of such a rapid radiation African cichlids have proven to be useful [Stiassny and Meyer, 1999], which have in Lake Victoria, to give an example, evolved from 2 founder species into at least 500 endemic species in just 100,000 years [Verheyen et al., 2003]. This is an outstanding example of fish evolution, where many hypotheses on evolution and ecology can be tested, but these scenarios require probably special explanations (for instance the special anatomy of Cichlidae, having two sets of jaws which can change their morphology even within a lifetime of an individual [Meyer, 1990], and therefore may enable the occupation of different ecological niches and so promoting speciation) not adoptable to other evolutionary scenarios leading to fish species radiation [Salzburger and Meyer, 2004]. Not only this extreme example of fish evolution makes clear that research on the genetic diversification in fishes is essential for ecological, economical and scientific reasons. Therefore further analyses on molecular data in different fish species will be essential that may provide detailed information on genes, their sequences, and regulation. Numerous experiments were performed in the past years on various fish species: experiments on evolutionary aspects using the already mentioned East African cichlids, research on tumorigenesis often uses platyfish and swordtails (Cypridontiformes, mostly *Xiphophorus*; [Schartl, 1995]), studies of ecology and behaviour are done on sticklebacks [McKinnon et al., 2004], for investigating sex determination *Xiphophorus* and medaka [Kondo et al., 2002, Schartl, 2004] are used, for genetical and developmental studies zebrafish and medaka are widely used (see below) and whole-genome experiments on pufferfish genomes are also further described in subsequent sections. Meanwhile several genome projects on fishes like fugu [Aparicio et al., 2002], zebrafish

[Wilming et al., 2008], Medaka [Kasahara et al., 2007] and stickleback [Kingsley et al., 2004] are finished. This chapter will illustrate experiments done on three fish systems, *Danio rerio* (zebrafish), *Oryzias latipes* (Japanese ricefish, medaka) and the two pufferfishes, *Tetraodon fluviatilis* (Freshwater pufferfish) and *Takifugu rubripes* (Japanese pufferfish, fugu) to highlight the endless opportunities provided for research on fishes.

Bony fishes build their own evolutionary group that split from land vertebrates around 450 Myr ago [Kumar and Hedges, 1998]. Teleosts are in evolutionary terms the highest evolved group within the bony fishes. Medaka and pufferfishes belong to the superorder Acanthopterygii, the most numerous and a relatively evolutionarily advanced group within the Teleostei. The zebrafish (belonging to the order Cypriniformes), in contrast, belongs to the superorder Ostariophysi (contains also carp and goldfish), a relatively old group within the Teleostei [Ishikawa, 2000]. The evolutionary distance between medaka and fugu is estimated to be around 60 Myr, which is as close as the fast-evolving *Drosophila* species, *D. melanogaster* and *D. hydei*, whereas zebrafish and medaka separated already around 110-160 million years [Chen et al., 2004] ago. Further results show that medaka (Beloniformes) and cichlids (Perciformes) appear to be more closely related to each other than either of them is to pufferfish (Tetraodontiformes) [Chen et al., 2004], from which the following phylogenetic data was taken (see fig. 1.1). Newer data stated that the last common ancestor of medaka and zebrafish lived more than 314-332 Myr ago and Medaka and fugu are 184-198 Myr apart [Yamanoue et al., 2006]. However, uncertainties concerning the divergence time of the different teleost lineages remain.

1.1.1 Defining a model organism

Biological research has in many cases led to the establishment of model organisms, which facilitate experimental laboratory research by large extent. However they represent only a small fraction of the biodiversity found on Earth. In past decades, the term model organism has been narrowly applied to species such as mouse or *Drosophila* that are characterised by their small size, short generation time and easy maintenance and breeding in the laboratory. In the past decade this definition has broadened and does now also include model systems which are useful for a special area of research like pufferfishes for comparative genomics [Hedges, 2002]. Within that context the teleosts zebrafish and medaka can be viewed as classical vertebrate model organisms in modern biology [Ishikawa, 2000], combining the power of genetics with experimental embryology and molecular biology [Wittbrodt et al., 2002].

Zebrafish was introduced into genetical research by George Streisinger who started around 1970 [Stahl, 1985] to develop a vertebrate model system with lots of useful tools he was used to apply for his research on the T4-bacteriophage. According to [Laale, 1977] the zebrafish was already used in several biological studies from the nineteen thirties onwards because of its short generation time and its easy breeding and maintenance and he describes in his review studies on the embryonic and larval development, physiological functions, behavioural patterns, taxonomy and systematics,

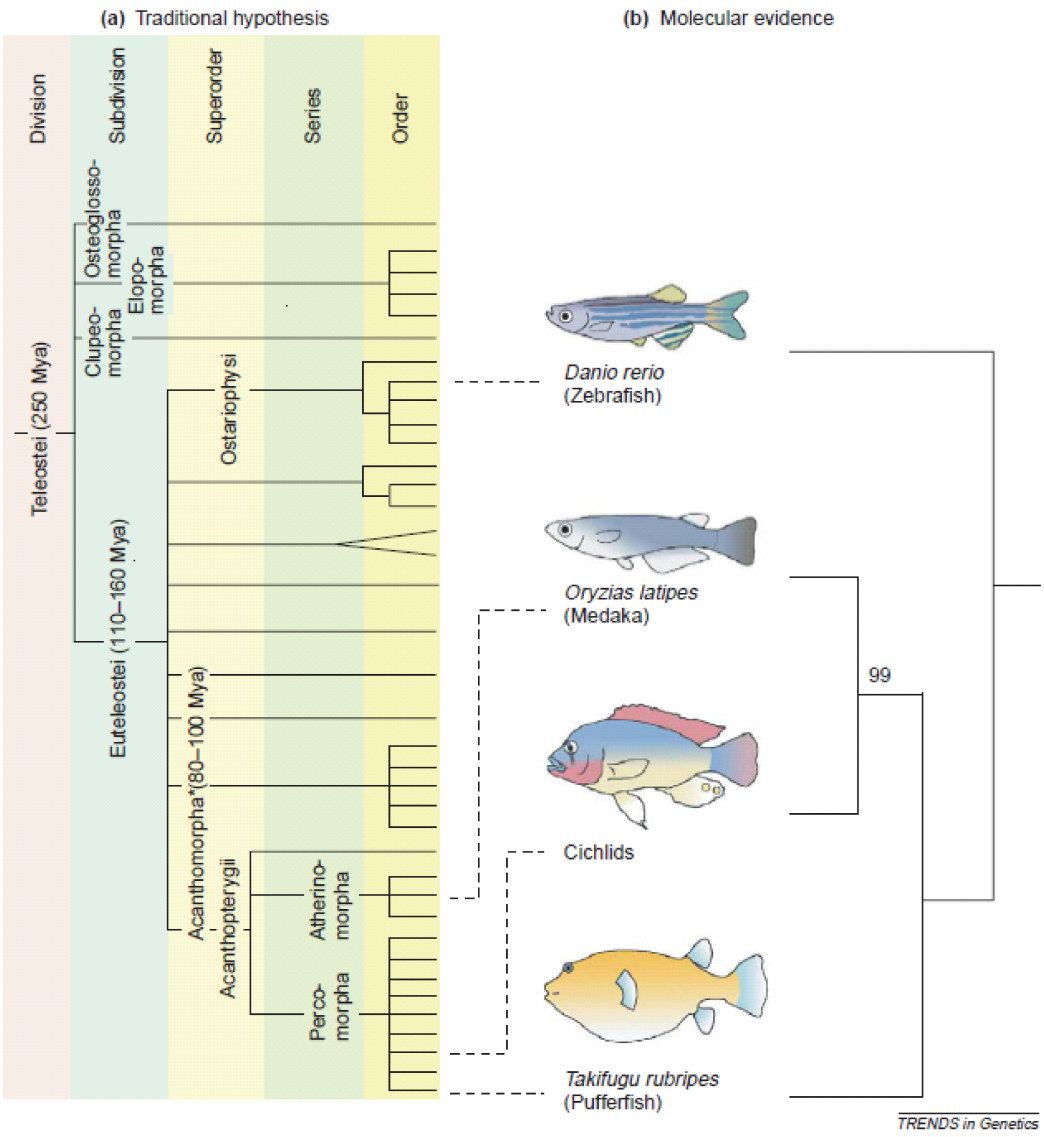


Figure 1.1: Phylogenetic trees depicting the evolutionary relationships between four fish models. (a) Traditional hypothesis. Cichlids and pufferfish are closer related to each other than either of them is to Medaka. (b) Molecular (new) data. Cichlids are closer related to Medaka than to pufferfish; taken from [Chen et al., 2004].

for example developmental stages of zebrafish were already described in the nineteen fifties in [Hisaoka and Battle, 1958, Hisaoka and Firlit, 1960].

Medaka research can look back on a still longer history as a biological model, as introduced into Japanese laboratories already at the beginning of the last century, but it has been bred for several centuries for amusement reasons. This fish was for the first time scientifically described by Philipp Franz von Siebold, a physician and naturalist, who recorded his findings on Japanese fauna in a book edited by Siebold and published in Leiden, the Netherlands [Shima and Mitani, 2004, Temminck and Schlegel, 1846].

In 1993 a different model system was suggested by [Brenner et al., 1993] in using the small genome of *Takifugu rubripes* as a genome model. Experimental work on pufferfishes is very difficult, but the 365 Mb genome provides a good opportunity for exploring the basic genetic content of all vertebrates, because of its similarities to other vertebrate genomes but its reduced length of non-coding DNA like introns or repetitive sequences [Brenner et al., 1993]. The sequence of two pufferfish genomes has already been published, in 2002 the fugu genome [Aparicio et al., 2002] and 2004 the genome of *Tetraodon* [Jaillon et al., 2004].

Using fish as a model system is a challenging task which appeared with the identification of several duplicated fish genes as compared to the gene numbers of higher vertebrates, especially mammals [Aparicio et al., 1997, Amores et al., 1998, Wittbrodt et al., 1998]. To evaluate the evolutionary scenarios leading to these findings and also to avoid upcoming difficulties arising, different fish model systems have to be compared [Furutani-Seiki and Wittbrodt, 2004].

1.1.2 *Danio rerio* (zebrafish) - the popular model

Species belonging to the genus *Danio* cover Burma, Thailand, Indochina, the Malay Peninsula, Sumatra and Borneo, only *D. rerio* is found in India and Pakistan. Genetic analyses of the zebrafish wild population have been rather rarely accomplished, for example by [Wright et al., 2003, McCune et al., 2004]. The Zebrafish Information Network (ZFIN, http://zfin.org) lists in July 2004 18 different domesticated strains, but there are still no inbred strains available, which may have complicated experiments like mapping experiments, but which were circumvented by approaches like gynogenesis in genetic mapping [Streisinger et al., 1981], also used to carry out genetic mapping and complementation analyses [Streisinger et al., 1986].

The biology of zebrafish provides researchers with several advantageous features. Zebrafish is an oviparous species and therefore artificial fertilisation is fairly easy to achieve. Zebrafish generation time is similar to that of the medaka, although embryonic development is faster compared to medaka, with embryos hatching between 2 to 3 days after fertilisation at 28 °C. Chorions of zebrafish embryos are soft and transparable, and blastomeres at early cleavage stages are larger than those of the medaka. Therefore, the manipulation of embryos is easier ([Fishman et al., 1997]). One drawback in genomic research is the big genome size of 1700 Mb divided in 25 chromosomes (haploid chromosome set). Zebrafish research on sex determination is also limited, as none of the many

Figure 1.2: Zebrafish. Image from Max Planck Institute for Developmental Biology.

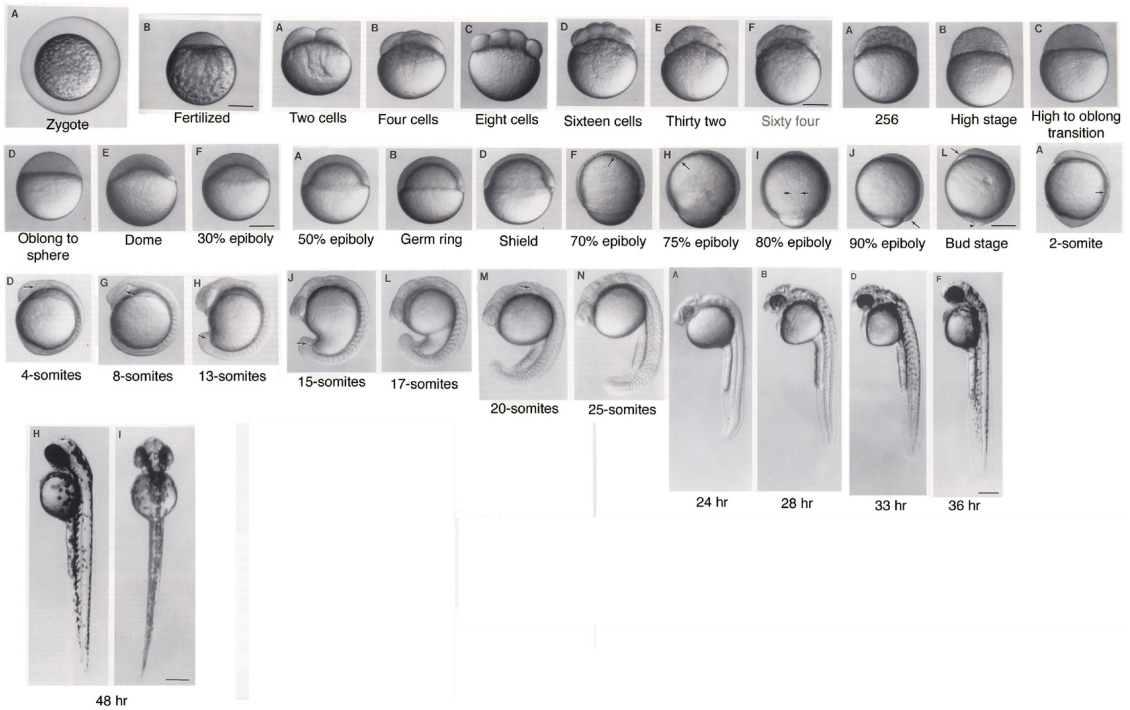


Figure 1.3: Embryonic development of the zebrafish. Picture taken from [Kimmel et al., 1995].

markers of the zebrafish genetic map have been found to be sex linked [Wittbrodt et al., 2002].

After establishing several genetical tools, research on developmental aspects was promoted, one milestone being the establishment of the zebrafish fate map published in [Kimmel et al., 1990]. The big breakthrough of research on developmental biology of the zebrafish was initiated in 1993 with two big mutagenesis screens and concluded two years later with identifying around 6,647 developmental mutant phenotypes [Haffter et al., 1996, Driever et al., 1996]. One limitation of large-scale mutant screening in fish was gene redundancy as some classes or groups of mutations might not have been found due to overlapping functions of multiple genes in the same developmental process [Haffter et al., 1996, Driever et al., 1996]. Possible solutions to such problems are stated in subsection 1.1.5.

For identification of genes which became mutated, genome resources like linkage maps were needed in zebrafish, which wholly depended, because of the lack of inbred strains, on Streisingers

gynogenesis methods, as segregation of polymorphic DNA markers was assessed among sibling haploid progeny of Darjeeling and Oregon AB strains [Postlethwait et al., 1994, Johnson et al., 1996] or by the means of bulk segregation analysis [Postlethwait and Talbot, 1997] as an initial mapping approach. Also in February of 2001 the Wellcome Trust Sanger Institute started sequencing the zebrafish genome, which is still being improved. In July 2007 the 7th assembly was released, where 1.4 gigabases of sequence were assembled and 400 megabases remaining to be sequenced. The last genebuild from this assembly (July 2008) contains 17,330 known protein-coding genes, 1,194 projected protein-coding genes, 2,798 novel protein-coding genes and 35,653 gene transcripts, described at Ensembl. Already a Zv8 pre-assembly is available from December 2008 combining physical and genetic maps further, because until then 10% of the draft genome was misplaced compared to meiotic maps [Ekker et al., 2007]. It is planned to get a complete, finished sequence by the end of 2009 [Ekker et al., 2007].

1.1.3 Tetraodontiformes (pufferfishes) - compact genomic models

Among fishes there are several 'small genome' fish including pufferfish, sticklebacks, seahorses, flatfishes, blue-striped grunt and Siamese algae eater [Arai et al., 1988, Hinegardner and Rosen, 1972]. For *Tetraodon fluviatilis* (Freshwater pufferfish) a haploid gene content of 0.38 pg was estimated, which corresponds to a genome size of approximately 380 Mb [Hinegardner and Rosen, 1972] and [Hinegardner, 1976], about one eighth of the human genome. These measures show that *Tetraodon* possesses the smallest genome among vertebrates [Grützner et al., 1999, Fischer et al., 2000], which also applies similarly to other pufferfishes. Therefore another member of tetraodontoid fishes, *Takifugu rubripes* (Japanese pufferfish), was proposed in 1993 as an ideal vertebrate model, because it shows despite its small genome still sufficient homologies to the human genome [Brenner et al., 1993] and this way it became the second vertebrate genome to be sequenced completely [Aparicio et al., 2002]. The pufferfishes have to be viewed as model genomes, but not as model organisms, as experimentation on them is difficult, e.g. Fugu shows a complicated life cycle as it becomes sexually mature in three years, and it is difficult to maintain using traditional methods. The Japanese pufferfish (*Takifugu rubripes*) and Freshwater pufferfish (*Tetraodon nigroviridis*) belong to the order Tetraodontiformes. The natural habitat of *Takifugu* spans the Sea of Japan, the East China Sea, and the Yellow Sea. The fish *Tetraodon nigroviridis* lives in the rivers and estuaries of Indonesia, Malaysia and India.

According to the draft version of the fugu genome, generated using a whole genome shotgun approach by the International Fugu Sequencing Consortium (assembly version 3 has been available since August 2002), the size of the fugu genome was estimated with 365 Mb [Aparicio et al., 2002] and 31,059 gene loci were predicted. To finally produce a complete and finished fugu genome sequence a physical map integrating sequenced scaffolds is currently under construction (MRC HGMP-RC, fugu genomics project site at http://fugu.hgmp.mrc.ac.uk). The compactness of the fugu genome is explained with its smaller introns (75% of introns are less than 425 bp in length,

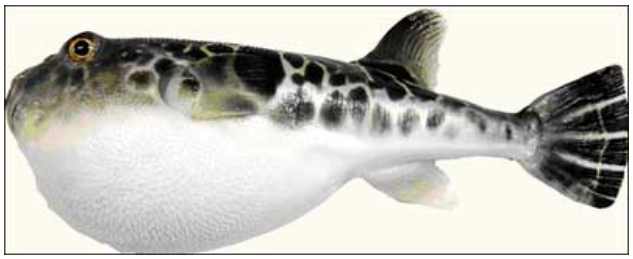


Figure 1.4: *Takifugu rubripes*. Photos by http://www.st-andrews.ac.uk.

compared to human with 75% of introns smaller than 2609 bp) and intergenic regions, as well as its small amount of repetitive sequence (less than one sixth of the genome) [Aparicio et al., 2002], which enlarges the proportion of coding material to a large extent compared to other vertebrate genomes. Besides that the fugu genome shows remarkable homologies to the human sequence, as intron-exon structures are conserved and 75% of the predicted human proteins have a strong match to fugu. Interestingly, both gain and loss of fugu introns have been shown compared with the intron-exon structure of human genes, even 571 gene loci were found that were 1.3 times or greater in size than their human homologues [Aparicio et al., 2002]. There are also examples where introns are of similar size in fugu and mammals, as described for *Hoxb-4* in [Aparicio et al., 1995].

The characteristics of the pufferfish genomes make them also valuable tools for comparative genome analysis. Experiments were undertaken using human-fugu comparisons, like the prediction of almost 1000 novel genes in the human genome [Aparicio et al., 2002] or to solve evolutionary tasks [Vandepoele et al., 2004]. The large similarities between te fugu and human genome were also used for comparative mapping and positional cloning of human disease genes as described in [Trower et al., 1996]. But such approaches are difficult because despite similar homologous genome regions considerable differences in the gene order were described for examples for human chromosome 12 [Montpetit et al., 2003] and for human chromosome 20 [Smith et al., 2002] in comparison to their fugu counterparts. There lie also great possibilities in the comparison of medaka and fugu genome as their evolutionary distance is estimated to be around 60 Myr, which is as close as the fast-evolving *Drosophila* species, *D. melanogaster* and *D. hydei*, that have been successfully used to establish conserved genomic features in Diptera [Marquart et al., 1999].

Tetraodon nigroviridis, the Freshwater pufferfish, is like the Japanese pufferfish not used as an experimental system. The primary value is also its exceptionally small genome size, which is a tremendous advantage in genome analysis. The *Tetraodon* genome project was started at Genescope (Centre national de séquençage) in 1997. The general strategy follows a Whole Genome Shotgun (WGS) approach, where single read sequences are generated from the ends of cloned genomic DNA fragments. In June 2001, The Whitehead Institute for Genomic Research (WIGR) joined Genoscope in this project, to accelerate the sequencing of the Tetraodon genome. In 2004 a draft sequence of the *Tetraodon* was published [Jaillon et al., 2004], covering 342 Mb (including gaps), still smaller than the Fugu genome (around 365 Mb).

Figure 1.5: *Tetraodon nigroviridis*. Photos by Dr. R. Sprackland.
http://www.curator.org/LegacyVMNH/WebOfLife/Kingdom/P_Chordata/ClassOsteichthyes/Class Osteichthyes/figure_8_puffer.htm

1.1.4 *Oryzias latipes* (Medaka) - small fish with big eyes

The first comprehensive record of experimental work applying medaka was published in 1975 by T. Yamamoto and his colleagues [Yamamoto, 1975], but early experiments already showed the suitability of Medaka as a model organism, like the demonstration of sex chromosome-dependent sex determination [Aida, 1921], the first description of early development in [Kamito, 1928], which was recently refined in [Iwamatsu, 2004], and the analysis of Mendelian characters in Medaka [Toyama, 1916, Goodrich, 1926], proving that Mendelian laws also apply to fish. Genetic studies in Medaka have in early years focused on pigmentation and sex determination. Sex differentiation and sex determination in fishes show an amazing variety of mechanisms, as reviewed in [Schartl, 2004] and therefore fishes have been an interesting object of research. Medaka was the first animal in which complete hormone-induced reversal of sex differentiation in both directions was successfully achieved [Yamamoto, 1958]. In contrast to Medaka the mechanisms of sex determination in zebrafish are still unravelled. The same applies for *Tetraodon nigroviridis* and *Takifugu rubripes*, in which no results on the mechanisms of sex determination were obtained or sex-linked markers were found [Li et al., 2002].

Big advance of genetical work in Medaka was made by the discovery of genetic polymorphisms between the Northern and Southern Japanese populations obtained by different approaches (allozymic analysis described in [Sakaizumi et al., 1983] and using DNA fingerprints with arbitrarily primed PCR [Kubota et al., 1992, Kubota et al., 1995, Shimada and Shima, 1998]). These polymorphisms founded the first genome-wide genetic linkage map [Wada et al., 1995] of the Medaka genome, which was improved later [Naruse et al., 2004]. Further analyses of genetic differences in

Figure 1.6: Medaka fish. Image taken from http://www1.kyoto-be.ne.jp/ed-center/gakko/zyoho/16leader/higasi/img/medaka.jpg.

Medaka showed that wild populations of *Oryzias latipes* are divided into four groups: Southern Japanese, Northern Japanese, East Korean and China/West Korean population [Sakaizumi and Jeon, 1987, Takehana et al., 2004]. The Northern Japanese population, which is distributed along the coast of the Sea of Japan, is genetically homogenous, as very few genetic variations of proteins and mitochondrial sequences have been observed [Sakaizumi et al., 1983, Takehana et al., 2003].

The Southern Japanese population is distributed along the Pacific coast and is genetically quite variable and therefore within that population region-specific genetic variations in both proteins and mitochondrial sequences were detected [Sakaizumi et al., 1983, Takehana et al., 2003]. Northern and Southern populations are closest related among the four ricefish populations. Nucleotide divergence of the mitochondrial *cytochrome b* sequence is about 10% between the Northern and Southern populations [Takehana et al., 2003] and experiments applying various orthologous sequences showed that single base-pair polymorphisms between Northern and Southern populations occur at a frequency of 3% in introns and 1% in exons [Wittbrodt et al., 2002]. The East Korean population extends from the Sea of Japan side of Korea to the southern part of the Korean peninsula [Sakaizumi and Jeon, 1987]. The China/West Korean population is widely distributed, ranging from Yunnan to the western coast of the Korean peninsula. Korean strains still show more differences to Japanese populations [Sakaizumi and Jeon, 1987, Takehana et al., 2004]. These large genetic differences can be advantageous in genome mapping studies. Despite the differences of wild populations in many morphological, behavioural and genetical characteristics, intercrosses breed normally producing hybrid offspring. Several inbred strains were derived from three different wild populations [Hyodo-Taguchi and Egami, 1985]. Information on origin and features of all strains are summarised in [Naruse et al., 2004] and table 1.1. Research on medakafish is not only necessary for scientific, but also for ecological reasons. Wild populations of medaka in Japan have been reduced by loss of habitat, including irrigation canals, swamps, marshes, and ponds [Hawkins et al., 2001]. Therefore in 1999 the Environmental Agency of Japan listed the Japanese ricefish as an endangered species.

Strain	Genetic background	reference	experiments done or in progress / available resources
HO4C	Southern population	[Hyodo-Taguchi, 1996]	
HO5	Southern population	[Hyodo-Taguchi, 1996]	
HB32C	Southern population	[Hyodo-Taguchi, 1996]	
HB32D	Southern population	[Hyodo-Taguchi, 1996]	
HB12A	Southern population	[Hyodo-Taguchi, 1996]	
HB11A	Southern population	[Hyodo-Taguchi, 1996]	
HB11C	Southern population	[Hyodo-Taguchi, 1996]	
Hd-rR	Southern population	[Hyodo-Taguchi, 1996]	[Matsuda et al., 2001] BAC library; [Zadeh Khorasani et al., 2004] physical map; (http://dolphin.lab.nig.ac.jp/medaka) genomic sequencing
Hd-rr	Southern population	[Hyodo-Taguchi, 1996]	
d-rR	Southern population	[Yamamoto, 1958]	first demonstration of sex reversal in fish
Cab	Southern population	[Loosli et al., 2001]	[Wittbrodt et al., 2002] BAC library; [Zadeh Khorasani et al., 2004] physical map; [Furutani-Seiki et al., 2004] mutation screens
AA2	Southern population	[Shimada and Shima, 1998]	
HNI-I	Northern population	[Hyodo-Taguchi, 1996]	genetic map (HNI vs. AA2)
HNI-II	Northern population	[Hyodo-Taguchi, 1996]	[Kondo et al., 2002] BAC library
Kaga	Northern population	[Loosli et al., 2001]	
HSOK	East-Korean population	[Hyodo-Taguchi, 1996]	

Table 1.1: Important experiments and resources available for *Oryzias latipes*.

Like zebrafish the medakafish possesses several biological characteristics, which make that fish a perfect model organism for developmental experiments and several tools were developed to successfully exploit that biological system. Adult ricefish are about 3 cm long, which they achieve within 2-3 months after hatching under proper conditions [Shima and Mitani, 2004], weigh 0.3 g and as eurythermal fish can tolerate a wide range of temperatures (4-40°C) and are less susceptible to disease compared to other fish models. Since they inhabit stagnant waters, aeration and thermostability are not necessary when breeding. These characteristics make it quite easy to maintain medaka and it is possible to breed zebrafish and medaka in one aquatic system. The male medaka can be easily distinguished from the female by its external morphology (dimorphic anal fins). Spawning is under strict control of light (14 h light and 10 h dark), temperature and food, but when applying proper conditions the female lays a cluster of eggs (10 to 40) every day. Embryos stay inside the tough chorion until they hatch as young fish 7 days after fertilisation at 25 °C. The embryonic development of medaka embryos (fig. 1.7) can be easily visualised through the transparent egg membrane under a dissection microscope. Using testis as material, the precise chromosome number in *O. latipes* was established by Iriki in 1932, who concluded that the haploid set of chromosomes was 24 [Iriki, 1932] which was confirmed in [Katayama, 1937]. The genome size is estimated to be around 650-1000 Mb [Hinegardner and Rosen, 1972, Lamatsch et al., 2000].

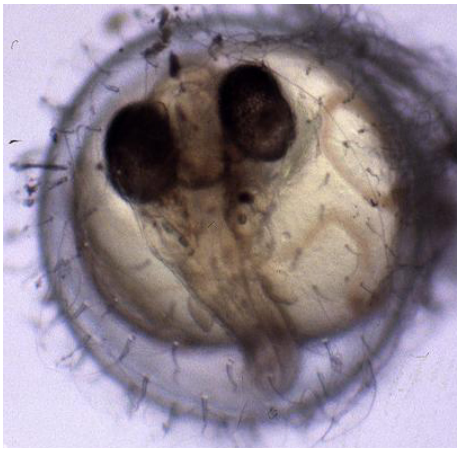
Both fish systems, zebrafish and ricefish, show many similarities in their development. They are characterised by a similar short generation time (medaka between 6 and 8 weeks, zebrafish between 8 and 10 weeks) and female fishes lay approximately same amounts of eggs per week (medaka spawns 10 to 40 eggs every day and zebrafish lays about one hundred eggs once in 1-2 weeks). In early development a different timing is visible comparing both fish systems: medaka brain morphogenesis and most of organogenesis occur early relative to somitogenesis, in contrast to zebrafish. There are more differences in early development as medaka eggs and embryos are surrounded by a tough chorion and hatch only as young fish, whereas zebrafish hatch already as swimming larvae and embryos are only surrounded by a thin chorion. Therefore all technical manipulations concerning the medaka embryo start with the removal of the chorion, but which is easily achieved with hatching enzyme [Ishida, 1944a, Ishida, 1944b], and so most standard experimental procedures can be applied to zebrafish and medaka [Furutani-Seiki and Wittbrodt, 2004], which include observation of embryos, gynogenesis, sperm freezing, *in vitro* fertilisation, cell transplantation, RNA and DNA injection, *in situ* hybridisation, production of transgenic fish by various methods, e.g. by injecting DNA into germinal-stage oocytes [Ozato et al., 1986] or by cytoplasmic injection into one-cell stage embryos [Winkler et al., 1994] and for various reasons (ectopic expression of genes [Grabher and Wittbrodt, 2004], disruption of genes by random transgene insertion, insertion of reporter genes [Grabher et al., 2003], use of GFP reporter constructs and morpholino based knock down experiments), summarised in [Naruse et al., 1994, Ishikawa, 2000, Wittbrodt et al., 2002, Shima et al., 2003, Furutani-Seiki and Wittbrodt, 2004].

The development of mutagenesis protocols for Medaka (Shima and Shimada, 1991) led to the

(a) Two-cell stage embryo: lateral view.

(b) Mid gastrula embryo: blastoderm covers 50% of yolk.

(c) Late neurula: two divisions of the brain are evident.



(d) Organogenesis.

Figure 1.7: Medaka fish embryonic development. Pictures taken from http://biog-101-104.bio.cornell.edu/BioG101_104/tutorials.

first systematic mutagenesis screens for developmental phenotypes [Loosli et al., 2000]. By the same method adopted by [Haffter et al., 1996] and [Driever et al., 1996] mutant screenings were successfully carried out for medaka using ENU- or X-ray-mutagenesis followed by three-generation crossings and many mutants showing specific patterning defects were identified. One surprising result were mutants, which were not found in large-scale zebrafish screenings, for example the *Oot* (*One-sided optic tectum*) mutant in medaka shows defects in mirror symmetry or bilateral symmetry only in the developing optic tectum and both its hemilobes exhibit the same morphology on both sides, but none of the phenotypes, in which the bilateral symmetry is broken and one-sided morphology is duplicated on both sides, were found in the 6647 zebrafish mutants described by

[Haffter et al., 1996] or [Driever et al., 1996]. To describe patterns obtained from *in situ* hybridisations or mutagenesis experiments a catalogue of medaka anatomy and development was established [Quiring et al., 2004] and is provided publicly via the OBO database (Open Biological Ontologies; http://obo.sourceforge.net).

This summary of tools available for medaka make it obvious that the key technologies that have made the zebrafish such a successful model species are fully applicable to the medaka. Additionally, genomics in the medaka offers several advantages, such as the availability of divergent, perfectly inbred strains and a genome of only 800 million bases [Hinegardner and Rosen, 1972, Lamatsch et al., 2000], half the size of the zebrafish genome. In recent years large scale EST and gene mapping projects were initiated in Japan and the genome sequencing project was started in mid 2002 with the Hd-rR strain, established from a Southern Japanese population, using the whole genome shotgun approach (http://dolphin.lab.nig.ac.jp/medaka) at the National Institute of Genetics and the University of Tokyo and Riken Institute. A first version was made publicly available in summer 2004 (http://medaka.utgenome.org), providing a genome coverage of 91% to 99% and improved in summer 2005, assembling around 763 Mb. At the same time the construction of a physical map of medaka based on bacterial artificial chromosome (BAC) clones was published [Zadeh Khorasani et al., 2004]. In addition to providing templates for the genomic sequencing of difficult regions, BACs can be used for the positional cloning of genes that cause particular mutant phenotypes and also for the rescue of mutants once candidate genes have been mapped to them. In 2007 the draft sequence of the Medaka genome was published [Kasahara et al., 2007], where 700 megabases of sequence were assembled and 20,141 genes were predicted, using 5'-end SAGE analysis.

1.1.5 Redundant fish model systems?

Teleosts were chosen as vertebrate models because of their relative ease of applying forward genetics and their relatively small genome size. Lots of efforts were involved to establish several fish model systems in parallel. Still in recent research it became increasingly clear, especially after publication of the fugu genome [Aparicio et al., 2002], that only a combination of research on different fish systems will provide us with the answers to all the problems which arise from the high amount of gene duplication in fishes. The high amount of duplicated genes clearly complicates molecular work on fish models, but it does also raise lots of interesting questions as to which were the evolutionary scenarios that took place to produce the fish genomes, which may have provided the foundations for the high adaptation rates of fishes to their environment.

There are three different explanations [Van de Peer et al., 2003] for the origin of duplicated fish genes. Duplicates may have been produced during a fish-specific whole-genome duplication around 350 million years ago (as proposed by [Amores et al., 1998] and [Wittbrodt et al., 1998]) and subsequent degeneration of some genes. The second explanation is an increased rate of independent gene duplications in fish [Robinson-Rechavi et al., 2001a, Robinson-Rechavi et al., 2001b]. A third

possibility is that after gene duplication events in the common ancestor of fish and tetrapods, the latter lost more genes [Van de Peer et al., 2003]. All described processes will lead to the present status of fish genomes, where we find different sets of genes duplicated in different fish species, in comparison to higher vertebrates like mammals. These scenarios can be theoretically traced back by calculating phylogenetic trees which should provide a different tree for different processes (see fig. 1.8). But this approach is hampered by the secondary loss of genes or whole chromosomes, additional independent duplications and different evolutionary rates of duplicate gene copies [Venkatesh, 2003]. Also mapping approaches as an alternative are difficult because of possible chromosomal rearrangements [Venkatesh, 2003].

By comparing human and fugu genomes [Vandepoele et al., 2004] it was found that most paralogous genes of fugu are the result of three complete genome duplications. They analysed more than 150 block duplications in the fugu genome, which in their opinion clearly supported a fish-specific genome duplication around 320 million years ago that coincided with the vast radiation of most modern ray-finned fishes. Similar results were obtained by [Christoffels et al., 2004], which estimated the timepoint of fish-specific genome duplication at 350 Myr, or with the comparison of the two pufferfish genomes by [Van de Peer, 2004]. Using early-branching Actinopterygii the timepoint of fish-specific genome duplication was dated to between 335 and 404 million years, taking place between the split of the Semionotiformes (*Lepisosteus platyrhynchus*) from the fish stem lineage and the origin of the Osteoglossomorpha [Hoegg et al., 2004], consistent with the finding that bichirs, a representative of the most basal extant ray-finned fish lineage, have only a single HoxA cluster [Chiu, 2004]. Comparative fish gene mapping provided additional evidence for a whole genome duplication preceding the teleost radiation, like [Postlethwait et al., 1998] localized genes in zebrafish and then compared the arrangement of zebrafish genes with their mammalian orthologues, where paralogous genes were mapped onto separate chromosomes in zebrafish. Paleontological evidence suggest that modern teleosts first appeared around 220 million years ago and underwent rapid diversification during the Jurassic and Cretaceous periods (205 to 135 Myr) [Maissey, 1996], thus, the whole-genome duplication appears to have occurred before the origin of modern teleosts.

In contrast to the above results others [Robinson-Rechavi et al., 2001a] published results were they identified several duplicated genes, but stated that most of these were the products of lineage-specific duplication events and not of an ancient duplication event. But this work was largely critisised in [Taylor et al., 2001], where it was shown that too few genes were used, which makes it difficult to interpret the data reliably. In addition seven gene families where found, showing a phylogenetic pattern which indicates their probable ancient duplication within a whole-genome duplication event. A second criticism noted comes from the use of tetraploid fishes, which have recently duplicated their genomes [Phillips and Rab, 2001], which introduces bias into the analysis of time point of duplication events. In the same publication [Taylor et al., 2001] and also later [Van de Peer et al., 2002] the authors showed that results in [Robinson-Rechavi et al., 2001a,

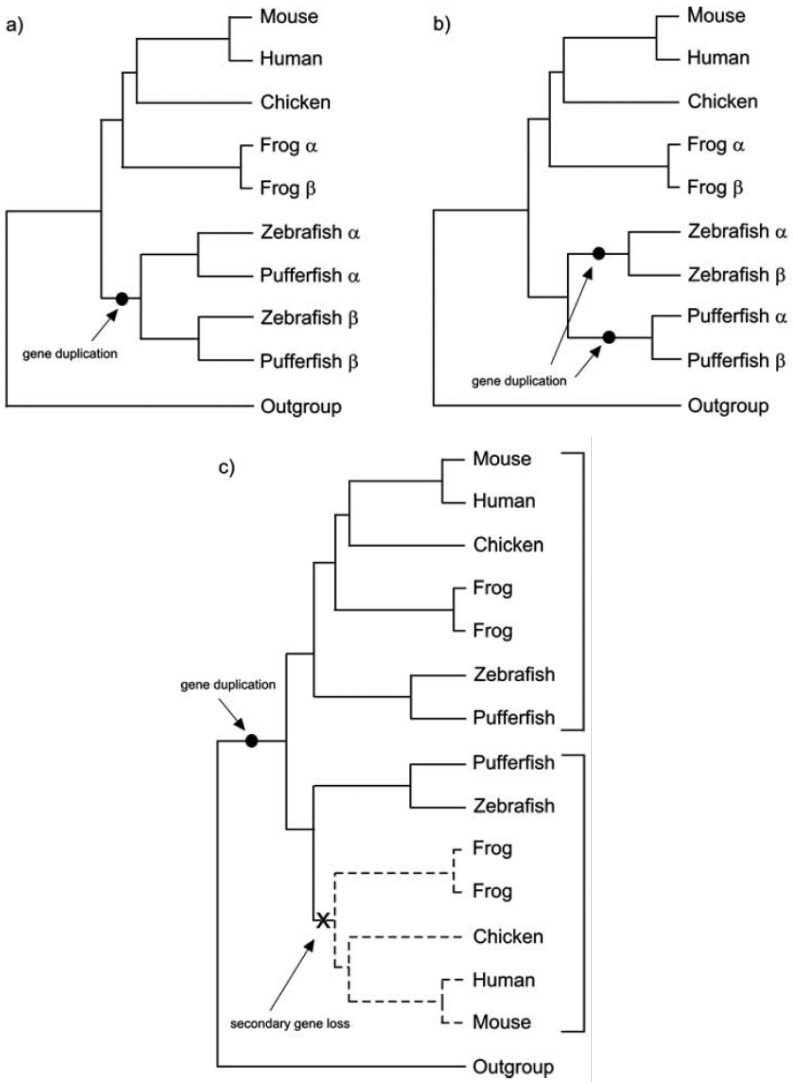


Figure 1.8: Different scenarios and expected inferred tree topologies to explain the presence of more genes in fish. (a) Duplicated fish genes are the result of a gene/genome duplication that preceded the divergence of different fish species. (b) Duplicated genes are formed by independent gene duplications. (c) This topology is expected to be inferred when genes produced during a duplication event in the ancestor of Actinopterygii and Sarcopterygii have been secondarily lost in the sarcopterygian lineage after the split of these two lineages. Taken from [Van de Peer et al., 2003].

Robinson-Rechavi et al., 2001b] were derived from erroneous phylogenetic trees and can be corrected when mutual saturation is taken into account. Mutual saturation in DNA and therefore in protein sequences occurs when sites have undergone multiple mutations causing sequence dissimilarity to no longer accurately reflect the 'true' evolutionary distance, and therefore the number of observed differences no longer increases with increasing evolutionary distance. After removal of these sites the authors obtained phylogenetic trees which supported a fish-specific genome duplication and they proposed that the removal of saturated sites will not bias their results in favor of their inferred tree topology [Van de Peer et al., 2002].

Even if the scenarios leading to massive gene duplication rates in fishes will be unravelled, there still remains the question why these processes took place. Already [Ohno, 1970] proposed a major role of gene duplication during the evolution of genomic and possibly phenotypic complexity. But in most cases one of the duplicates is inactivated, most probably by a mutation, and gets lost [Li, 1980]. The half-life of a duplicated gene was estimated with around 4.0 million years [Lynch et al., 2001].

Gene redundancy provides fishes with the chance of taking different evolutionary ways. These possibilities are non-, neo- and subfunctionalisation. One copy of a duplicated gene could be silenced by degenerative mutations (nonfunctionalisation) without any disadvantages to the fish and therefore occurs in as much as 50-90% of all duplicated gene pairs [Nadeau and Sankoff, 1997]. In the case of neofunctionalisation one copy evolves a new beneficial function that permanently preserves it in the population or alternatively both copies may be reciprocally preserved through the fixation of complementary loss-of-subfunction mutations, which results in a partitioning of the task of the ancestral gene [Lynch et al., 2001]. According to the findings of several examples of subfunctionalisation the DDC ('duplication-degeneration-complementation') model was proposed [Force et al., 1999, Lynch and Force, 2000], which predicts that the likelihood of preservation of duplicated genes is correlated with the number of 'subfunctions' that can be ascribed to a gene. Table 1.2 summarises possible fates of duplicated genes and lists examples and respective references.

Redundancy may be directly advantageous as a mechanism to compensate phenotypic effects of mutations or developmental accidents [Lynch et al., 2001], but fishes would not specifically benefit from that because of their high reproduction rate [Venkatesh, 2003]. So the biggest advantage could be, that the high number of genes allows fishes to acquire new or different functions permitting faster adaptation and evolution, but there is also the notion that polyploid amphibians and reptiles lack the high diversity observed in fishes, which indicates that genome duplication alone is not sufficient to drive species diversity [Venkatesh, 2003]. Another model, 'divergent resolution', was proposed by [Lynch and Force, 2000, Lynch and Conery, 2000] to explain the abundance of fish species. It occurs when different copies of a duplicated gene are lost in geographically separated populations and could genetically isolate these populations. Therefore, large-scale gene duplications and rapid speciation of organisms might be correlated.

It becomes clear that only the application of experiments to different fish models will provide

fate	gene/gene family	reference
neofunctionalisation	antifreeze protein in Antarctic fishes evolved from a protease gene	[Cheng and Chen, 1999]
neofunctionalisation	teleosts have two copies of the human estrogen receptor gene *ESR2*; *esr2b* evolved very rapidly after duplication, compared to *esr2a*, suggesting the emergence of a novel function	[Hawkins, 2000]
subfunctionalisation	retina specific homeobox gene Rx	[Loosli et al., 2001] [Loosli et al., 2003]
subfunctionalisation	in zebrafish *en1a*, expressed in limb bud, and *en1b*, active in hindbrain, were found, whereas mouse *En1* is expressed in both structures	[Force et al., 1999]
subfunctionalisation, neofunctionalisation	*Sox9* duplicates (*Sox9a* and *Sox9b*) in zebrafish and stickleback: their combined pattern is similar to expression of a single mammalian ortholog, additional expression of one gene duplicate in the ovary, not seen in mice	[Cresko et al., 2003]
subfunctionalisation - spatial subfunction, neofunctionalisation	*mitf* genes: in mammals and birds different isoforms encoded by a single gene were identified, but in zebrafish different genes for at least two isoforms were found; one gene duplicate is expressed in epiphysis of fish, but not in mouse ephiphysis	[Altschmied et al., 2002]
subfunctionalisation - temporal subfunction	two *hoxb1* in zebrafish show different temporal expression according to loss of the respective regulatory sequences	[McClintock, 2002]

Table 1.2: Evolutionary fates of duplicated fish genes.

new information. Therefore lots of efforts concentrated on comparative studies. As shown for human and mice, comparison of genome structure can be used for the annotation of previously undefined genes, the identification of large (80-1000 bp) functional gene-regulatory elements and detailed characterisation of transcription factor binding sites ('phylogenetic footprints') present in larger conserved non-coding regions, as reviewed in [Nobrega and Pennacchio, 2004]. In case of slowly evolving evolutionary changes, far distant phylogenetic relationships need to be be used for comparative analysis, which is valid for fish and humans, separated from their common ancestor by 450 million years [Kumar and Hedges, 1998]. A first example for using a fish genome in comparative genome analysis was the annotation of conserved sequences between the human and *Fugu rubripes* genomes, which led to the rapid identification of more than 1000 previously unidentified human gene candidates [Abrahams et al., 2002, Aparicio et al., 2002]. By unravelling gene and whole genome structure, researchers will also be provided with insights into genetic and biochemical networks controlling development and the plasticity of these regulatory networks [Furutani-Seiki and Wittbrodt, 2004], as could be shown on the retina specific homeobox gene *Rx* [Loosli et al., 2001, Loosli et al., 2003] as an example for subfunctionalisation of gene function. Finding mutants which only affect one subfunction or knocking out one subfunction could be probably only done in fishes when defects in whole function will cause a too severe phenotype in mammals to be identified (unless using conditional mutagenesis). The same can be true for different fish species, as different subsets of duplicated genes are found in different fishes. Combination of mutagenesis approaches will in part also overcome differential mutabilities due to potential differences in the genomic context (e.g. the presence of mutation hotspots, differences in region specific DNA repair, or chromatin structure) [Furutani-Seiki and Wittbrodt, 2004]. It will also be advantegous to apply comparative experiments between different fishes, as it was also shown that important genes like the *b* gene (maybe coding for a sugar-transport protein; first successfully positionally cloned medaka mutant; [Fukamachi et al., 2001]), or *eyeless* [Loosli et al., 2001] or *rs3* [Kondo et al., 2001] could not be found in large-scale zebrafish mutagenesis screens, illustrating the complementarity of the two fish systems.

1.2 Analysing the transcriptome

In order to unravel the scenarios in a living cell it is essential to know the spectrum of genes expressed at a given time or under certain conditions, that means to gain knowledge of the transcriptome of an organism. Whole-genome projects in principle provide all the genetic material of one cell, but de-novo prediction of genes from genomic DNA is still incorrect to a large extent [Reese et al., 2000]. This applies especially to eukaryotic genomes, where only a small portion of the genome is actually coding for proteins (3-5%, [HGS Consortium, 2001]). Many coding sequences are therefore not identified by gene prediction programs or furthermore if genes are recognized, splicing sites or even different splice variants are even more difficult to compute. On the other

hand lots of false positives may be calculated. Therefore identifying new genes on the mRNA level has several advantages compared to *in silico* prediction. Also, many species have not yet been taken into account for large-scale genome sequencing projects. In case the genomic sequence becomes finally availabe, a collection of cDNAs provides the best tool for identifying genes within the genomic DNA sequence, which may later also facilitate the annotation and assembly of that genome.

Messenger RNAs present in one cell represent active genes in a certain stage of that cell. These transcripts can be isolated, a complementary DNA (cDNA) can be synthesised and cloned into high copy number bacterial plasmids and therefore kept in large cDNA libraries. These cDNAs need then to be analysed and assigned back to their corresponding genes, thereby creating the gene catalogue of a certain tissue or organism at a certain point of development, disease or others. Clones obtained do contain the right splice variants and therefore the correct protein products. Collecting of genes becomes especially difficult for rare transcripts, which can appear with as little as 1 copy in 10 million mRNA molecules of a given cell [Velculescu et al., 1995]. To ensure subsequent high-quality experiments (e.g. multiparallel determination of transcription levels) a non-redundant set of well-defined cDNAs or oligonucleotides representing different genes is needed. Only such a non-redundant set of cDNA clones allows the efficient production of species-specific unigene functional tools.

How many genes do we have to collect? In the release of the public human genome issue [HGS Consortium, 2001] the number for human protein-coding genes was given with 34,000 (new numbers are lower, 20,000-25,000 genes; [HGS Consortium, 2004]), but according to different splice forms, approximately 2-3 per average gene [HGS Consortium, 2001], a lot more different transcripts are expected. Even that the number of ESTs exceeds some million for human and mouse we are still not sure that everything is detected. One example is the work of [Aparicio et al., 2002] where 1000 novel candidate genes were identified after comparison to the genomic sequence of fugu. The number of fish genes is given for the Ensemblv30 sets (zebrafish v4, fugu v2, tetraodon v7; April 2005; [Hubbard et al., 2005]) with around 30,000, but besides fugu this picture is still incomplete. For the 365 Mb fugu genome 31,059 gene loci were predicted by [Aparicio et al., 2002]. A crucial step in establishing a gene catalogue is the collection of full-length cDNA clones which has already been extensively done for zebrafish [Rasooly et al., 2003] within the ZGC (Zebrafish GeneCollection) of the NIH zebrafish initiative, as one example. In the following subsections methods of transcriptome analysis will be compared and the strategy used in this thesis to obtain a highly representative gene catalogue for *Oryzias latipes* is explained.

1.2.1 Methods of transcriptome analysis

Different methods of transcriptome analysis can be distinguished into methods that specifically compare two populations of mRNA (subtractive methods), methods that precisely count the amount of molecules of a particular species present in a sample (tagging methods), and methods

that read the mRNA contents from the strength of a hybridisation signal (hybridisation methods), as it is reviewed in [Vingron and Hoheisel, 1999].

Applying the standard protocol for cDNA library production, mRNA is usually extracted from the tissue of interest and all available material is transcribed into cDNA [Venter, 1993], ligated into a plasmid vector and then transformed into bacterial clones which can be collected in a cDNA library. These are called non-normalised cDNA libraries, because the amount of synthesised and cloned cDNA correlates to the occurrence of mRNA-templates inside cells of tissues used, and therefore highly expressed mRNA will represent the largest part of the cDNA libraries. The disadvantage in using such libraries is the high redundancy in respect to house-keeping genes and on the other hand, low abundant mRNAs will be rarely detected. To avoid highly expressed mRNAs their appearance can be reduced by the means of different techniques to produce so-called normalised libraries. These approaches facilitate the discovery of weakly expressed genes [Bonaldo et al., 1996, Soares et al., 1994]. Disadvantages of this approach is the technically demanding procedure of producing normalised libraries and also the fact that the information regarding the expression level of all mRNAs is definitely lost.

During subtractive methods two mRNA populations will be compared and molecules occuring only in one population will be transcribed into a cDNA library, which will then be enriched with the transcripts of interest, for example those which may be expressed only under certain conditions [Carninci et al., 2000]. Two examples for subtractive methods are representational difference analysis [Hubank and Schatz, 1994] and differential display [Liang and Pardee, 1995, Liang and Pardee, 1998]. Also microchip hybridisation methods applying two differently labeled probes [Schena et al., 1995] may be viewed as subtractive methods as a comparison between two mRNA populations can be derived by this approach. Subtractive methods increase the chance to identify new genes needed for a certain biological process or in a tissue of interest.

By using tagging methods on non-normalised cDNA libraries the occurence of certain mRNA molecules can be counted. Such a tag is obtained by various methods. One approach is partial sequencing of randomly chosen cDNA clones to produce the so-called expressed sequence tags (ESTs) [Adams et al., 1991, Wilcox et al., 1991, Adams et al., 1992, Khan et al., 1992], which would rapidly provide evidence for the transcription of corresponding genes in any organism. These ESTs may then be checked against existing sequences in the relevant databases and if any sequence displayed a high degree of similarity to an existing database entry (highly valuable are genomic sequences or mRNA sequences), that would indicate the presence of the corresponding gene. The big advantage of EST projects is the lower price for gene discovery compared to genomic sequencing, as for the same amount of DNA sequenced EST sequencing would detect a greater number of genes. Gene identification becomes easier if full-insert sequences of cDNAs are available. The alignment of several shorter ESTs can yield full length cDNAs, which represent the entire coding sequence of the corresponding gene.

Another tagging procedure is SAGE (serial analysis of gene expression) [Velculescu et al., 1995],

where short tags are concatenated prior to sequencing and sequence data can be processed like in EST collection projects. Also the oligonucleotide fingerprinting (OFP) technique may be viewed as a tagging method (see fig. 1.9), where the hybridisation of a clone with a set of oligonucleotides yields a vector of hybridisation signals which can be seen as a tag. Identical tags are then clustered and different groups should represent different transcriptional units as summarised in 1.2.2. All tagging methods allow the detection of new genes or splice variants by comparing the obtained sequence information with sequence databases, but additionally to gene identification, expression levels can be estimated by the number of identical tags counted [Audic and Claverie, 1997, Megy et al., 2002]. The data reliability of estimated expression levels depends strongly on the method for selecting the tags. In all tagging methods there is a chance that identical mRNAs will not produce the same fingerprint due to experimental noise or that identical tags are derived from different mRNAs. One advantage of oligonucleotide fingerprinting is that it produces information from the entire length of the clone which is not the case with SAGE or random EST sequencing, and from this point of view OFP is more sensitive in gene discrimination.

Applying hybridisation methods on transcriptome analysis requires either the immobilisation of cDNAs or PCR products on arrays and subsequently the hybridisation with the mRNA population [Schena et al., 1995] or high density arrays of oligonucleotides are used for hybridisation with mRNA [Lockhart et al., 1996]. These approaches provide semi-quantitative data by correlating hybridisation signal strength to the amount of mRNA in a population. The hybridisation signals are not easily reproducible, and can be affected by many unknown properties such as the cDNA library complexity, as well as clone and sequence specific features (e.g. insert size, nucleotide composition, presence of repeats, secondary structure, triple helix interaction) and require therefore repetitions of each experiment and multiple standardisation and calibration procedures to allow the meaningful comparison of hybridisation patterns obtained from various sources like different tissues or different experiments.

1.2.2 OFP normalisation prior to sequencing

As pointed out in the previous section in the case of non-subtractive and non-normalised libraries a subsequent EST sequencing approach will lead to a high redundancy of obtained sequences. Therefore a hybridisation based tagging method, oligonucleotide fingerprinting (OFP), was used in this work to reduce the redundancy in cDNA clone representation [Drmanac et al., 1996], [Meier-Ewert et al., 1998] and [Poustka et al., 1999] within one library, prior to production of ESTs. With the help of this method partial sequence information is obtained via a characteristic hybridisation pattern, the fingerprint, along the entire insert length of one cDNA clone (this method may also be characterised as a sequencing by hybridisation approch, SBH). According to their fingerprints, cDNA clones are grouped into clusters with similar fingerprints and only representatives of such clusters or clones with unique fingerprints are chosen for subsequent experiments (see fig. 1.9).

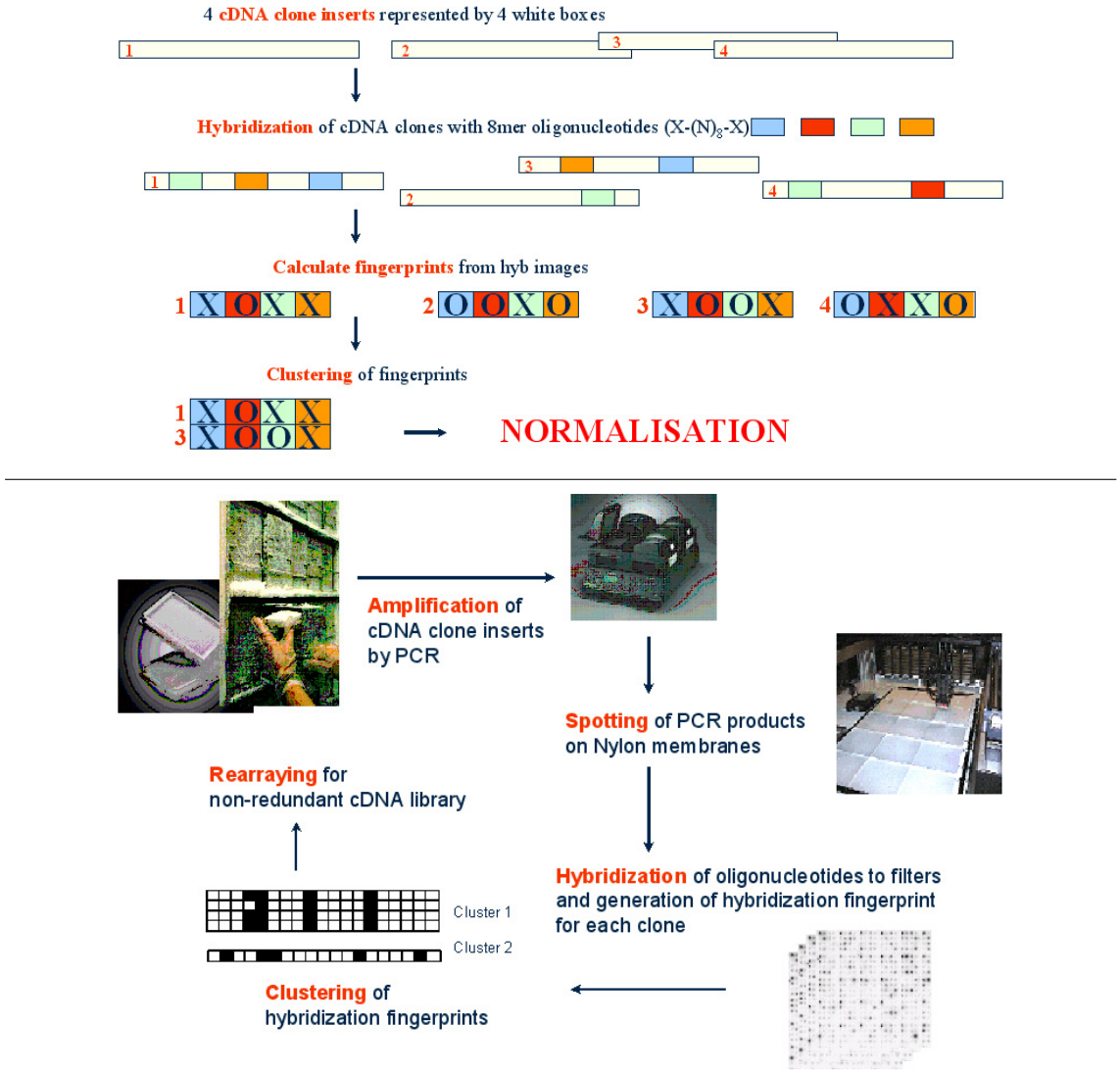


Figure 1.9: Principle and workflow of OFP.

The size of clusters obtained itself gives information about the abundance distribution of transcripts in the subjected tissue or developmental stage and it is therefore possible to identify differentially expressed genes between different libraries [Meier-Ewert et al., 1998]. The number of different clusters reflects the number of different expressed genes in the library under subject.

The fingerprint for one clone is obtained by hybridising 8mer oligonucleotides to PCR products of cDNA clone inserts immobilised on Nylon membranes. Choosing appropriate parameters hybridisation of oligos down to 6 bp can be sequence specific [Drmanac et al., 1990], but to obtain reliable hybridisation signals a mixture of all possible 10mers containing a common 8mer core sequence were hybridised. The sequence of hybridised oligonucleotides were chosen from a standard set established in previous experiments (see [Meier-Ewert et al., 1998, Clark et al., 2001, Herwig et al., 2002]) or specifically calculated from available Medaka EST data [Herwig et al., 2000].

By application of robotic technology [Maier et al., 1994] thousands of cDNA clones can be analysed in parallel (see fig. 1.9). Inserts of cDNA clones will be replicated by PCR and spotted on 22 x 22 cm Nylon membranes in duplicates. These membranes are subjected to hybridisation analysis and hybridisation images are scanned and automatically analysed. Hybridisation data are collected in a database and fingerprints are calculated and clustered. The clustering output file contains all clustered cDNA clones, ranked within one cluster according to the similarity of their fingerprint to the calculated consensus cluster fingerprint, and clones with unique fingerprints.

OFP analysis has been employed to several applications, like the identification of overlapping clones in order to construct ordered clone libraries (e.g. for the Herpes simplex virus type I genome, creating a physical map [Craig et al., 1990]), the establishment of clone maps from large insert clones [Hoheisel et al., 1991, Hoheisel et al., 1994] or the pre-selection of shotgun clones to reduce the redundancy within large-scale genomic sequencing projects [Radelof et al., 1998]. But the most significant field of application is the generation and characterisation of normalised cDNA libraries and parallel to this expression profiling studies by employing this strategy to different tissues or conditions [Clark et al., 2001, Meier-Ewert et al., 1998, Eickhoff et al., 2000, Poustka et al., 1999, Poustka et al., 2003, Herwig et al., 2002] or the elucidation of different gene families like the human olfactory receptor genes [Fuchs et al., 2002]. OFP-based unigene sets were so far created for mouse [Meier-Ewert et al., 1998], zebrafish [Clark et al., 2001], sugar beet [Herwig et al., 2002], sea urchin [Poustka et al., 2003] and amphioxus [Panopoulou et al., 2003].

1.2.3 EST sequence analysis

EST sequence projects provide lots of data which have to be analysed in more or less automated ways. Analysing EST sequences has to consider their low quality in every processing step which requires to estimate base quality values and to clean sequences from contaminating vector or even bacterial chromosomal DNA. Further processing steps involve the masking of repetitive sequences, clustering of sequences and the annotation of obtained clusters and sequences left as singletons. Obtained sequences establish a powerful resource for further experiments like expression profile analyses, genetic mapping experiments or phylogenetic studies.

1.2.3.1 EST clustering

Clustering of related EST sequences is important to extract useful information from these sequences that show high redundancy, low quality and too short length [Malde et al., 2003] and therefore to increase quality of subsequent experiments and to analyse alternative splicing variants [Bouck et al., 1999]. False results lead to underclustering where ESTs actually belonging to the same gene will be split into different clusters (generated by too short ESTs that do not cover all of the gene, therefore ESTs of the same gene may not overlap) or overclustering when ESTs that originated from different genes (most likely of the same gene family) are put into one group. Both cases lead to a wrong estimation of gene numbers.

The first sequence clustering programs were originally programmed for assembling shotgun DNA sequence data, like PHRAP ("phragment assembly program") [Staden et al., 2000], the TIGR assembler [Liang et al., 2000] and CAP3 [Huang and Madan, 1999]. These programs are more or less adaptable to cluster EST sequence data [Liang et al., 2000], caused by the difference in obtaining the DNA sequence during genome shotgun approaches compared to EST sequencing efforts. Genomic shotgun sequencing produces several reads for a single interval, which should be highly identical (sequences showing less than 98% identity can be assumed to come from different copies of a repetitive element), but EST data are derived from a wide variety of sources representing the spectrum of polymorphisms in the original samples, like a relatively high rate of insertions and deletions, contamination by vector and linker sequences, non-random distribution of sequence start sites in oligo(dT)-primed libraries and every clone is usually sequenced only once, in best cases there are 5' and 3' ends sequenced. For all these reasons the degree of identity in overlapping sequences from the same gene will often be lower than in genomic projects and additionally patterns of overlapping sequences caused by alternative transcripts are different from that observed in a genomic shotgun project. Further, in working with publicly available sequences the base quality values are often not provided, which further complicates the clustering.

In comparing TIGR Assembler, PHRAP and CAP3 [Liang et al., 2000], it was found that CAP3 produces the fewest high-quality assemblies from single genes while being tolerant to random errors yet maintaining the ability to discriminate between related genes. With the help of PHRAP calculated consensus sequences contained a large number of insertions and deletions (when several input sequences disagree, the underlying algorithm often resolves the problem by inserting two different bases in the final consensus, producing an insertion error), it also over-assembles sequences, combining ESTs from distinct transcribed genes into one single consensus sequence. The TIGR assembler splits sequences into singletons or separates contigs as the sequence quality decreases, which is not observed with PHRAP or CAP3. The authors also made the observation that in case of highly expressed genes where lots of ESTs were available, PHRAP and TIGR assembler have a strong tendency to split these sequences into several contigs [Liang et al., 2000].

Careful preclustering prior to assembly, done by all-versus-all pairwise similarity searches, is essential to produce faithful clusters [Pertea et al., 2003]. There are two tools implemented involving preclustering and using CAP3 or a modified CAP3 as clustering tool. One is Paracel Transcript Assembler (PTA) from Paracel (www.paracel.com), which modified the CAP3 algorithm to bring up a commercial version, CAP4, as clustering tool. It is a whole package taking raw trace files as input and calculating EST clusters from cleaned and processed sequence data, which is now also used from the TIGR institute to produce its TIGR gene indices (http://www.paracel.com/sas/pta.htm). The TIGR insitute established earlier a publicly available pipeline TGICL [Pertea et al., 2003], which performs clustering of large EST data sets by preclustering with mgblast, a modified version of megablast [Zhang et al., 2000], and then the assembly and consensus building of each obtained group is done with CAP3. Base quality values produced by PHRED [Ewing et al., 1998] are used

in computation of multiple sequence alignments of reads, construction of overlaps between reads, and generation of consensus sequences. In this work the TGICL pipeline is used for assembly of EST data.

A completely different approach to create a gene index is UniGene [Pontius et al., 2003] from NCBI. It utilises ESTs and properly annotated mRNA sequences that are derived from the dbEST database [Boguski et al., 1993] and GenBank [Benson et al., 2002]. These sequences are compared with each other and all sequences that have a significant overlap are placed into a single group. No consensus sequence is determined for each cluster. The advantage lies in the non-stringent clustering parameters that allow alternatively spliced transcripts to be incorporated into the same cluster, thereby simplifying gene identification. Also to minimise the frequency of multiple clusters being identified for a single gene, UniGene clusters are required to contain at least one sequence carrying readily identifiable evidence of having reached the 3' terminus, like the presence of a poly(A) tail or at least two ESTs that were generated using the 3' sequencing primer.

1.2.3.2 Sequence annotation

After the assembly of ESTs a major task is to predict the number of transcripts expressed among computed consensus or singletons and their functions, which is mostly done by running similarity searches against known databases using the blast-algorithm [Altschul et al., 1990]. A nice way to present these data uses the concept of ontologies [Stevens et al., 2000, Schulze-Kremer, 2002], where genes and gene products are consistently presented in a directed acyclic graph to standardise classifications for sequences and sequence features [Ashburner and Lewis, 2002, GO Consortium, 2004, GO Consortium, 2000]. The ontology includes a vocabulary of around 16,000 terms describing proteins in three different, non-overlapping, categories: molecular function, biological process and cellular component. The GO-term "molecular function" represents activities, "biological process" describes the biological goal accomplished by the protein and "cellular component" describes the locations of proteins, at the level of subcellular structures and macromolecular complexes. All these terms are freely available via http://www.geneontology.org. A centralised public resource was provided at the same web location allowing not only access to the ontologies and annotated data sets but also different software tools were developed to extract information about certain protein functions and the proteins belonging in this category. The GO system is changing rapidly as the state of biological knowledge of what genes and proteins do is very incomplete and still changing. New GO annotations must always be attributed to a source, which may be a literature reference, another database or computational analysis [GO Consortium, 2004]. Therefore in obtaining GO annotations one has to take into account the different quality of annotations, as there are high-quality GO annotations based on curatorial review of published literature and supported by experimental evidence in contrast to annotations based on automated methods, called IEA annotations ("inferred from electronic annotation"). It was also shown that knowledge of the biological role of proteins in one organism can be transferred to other organisms, e.g. in [Hennig et al., 2003].

The GOblet platform [Hennig et al., 2003, Groth et al., 2004] uses this approach. This system allows the uploading of a batch of sequences for which similarity searches against known and GO annotated sequences will be run and therefore the queries will be assigned with gene ontology terms and presented in a GO tree. This approach allows easily to compare different sets annotated by gene ontology.

1.2.4 Identifying single nucleotide polymorphisms

SNPs are the most abundant polymorphisms found in a species. In the case of cichlids were species are so fast evolving that even their mitochondrial genomes evolve too slowly to identify markers, SNPs are the only data to understand the phyletic relationships within arrays of closely related cichlid species [Salzburger and Meyer, 2004].

Base substitutions are distinguished into transitions and transversions. If purines (adenine or guanine) are replaced by pyrimidines (cytosine or thymin) or vice versa then one talks of transversions. In case of transitions replacing of bases happens just between different pyrimidines or between different purines. Frequency of such base polymorphisms is different in any genome.

For linkage mapping experiments, polymorphisms between different strains, preferrably inbred strains, are investigated [Kruglyak, 1997]. SNPs are advantegous tools for linkage experiments because of their high frequencies within the genome and their low mutation rates and several methods of automated SNP detection of known SNPs (summarised in [Syvänen, 2001]) were developed.

In addition to being a source for a gene catalogue, EST sequences can be useful for SNP discoveries [Brentani et al., 2003, He et al., 2003, Schmid et al., 2003] by aligning ESTs from different strains or organisms from one species and searching for single nucleotide polymorphisms in their alignments. Therefore EST sequences from different sources have to be obtained, assembled using EST clustering tools and then examined for SNPs. There are mostly only semi-automated systems available [Lehnert et al., 2001] for identification of SNPs in EST alignments which display the alignment and label the polymorphism with different colours, but candidate SNP positions have to be visually scored.

1.2.5 Comparing transcriptomes

In cases where complete genomes are available the full sets of predicted or known proteins of different species can be compared to each other, like it was done for *Caenorhabditis elegans* and *Saccharomyces cerevisiae* [Chervitz et al., 1998]. On one hand they found highly conserved proteins with mostly one copy in worm and one copy in yeast. These carry out mostly the core biological processes, such as intermediary metabolism, DNA and RNA metabolism, protein folding, trafficking, and degradation. These findings also suggest, that the core biological processes are carried out by a similar number of proteins. Of course on the other hand the greater complexity of the worm is also visible by distinct subsets of proteins, which have no counterpart in yeast.

To compare fish transcriptomes is not an easy task. Protein sets of fishes are still not complete and are to a large extent *in silico* annotated. Also the redundancy of fish genes and their protein products provides certain complications, which makes it hard to distinguish between orthologs, which have evolved by vertical descent from a common ancestor, and paralogs, that arise by duplication and domain shuffling within a genome.

1.3 Objective

The primary goal of this thesis included to gather information about the transcriptome of the Medaka fish, *Oryzias latipes*. The availability of the gene set of an organism is a very helpful tool for all genetic experiments on that organism. Even in case of a whole genome sequence (genome project was just at the beginning at the start of this project's work) experiments on transcriptome level are necessary for elucidation of all genes available. EST analysis as a method of transcriptome analysis is still a cost-intensive approach identifying unknown genes. Therefore Oligonucleotide fingerprinting (OFP) analysis was chosen for prior normalisation of used cDNA library before the production of ESTs because of its reduced costs compared to EST sequencing alone.

Four Medaka cDNA libraries were provided covering three embryonic stages, Gastrula, Neurula and Organogenesis and one adult tissue, the ovary. These libraries should be subjected to OFP analysis and subsequent EST production. Besides normalisation of cDNA libraries, OFP analysis may apply information on differential expression of OFP cluster by comparing the appearance of cDNA clones in an OFP cluster from different fish stages or tissues.

A further part of this work will be firstly to provide high-quality sequence data as a valuable resource for different cooperation partners and secondly the *in silico* evaluation of sequence information like assigning a function to sequence data.

Chapter 2

Materials and methods

2.1 Materials

2.1.1 Laboratory equipment

PCR machines, PTC 225	MJ Research Inc., Watertown, USA
hybridisation oven	Appligene Oncor, Illkirch Graffenstaden
phosphor-imaging system: Phosphorimager 445SI, version 4.0 and phosphor storage screens	Molecular Dynamics, Sunnyville, CA
filmprocessor	Curix 60, Agfa-Gevaert, N.V., Morstel, Belgium
spotting robot	QBot, Genetix GmbH, München-Dornach or custom-built
rearraying robot	custom-built
robot for filling 384-well microtitre plates	Qfill, Genetix GmbH, München-Dornach
UV crosslinker	UV stratalinker 2400, Stratagene, La Jolla, USA

2.1.2 Chemicals and reagents

Agarose	Invitrogen, Groningen, Netherlands
Ampicillin-Na-salt	Sigma-Aldrich Chemie GmbH, Steinheim
Bacto Tryptone	Difco Laboratories, Detroit, USA
Bacto Yeast Extract	Difco Laboratories, Detroit, USA
Betaine	Sigma-Aldrich Chemie GmbH, Steinheim
Bromphenol blue	Sigma-Aldrich Chemie GmbH, Steinheim
Cresol red	Sigma-Aldrich Chemie GmbH, Steinheim
dNTPs	Amersham Pharmacia Biotech Europe GmbH, Freiburg (now: GE Healthcare)
EDTA	Merck Eurolab GmbH, Darmstadt
Glacial acetic acid	Merck Eurolab GmbH, Darmstadt
Ethanol	Merck Eurolab GmbH, Darmstadt
Ethidium bromide	Sigma-Aldrich Chemie GmbH, Steinheim
Glycerol	Merck Eurolab GmbH, Darmstadt
HEPES	Sigma-Aldrich Chemie GmbH, Steinheim
Isopropanol	Merck Eurolab GmbH, Darmstadt
Potassium chloride	Merck Eurolab GmbH, Darmstadt
Magnesium chloride	Merck Eurolab GmbH, Darmstadt
Sodium chloride	Merck Eurolab GmbH, Darmstadt
Salmon sperm DNA	Sigma-Aldrich Chemie GmbH, Steinheim
SSarc, Sarcosyl (Sodium-N-Lauroyl-sarcosin-Na-salt)	SERVA Electrophoresis GmbH, Heidelberg
TrisBase	Merck Eurolab GmbH, Darmstadt
TrisHCl	Merck Eurolab GmbH, Darmstadt
Tween20	Sigma-Aldrich Chemie GmbH, Steinheim
2 x YT agar	BIO 101, Vista, CA, USA
2 x YT broth	BIO 101, Vista, CA, USA

2.1.3 Radiochemicals

- [γ-33P]-dATP for end-labeling of oligonucleotides (Amersham Pharmacia Biotech Europe GmbH, Freiburg)
- [α-33P]-dCTP or [α-32P]-dCTP for incorporation into randomly synthesised probes (Amersham Pharmacia Biotech Europe GmbH, Freiburg)

2.1.4 Buffers and Solutions

- 10 x PCR buffer (LEO's buffer):
  500 mM KCl
  350 mM TrisBase
  150 mM TrisHCl
  15 mM $MgCl_2$
  1% v/v Tween 20,
  adjusted to pH 8.3 and autoclaved;
  Cresol red was filtrated and then added into the buffer to a final concentration of 150 mM.

- 5 M betaine: dissolved in water by heating slightly

- 6 x DNA loading buffer:
  0.2% Bromophenol blue
  60% Glycerol
  60 mM EDTA

- 50 x TAE buffer:
  2 M TrisBase
  1 M acetate
  100 mM EDTA (pH 8.0),
  adjust final pH to 8.2

- 1 x TE buffer:
  10 mM TrisHCl
  1 mM EDTA,
  adjusted to final pH 8

- SSarc hybridisation buffer:
  4 x SSC
  7.2 M Sarcosyl (30%)
  4 mM EDTA (pH 8.0)

- SSarc stripping solution:
  dilute SSarc hybridisation buffer 1:10,
  add 2 ml of 0.5 M EDTA per 1 litre of stripping solution

- 20 x SSC stock solution:
  3 M NaCl (pH 7.5)
  0.3 M Na_3-citrate

- Modified Church, hybridisation buffer:
  5% SDS
  0.25 M Na_2HPO_4 (pH 7.2)
  1 mM EDTA

- WashI solution:
  2 x SSC
  0.1% SDS

- WashII solution:
  0.1 x SSC
  0.1% SDS

- Hexanucleotides (Amersham Biosciences):
  pdN6 (45 OD), 1:8, dissolved in
  1 mM TrisHCl (pH 8.0)
  1 mM EDTA (pH 8.0)

- Labeling solution (for labeling with Klenow fragment polymerase):
  1 M HEPES buffer (pH 6.6)
  250 mM TrisHCl (pH 8.0)
  25 mM $MgCl_2$
  50 mM β-mercaptoethanol
  6 A_{260} units/ml of hexanucleotides

- Stripping solution:
  0.1% SDS
  2 mM EDTA

- Ampicillin: 50 μg/ml dissolved in 70% Ethanol

- Salmon sperm DNA: 600 ng/ml dissolved in 0.4 M NaOH

2.1.5 Media

- LB medium:
  10 g Bacto tryptone
  5 g Bacto yeast extract
  10 g NaCl,
  adjust to pH 7.0 with 1 M NaOH,
  adjust the final volume to 1 litre with H_2O

- LB agar:
  LB medium
  15 g/l Bacto agar

- 2 x YT medium:
  16 g Bacto tryptone
  10 g Bacto yeast extract
  5 g NaCl,
  adjust to pH 7.0 with 1 M NaOH,
  adjust the final volume to 1 litre with H_2O

- 2 x YT agar:
  2 x YT medium
  15 g/l Bacto agar

- 10 x HMFM (Hogness Modified Freezing Medium):

Solution A:	0.9 g $MgSO_4 \cdot 7\ H_2O$
	4.5 g Na_3-citrate$\cdot 2\ H_2O$
	9 g $(NH_4)_2SO_4$
	440 g Glycerol
	adjust the final volume to 800 ml with H_2O
Solution B:	18 g KH_2PO_4
	47 g K_2HPO_4
	adjust to pH 7.0 with 1 M NaOH
	adjust the final volume to 200 ml with H_2O
A and B were mixed shortly before use.	

All media were autoclaved immediately after preparation. To select for clones with proper clone inserts media and agar were supplemented with antibiotics, ampicillin in case of all cDNA libraries.

2.1.6 Size standards

To estimate the size of DNA fragments after their separation by agarose electrophoresis the ϕX174 DNA(HaeIII digest) marker (NEB Inc., Beverly, MA, USA) was used, which includes fragments ranging in size from 72 to 1353 base pairs (fig. 2.1).

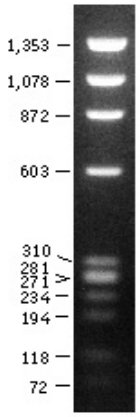


Figure 2.1: ϕX174 DNA(*HaeIII* digest) marker separated on a 1.7% agarose gel. Picture taken from NEB Inc.

2.1.7 Enzymes

- DNA polymerases (home-made): *Taq* and *Pfu* polymerases, supplied at a concentration of 10 U/μl.
- T4 polynucleotide kinase (NEB, Inc., Beverly, MA, USA): supplied in a concentration of 10 U/μl with corresponding buffer.

- Klenow fragment of DNA polymerase I of *Escherichia coli* (USB Corporation, Cleveland, Ohio, USA): obtained with a concentration of 5 U/μl

2.1.8 cDNA libraries

The cDNA clones subjected to OFP analysis were produced by poly-dT-priming in the group of M. Furutani-Seiki (Japan Science and Technology Agency, Kyoto, Japan) and cDNAs were ligated into a modified pCS2+(SfiIA-B) vector (see details of vector design and cloning in appendix B). Clones were picked into cDNA libraries at the RZPD. Unfortunately during the cloning procedure the ampicillin resistance gene was negatively affected. Therefore bacteria containing these cDNA clones have to be raised with a lower ampicillin concentration of 50 μg/ml and still grow comparatively poorly. Description of libraries and the number of cDNA clones used is summarised in table 2.1.

library name (RZPD ID)	clone number	RNA source	vector	produced by
gastrula (Med1015)	63,744	gastrula stage, CAB male and female	pCS2+(SfiIA-B)	M. Furutani-Seiki (Japan Science and Technology Agency, Kyoto, Japan)
neurula (Med1028)	11,520	neurogenesis (stage 17-23), CAB male and female	pCS2+(SfiIA-B)	M. Furutani-Seiki
organogenesis (Med1030)	23,424	organogenesis (stage 24-33), CAB male and female	pCS2+(SfiIA-B)	M. Furutani-Seiki
ovary (Med1029)	20,352	ovaries from CAB strain females	pCS2+(SfiIA-B)	M. Furutani-Seiki

Table 2.1: Characteristics of cDNA libraries subjected to OFP analysis

2.1.9 Oligonucleotides

2.1.9.1 PCR primers

For amplification of cDNA clone inserts in pCS2 vector primers Med1 and Med2 were used. For sequencing of PCR products inserts were amplified with Med1 and Med9. Sequences of all primers can be found in table 2.2.

2.1.9.2 Sequencing primer

For sequencing of individual cDNA clones in pCS2 vector, inserts were amplified with Med1 and Med9 and 5' sequencing reaction was run with Med3 and 3' reaction with Med2 (see table 2.2). In case of large scale sequencing, inserts were amplified by a different mechanism (see 2.2.2) and then sequenced from their 5' end with Med3.

Oligonucleotide	Sequence (5'-3')	cDNA cloning vector	Annealing temperature
Med1	CTTGTTCTTTTTGCAGGATCCCATCG	pCS2	65 °C
Med2	GGATCTACGTAATACGACTCACTATAG	pCS2	65 °C
Med3	GCAGGATCCCATCGATTCGAATTC	pCS2	65 °C
Med9	TGTCTGGATCTACGTAATACGACTC	pCS2	65 °C

Table 2.2: Sequences of PCR primers used to amplify certain cDNA inserts.

2.1.9.3 Oligonucleotides used for OFP

Sequences of oligonucleotides needed for oligofingerprinting analysis were taken from a standard set [Meier-Ewert et al., 1998, Clark et al., 2001, Herwig et al., 2002] or computed as described in 2.2.1.3 specifically for publicly available Medaka sequences. Oligonucleotides from the standard set were provided from M. Janitz group, synthesised by TIB MOLBIOL, Berlin. Computed oligonucleotides were ordered from MWG Biotech, or BioTeZ, Berlin-Buch GmbH. It is crucial to obtain HPLC-purified probes as unspecific binding of probes cleaned by other methods like the HPSF system (high purity salt free purification) from MWG were noticed. Sequences of probes used and also different sets of probes hybridised during the two different fingerprinting rounds can be found in the appendix A.

2.2 Methods

2.2.1 Oligonucleotide fingerprinting approach

2.2.1.1 Generation of PCR products

The hybridisation of short oligonucleotides requires the production of PCR products from cDNA clone inserts. PCR amplifications were carried out in 384-well microtitre plates (Thermo-Fast 384, ABgene). Using disposable 384-pin inoculation devices (Genetix) a small amount of bacterial suspension was spotted onto 22 x 22 cm agar dishes (Nunc) containing ampicillin (50 μg/ml) and bacteria were grown overnight at 37 °C. With the help of the same inoculation devices, bacteria were added to 25 μl PCR-reaction containing 1 x PCR buffer (see 2.1.4), 200 μM of each dNTP, 0.5 μM betaine, 2.5 units *Taq* polymerase and 0.125 units *Pfu* polymerase, 7.5 pmol of each PCR primer (for sequences see 2.1.9). PCRs were performed for 30 cycles: 30 sec 94 °C, 30 sec 65 °C and 3 min at 72 °C in 384-well PCR machines.

2.2.1.2 Arraying of PCR products

High density filter arrays of PCR products were generated robotically as described [Maier et al., 1994]. Each 22 x 22 cm Nylon membrane (Amersham Biosciences) carries up to 27,648 different cDNA inserts spotted in duplicate and in addition 2304 spots of genomic salmon sperm DNA (600 ng/ml

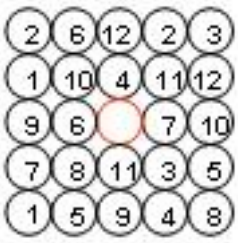


Figure 2.2: PCR products were spotted with a 5 x 5 pattern which simplifies the identification of positive clones and the consequent calculation of clone coordinates to a large extent. For example, the two spots on position "1" correspond to one PCR product. To count as positive, the PCR product has to show positive hybridisation results on both spots. One guide dot is spotted in the center of each 5 x 5 field.

in 0.4 M NaOH), which hybridise with each oligonucleotide and were therefore suitable as guide dots for automated image analysis. Up to 30 filters were prepared from one set of PCR plates containing 25 μl of PCR reaction. For spotting a 5 x 5 pattern (see fig. 2.2) was used. One filter is divided into 6 big fields, where on each field 12 plates with 384 wells are spotted in duplicates.

This pattern is achieved by the aid of a robot, which uses a 384-pin head to carry the liquid from the 384-well PCR plate to the filter. A 250 μm pin transfers on average 2 nl of DNA solution. In case the DNA concentration of PCR products is around 100 ng/μl and each PCR product is spotted 10 times on one position then 2 ng of cDNA insert were transferred to the membrane. DNA was spotted on membranes soaked in 0.4 M NaOH. After spotting filters were washed for 2 min in 0.4 M NaOH, and then for another 2 min in 5 x SSC. Afterwards membranes were dried on Whatman paper. To fix more DNA, filters were also baked at 80 °C for 30 min and crosslinked with UV-light. After these procedures the membranes were ready to be used.

2.2.1.3 Oligonucleotide probe selection

A list of 8mer oligonucleotide probes were computed according to [Herwig et al., 2000] using `oligo_design`, that was optimised according to a training set of medaka DNA sequences, assuring that selected probes would hybridise with sufficient frequency but will also be dissimilar enough to differentiate the training sequences. This will be crucial as obtained fingerprints have

to be clone specific so that different genes can be distinguished by their fingerprints and are put into different groups by the clustering algorithm. The quality of one probe is measured by Shannon entropy [Shannon, 1948] of the partition, which reaches its maximum value in the case the probe distinguishes a set of sequences into exactly two sets. It was shown that the clustering depends strongly on the probe set used and it is advisable to design a species-specific set [Herwig et al., 2000]. As training set, EST data from ricefish publicly available in May 2002 (GenBank/EMBL) were size selected for reads of 300-2000 bp (average length 600 bp) and clustered with PHRAP (7897 sequences used as input data; Green, unpublished; http://www.phrap.org). In a later run, the TIGR unigene index plus GenBank/EMBL data (January 2003; in total 12,771 sequences with average length of 685 bp) clustered with PHRAP were used as input sequences. Octamer probes matching vector sequence of medaka cDNA clones were excluded. The program was run with default parameters. Additionally, a standard set of oligonucleotides was chosen that showed good hybridisation results in previous experiments on different organisms, like zebrafish or sugar beet [Clark et al., 2001, Herwig et al., 2002]. Since 10mers hybridise more reliably than 8mers each hybridisation probe comprises a pool of all sixteen 10mers sharing the same 8mer core sequence (`NXXXXXXXXN`). The sequences of all oligonucleotide probes can be found in the appendix A.

2.2.1.4 Oligonucleotide hybridisation

Oligonucleotides (100 pmol) were labeled at the 5' end by a kinase reaction using 30 μCi [γ-33P]-dATP and 20 units T4 polynucleotide kinase. Hybridisations were carried out in hybridisation bottles as described in [Meier-Ewert et al., 1998] at 4°C with the probe diluted in 10 ml of SSarc buffer. After hybridisation and washing the membranes twice for 20 min in 1 litre of cold SSarc buffer, filters were exposed to a phosphor-imaging system and scanned after 1 to 3 nights of exposure, where this time depends on the quality of the PCR-product, mainly on the DNA concentration produced and therefore transferred to the Nylon membrane. The phosphor imager scans with a resolution of 176 μm/pixel. To correlate filters with their hybridised probes these results were reported in a database, interfaced by a script `rename_images`, provided by the bioinformatics group of R. Herwig. Nylon membranes can be stripped with 1:10 SSarc solution, meaning to remove the probe bound to cDNA clone inserts, and used again up to 30 times.

2.2.1.5 Image analysis

Scanned filter images were analysed automatically with in-house analysis software by the bioinformatics group of R. Herwig [Steinfath et al., 2001], which positions a grid on top of the image by calculating filter edges and positions of positives on the image according to the spotting pattern, mostly with the help of guide dots, which were spotted in the middle of each block. These guide dots consist of salmon sperm DNA which is likely to produce a signal with all oligonucleotide probes. According to their grid position intensity values will be assigned to the identified spots

proportional to the amount of probe hybridised to this clone. The most frequent intensity outside the spotting area is taken as the background intensity which is therefore subtracted from all other signal intensities. Clones are spotted in duplicates to judge the hybridisation reproducibility. The correlation between the intensities of these duplicates is used as a quality measure for each hybridisation, which is expressed for the whole array as a correlation coefficient. In case of a low correlation coefficient images are excluded from further analysis. The results of hybridisations and grid finding were visually checked with Xdigitise [Wruck et al., 2002].

2.2.1.6 Normalising and clustering fingerprinting data

The raw fingerprint data were normalised in R. Herwig's group in order to eliminate experimental errors across all filters and all probes. Normalisation involves two steps: The first step is normalisation within each filter for all clones (variations arise as a result of the specific amount of clone material derived from PCR amplification, as a result of the specific amount of transferred clone material introduced by the spotting procedure, and as a result of specific position of the spots on the filter membranes) and the second step is normalisation across all probes for each clone (variations arise as a result of specific behaviour of the probes when using the same hybridisation conditions, because of the filter material that can be of different quality and because of differences in radioactive labeling of the probes) [Herwig et al., 1999].

The normalised data were grouped applying a sequential k-means clustering algorithm that uses mutual information as a pairwise similarity measure [Herwig et al., 1999]. The sequential k-means approach has been introduced by [MacQueen, 1967] and further developed by [Mirkin, 1996]. The advantage of this method to conventional k-means algorithms, where the number of centroids has to be fixed before, lies in the fact that it finds the number of different clusters from the data itself. The advantage of using mutual information instead of metric distances are that on one hand it takes into account the total number of matched similarities and on the other hand uninformative fingerprints (matching only to a few probes) are simply ignored by assigning very low similarity value to those. A consensus fingerprint was calculated from all fingerprints in a cluster, allowing the ranking of individual clones within a cluster according to the similarity of their fingerprint to the consensus fingerprint.

Singletons and representatives of clusters were re-arrayed using robotic devices [Maier et al., 1994]. Representatives were chosen which showed highest similarity to the calculated consensus fingerprint for each cluster.

2.2.1.7 Hybridisation of cDNA control clones

To control the efficiency of oligofingerprinting analysis, PCR products of randomly chosen cDNA clones representing the consensus fingerprint of one OFP cluster and additionally clones left as OFP singletons were hybridised to colony filters (in the case of gastrula clones) or to the same cDNA clone insert filters used for OFP analysis. Approximately 50 ng of each PCR probe were

labeled in a random hexamer priming reaction [Feinberg and Vogelstein, 1983] using 30 μCi [α-33P]-dCTP or [α-32P]-dCTP and 10 units Klenow polymerase. Hybridisations were performed overnight at 65°C in hybridisation bottles containing 10 ml of modified Church buffer. Filters were washed once in Wash I for 20 min and for another 20 min in Wash II at 65 °C. Intensities of the hybridisation signals were measured via autoradiography on X-ray films (Kodak) or with a phosphor storage system.

Hybridising one cDNA clone insert to other cDNA clones should only highlight clones representing the same gene, and all such copies should be found. Therefore these results were taken as the true appearance of copies from the same gene within the cDNA libraries and all these positive clones should be clustered together applying the fingerprinting analysis. The splitting of individual true gene clusters during the OFP approach can be numerically evaluated by calculating the diversity index δ using entropy as described in [Herwig et al., 1999]. Suppose that gene g_i is present in the library with N_i copies and suppose that these copies are split in K different clusters with frequencies $n_1, ..., n_k$ $(n_1 + ... + n_k = N_i)$, then the diversity can be computed as

$$\delta(g_i) = \frac{-\sum_{j=1}^{K} \frac{n_j}{N_i} \log_2 \frac{n_j}{N_i}}{\log_2 N_i}.$$

The diversity is maximal ($\delta(g_i) = 1$) if all copies belong to different cluster, it is minimal ($\delta(g_i) = 0$) if all copies belong to the same cluster.

2.2.2 cDNA sequencing

The sequencing of cDNA clone inserts was done in Prof. Shimizu's lab at Keio Medical School, Tokyo. Bacterial plasmids were amplified with a rolling circle amplification technique (RCA; Amersham Biosciences). The diluted plasmid suspension was subjected to sequence analysis using BigDye-terminator chemistry. Sequencing reaction was then cleaned with MultiScreen$_{384}$-PCR plates (Millipore) and 0.1 mM EDTA as washing solution and loaded on an ABI3730xl DNA sequencer (Applied Biosystems). All reactions were done with the help of robotics in a 4 x 96-well or 384-well format. Processing of obtained ABI trace files is described in subsection 2.3.5

2.2.3 Identification of differentially expressed OFP cluster

The oligofingerprinting result provides the number of clones in an OFP cluster, which derive from a certain library. Based on the number of clones fingerprinted from each library differentially expressed clones can be identified using statistical tests. For every OFP cluster the null hypothesis, that a putative transcript is equally represented in all libraries, is approved or rejected by applying a log likelihood ratio test, the R-test (Stekel, 2000). This ratio gives a measure of the extent to which the differences in gene expression correspond to heterogeneity of the libraries as opposed to random sampling variability. The statistic, denoted R_j for gene j is given by the expression

$$R_j = \sum_{i=1}^{m} x_{i,j} \log\left(\frac{x_{i,j}}{N_i f_j}\right),$$

where m is the number of cDNA libraries, $x_{i,j}$ is the number of transcript copies of gene j in the ith library and N_i is the total number of cDNA clones sequenced in the ith library. f_j is the frequency of gene transcript copies of gene j in all of the libraries, given by the formula

$$f_j = \frac{\sum_{i=1}^{m} x_{i,j}}{\sum_{i=1}^{m} N_i}.$$

In a library in which there are no observed copies of the gene, that is, $x_{i,j} = 0$, its contribution to R_j is zero. The formula is only valid if at least 50 ESTs have been sequenced from each library, and no single gene contributes >20% of the ESTs in a library.

The significant cut-off value for the R-test was estimated from the data itself, calculating R-values for all OFP clusters as described and then the number of OFP clusters for a given value of the test statistic R is plotted as a function of R. In case of random data the number of clusters achieving levels of the statistic R should fall exponentially as a function of R. If there are more clusters than predicted by this exponential decline, then this would be an indication that these OFP clusters represent true effect. Additionally, it is known that log likelihood values (2R) approximate a Chi-square distribution and therefore additional candidates of sizes greater than 50 clones were identified at a significance value of 0.001 (chi-square critical value is 16.266 at 3 degrees of freedom, because of clones coming from four different cDNA libraries).

2.3 Computational resources and methods

2.3.1 Publicly available resources applied

Publicly available medaka EST data (149,697 mRNAs and ESTs in Spring 2005) from GenBank/Ensembl were accessed via the GCG package from locally provided EST databases. Protein sequences were available by a local non-redundant database (nrprot). Later this database was substituted by a locally downloaded UniRef (UniProt Non-redundant Reference) database from EMBL-EBI (ref. http://www.ebi.ac.uk/uniref). Uniref100 contains 2,151,386 entries, redundancy of this database is further reduced in UniRef90 and UniRef50, where no pair of sequences shows $> 90\%$ or $> 50\%$ sequence identity. UniRef databases are built from UniProt (Universal Potein Resource), which is a repository of protein sequence and function created by joining information contained in Swiss-Prot, TrEMBL, and PIR.

TIGR gene indices [Quackenbush et al., 2001] are an attempt from The Institute for Genomic Research (TIGR), Rockville, MD, USA to identify and classify transcribed sequences in several species using publicly available EST and gene sequence data [Quackenbush et al., 2001]. Individual databases are updated and released three times yearly if the number of ESTs for that species has increased more than 10% or 25,000 items, whichever is fewer, available at http://www.tigr.org/tdb/tgi. For clustering of ESTs mgblast is used for preclustering and PTA for contig building (see 1.2.3.1) to produce a set of unique, virtual transcripts, so-called tentative consensus (TC) sequences. Further annotation of TCs are provided within the TC report. Sequences of all available fish sets and

Species version No.	Unigene set ID	No of unigene seqs	No of GO annotated seqs
Oryzias latipes 5.0	OlGI	26,689	2,289
Takifugu rubripes 2.0	FGI	11,112	534
Danio rerio 16.0	ZGI	93,442	6,423
Ictalurus punctatus 6.0	CfGI	23,262	1,112
Haplochromis chilotes 1.0	HchGI	6,177	352
Fundulus heteroclitus 2.0	FhGI	18,536	648
Oncorhynchus mykiss 4.0	RtGI	50,773	3,562
Haplochromis sp. "red tail sheller" 1.0	HsGI	6,305	279
Salmo salar 2.1	AsGI	31,341	1,623
Astatotilapia burtoni 1.0	AbGI	2,717	67

Table 2.3: Unigene sequences and GO annotations available for different fish systems in the TIGR unigene data sets.

their annotation (see table 2.3) were used to annotate EST sequences obtained in this work, and additionally other vertebrate model systems were included into annotation analyses (table 2.4).

To obtain full protein sets of various organisms, data was downloaded from Ensemblv30 in April 2005 ([Hubbard et al., 2005]).

2.3.2 Perl scripts

Different Perl Scripts were developed to automate analyses, either by starting programs in batch mode or by simplifying the analysis of outputs from different applied computer programs, e.g. little scripts were written to extract BLAST outputs. Also the development of the MedakaDB and its publication on the internet was done with Perl/CGI Scripts. Important Scripts applied during this project are summarised in table 2.5.

2.3.3 GCG and EMBOSS software packages

GCG (Accelrys Inc., subsidiary of Pharmacopeia, Inc.) and EMBOSS (European Molecular Biology Open Software Suite; developed at the MRC UK HGMP Resource Centre, Hinxton, Cambridge, UK [Rice et al., 2000]) are two large software packages, containing several programs for DNA and protein analysis and much more. GCG was mostly used as an interface to locally provided sequence databases, like emest1 (EST sequences) and emest2 (EST sequences) and nrprot (non-redundant protein sequence set compiled from databases Swiss-Prot, TrEMBL, and PIR). The EMBOSS package was mainly used for EST analysis, where programs of that package were started within self-made perl scripts.

Species version No.	Unigene set ID	No of unigene seqs	No of GO annotated seqs
Bos taurus 11.0	BtGI	108,743	7,037
Caenorhabditis elegans 9.0	CeGI	30,919	3,518
Drosophila melanogaster 10.0	DGI	36,289	7,681
Gallus gallus 10.0	GgGI	113,951	7,563
Homo sapiens 15.0	HGI	835,626	34,452
Mus musculus 14.0	MGI	777,505	24,280
Rattus norvegicus 13.0	RGI	147,056	10,912
Xenopus laevis 9.0	XGI	77,599	7,748
Sus scrofa 11.0	SsGI	104,327	6338

Table 2.4: Unigene sequences and GO annotations available for different model systems in the TIGR unigene data sets.

Name	Function
`Ace.pl`	Edit of `*.ace` files containing the `cap3` clustering result. Provides contig identifier and clones belonging to a cluster.
`Calc_Divindex.pl`	Calculates the div index for a list of files containing back-hybridisation results.
`Calc_Rearray.pl`	Calculates clones which may be rearrayed for further experiments. It avoides taking clones from clusters which were already sequenced (these are provided in an additional file).
`Cluster_ESTs.pl`	May be used to cluster a small amount of ESTs. Sequence data and sequence quality data is extracted from `*.abi` files with GCG `extract_seq`. `Crossmatch` is run to clip vector sequence. All sequences are blasted against each other by building a BLAST searchable database from all fasta files using `formatdb`. Blastoutput is parsed to identify groups showing similarity to each other. These groups are then clustered with `cap3`.
`Edit_acefiles.pl`	Same as `Ace.pl` but does additionally calculate the proportion of clone length to the length of the contig they belong to. Therefore some clones were exluded from the `cap3` clustering by these means.
`Edit_CloneAC.pl`	Compares two different clustering result files and finds contigs which were not found comparable in both files.
`Run_CAP3`	A batch of `*.fasta` files are clustered with `cap3`, providing all `cap3` output files for all input `*.fasta` files.

Table 2.5: Perl Scripts used during this project.

2.3.4 BLAST algorithm

Searching for similarities of sequences against sequence databases was done with the BLAST algorithm [Altschul et al., 1990]. BLASTX was used for querying nucleotide sequences against protein databases like nrprot or UniRef100, BLASTN was developed [Altschul et al., 1990] to run the blast-algorithm for nucleotide sequences against nucleotide sequence databases. Sets of nucleotide or protein sequences may also be compiled into a blast-searchable database, which was done by using formatdb with the option "`-p F`" for DNA and "`-p T`" for protein sequences. As an output three files are produced, which together build the database. Blast outputs were processed with different perl scripts depending on the task like extracting only the best hit.

2.3.5 Sequence analysis

ABI trace files were base called using PHRED [Ewing et al., 1998, Ewing and Green, 1998], sequence and base quality files were trimmed [Poustka et al., 2003] for base quality values lower than 25 on average and for vector sequences using CrossMatch (Green, unpublished; http://www.phrap.org), calculated by S. Hennig as described in [Poustka et al., 2003]. Also sequences shorter than 60 bp were excluded from further analyses. Prior to clustering, sequences were scanned for the presence of known repeat sequences from medaka, fugu and zebrafish with RepeatMasker (Smit and Green, unpublished; http://www.repeatmasker.org).

2.3.6 EST clustering

The efficiency of OFP clone selection may be evaluated by EST clustering. The EST clustering pipeline tgicl [Pertea et al., 2003] was used with default parameters [Pertea et al., 2003], running a modified megablast [Zhang et al., 2000] and CAP3 [Huang and Madan, 1999] for assembly of ESTs. As input data besides sequence data also base quality data may be provided, where both files are in FASTA format. Sequence input has to be trimmed for low quality bases and vector sequences before running tgicl. The result is written to several files containing singleton and cluster sequences, the base quality values for calculated consensus sequences of clusters and also a description of the sequence alignment in *.ace files. The latter can be visually inspected by clview [Pertea et al., 2003], also available from TIGR institute (see figure 3.7). The tgicl result was further changed by repeating the CAP3 clustering of clusters originated from the same precluster. Additionally blastn of all against all clusters and singletons was run, identifying groups with high similarities. These groups were again clustered with CAP3. Applying these two methods additional clusters were merged.

2.3.7 Sequence annotation

High-quality ESTs were, if possible, assigned functions by gene ontology (GO; [GO Consortium, 2000] and [Ashburner and Lewis, 2002]). Firstly, a BLASTX was run with an e-value of $1.0e^{-20}$ against

a subset of the SP-TREMBL database of GO annotated entries that had been downloaded at the GO website (www.geneontology.org). As this database contained only 6,423 entries from zebrafish and 2,289 for medaka (2005), the GO annotations from TIGR for all available fish species (medaka, fugu, zebrafish, catfish, *Haplochromis chilotes*, killifish, rainbow trout, *Haplochromis sp.*, Atlantic salmon and *Astatotilapia burtoni*; see 2.3.1) were included in the GO analysis of our ESTs. Sequences not assigned to certain GO categories using the first approach were then blasted to GO annotations for other species (cattle, *Caenorhabditis elegans*, *Drosophila melanogaster*, chicken, human, mouse, rat, *Xenopus laevis*, pig; see 2.3.1) contained in the TIGR unigene data sets. All GO results could then be easily viewed by the GOblet system [Groth et al., 2004], developed at the MPI for Molecular Genetics. For all remaining sequences, BLASTX against a local non-redundant protein set (nrprot or UniRef100) constructed from UniProtKB/SWISS-PROT, UniProtKB/TrEMBL and PIR [Bairoch et al., 2005] was run with an e-value of $1.0e^{-05}$ and the best hits were extracted. Sequences with GO-annotated UniProt sequences were identified and added to the GO-annotated sequences. Also, blastn runs against all remaining TIGR fish sequences without GO annotation was performed.

2.3.8 Identification of alternative splices

Candidate sequences were aligned to the draft sequence of the Medaka genome using the UT Genome Browser (Medaka) at http://medaka.utgenome.org or simple BLASTN against the medaka scaffolds at http://dolphin.lab.nig.ac.jp/medaka. From the UT Genome Browser view additional publicly available data was collected and included in further analyses. These EST evidences were again aligned to the respective Medaka scaffold using EMBOSS est2genome (based on est_genome; [Mott, 1997]) and PipMaker [Schwartz et al., 2000]. Est2genome calculates the alignment by considering not only nucleotide similarity between query and subject but also possible splice sites. It strictly expects introns to start with GT and to end with AG (or CT and AC if the splicing direction is reversed), otherwise penalty costs are calculated. This is based on the finding that within the human genome of 53,295 introns, 98.12% use GT at the 5' splice site and AG at the 3' site, whereas only 0.76% use GC-AG and 0.10% use AT-AC (International Human Genome Sequencing Consortium, 2001). Exons obtained from est2genome were converted into PipMaker exon description by a Perl-Script provided by D. Groth (unpublished, 2003) and then provided as underlay-files to PipMaker, which highlights exons with colour. Data are submitted to the PipMaker WebServer and then the percent of identity plot (pip plot) is calculated and depicted. This gives the opportunity to view the quality of the mapping result and alternative splices may easily be detected. ESTs were only kept as coding for the same gene if at least one exon or intron was recorded by est2genome with the same splice sites. Candidates were only recorded within this project's data set.

2.3.9 Setting up a project database

All data was collected in a relational database to handle the large amount of it. The database was built using SQLite (www.sqlite.org), which allows to save the database in a single flat file and therefore to avoid the need for a database server. By this method data access is also greatly enhanced and data can be queried by SQL. With the help of the DBD::SQLite module (www.CPAN.org) and DBI Perl-modules the database building and loading of data from various text files was done with Perl. To provide an easy readable output a CGI-script was written with the help of Perl to create a web-based interface to this database.

Chapter 3

Results

3.1 Creating a medaka gene catalogue

Research on genes active during a certain process, e.g. development of an organism, or during some event in an organism's life, e.g. tumorigenesis, requires certainly the knowledge of all responsible genes. One step towards elucidation of such scenarios includes to establish a gene catalogue for the whole organism, or some tissue at a certain developmental or adult stage.

At the time this poject was started at the end of 2001 only a limited number of ESTs was available and only at the beginning of 2002 the TIGR institute started to calculate a unigene set for *Oryzias latipes* (OlgI), which included in its first version 41,426 ESTs from dbEST clustered into 14,024 groups and singletons [Boguski et al., 1993]. The advance of ongoing medaka, zebrafish and pufferfish EST sequencing efforts is summarised in figures 3.1 and 3.2. It is clearly visible that there is a strong increase of work done on EST projects in medaka, but on still a much smaller scale compared to zebrafish. In table 3.1 the different *Oryzias* EST projects are summarised according to their developmental stages, the strains used and also the number of submitted sequences as an indication for the scale of a certain project.

Libraries subjected to OFP analysis were already provided by our cooperation partners so no influence could be made into improvement of cDNA library production. All libraries were oligo dT primed, which is an advantage compared to working with randomly primed cDNA libraries [Poustka, 2000], because of the unique end point in most cDNA clones of the oligo dT primed library, which simplifies the OFP clustering process.

Material chosen in that project provided access to mRNA from three different embryonic stages (gastrula, neurula and organogenesis) and ovary as one adult tissue. These stages were chosen because of their high relevance towards experiments on developmental problems. Our cooperation partners are mostly involved in elucidation of developmental problems like the group of Jochen Wittbrodt at the EMBL in Heidelberg works on development of the vertebrate eye or Franck Bourrat, working in Gif-sur-Yvette, France, on development and morphogenesis of the Medaka

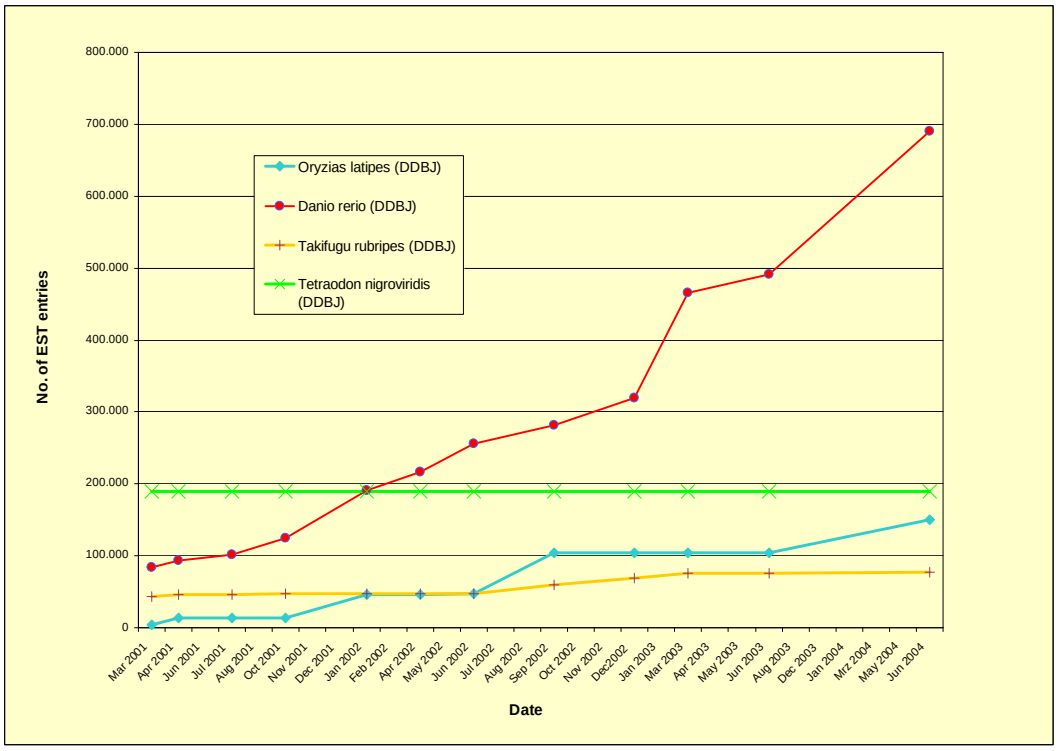


Figure 3.1: Comparison of EST sequences available for different fish model organisms in DDBJ.

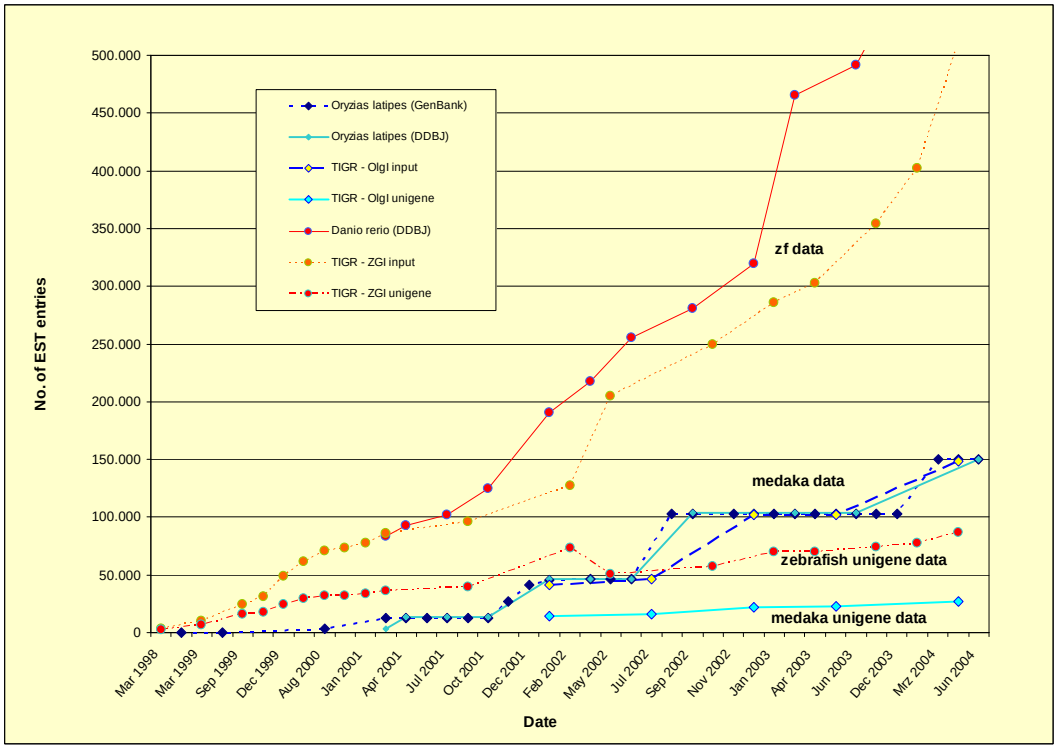


Figure 3.2: Comparison of available EST sequences for zebrafish and Medaka.

brain (see [Deyts et al., 2005] or [Candal et al., 2005]). Gastrulation is a very complicated process were by diverse movements of cell material the three germ layers are produced and determined. The eye development is initiated at late gastrula stage and at the end of gastrulation the eye field is largely determined. During neurula stage were the neural tube is built, also the evagination of optic vesicles happens. Finally during the stage of organogenesis the organs are developed. Therefore these three embryonic stages provide us with specific transcripts of great importance for the early development of the medaka embryo. On the other hand the ovary as an adult tissue provides not only information about an adult organ but also additional transcripts because of the accumulation of maternal information (RNAs and proteins) in oocytes.

AC Numbers	NO	Year	library name	strain	developmental stage	tissue	reference
AV669110, AV670 – AV671	1890	2000	cell-line cDNA library	HNI		cell line	Naruse K., Tanaka M., Shima A., Mitani H., unpublished
AU167 – AU172	5280	2001	Ol-br-ad cDNA	HNI	adult	brain	Mita K., Ishikawa Y., Yamauchi M., unpublished
AU178941 – AU180213	1273	2001	Medaka liver cDNA library (OLe)	HNI	adult	liver	Naruse K., Mitani H., Tanaka M., unpublished
BJ	28697	2001	MF01SSA cDNA	Hd-rR	stage 20-25	whole embryo	Kohara Y., Shin-i T., Kimura T., Narita T., Jindo T., Takeda H., unpublished
AU240 – AU241767	1400	2002	UV irradiated OLHNI cell line cDNA library (Olc)	HNI		cell line	Naruse K., Mitani H., Tanaka M., unpublished
AU241768 – AU242851	1084	2002	Medaka ovary cDNA library (OLd)	HNI	adult	ovary	Naruse K., Mitani H., Tanaka M., unpublished
AU242852, AU243, AU244	1634	2002	Medaka eye cDNA library (SNK01)	WT	adult	eye	Sanaka E., Hori H., Naruse K., Mitani H., Tanaka M., unpublished
BJ487 – BJ517 – BJ518596	31592	2002	MF01FSA cDNA	d-rR	fry stage 40	whole embryo	Kohara Y., Shin-i T., Kimura T., Narita T., Jindo T., Takeda H., unpublished
BJ518597 – BJ543	25196	2002	MF01SSB cDNA	Hd-rR	segmentation stage 20-25	whole embryo	Kohara Y., Shin-i T., Kimura T., Narita T., Jindo T., Takeda H., unpublished

continued on next page

AC Numbers	NO	Year	library name	strain	developmental stage	tissue	reference
BJ704106 – BJ727005	22900	2004	MF01FFA cDNA	Hd-rR	fry stage 40	whole embryo	Kohara Y., Shin-i T., Kimura T., Narita T., Jindo T., Takeda H., unpublished
BJ727006, BJ73 – BJ75	23697	2004	MF015DA cDNA	Hd-rR	organogenesis, stage 35	whole embryo	Kohara Y., Shin-i T., Kimura T., Narita T., Jindo T., Takeda H., unpublished
BJ000001– BJ028697, BJ487005– BJ543792, BJ704106– BJ750702	132,082	2004	SSA, FSA, SSB, 5DA, FFA	d-rR and Hd-rR	stage 23, 35 and 40	whole embryo	[Kimura et al., 2004]

Table 3.1: Summary of EST projects on *Oryzias latipes* which submitted more than 1000 ESTs to dbEST database. AC numbers - range of accession numbers given to that project. NO - number of sequences submitted.

3.1.1 Normalisation of 26,880 medaka gastrula clones by OFP

One filter set containing PCR products of 26,880 medaka gastrula clones was prepared and subjected to oligonucleotide fingerprinting analysis which resulted in successful fingerprints for 22,848 cDNA inserts with hybridising 160 oligonucleotides, of which 73 had been calculated from known medaka sequences (2.2.1.3) and 87 were from a standard set, including 95 oligonucleotides also used in the zebrafish project (see appendix A for details; [Clark et al., 2001]). The remaining clones for which no fingerprint was calculated showed no or little hybridisation results, which is most probably caused by unsuccessful PCR or unsufficient transfer of PCR product to the Nylon membranes. The fingerprints were grouped into 1860 clusters, ranging in size from 2 to 465 clones. 5983 clone inserts were left as singletons (table 3.2). Thus, 74% of clones for which fingerprints were obtained fell in clusters, whereas 26% remained as singletons.

As previously estimated the majority of transcripts are expressed at a level as low as one copy per cell [Velculescu et al., 1999], but others state that genes with midrange profile, expressed at a high level in a subset of the tissues, and at a much lower level or not at all in other tissues, make up more than 50% of all expressed genes [Yanai et al., 2004]. Both results came from analyses of the human transcriptome. Within the Medaka project it was found that clusters with less than six clones encompassed 7216 transcripts corresponding to 92.0% of all clusters and included 9315 clones (41% of all clones). Only 25 large clusters were obtained that contain more than 70 clones per cluster. These clusters comprised 3990 clones (17.5% of all clones) but only 0.3% of all

Cluster size	Amount of OFP cluster
1	5983
2	738
3	240
4	139
5	116
6	79
7	74
8	49
9	48
10	37
>10	116
>15	65
>20	42
>25	19
>30	19
>35	16
>40	13
>45	5
>50	15
>60	5
>70	2
>80	7
>90	6
>100	3
>200	5
396	1
465	1

Table 3.2: Cluster size distribution of first fingerprinting experiments.

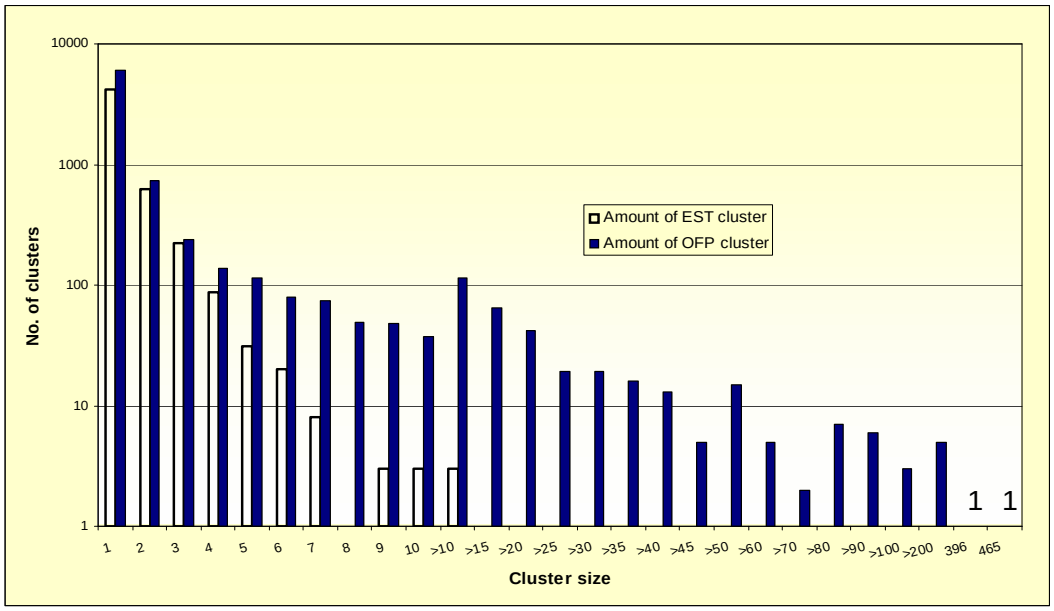


Figure 3.3: Cluster size distribution of first OFP clustering compared with EST clustering of 6909 sequences obtained from the fingerprinted set. The amount of clusters with a certain cluster size is depicted. OFP cluster size ranges from one clone (singleton) to one cluster containing 465 clones.

Gene (Clone)	Copies	Clustered (%)	Singletons (%)	Diversity index (δ)
MedGastrula_222G16	4	2 (50.00)	0 (0.00)	0.500
MedGastrula_222M21	314	289 (92.00)	3 (0.96)	0.085
MedGastrula_226B14	13	8 (61.54)	2 (15.38)	0.460
MedGastrula_228A10	1	1 (100.00)		0.000
MedGastrula_232H4	77	71 (92.21)	2 (2.60)	0.095
MedGastrula_234H5	11	7 (63.63)	2 (18.18)	0.484
MedGastrula_247K2	253	208 (82.21)	2 (0.79)	0.136
MedGastrula_265C21	7	3 (42.86)	3 (42.86)	0.758
MedGastrula_273P10	9	9 (100.00)	0 (0.00)	0.000
MedGastrula_275G23	5	2 (40.00)	1 (20.00)	0.828
MedGastrula_283I23	8	5 (62.50)	2 (25.00)	0.516
MedGastrula_284K13	11	10 (90.90)	0 (0.00)	0.127
MedGastrula_222G11	2	1 (50.00)	0 (0.00)	1,000
Total	715	616 (86.15)	17 (2.38)	0.384

Table 3.3: Statistics of backhybridisation experiments to control the quality of first fingerprint experiments. One cDNA clone, e.g. MedGastrula_222M21 is hybridised to the gastrula clone filter. This hybridisation resulted in 314 positive clones. Out of these clones 289 were clustered by OFP analysis together into one cluster, so 92% of copies were really clustered and the other copies were falsely put into different clusters, and 3 clones were falsely assigned as singletons. A diversity index of 0.384 was calculated after hybridisation of 13 cDNA control clones.

transcripts. 602 medium-sized clusters (7.7% of transcripts) with 6-70 members contained 9543 clones (42% of all clones). This is depicted in figure 3.3, which shows the cluster size distribution of all OFP clusters. The largest clusters were normalised 160 times, but all clones on average only 2.9-fold.

Clustering results were quality-controlled. Randomly chosen cDNA clones were labeled and hybridised to colony clone filters at stringent conditions. Clones that gave positive signals with a given cDNA probe were considered as a true gene cluster. These groups were compared to the OFP clustering. A total of 13 cDNA clone families containing 715 cDNAs with sizes varying from 1 to 314 members were detected by these experiments (tab. 3.3).

A fraction of 86% of hybridised cDNA fragments fall into pure clusters, where a cluster is defined as pure if its cDNA clones originate to a large extent from only one OFP cluster and only some outlier come from different OFP clusters. In 17 cases (2.4% of all clones), clones had falsely been assigned as singletons. As a quality measure for OFP clustering, a diversity index δ ranging from 0 to 1 was calculated as described in [Herwig et al., 1999] and subsection 2.2.1.7. This statistic evaluates the splitting of gene clusters. The diversity equals 1 if all cDNA clones, positive in one hybridisation with a cDNA probe, had been placed into different OFP clusters and a diversity index of 0 is obtained when all these clones belong to the same OFP cluster. For the clustering of gastrula clones an average diversity index of 0.384 (table 3.3) was calculated.

length of sequences (bp)	NO of sequences
60-99	84
100-149	46
150-199	81
200-249	103
250-299	128
300-349	179
350-399	205
400-449	232
450-499	268
500-549	301
550-599	345
600-649	419
650-699	620
700-749	1102
750-799	1517
800-849	1035
850-899	219
900-963	25
mean length of all 6909 traces: 655bp	

Table 3.4: Distribution of sequence read length obtained after the removal of low quality sequence stretches and vector sequence from the 6909 ESTs obtained from clustering experiments.

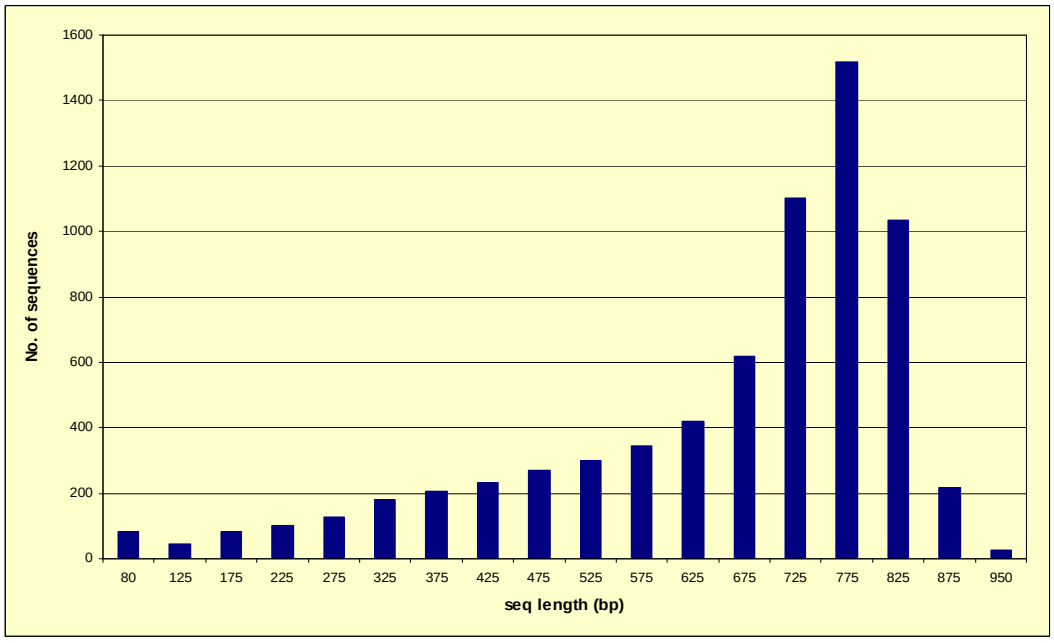


Figure 3.4: Distribution of sequence read length obtained after the removal of low quality sequence stretches and vector sequence from the 6909 ESTs obtained.

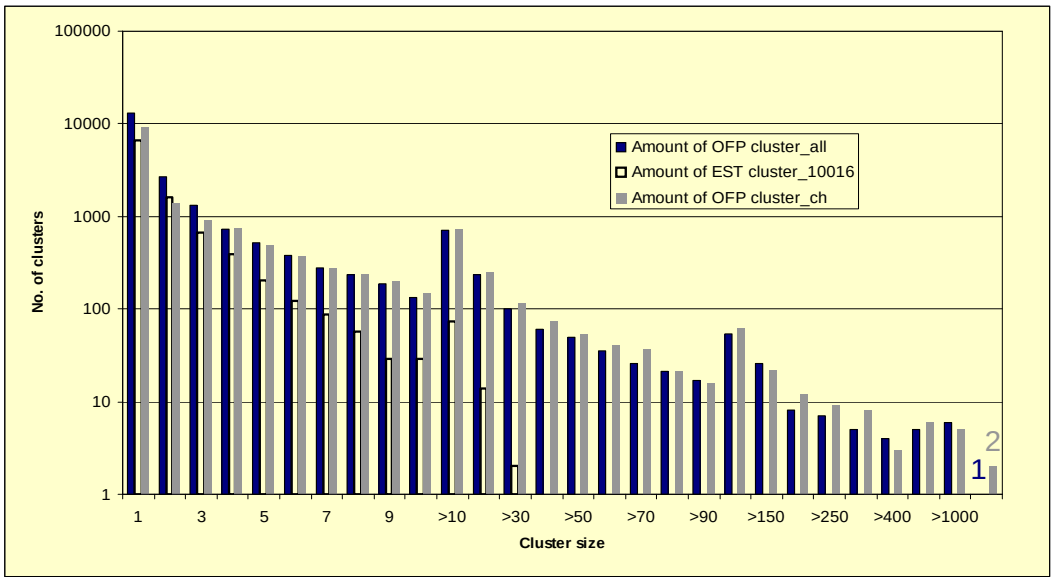


Figure 3.5: Cluster size distribution of OFP clustering of all clones (OFP cluster_all) compared with EST clustering of high quality sequences obtained from the fingerprinted and rearrayed set (EST cluster_10016). The amount of clusters with a certain cluster size is depicted. The OFP clustering result was changed for further experiments according to the EST clustering results (OFP cluster_ch).

7680 cDNA clones were chosen and re-arrayed for sequencing, each representing either one of 5324 singletons or one of 2356 OFP clusters. As representative of one fingerprinting cluster the clone showing highest similarity to the calculated consensus fingerprint was selected for sequencing. Such a clone was chosen because it represented most likely the corresponding gene and it reduced the risk to sequence a clone that had been falsely assigned to a cluster, since such clones were found to be rather distant from the average fingerprint.

A total of 6909 high-quality 5' EST sequences were obtained with an average length of 655 bp (see table 3.4 and figure 3.4) after base calling with PHRED and trimming of vector and low quality sequence. These ESTs were subjected to EST clustering with the tgicl-package [Pertea et al., 2003], resulting in 4210 singletons and 1006 clusters containing 2699 reads (fig. 3.3). Therefore, 5216 (75%) of the EST sequences represent different genes and one gene is on average represented by 1.3 cDNA clones. The redundancy of the library was reduced 5-fold (26,880 clones subjected to OFP analysis vs. 5216 unique sequences obtained) by OFP normalisation and EST clustering. 63% of the OFP singletons for which high-quality ESTs were produced, resulted in unique EST sequences.

3.1.2 Combined EST analysis of 119,040 medaka cDNA clones

A particular advantage of the OFP technology is the possibility of normalisation across different libraries from the same species. OFP data sets from different cDNA libraries can be merged, such that genes expressed at high level in many tissues and therefore present in any cDNA library,

stage/library	ovary	gastrula	neurula	organo-genesis	total
Clones analysed	20,352	36,864 and 26,880	11,520	23,424	119,040
Clones successfully fingerprinted	20,219	30,434 and 22,628	8,595	18,784	100,660
Clones in OFP clusters	18085	45434	7871	16785	88,175
Clones remaining as singletons before EST clustering (% of successfully fingerprinted clones)	2134 (10.6%)	7628 (14.4%)	724 (8.4%)	1999 (10.6%)	12,485
Clones selected for sequencing	2857	7680 and 6741	1086	2756	21,120
Number of high-quality reads	2389	6909 and 5663	953	2464	18,378
Clones remaining as singletons after all EST clustering experiments (% of successfully fingerprinted clones)	1025 (5.1%)	4281 (8.1%)	374 (4.4%)	1030 (5.5%)	6710

Table 3.5: Number of clones per library subjected to OFP analysis.

e.g. house keeping genes, are efficiently eliminated. This enlarges the probability to identify also rare transcripts. Additionally, the analysis of different libraries provides the opportunity to compare transcription levels in different tissues or developmental stages under subject. Therefore additionally to the gastrula clones from the pilot experiment described above further clones of various cDNA libraries were included in the OFP approach: 36,864 gastrula clones (in total 63,744 gastrula clones), 11,520 neurula (stage 17-23) clones, 23,424 clones from organogenesis (stage 24-33) and from the adult tissue ovary 20,352 clones were included (tab. 3.5). All clones were arrayed on 4 additional filter sets each containing 72 times 384-well plates. Filters of the first round were also included, so in total 119,040 medaka cDNA clone inserts were subjected to OFP analysis.

Probably because of differences in sequence complexity in different libraries only 124 oligonucleotides (85 coming from the standard set, of these 70 also used for zebrafish, 39 medaka specific; see appendix A) were successfully hybridised to all 5 filter sets. OFP analysis resulted in successful fingerprints for 100,660 clones which were grouped into 7826 clusters of sizes between 2 and 1648

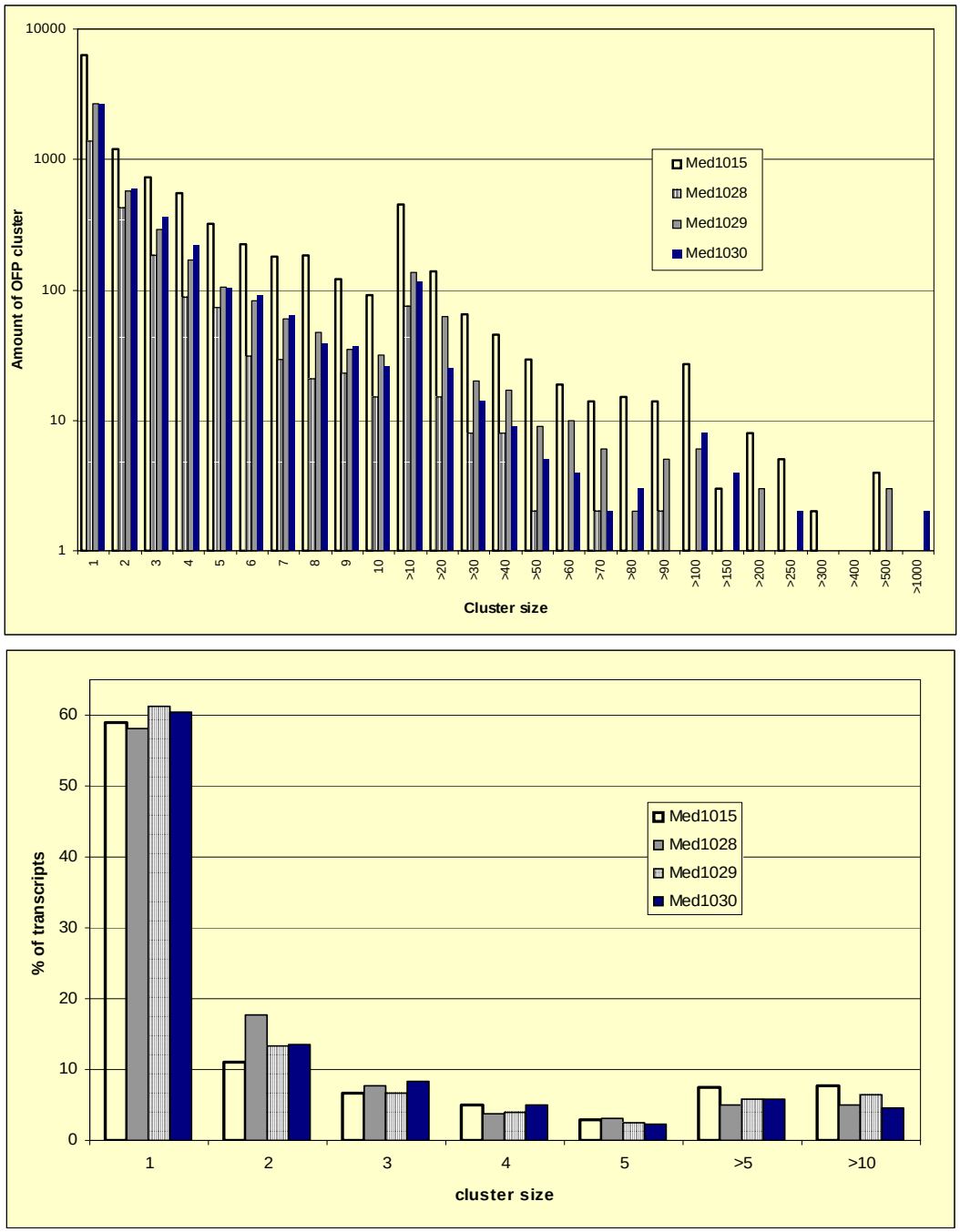


Figure 3.6: Cluster size distribution between the four different libraries. (a) Total number of OFP cluster is depicted. (b) Same data as in (a), but the amount of clusters is compared to the number of transcripts included from each of the four libraries.

clones and 13,117 clone inserts left as singletons, together representing 20,943 transcripts. This number of clusters represented the optimal solution in high- and low-stringency clustering runs. Most clusters were small: 18,354 OFP clusters (87.6%) were of a size smaller or equal to five; only 180 OFP clusters (0.9% of clusters) were of sizes greater than 70 (see cluster distribution in figure 3.5), mostly coding for ribosomal proteins, globins, parvalbumin, Gamma-crystallin, ubiquitin or keratin (see table 3.8).

Size distribution was also compared between the four different libraries, where it seems like that in neurula big clusters are missing, but this is caused by the small number of neurula clones included in the analysis. In the case where the cluster sizes are depicted in comparison to the percentage of transcripts included, there are no significant differences between the four libraries noticeable (fig. 3.6).

The quality of clustering results was examined by hybridisation of 26 cDNA clones against the fingerprinted libraries (tab. 3.6). This time only 59% of positively hybridised clones fell into pure clusters, meaning most clones originated from one OFP cluster, and 112 clones (4%) were falsely assigned as singletons. The diversity index was calculated as 0.481, indicating that genes were represented on average by 2 OFP clusters in this result.

Because of low numbers of hybridised oligonucleotides, choosing new clones for sequencing, the OFP clustering program was run with different parameters concerning stringency of clustering. Numbers of clusters and singletons represent optimal solution for clustering. For rearraying and subsequent sequencing of singletons, clones were chosen which were left as singletons in high-stringency runs as well as in low-stringency runs. Only representatives of clusters were chosen which were comparably clustered in other runs, i.e. containing on average the same clones. Also any clones already been sequenced in the medaka gastrula pilot experiment described above were not re-sequenced. From the remainder of clusters and singletons, 13,440 cDNA clones were chosen for sequencing, including 6089 representatives of OFP cluster and 7351 singletons. For these 11,469 high-quality traces were obtained successfully.

These 5' ESTs were again vector and quality trimmed to an average size of 511 bp (tab. 3.7) and then clustered with the tgicl package [Pertea et al., 2003]. Clustering of only the new sequences gave 1876 clusters and 5852 singletons, such that one gene is on average represented by 1.48 clones in the data set. Out of 9565 fingerprinting singletons for which successfully traces were obtained, 4382 (46%) are not clustered with any other sequences and therefore remain as EST singletons. Finally, clustering all 18,378 high-quality sequences (average length of 565 bp) together, 3268 clusters and 7274 singletons were obtained, resulting in 10,542 unique sequences (57%). The corresponding cDNA clones were rearrayed into the Medaka Rearray2 library.

Still observing redundancy within the data set another clustering step was included by grouping clusters and singletons according to high sequence similarity using the BLASTN algorithm (e-value 1.0e-20) and clustering each of these groups with CAP3. Additionally the clustering of preclusters calculated within the tgicl-package was repeated by clustering the different contig consensus

Gene (Clone)	Copies	Clustered (%)	Singletons (%)	Diversity index (δ)
1015_295C19	63	30 (47.62)	4 (6.35)	0.472
1015_322A17	54	9 (16.67)	5 (9.26)	0.740
1015_323I3	11	6 (54.54)	1 (9.09)	0.592
1015_329H23	78	34 (43.59)	1 (1.28)	0.401
1015_354F3	6	1 (16.67)	1 (16.67)	1.000
1015_354O7	262	77 (29.39)	8 (3.05)	0.497
1028_08H14	9	8 (88.89)	1 (11.11)	0.159
1028_18D17	196	155 (79.08)	7 (3.57)	0.216
1028_18G9	107	79 (73.83)	2 (1.87)	0.286
1028_18L12	17	11 (64.71)	0 (0.00)	0.424
1028_22B24	26	6 (23.08)	5 (19.23)	0.775
1028_24A14	625	488 (78.08)	16 (2.56)	0.217
1028_26E1	39	13 (33.33)	4 (10.26)	0.718
1029_01L3	197	34 (17.26)	8 (4.06)	0.584
1029_04C17	57	27 (47.37)	6 (10.53)	0.553
1029_13M1	15	5 (33.33)	0 (0.00)	0.687
1029_17D2	102	69 (67.65)	4 (3.92)	0.345
1029_42A2	6	3 (50.00)	0 (0.00)	0.693
1029_43P7	231	195 (84.42)	7 (3.03)	0.171
1029_47P23	86	48 (55.81)	7 (8.14)	0.447
1029_59M14	127	114 (89.76)	5 (3.94)	0.113
1030_03D24	63	50 (79.37)	3 (4.76)	0.251
1030_19L21	17	10 (58.82)	2 (11.76)	0.522
1030_22H9	123	23 (18.70)	4 (3.25)	0.601
1030_22M18	134	87 (64.93)	8 (5.97)	0.318
1030_44E7	50	15 (30.00)	3 (6.00)	0.728
Total	2701	1597 (59.13)	112 (4.15)	0.481

Table 3.6: Statistics of backhybridisation experiments. See table 3.3 for details. The diversity index for the fingerprinting across all libraries was estimated with 0.481 according to hybridisation of 26 cDNA control clones. Legend for clones: 1015 - gastrula clones, 1028 - neurula, 1029 - ovary, 1030 - organogenesis.

length of sequences (bp)	NO of sequences
80-99	88
100-149	194
150-199	276
200-249	361
250-299	451
300-349	512
350-399	582
400-449	676
450-499	1015
500-549	1563
550-599	2046
600-649	2020
650-699	1189
700-749	390
750-799	97
800-845	9
mean length of all sequences: 511bp	

Table 3.7: Distribution of sequence read length.

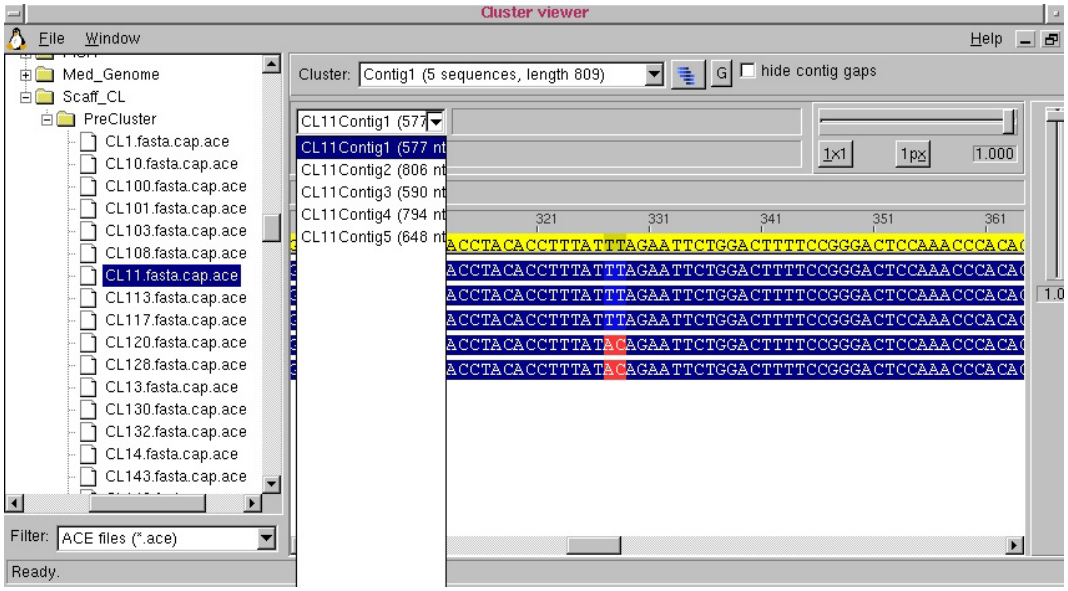


Figure 3.7: CAP3 alignment of precluster CL11Contig1, CL11Contig2, CL11Contig3, CL11Contig4 and CL11Contig5 viewed with clview [Pertea et al., 2003].

OFP cluster	OFP cluster size	EST contig	EST contig size	Normalisation	Annotation
OFP7	1855	CL1Contig6	37	50.1	similar to SP|P02383|RS26_HUMAN 40S ribosomal protein S26
OFP3	1434	CL2Contig12	31	46.3	similar to UniRef100_Q9PTQ5 Hypothetical protein [Oryzias latipes]
OFP4	1254	CL2Contig1	24	52.3	similar to UniRef100_Q9W7D3 Hypothetical protein [Oryzias latipes]
OFP5	1239	CL77Contig1	12	103.3	similar to UP|Q8AYQ8 Alpha-type globin
OFP8	903	CL153Contig1	9	100.3	similar to UP|Q8UUS2 Parvalbumin
OFP9	824	CL15Contig3	16	51.5	similar to UP|Q7ZTS2 Cldni protein
OFP12	663	CL6Contig3	15	44.2	similar to UP|Q31555 Gamma-crystallin M2-1
OFP13	594	CL9Contig1	29	20.5	similar to UP|Q9PT73 Apolipoprotein E (Fragment)
OFP16	517	CL131Contig1	11	47.0	similar to UP|Q8AYQ6 Beta-type globin
OFP14	491	CL4Contig17	3	163.7	similar to UP|Q9PT09 Ubiquitin
OFP15	433	CL3450Contig1	2	216.5	similar to SP|Q8IUB9|K191_HUMAN Keratin associated protein 19-1 (High tyrosine-glycine keratin associated protein 19.1) (Fragment)
OFP30	401	CL33Contig1	19	21.1	similar to UP|Q8JFQ5 Keratin 12 (Fragment)
OFP17	393	CL135Contig1	9	43.7	similar to SP|Q9YGF2|RS6_ONCMY 40S ribosomal protein S6
OFP19	387	CL31Contig1	19	20.4	similar to UP|O57691 Fatty acid binding protein H6-isoform
OFP20	381	CL17Contig1	27	14.1	similar to SP|P25111|RS25_HUMAN 40S ribosomal protein S25

Table 3.8: The largest 15 OFP clusters are listed together with their corresponding EST contig. The normalisation success was calculated for each cluster.

Contig	Contig size	Annotation
CL3Contig5	46	similar to GB|AAA79835.1|1036735|HSB1ITGB4 beta 1 integrin isoform A *Homo sapiens*
CL1Contig6	37	similar to SP|P02383|RS26_HUMAN 40S ribosomal protein S26
CL2Contig12	31	similar to UniRef100_Q9PTQ5 Hypothetical protein [Oryzias latipes]
CL8Contig1	30	similar to UP|CYB_SALSA (Q35925) Cytochrome b
CL9Contig1	29	similar to UP|Q9PT73 Apolipoprotein E (Fragment)
CL11Contig1	28	similar to UP|Q7VCS4 Predicted protein
CL10Contig1	27	similar to ferritin H3 [Oryzias latipes]
CL17Contig1	27	similar to SP|P25111|RS25_HUMAN 40S ribosomal protein S25
CL12Contig1	26	similar to UP|Q9DFT6 Beta tubulin
CL13Contig1	26	similar to UP|Q8C962 Odd Oz/ten-m homolog 1 (Fragment)
CL20Contig1	25	similar to SP|Q9YIC0|EF1A_ORYLA Elongation factor 1-alpha (EF-1-alpha). Oryzias latipes;
CL2Contig1	24	similar to UniRef100_Q9W7D3 Hypothetical protein [Oryzias latipes]
CL18Contig1	23	similar to UP|NU1M_BRARE (Q9MIZ0) NADH-ubiquinone oxidoreductase chain 1
CL19Contig1	23	similar to UP|Q90W75 Type II keratin E2
CL1Contig7	23	similar to UP|AAP20215 40S ribosomal protein S24
CL23Contig1	22	similar to GB|AAR22526.1|38503465|AY459292 keratin 17n Mus musculus
CL35Contig1	22	NA

Table 3.9: The largest EST cluster, containing more than 20 ESTs, are listed together with their assigned functions.

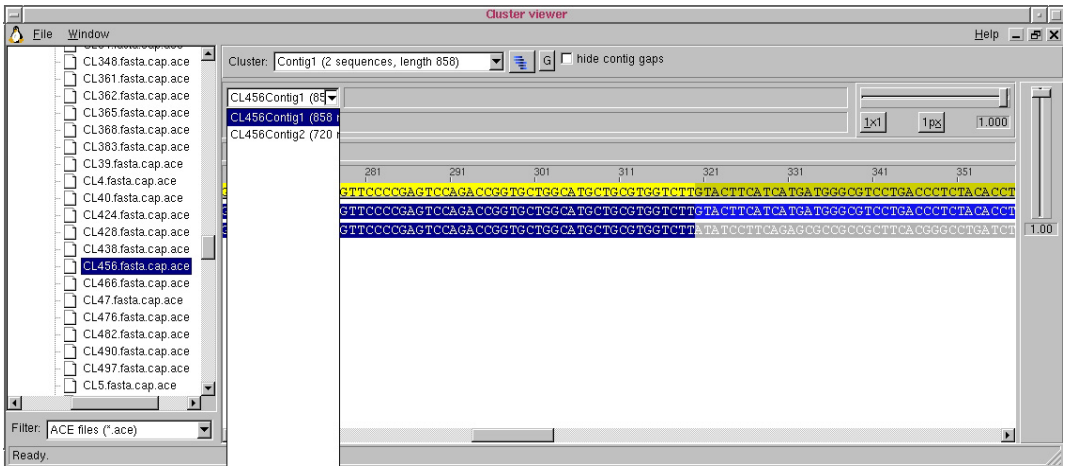


Figure 3.8: CAP3 alignment of cluster CL456Contig1 and CL456Contig2 viewed with clview [Pertea et al., 2003].

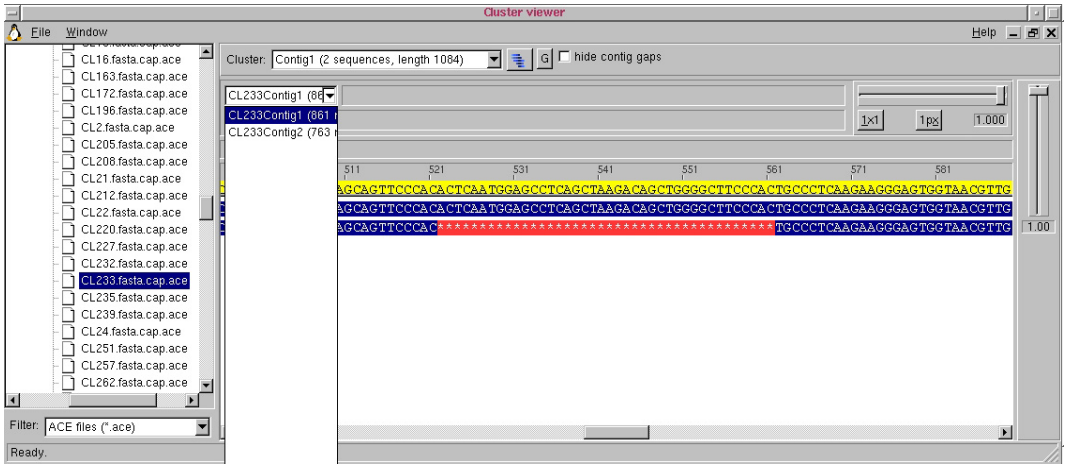


Figure 3.9: CAP3 alignment of cluster CL233Contig1 and CL233Contig2 viewed with clview [Pertea et al., 2003].

sequences within these preclusters. All calculated CAP3 contigs were visually inspected with the help of clview [Pertea et al., 2003]. Figures 3.7, 3.8 and 3.9 show different possibilities which came up after clustering different preclusters. In cases where the CAP3 contig shows perfect agreement, depicted in figure 3.7, these preclusters were joined into one cluster. This was not the case in the two other examples. In figure 3.8 the two preclusters have only a part of sequence in common and in figure 3.9, where a big gap is visible, which is maybe due to an alternative splicing event, the preclusters were also not merged into one contig. The same controls were applied to all other CAP3 alignments. Unperfect alignments were provided for further experiments as candidates for alternative splicing events. By these means the first set was further reduced to 10,016 unique sequences (3306 clusters and 6710 singletons) with a contig size of up to 46 ESTs in one cluster (see figure 3.5 and table 3.9).

These results for EST clustering were applied to correct the OFP clustering. The OFP cluster

size distribution was changed, obtaining bigger clusters. The number of singletons and clusters with 2 and 3 clones was significantly reduced. The size of bigger clusters stayed approximately the same or was only slightly enlarged (fig. 3.5). Reminding that 100,660 cDNA clones were successfully subjected to OFP analysis and 10,016 unique sequences were left after EST clustering, the redundancy of the four unnormalised libraries was successfully reduced almost 10-fold. Considering the 100 superprevalent transcripts, which were represented by OFP cluster of sizes from 122 to 1855 clones, the occurrence of these genes was reduced up to 216-fold (OFP15 cluster, probably representing a gene for Keratin, contains 433 clones and its corresponding EST cluster, CL3450Contig1, contains 2 ESTs; see table 3.8). Noting that the OFP analysis provided 18,378 potentially different transcripts, but which were grouped into 10,016 unique sequences, it may be estimated that only 54% of obtained OFP clusters comprise different genes. This result is comparable to the OFP analysis of zebrafish cDNA libraries (68% gene diversity after OFP analysis; Clark, 2001), but better results were already obtained for sugar beet (89%; Herwig, 2002) and sea urchin (92%; Poustka, 2003).

3.2 Annotation of ESTs and EST contigs

A crucial step for further experiments is the assignment of function to the obtained sequences by similarity searches. Additionally sequences covered by repetitive sequence stretches have to be identified.

3.2.1 Repeat content

All EST sequences were searched for repetitive elements, general and fish specific. Identified repeats were divided into three groups, regions of low complexity (poly A tails were not taken into account), microsatellites and transposable elements. In 4.5% (449 of 10016 sequences) of sequences, stretches of low complexity were identified and 3.5% (349 out of 10016) contained microsatellites, meaning repeats of 1 to 6 nucleotides. The five most common microsatellites were $(T)_n$, $(TG)_n$, $(CAG)_n$, $(A)_n$ and $(CA)_n$, making up 45% of all repetitive sequence (tab. 3.10).

As already known, fishes show a great diversity of transposable elements, which are classified into SINEs, Non-LTR retrotransposons, LTR-retrotransposons, Penelope-like retrotransposons (where LTRs are optional) and DNA transposons (see [Aparicio et al., 2002] or [Volff et al., 2003]). It was estimated that Fugu and *T. nigroviridis* contain around 5600 and 3600 reverse transcriptase (pseudo)genes, respectively, which is much smaller than the around 44000 reverse transcriptase (pseudo)genes identified in the human genome [Volff et al., 2003]. Despite the lower copy number, transposable elements are much more diverse in teleosts than in the mouse and human genome.

From all 10016 sequences 171 showed similarities to transposable elements described before in the literature. These hits come from all different classes of TEs, the most common sequence found were DNA transposons covering 20872 nucleotides, followed with 10362 bp covered by non-LTR

microsatellite	copy no	bp covered
$(\mathtt{T})_n$	63	1836
$(\mathtt{TG})_n$	31	1355
$(\mathtt{CAG})_n$	23	1320
$(\mathtt{A})_n$	44	1200
$(\mathtt{CA})_n$	25	1063
$(\mathtt{CAT})_n$	13	748
$(\mathtt{ATG})_n$	16	634
$(\mathtt{TTC})_n$	6	613
$(\mathtt{GGA})_n$	20	599
$(\mathtt{GAA})_n$	15	512
$(\mathtt{TCTCTG})_n$	1	476
$(\mathtt{CTG})_n$	9	347
$(\mathtt{CCA})_n$	4	287
$(\mathtt{CAGAGA})_n$	3	267
$(\mathtt{GGGAGA})_n$	2	251
$(\mathtt{TCC})_n$	9	233
$(\mathtt{TCTA})_n$	1	179
$(\mathtt{TA})_n$	5	161
$(\mathtt{TAAA})_n$	3	159
$(\mathtt{CACG})_n$	2	146
$(\mathtt{CAA})_n$	2	140
$(\mathtt{TTA})_n$	4	137
$(\mathtt{TCCA})_n$	4	135
$(\mathtt{TCG})_n$	2	118
$(\mathtt{CGA})_n$	1	114
$(\mathtt{CACCAT})_n$	1	113
$(\mathtt{TTG})_n$	2	101
$(\mathtt{CAAA})_n$	2	100
$(\mathtt{TGG})_n$	2	99
$(\mathtt{TAA})_n$	3	86
$(\mathtt{CAACG})_n$	1	85
$(\mathtt{CAGA})_n$	2	77
$(\mathtt{CCCGG})_n$	1	75
$(\mathtt{CGTCG})_n$	1	75
$(\mathtt{TATATG})_n$	1	72
$(\mathtt{TCTG})_n$	1	71
$(\mathtt{TGGA})_n$	2	63

Table 3.10: Microsatellites identified within the unique sequence data set covering more than 60 base pairs of sequence.

transposons, then come SINEs with 5616 nucleotides and LTR retransposons with 3073 bp and finally the small class of Penelope-like transposons are found to cover 1470 nucleotides of sequence. Transposable elements showing similarities to the new sequences are summarised in table 3.11.

	Transposable element	bp covering	no of copies
SINE			
	SINE_AFC	1452	7
	SINE_DR2	115	2
	SINE_FR1A	85	1
	SINE_FR1C	162	2
	SINE_FR1D	205	2
	SINE_FR2	1321	18
	AluJo	309	1
	AluYc5	71	1
	HE1_DR1	1896	25
Non-LTR			
Restriction enzyme-like endonuclease	DONG_FR	395	3
	DONG_FR2	241	2
Apurinic/apyrimidinic endonuclease	MAUI	1266	2
	REX1_FURC	1847	3
	REX1-1_DR	1721	5
	SWIMMER1	1131	2
	L1M2	33	1
	L1MD	184	1
	L1MD3	75	1
	I-1_DR	67	1
	L1P3	435	1
	L1PA17	37	1
	L1-1_DR	161	1
	L1-10_DR	322	1
	L1-3_DR	145	1
	CR1-1_DR	33	1
	CR1-4_DR	297	1
	L2	77	1
	L3	38	1
	EXPANDER	324	2
	EXPANDER1_DR	561	3
	EXPANDER2	48	1
	REX3	442	2
	L4	90	1
	gi|19570857|dbj|AB081572.1|	297	3
	LINE_DR	95	2

continued on next page

	Transposable element	bp covering	no of copies
LTR			
retroviral	VIRDR1	350	1
	PRIMA4-int	70	1
Gypsy/Ty3	RONIN2_I	44	1
	GYPSYDR1	92	1
	SUSHII	688	2
	SAMURAI_I	269	1
Copia/Ty1	KOPI2_I	445	1
BEL/PAO family	CATCH2I_DR	547	1
No further classification	THE1C	318	1
	THE1C-int	250	1
Penelope-like			
	BRIDGE1_FR	194	1
	BRIDGE2_FR	1276	4
DNA			
TC1/Mariner: TC1 group	TZF28	57	1
	TC1DR2	89	1
	TC1DR3	1384	6
	TC1L_SS	369	2
TC1/Mariner: pogo group	TIGGU1B_FR	160	1
	TC1_FR2	524	2
	TC1_FR3	688	4
	TC1_FR5	375	2
	TC2_FR1	295	1
	TC2_FR1A	178	1
	TC2_FR2	872	3
	TC2_FR2A	140	2
	TC2_FR4	111	1
Hobo-Activator-Tag 1	Charlie3	212	1
	Charlie8	177	1
	CHAPLIN3_FR	171	2
	CHAPLIN4_FR	292	2
	CHAPLIN6_FR	723	9
	TRILLIAN1	127	1
	TOL2_OL	608	1
	FUROUSHA2	576	1
	MER5B	163	1
	MER6	11284	48
	MER6A	256	1
	Tigger2	527	1
	Tigger7	110	2
	TDR19	49	1
PiggyBac	PIGIBAKKU	124	1
IS5/Harbinger	SENKUSHA1	231	2

Table 3.11: Transposable elements identified within the unique sequence data set.

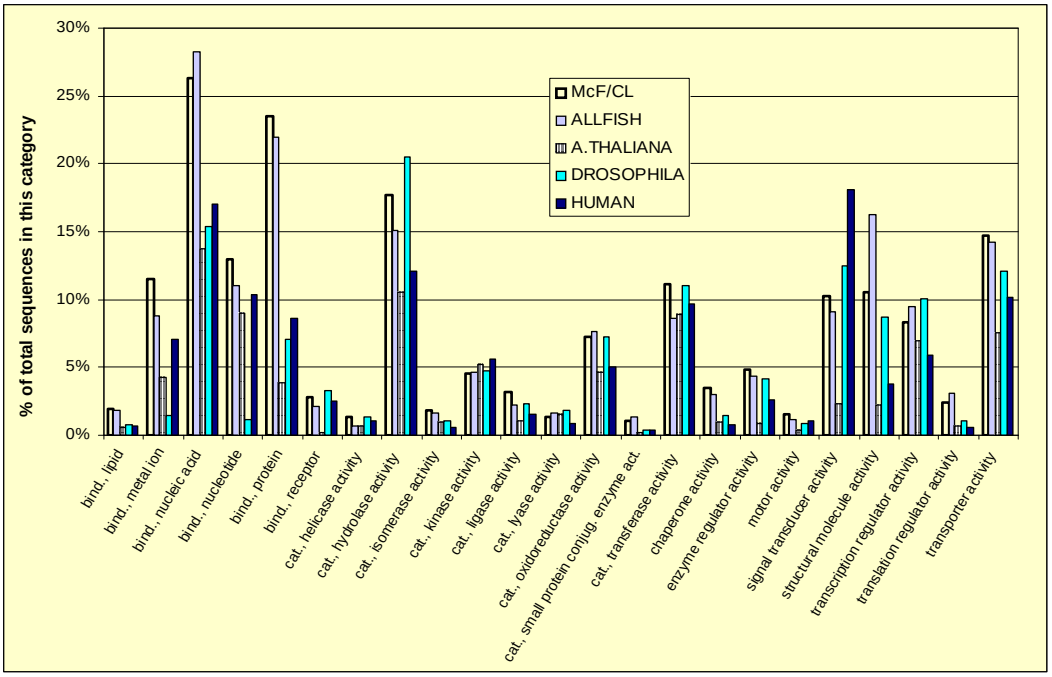


Figure 3.10: Fraction of GO-annotated sequences in different categories of molecular function. McF/CL - this project' s data; ALLFISH - TIGR fish unigene sets; A.THALIANA, DROSOPHILA and HUMAN - GO consortium; see text for ref.

3.2.2 Annotated functions represented by gene ontology

Significant blast results can be most easily viewed as a gene ontology tree containing three main subjects: molecular function, biological process and cellular compartment (see subsection 1.2.3.2). In this way the annotations can be easily compared between different EST projects. Using the GO annotations from SP_TREMBL, from TIGR institute and UniProt (see 2.3.7) to 5590 (2184 cluster and 3406 singletons; in total 56%) of unique sequences functions were assigned. Of these, 3390 were annotated in all three main categories (subcategories like "unknown" and obsolete functions were not counted).

All GO annotations were integrated into one database and a GO tree was created which was compared with a TIGR GO database, containing all GO annotations for all TIGR fish unigene sets (see 2.3.7), and annotations from the geneontology consortium for human, Drosophila and Arabidopsis [Groth et al., 2004]. The tool used for this comparison was GOblet [Groth et al., 2004]. Some categories comprising a significant number of genes are depicted in figures 3.10, 3.11 and 3.12), to compare the composition of GO annotated data sets from different species. Such a comparison of different sets obtained for different organisms is made difficult because of incomplete sets (this project' s data and probably other fish data) or not fully GO annotated sets, as there are until now only a restricted number of sequences GO annotated (see 2.3.7). But from the graphical presentation obviously comparable proportions of genes are found in this data and the TIGR collection of all GO annotated fish data. In evaluating GO data one has to take into account that

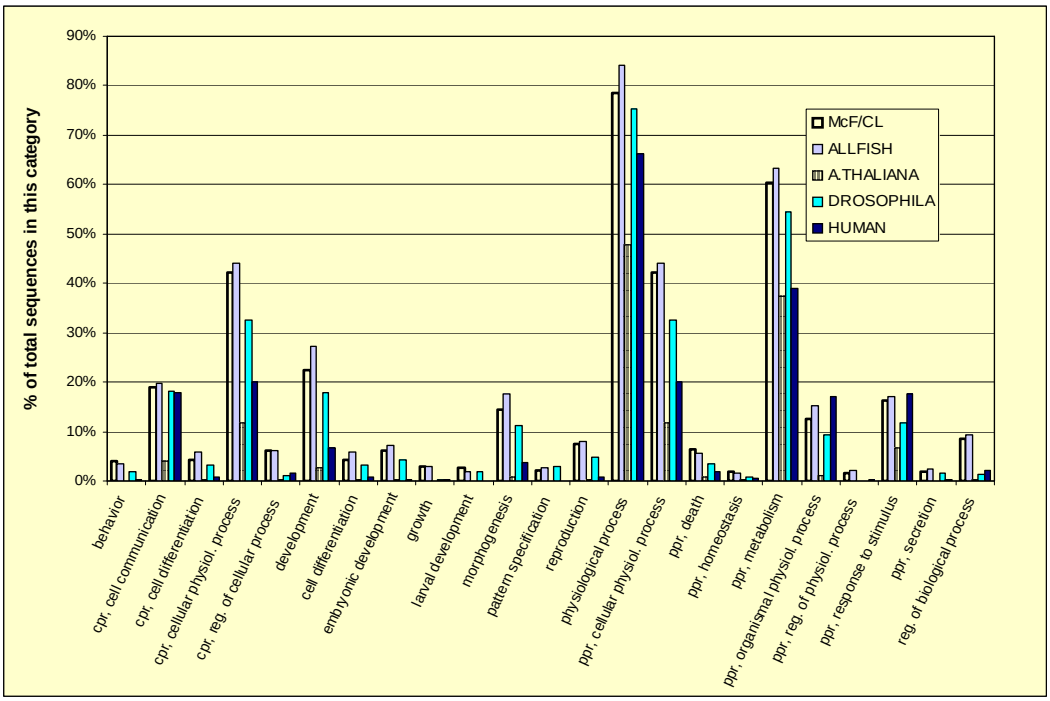


Figure 3.11: Fraction of GO-annotated sequences in different categories of biological process. McF/CL - this project' s data; ALLFISH - TIGR fish unigene sets; A.THALIANA, DROSOPHILA and HUMAN - GO consortium; cpr - cellular process; ppr - physiological process; see text for ref.

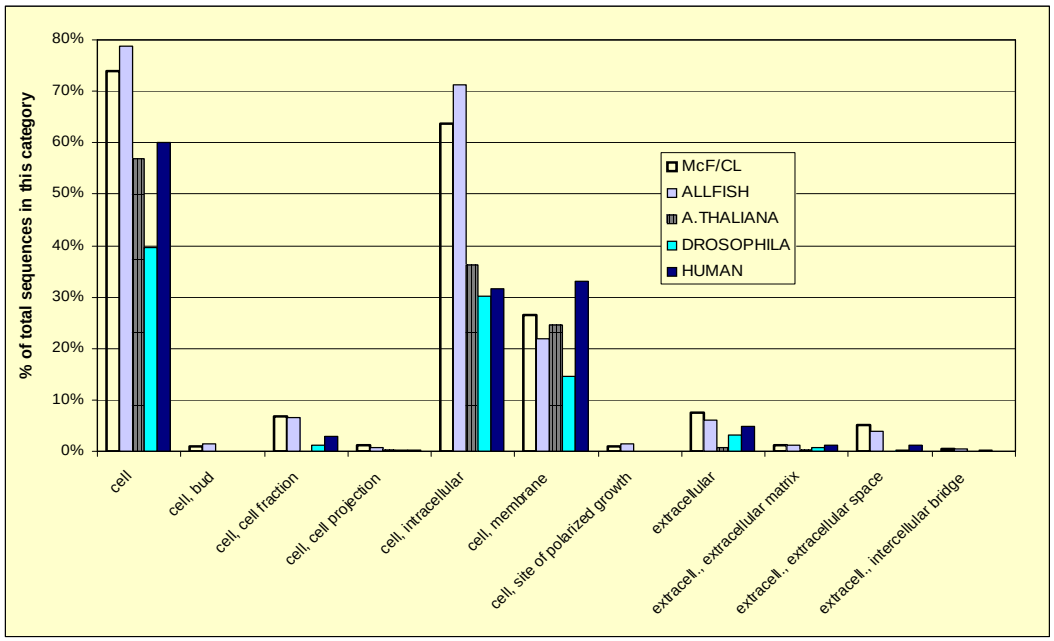


Figure 3.12: Fraction of GO-annotated sequences in different categories of cellular component. McF/CL - this project' s data; ALLFISH - TIGR fish unigene sets; A.THALIANA, DROSOPHILA and HUMAN - GO consortium; see text for ref.

	sequences	sequences with repeats		length of repetitive sequence	
total 10016 unique sequences	10016	921	9%	79775bp	8bp repeat/ seq
non annotated	1861	263	14%	16777bp	9bp repeat/seq

Table 3.12: Repeat analysis of ESTs, which were not annotated by any means, showed only little differences to the same analysis of the total set of ESTs.

usually the sum of all numbers of for example all functions may exceed the total number of all contigs and singletons annotated, e.g. many proteins may be annotated by different nodes like "binding" and "catalytic activity" within the "molecular function" category. ESTs or EST contigs showing important functions in embryonic development of the nervous system were identified and summarized in appendix D.

3.2.3 Further annotation by BLAST searches

A great part of known proteins is still not GO annotated (TIGR institute, 6% of all fish unique sequences, see 2.3.7 for version numbers; SPTREMBL lists 22,728 mostly electronically derived GO annotations for zebrafish). Therefore additionally the BLASTX algorithm was run against a locally available non-redundant protein set (nrprot and uniref100) and blastn runs against all TIGR sequences, which are not annotated by gene ontology. For 31% (1393/4426) of ESTs without GO annotation similarities to protein-coding sequences were found and 26% (1172) of these showed only hits to ESTs or hypothetical proteins.

Still for 19% (1861) of sequences no function could be assigned even that the amount of short sequences (proportion of sequences $<$ 200 bp is 8% compared to 3% in the total set) or of repeats (14% of sequences are totally covered by repeats compared to 9% in the total set; the amount of repetitive sequence motifs did not change; tab. 3.12) was only slightly enlarged in this subset, so this non annotated data represented sequence which must have been polyadenylated and therefore synthesised into cDNA. Possible explanations may be that lately a large amount of transcription (like functional non-coding RNAs or polyadenylated antisense transcription) was found outside gene boundaries, summarised in [Johnson et al., 2005], but this may also be caused by biological artefacts like non-specific polyadenylated transcripts, but also experimental artefacts like contamination from unspliced RNAs.

Possible genomic contamination of RNA samples was analysed by blasting 50 randomly selected public full-length mRNAs against the medaka genome using the UT Genome Browser (Medaka) at http://medaka.utgenome.org. This was compared to the BLAST analysis of 50 randomly selected EST singletons and 50 EST contigs (tab. 3.13). Surprisingly a higher amount of full-length coding sequences were not aligned to the Medaka genome, but this may be caused by the incomplete genome sequence. In cases where public data matched the Medaka genome, always additional public evidence or at least some genscan predicted information was found. This was not the case

	public mRNA	public mRNA (%)	EST contigs	EST contigs (%)	EST singletons	EST singletons (%)
no alignment	15	30	9	18	11	22
full alignment	35	70	41	82	39	78
in case of full alignment:						
no public evidence, no prediction	0	0	6	12	7	14
no public evidence, no prediction, in proximity to other data (public or predicted)			3 (< 500bp); 2 (< 1000b); 1 (< 5000bp)		2 (< 500bp); 2 (< 1000b); 3 (< 5000bp)	
genscan prediction, introns and/or exons	5	10	11	22	24	48
genscan prediction, at least two exons	14	28	7	14	1	2
public evidence, introns and/or exons	0	0	2	4	2	4
public evidence, single exon ESTs	2	4	4	8	3	6
public evidence, at least two exons	14	28	11	22	2	4

Table 3.13: Estimate of genomic contamination by alignment of sequence data to public available Medaka genome scaffolds. Results for EST contigs and EST singletons were compared to the alignment of 50 full-length coding sequences (public mRNA) to the same scaffolds. In the case of full alignment to the Medaka genome that region was searched for further public sequences aligned to the same place or at least some genscan prediction. It was analysed if these hits include the same exons and introns (contamination of cDNAs with intronic sequences is quite frequent). In the case where no other sequences were overlapping the query sequence, the close proximity was searched for gene evidence with three intervals from 500 to 5000 bp. Maybe such clones contain UTRs.

Medaka_10016, 10016 seqs - 100%					
Medaka_public	Danio	fugu	tetraodon	drosophila	yeast
7259	5377	6010	6092	3705	1750
72%	54%	60%	61%	37%	17%
Medaka_public_27109, 27109 seqs - 100%					
Medaka_10016	Danio	fugu	tetraodon	drosophila	yeast
9781	11398	12544	12533	7357	3226
36%	42%	46%	46%	27%	12%

Table 3.14: BLAST results (e-value cutoff 1.0e-10) approximate the similarity of different fish protein sets to each other. Medaka_10016: x% of 10016 unique sequences showed significant blast results to the other protein sets, e.g. 72% showed similarities to Medaka public data. Medaka_public_27109: 36% showed similarities to this project's data, Medaka_10016.

for 14% of this project's data. EST contigs show similar results as for public full-length coding sequences, but EST singletons maybe show some contamination as the amount of hits to public evidences is clearly reduced (tab. 3.13).

3.2.4 New sequences - new information

For further sequence analysis it is important to know to which extent the non-redundant set of 10,016 ESTs represented new genes, not yet found in medaka. For this task, public medaka EST data (149,697 ESTs from GenBank/Ensembl in April 2005) was clustered with the non-redundant data set, which resulted in 17,418 singletons and 13,852 EST contigs. Of the unique sequences 4392 (814 of EST cluster and 3578 of single ESTs) were not clustered with public medaka EST data. To 3080 (70%) of these sequences functions were assigned, compared to the assignment of protein function to 90% (5075) of the ESTs, which were already known in medaka. This analysis may have been changed by now, as there are at the beginning of 2009 already 616,739 ESTs available at dbEST (dbEST release 013009, Jan. 2009) and also the Medaka draft genome was completed, including over one million 5'-end serial analysis of gene expression tags [Kasahara et al., 2007].

To compare the new data with Fugu (FUGU2, 33,003 seqs), zebrafish (ZFISH4, 32,062 seqs) and Tetraodon (TETRAODON7, 28,005 seqs), Ensembl data was downloaded (Ensemblv30 from April 2005), together with protein sets for Drosphila (assembled from BDGP 3.2.1, 19,177 seqs) and Yeast (assembled from SGD 1, 6,680 seqs), to get an overview of the amount of proteins common to all eukaryotes (see tab. 3.14). BLASTX (e-value cutoff 1.0e-10) was run to identify similarities of this project's data to Ensembl protein sets. All data was also compared to the tgicl-clustered public Medaka data (149,697 sequences from GenBank/Ensembl, clustered into 27109 unique sequences) by using TBLASTX (e-value cutoff 1.0e-10). The same analyses were done using the 27109 unique sequences from public Medaka data as query sequences. The evolutionary distance between zebrafish, belonging to one of the older fish families, and the pufferfishes and

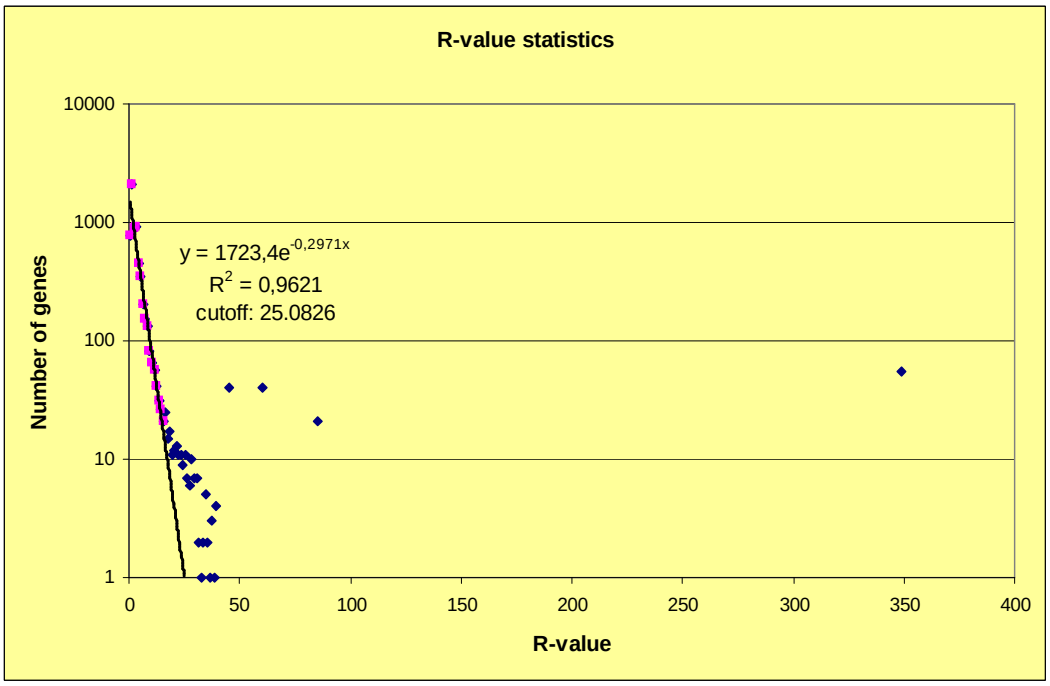


Figure 3.13: R-values of all OFP cluster are plotted. An OFP cluster behaves differently if its R-value lies above the exponential curve (blue data points). The cut-off value was calculated with 25.0826, and OFP cluster with values greater than the cutoff value are called differentially expressed.

Medaka becomes not visible in the blast results. These fishes seem to have a big reservoir of proteins showing high similarities to each other.

3.3 Differential gene expression

In the OFP run, fingerprints were successfully obtained for 53,062 gastrula clones, 8595 neurula clones, 18,784 ovary clones and 20,219 organogenesis clones. Biased numbers of clones from each library in a cDNA cluster indicate differential expression of the respective genes. The OFP result was firstly changed according to the EST clustering results by merging different OFP clusters which contained representatives grouped into the same EST cluster. To find differentially expressed clusters, the R-test [Stekel et al., 2000] was applied. The cut-off value was estimated with 25.0826 (fig. 3.13) as described in 2.2.3, and approved significant for 218 fingerprinting cluster (tab. 3.15 and appendix C). In cases of OFP clusters containing more than 50 clones where the R-value was insignificant, the approximation of 2R to the chi-square distribution was used to identify additional 19 candidates for differentially expressed putative transcripts. For 151 of these putative transcripts EST sequences were obtained, from which the majority were annotated as functional proteins, 13 only showed similarities to hypothetical proteins. Candidates for differentially expressed genes were roughly categorized into groups, being *cell signaling and signal transduction*

(Cellular retinoic acid-binding protein, Pleiotrophic factor-beta-2 precursor), *cell cycle* (Cyclin B2, RAN protein, beta 1 integrin isoform A), *cell adhesion and cytoskeleton* (keratin, Zona pellucida glycoproteins, Fast muscle troponin I, Claudin-like protein ZF-A89), *transcription* (Small nuclear ribonucleoprotein D1, CCAAT-binding transcription factor subunit, PP2A inhibitor), *translation and protein metabolism* (Ubiquitin, Choriolysin H 1, Elongation factor 1-alpha, ribosomal proteins), *immune response* (thymosin beta b, Px19-like protein) and *general metabolism* (Cytochrome, Aldolase A fructose-bisphosphate, Apolipoprotein E). Most of the genes are differentially expressed in the gastrula stage, which depicts to some amount also the larger amount of gastrula cDNA clones subjected to OFP analysis. But also candidates were found differentially expressed in all other stages (tab. 3.15 and appendix C).

OFP cluster	OFP cluster size	R-value	stage overexpressed	stage underexpressed	EST cluster	Annotation
cell signaling and signal transduction						
OFP53	369	147.517	gastrula		CL7Contig2	similar to UniRef100_Q91ZW7 CD209 antigen-like protein E
OFP127	107	115.419	ovary		CL2483Contig1	similar to UP|AAR82892 Cellular retinoic acid-binding protein
OFP158	94	30.397	gastrula		CL212Contig1	similar to SP|Q98SJ7|TCTP_LABRO Translationally-controlled tumor protein (TCTP)
OFP436	29	28.172	ovary		McF02L02	similar to SP|P48533|PTB2_XENLA Pleiotrophic factor-beta-2 precursor (PTF-beta-2)
cell cycle						
OFP273	79	76.198	organo		CL107Contig1	similar to AAH66507 Cyclin B2
OFP118	110	70.432	gastrula		CL869Contig1	similar to UniRef100_P39963 G2/mitotic-specific cyclin B3
OFP26	277	34.957	gastrula		CL3Contig3	similar to PIR|A45941 histone H3

continued on next page

OFP cluster	OFP cluster size	R-value	stage overexpressed	stage underexpressed	EST cluster	Annotation
OFP70	180	34.851	gastrula		CL43Contig1	similar to UP|Q9YGC0 RAN protein
OFP141	94	31.936	organo		CL1593Contig1	similar to UniRef100_Q9CQJ7 Securin
OFP304	51	28.771	gastrula		CL502Contig1	similar to SP|P02264|H2AG_ONCMY Histone H2A gonadal
OFP351	42	25.982	organo		McF18O06	similar to UP|Q98TZ9 Histone H2A
OFP1159	55	24.705	gastrula		CL3Contig5	similar to GB|AAA79835.1 |1036735|HSB1ITGB4 beta 1 integrin isoform A
cell adhesion and cytoskeleton						
OFP15	433	644.096	organo	gastrula	CL3450Contig1	similar to SP|Q8IUB9|K191_HUMAN Keratin associated protein 19-1 (High tyrosine-glycine keratinassociated protein 19.1) (Fragment)
OFP58	170	272.436	organo		CL244Contig1	ssimilar to UP|Q8AYK7 ZPC5
OFP43	276	136.527	gastrula		CL23Contig1	similar to GB|AAR22526.1 |38503465|AY459292 keratin 17n
OFP177	105	112.119	organo		CL106Contig1	similar to UP|Q9DG37 Zona pellucida glycoprotein 3 (Fragment)
OFP173	83	108.918	ovary		CL455Contig1	similar to UniRef100_UPI0000273AC1 UniRef100 entry
OFP331	41	68.828	organo		McF49E03	similar to UniRef100_Q8AYK9 ZPC3
OFP74	158	64.706	ovary	gastrula	CL426Contig1	similar to UP|Q71N42 Fast muscle troponin I

continued on next page

OFP cluster	OFP cluster size	R-value	stage overexpressed	stage underexpressed	EST cluster	Annotation
OFP149	101	60.879	gastrula		CL7Contig6	similar to UniRef100_Q8HY01 CD209 antigen
OFP30	401	51.471	gastrula		CL33Contig1	similar to UP|Q8JFQ5 Keratin 12 (Fragment)
OFP164	112	49.023	neurula		CL3Contig10	similar to UP|AAH63955 Ckii protein
OFP207	71	43.994	organo		CL689Contig1	similar to SP|Q9YH91|CLDY_BRARE Claudin-like protein ZF-A89
OFP211	85	31.362	gastrula		CL7Contig1	similar to UniRef100_Q8HY02 CD209 antigen
OFP349	45	28.813	gastrula		CL1573Contig1	similar to UP|Q90XR9 Claudin e
OFP1182	17	28.539	organo		CL593Contig1	similar to UP|Q8AYL0 ZPC1
OFP853	24	25.968	organo		CL781Contig1	similar to UniRef100_UPI00002F978D UniRef100 entry
OFP329	71	19.908	gastrula		CL59Contig1	similar to UP|Q90W73 Type I keratin S8
OFP441	39	55.939	organo		CL1962Contig1	similar to UP|XLR1_FUGRU (Q9W6R5) Retinoschisin precursor (X-linked juvenile retinoschisis protein homolog)

continued on next page

OFP cluster	OFP cluster size	R-value	stage overexpressed	stage underexpressed	EST cluster	Annotation
transcription						
OFP29	284	111.872	gastrula, neurula		CL197Contig1	similar to UP|Q8JHH1 Small nuclear ribonucleoprotein D1
OFP52	178	80.926	neurula, ovary	gastrula	McF19K04	similar to SP|P18576|CBFB_RAT CCAAT-binding transcription factor subunit B (CBF-B) (NF-Y proteinchain A) (NF-YA) (CAAT-box DNA binding protein subunit A)
OFP36	224	67.975	ovary		CL114Contig1	similar to UP|Q8AWZ2 SI:dZ227P06.1 (Homeo box A3a)
OFP135	120	46.095	gastrula		CL54Contig1	similar to UP|O42468 PP2A inhibitor
OFP157	93	42.464	gastrula		CL532Contig1	similar to UP|Q90ZH5 SAP18
OFP84	108	18.508	gastrula		CL589Contig1	similar to UP|Q7ZZQ7 Nucleoside diphosphate kinase
immune response						
OFP197	72	72.873	ovary		CL3290Contig1	similar to PIR|A03270|GYRTI cysteine-rich intestinal protein
OFP102	134	35.213	neurula		CL50Contig1	similar to UP|Q9NDQ1 Fibrinogen-like protein
OFP391	44	35.115	neurula		CL85Contig1	similar to GB|AAH26200.1 |20073031|BC026200 immune associated nucleotide 3
OFP860	19	26.940	organo		CL2218Contig1	similar to GP|432442|AA antileukoproteinase precursor

continued on next page

OFP cluster	OFP cluster size	R-value	stage overexpressed	stage underexpressed	EST cluster	Annotation
OFP96	126	22.907	gastrula		CL335Contig1	similar to GB|BAA96493.1|8307710|AB028457 thymosin beta b
OFP249	58	21.287	organo		CL1589Contig1	similar to AAQ97836 Px19-like protein
OFP505	73	28.437	gastrula		CL45Contig1	similar to UP|Q9I886 (Q9I886) Natural killer cell enhancement factor

Table 3.15: Differentially expressed OFP cluster.

3.4 Identification of splice variants

With the advent of genome projects, the big difference between number of genes and number of encoded proteins became obvious. For the human genome a high rate of alternative splicing was estimated, with 35-59% of human genes showing evidence of at least one alternative splice form as reviewed in [Modrek and Lee, 2002]. ESTs and EST cluster show the right splicing events so they are a suitable subject to study alternative splicing by aligning ESTs against the respective genome sequence.

Good candidates for alternatively spliced genes are EST clusters originating from the same precluster calculated by mgblast during tgicl clustering (see 2.3.6). Additionally, during further EST clustering approaches, groups showing high similarities (using BLAST) were identified, but not completely clustered with CAP3. These sequences were aligned against the draft sequence of the Medaka genome using the UT Genome Browser (Medaka) at http://medaka.utgenome.org or simple BLAST against the medaka scaffolds at http://dolphin.lab.nig.ac.jp/medaka in 2005. For 58% (161) of the 276 candidate groups a full alignment (allowing for sequencing errors) of at least 2 sequences to the Medaka genome was obtained using strict mapping rules (more than 95% similarity over full sequence length, more than 90% at both ends of sequence). From the UT Genome Browser view additional publicly available data was collected and included in further analyses. These EST evidences were again aligned to the respective Medaka scaffold using EMBOSS est2genome (based on est_genome; [Mott, 1997]) and PipMaker [Schwartz et al., 2000]. The combination of both tools gives the opportunity to highlight candidate exons and their similarity to the Medaka genome separated by candidate splices as gaps in the EST-genome alignment. During this process 62 groups were excluded showing not sufficient similarity to the Medaka genomic sequence. Differences between PipMaker and est2genome come from the different approaches used, where PipMaker just

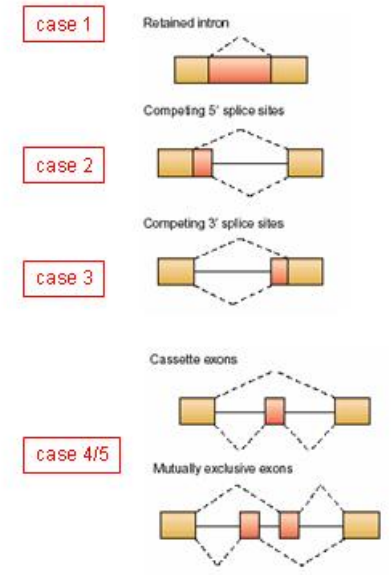


Figure 3.14: 1 Alternative splicing events are classified into five main types [Breitbart et al., 1987]: single and multiple cassette exons, mutually exclusive exons, retained introns, competing 5' (internal donor site) and 3' (internal acceptor site) splice sites. Figure changed from [Roberts and Smith, 2002].

shows similarities between medaka scaffold and the different ESTs, but est2genome calculates an alignment by also taking splice sites into account.

Identified splices were categorised into the five main types [Breitbart et al., 1987] of alternative splicing: single and multiple cassette exons, mutually exclusive exons, retained introns, competing 5' (internal donor site) and 3' (internal acceptor site) splice sites, depicted in figure 3.14. Because of uprising difficulties to distinguish cassette exons from mutually exclusive exons in cases where full coding sequences are not available, (probably not all possible transcripts were collected, yet) these two cases were counted as one. For all of these cases functional importances were reported, so all of them were recorded as candidates for alternative splicing in table 3.16, even that it is known that retained introns may be frequently caused by unsuccessful splicing (7-9% of cDNA

case	Number of cases
cassette exons and mutually exclusive exons	49
retained introns	7
competing 5' (internal donor site)	34
competing 3' (internal acceptor site)	32
total	122 (102 gene candidates)

Table 3.16: Number of alternative splicing candidates found in each category of alternative splicing events.

clones retain unspliced introns [Okazaki et al., 2002]).

3.5 SNP search

Searching SNPs in EST alignments can be tiresome because of visual inspection of computed alignments. There are approaches which are automatically calling SNPs like the PolyBayes program [Marth et al., 1999]. But this tool is based on a PHRAP clustering [Staden et al., 2000] that is not appropriate for clustering ESTs as summarised in 1.2.3.1.

For genetic mapping experiments SNPs between southern (Cab strain) and northern population (HNI strain) were searched. To identify these, all public available EST data (see 2.3.1) was clustered together with this project's EST data by applying the tgicl clustering program (2.3.6) and clusters which contained sequences from both strains were identified with a Perl script. TIGR clview (2.3.6) was used to visually identify SNPs in aligned sequences produced by CAP3. For calling candidate SNPs the following criteria were applied. Polymorphisms within the first or last 30 nucleotides of an alignment were not taken into account, around a SNP candidate there should be a 40 bp window of sequence which contains no other poymorphism. Only single base substitutions were noted, no insertion/deletion polymorphisms. In the case of more than one sequence available for one strain the polymorphism had to be apparent in all sequences so to avoid polymorphisms between different fishes. (There were polymorphisms noticed within the CAB strain, which were treated as experimental errors. In a later publication [Kimura et al., 2005] it was noted that the CAB strain is most likely not an inbred strain, because CAB fishes were polymorphic for many other genetic markers in contrast to the other Medaka inbred strains, where the polymorphism rates are about 1% or less.) The exact position of every SNP candidate within an alignment was recorded. With that strategy 195,712 nucleotides comprised by 446 alignments were inspected for polymorphisms, where in 74 nothing was found and in 372 alignments 918 candidate SNPs were identified. That means that 1 SNP occurs every 213 bp between the HNI strain and the Cab strain.

It was also checked to what extent there are single nucleotide polymorphisms between two southern populations, Cab strain EST sequences produced in this work and ESTs from Hd-rR [Kimura et al., 2004], also used for BAC mapping [Zadeh Khorasani et al., 2004], because the Medaka genome project is using the Hd-rR strain, but the physical mapping was done on Cab and Hd-rR. With the same approach used, but on a smaller scale, only 11 SNPs were identified within 41 alignments (only 5 clusters contained polymorphisms) which comprise 28,509 nucleotides, so that is just 1 SNP in 2591 bp (0.039% divergence) of southern populations as compared to 1 SNP in 213 nucleotides (0.47%) between northern and southern population under subject.

3.6 MedakaProjectDB

All data was collected in a relational database to provide it easily through a web-interface. Database tables were set up for clones with their OFP clustering result, their rearray coordinates and sequence information. Also the EST clustering result and the annotation of the unique data set was included.

3.6.1 Database model

The database was set up with 5 tables.

Table clone describes the entity clone. This entity is characterised at least by its clone name (`clone_name`) and the library (`library`) it was taken from, which is known for all clones subjected to OFP analysis. If one clone after successfully characterised by OFP was chosen as a representative of its OFP cluster then a rearray coordinate (`rearray1`) is also known for that clone. These clones were also sequenced and therefore obtained trace names (`trace_name`). In case of successful sequencing the table clone also included the 5' EST sequence (`clone_sequence`). After the EST clustering clones of the unique data set were again rearrayed and therefore these clones got another rearray coordinate (`rearray2`).

The SQL statement to create the table was given as follows:

```
CREATE TABLE clone (

    clone_name char(30) PRIMARY KEY,
    trace_name char(30),
    library char(20),
    rearray1 char(30),
    rearray2 char(30),
    clone_sequence VARCHAR)
```

Table trace_uniqseq describes the relation between traces and unique sequences. Every 5' EST, characterised by `trace_name`, was subjected to EST clustering and therefore grouped into an EST contig or left as singleton. In case of a singleton `trace_name` and `uniqseq_name` hold the same entries, but for EST contigs the clone may be identified by its trace name and `uniqseq_name` holds the name of the respective EST contig. The SQL statement was given as follows:

```
CREATE TABLE trace_uniqseq (

    trace_name char(30) PRIMARY KEY,
    uniqseq_name VARCHAR)
```

Table uniqseq describes the entity of a unique sequence, either represented by an EST cluster or a singleton. This entity is characterised by its contig size, which is "1" for singletons. Also the contig consensus sequence and its annotation are saved. The SQL statement was given as follows:

```
CREATE TABLE uniqseq (

    uniqseq_name char(30) PRIMARY KEY,
    size INTEGER,
    annotation VARCHAR,
    uniqseq_sequence VARCHAR)
```

Table clone_OFP describes the relation of clones to a certain OFP cluster. According to the EST clustering the fingerprint clustering was changed afterwards, so some clones at first put into different clusters or singletons were after EST clustering integrated into a fingerprint cluster. Therefore for every clone both OFP cluster are saved (`OFP_name_old` and `OFP_name_new`). Clones within one fingerprint cluster were ranked according to their similarity to the consensus fingerprint. This item `clone_rank` is only available for the original OFP clustering.

The SQL statement was given as follows:

```
CREATE TABLE clone_OFP (

    clone_name char(30) PRIMARY KEY,
    OFP_name_old VARCHAR,
    OFP_name_new VARCHAR,
    clone_rank INTEGER)
```

The table OFPcluster describes the entity of an OFP cluster. It contains a certain amount of clones from different library (e.g. `neurula_numb` states the number of neurula clones in this cluster). For every cluster the R-value (`R_stat`) was calculated across all libraries but also for each library separately (e.g. `neurula_expr`), which allows later to identify gene candidates differentially expressed in a certain developmental stage. The SQL statement was given as follows:

```
CREATE TABLE OFP (

    R_stat,
    OFP_name char(30) PRIMARY KEY,
    size INTEGER,
    neurula_expr INTEGER, neurula_numb INTEGER,
    gastrula_expr INTEGER, gastrula_numb INTEGER,
    organogenesis_expr INTEGER,organogenesis_numb INTEGER,
    ovary_expr INTEGER, ovary_numb INTEGER)
```

3.6.2 Database interface

The database was integrated into the project's web page and may be queried via the internet (http://www.molgen.mpg.de/~rodent/CloneInfo.html). Data can be read only. Figures 3.15, 3.16

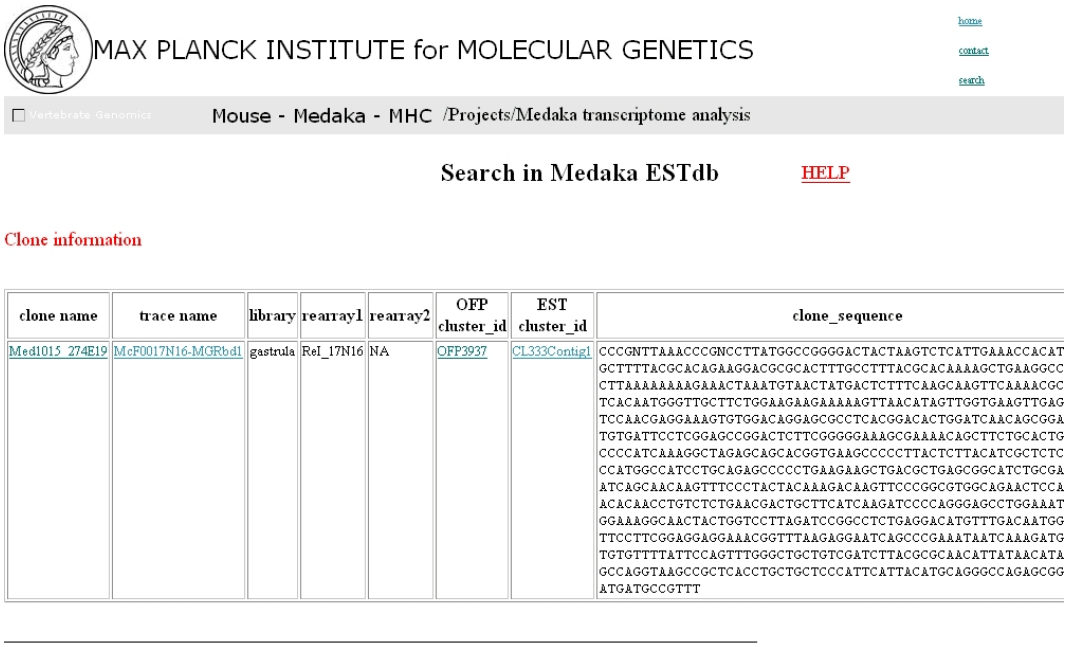


Figure 3.15: Attributes of cDNA clone Med1015_274E19. This clone was fingerprinted and sequenced successfully and then rearrayed once. It was not included into second rearray. It was found in OFP cluster 3937 and after EST clustering grouped into EST cluster CL333Contig1.

and 3.17 show screenshots of certain database queries. Data may be searched for various clone IDs (e.g. original clone ID, clone ID after rearraying, trace IDs), but also OFP clusters of a certain abundance profile may be searched for on site AbundanceProfile.html. With MedakaDB_help.html a help page is provided, explaining all available data. Source code is provided under supplemental material within the folder Perl\web. Database file fifth.sqlite is found in supplemental material in folder SQLiteDB.

Sequence cluster information

Cluster/Singleton name	Cluster size	GO/blastx SPTrembl/TIGR blastx nrprot blastn TIGR motif search	sequence
CL333Contig1	5	GO: GO:0003677 DNA binding; GO:0003700 transcription factor activity; GO:0005634 nucleus; GO:0006355 regulation of transcription, DNA-dependent; GO:0007275 development; GO:0045944 positive regulation of transcription from RNA polymerase II promoter; DE: similar to UP|O73784 (O73784) Fork head domain protein FKD8 ;UP|O73782 (O73782) Fork head domain protein FKD6 ;UP|FXGA_CHICK (Q98937) Forkhead box protein G1A (Forkhead-related protein FKHL2) (Transcription factor BF-2) (Brain factor 2) (BF2) (CBF-2) (T-14-6) ;UP|Q7T1C2 (Q7T1C2) Forkhead transcription factor i2 ;UP|O93613 (O93613) Brain factor 2 ;	AATCGGCTCCTATCACTGCTCAGAGCGTTTGGTAGAAGGTAATATGAAGGAAACGGCAT CATGATGCCGCTCTGGCCCTGCATGTAATGAATGGGAGCAGCAGGTGAGCGGCTTACCT GGCCCGGTATGTTATAATGTTGCGCGTAAGATCGACAGCAGCCCAAACTGGAATAAAAC ACAAATCCATCTTTGATTATTTCGGGCTGATTCCTCTTAAACCGTTTCCTCCTCCGAAG GAAGCTGCCATTGTCAAACATGTCCTCAGAGGCCGGATCTAAGGACCAGTAGTTGCCTT TCCCTGGATTTCCAGGCTCCCTGGGGATCTTGATGAAGCAGTCGTTCAGAGACAGGTTG TGTCTGATGGAGTTCTGCCACGCCGGGAACTTGTCTTTGTAGTAGGGAAACTTGTTGCT GATGAAGTCGCAGATGCCGCTCAGCGTCAGCTTCTTCAGGGGGCTCTGCAGGATGGCCA TGGTGATGAGAGCGATGTAAGAGTAAGGGGGCTTCACCGTGCTGCTCTGGCCTTTGATG GGGGGCTCAGTGCAGAAGCTGTTTTCGCTTTCCCCCGAAGAGTCCGGCTCCGAGGAATC ACACTCTTCCGCTGTTGATCCAGTGTCCGTGAGGCGCTCCTGTCCACACTTTCCTCGTT GGACTTCCTCAACTTCACCAACTATGTTAACTTTTTCTTCTTCCAGAAGCAACCCATTG TGAGAGGGCGTTTTGAACTTGCTTGAAAGAGTCATAGTTACATTTAGTTTCTTTTTTTT TAAGTTGCGGCCTTCAGCTTTTGTGCGTAAAGGCAAAGTGCGCGTCCTTCTGTGCGTAA AAGCGCAAATGTGGTTTCAATGAGACTTAGTAGTCCCCGGCCATA

Cluster	Clones
CL333Contig1	McF0017N16-MGRbd1
CL333Contig1	McF0001MGR-1A15bd1
CL333Contig1	McF0032K22-MGRbd1
CL333Contig1	McF0023K06-MGRbd1
CL333Contig1	McF0028B01-MGRbd1

Figure 3.16: Annotation and sequence of EST cluster CL333Contig1. This cluster comprises 5 EST sequences which are summarized below with an underlying link to the EST sequence.

MAX PLANCK INSTITUTE for MOLECULAR GENETICS

home
contact
search

Vertebrate Genomics Mouse - Medaka - MHC /Projects/Medaka transcriptome analysis

Search in Medaka ESTdb HELP

Fingerprint cluster information

Data for OFP3937:
OFP3937 contains 9 clones (without OL clones)

Abundance profile: (0 - no changes; 1 - overexpressed; -1 - underexpressed)
R statistics: R = 5.7625871415431
Neurula: 0 clones (0) Gastrula: 9 clones (0) Organogenesis: 0 clones (0) Ovary: 0 clones (0)

clone name	trace name	library	old OFP name	new OFP name	OFP clone rank
Med1015_003O04	NA	gastrula	OFP3937	OFP3937	3
Med1015_026P22	McF0023K06-MGRbd1	gastrula	Med1015_026P22	OFP3937	1
Med1015_217A21	McF0001MGR-1A15bd1	gastrula	OFP5051	OFP3937	3
Med1015_261F20	NA	gastrula	OFP5051	OFP3937	2
Med1015_274E19	McF0017N16-MGRbd1	gastrula	OFP5051	OFP3937	1
Med1015_314K04	McF0028B01-MGRbd1	gastrula	OFP3937	OFP3937	2
Med1015_338J03	NA	gastrula	OFP3937	OFP3937	4
Med1015_340P20	McF0032K22-MGRbd1	gastrula	Med1015_340P20	OFP3937	1
Med1015_357I03	NA	gastrula	OFP3937	OFP3937	1

Figure 3.17: Attributes of OFP cluster 3937. This cluster contains 9 clones which come in this case only from gastrula library. In the case of this OFP cluster there was no hint for differential expression.

Chapter 4

Discussion

This work provides valuable progress in elucidation of the Medaka transcriptome. A unique set of 10,016 Medaka expressed sequences at four developmental stages was established. These data did already enable various studies in the field of Medaka genomics, e.g. in physical mapping of the Medaka genome or on functional studies on gene regulation.

4.1 Analysis of a transcriptome

Obtaining the complete transcriptome is a very complex task keeping in mind the dynamic range of mRNA expression and the nature of mRNA itself, including alternative transcripts, promotor and termination sites, and the presence of noncoding RNAs. One needs to sample the full diversity of tissues and the full diversity of inducible states to obtain all genes encoded in a species' life. Work on the transcriptome level maybe advantageous against genome projects because of problematic *in silico* prediction of genes from the genome template, where an accurate structure of the genes is often not obtained or many transcripts are even missed [Hogenesch et al., 2001], [Hild et al., 2003].

A drawback of all experiments based on mRNA is clearly that the amount of mRNA in a cell at a given moment does not guarantee the precise prediction of amounts of subsequently produced protein and some studies have shed doubt on a quantitative correlation between mRNA and protein levels [Anderson and Seilhamer, 1997, Gygi et al., 1999]. In [Gygi et al., 1999] evidence was published that estimating mRNA levels is not sufficient for quantitative description of biological systems because of several posttranscriptional mechanisms controlling the protein translation rate or the half-lives of specific proteins or mRNAs. The authors found that correlation between mRNA and protein levels (the concentration of 150 yeast proteins was compared to SAGE results of the same strain under the same conditions) was insufficient to predict protein expression levels from quantitative mRNA data. Also one has to take into account that preparation of mRNA, cDNA, and other stages of library construction, can have unpredictable effects on EST distribution in the cDNA library used, and thus on subsequent analyses and conclusions. All results have to be taken

as preliminary until they will be verified *in vivo*.

Additionally there is still the danger that the mRNA sample used for cDNA library production may not contain a representative of a certain mRNA, from which one cannot conclude that this mRNA is not produced. During the EST project it seems like that gene is not expressed in the tissue or condition of subject [Vingron and Hoheisel, 1999], which may result in difficulties in analysing differential expression of gene products. But that problem of missing certain items stays the same in all approaches.

EST analysis provides only limited information on the derived amino acid sequence caused by the lack of full-length cDNA clones due to the reduced efficiency of the reverse transcription reaction, which usually cannot efficiently produce full-length first-strand cDNAs.

But still the transcriptome approach provides a useful approach to identify new genes and to provide hints on their expression levels and their location of expression. Within this project the transcriptome of three developmental stages, gastrula, neurula and organogenesis were analysed together with one adult tissue, the ovary. These libraries were taken into account because of their significance for early embryonic development. On the other hand the ovary provides information about an adult organ and on maternal RNA products accumulated in oocytes.

4.2 Oligonucleotide fingerprinting

4.2.1 Laboratory methods

Like all other methods of transcriptome analysis the production of cDNA libraries may influence the applied method by large extent. For OFP analysis it is advantageous to use oligo(dT) primed cDNA libraries [Poustka, 2000] to provide a unique end point for most clones and therefore to improve the fingerprint-based clustering of clones from the same gene compared to the usage of random-primed libraries. One disadvantage in using oligo(dT) primed libraries comes from the finding that ESTs created from such libraries show a lower amount of protein coding sequences [Poustka, 2000]. Such facts could not be taken into account as for this project four oligo(dT) primed cDNA libraries were provided, already subcloned and rearrayed.

It is known that the fingerprinting analysis improves with a higher average insert length of the library as longer fingerprints, meaning more information, are created from longer transcripts. But during the course of this project the provided libraries were of small average insert sizes, estimated by agarose gel electrophoresis of around 300 PCR products for each library with 1021 bp (tab. 4.1). This value was significantly lower in the neurula library with only 898 bp. This library also showed poor amplification results. The optimal insert length is published to be 1.5 kb [Poustka, 2000]. The percentage of vector-only clones was quite small with an average of 8% across all libraries, again the neurula library showed an outlying result with 19%.

The OFP normalisation success also strongly depends on length variation of cDNA clones [Herwig et al., 1999] which is due to the fact that reverse transcriptase stops at different end points

library name	No. of inserts checked	average length (bp)	standard deviation of insert length (bp)	vector only clones (%)
gastrula	299	1025	336	7.7
neurula	298	898	370	19.0
organogenesis	302	1134	290	4.0
ovary	310	1028	274	1.9
total	1209	1021	317.5	8.2

Table 4.1: Average length of cDNA inserts subjected to OFP analysis

when processing cDNA from mRNA during transcription. If because of high length variation, the sequences do not share enough probe matches, clustering of the respective copies of the same gene might fail. The clustering algorithm tolerates a length variation of up to 500 bp, but its quality decreases significantly if the variation is larger than 700 bp [Herwig et al., 1999]. This influence of length variation can only be reduced by using more hybridisation probes. During this project the standard deviation of insert sizes of all cDNA libraries subjected to OFP was calculated with 318 bp (tab. 4.1), which is tolerated by the clustering algorithm.

Obtaining high quality hybridisation results requires high concentration of PCR products, but no purification, so cDNAs should be ligated into high-copy plasmids. For this project's amplification of cDNA inserts a concentration of around 100 μg/μl of the PCR product were obtained. This was difficult to achieve for all PCR products as during production of these cDNA clones, the amipicillin gene was affected causing a low growth rate of bacterial colonies, which were directly used as template for the PCR reaction.

Synthesised PCR products were spotted in duplicates on high density filters. The process of spotting may influence by large extent the computational evaluation of hybridisation results. Different amount of material may be immobilised on the filter because of different concentration of PCR product or to a smaller extent because of different transfer efficiency of pins. Also the hybridisation of the filters may lead to variations in signal intensities caused by differences in the activity of labeled probes, or variations in hybridisation conditions and filter exposition on screens [Eickhoff et al., 2000]. The same hybridisation conditions were used for all oligonucleotides but they have different optimal annealing temperatures which also causes variation in signal strength. To avoid these problems all clones were spotted in duplicates. In case of failure in spotting and hybridisation, filters may show bad correlation of duplicates and were therefore excluded from further analysis, so the probability of wrong assignment of signals to clones is largely prevented. The identification of positive clones was also simplified by spotting salmon sperm on defined grid positions. These dots served in guiding the automated positioning of the grid on top of hybridised filters. The identification of positive clones depends mainly on the correct calculation of the grid

position.

In choosing oligonucleotide probes for hybridisation it is known that the most informative set of probes would have an average matching frequency of 50%, but in practice for 8mers that frequency is much lower. This was estimated by evaluating control clones of known sequences, where the range of positive hybridisations of the probes was found to be around 5-25% [Herwig et al., 1999]. An improvement to that problem would be to use shorter probes which increase matching frequencies statistically. It was shown that hybridisation of even 6 mers was reliable [Drmanac et al., 1990], but in this project's work it was decided even to use 10mers according to [Meier-Ewert et al., 1998], because of their reliable hybridisation results. There are projects underway to improve oligonucleotide fingerprinting analysis by applying shorter hybridisation probes and also by avoiding radioactively labelled probes (for avoiding over-shining effects) in using fluorescently labelled peptide nucleic acids (PNAs) [Guerasimova et al., 2001]. PNAs are molecules containing nucleobases attached to a neutral peptide-like backbone. They therefore hybridise to complementary RNA or DNA with higher affinity and specificity than conventional oligonucleotides and oligonucleotide analogues. But this approach is not yet applicable for large projects. To improve the information of oligonucleotide probes their sequences were partially taken from a standard set, which is known to hybridise well to different organisms, and partially calculated from already known medaka ESTs (2.2.1.3).

Usage of short hybridisation probes causes a high error rate, which can be circumvented by applying a high number of probes (around 200-300 probes). A large number of hybridisation probes would also improve separation and identification of partially overlapping genes such as splice variants and highly homologous genes [Herwig et al., 1999]. Within this project only 124 oligonucleotide probes were successfully applied to all cDNA clones because of too different hybridisation behaviour in different cDNA libraries.

4.2.2 Computational methods

Computational analyses during the course of oligonucleotide fingerprinting were started with automatic image analysis. This analysis includes the positioning of a grid to the filter image. This is done with the help of regular spots containing salmon sperm DNA, which should hybridise to every probe. Positioning the grid may still be complicated due to mechanical strains on the Nylon filters which cause the spots to deviate from their specified coordinates, but which is coped with by the implemented software [Steinfath et al., 2001]. This program is able to function properly even in cases of global distortion of the grid, missing of grid nodes, false positive spots and even local deviation of spots from its specified grid position. Analysing images may also be difficult because of over-shining effects where hybridisation signals of neighbours interfere with each other, which may only be avoided by a larger spot distance or by using fluorescently labelled probes. Many problems during automatic identification of positive clones are avoided by calculating the correlation of duplicated cDNA clones. In case of missing correlation, clones were not called positive in

this hybridisation. This approach should decrease false positive errors. False negative errors are hard to cope with. These may likely be caused by hybridisation errors which are hard to avoid in large-scale projects.

The high-quality information of all hybridisations was used to group all cDNA clones into clusters. The success of the applied clustering algorithm is largely affected by the irregular appearence of cDNA clones, where usually there is a small number of big gene clusters and a high number of singletons and small gene clusters [Meier-Ewert et al., 1998, Poustka et al., 1999]. Big clusters are easy to detect by clustering procedures, whereas small clusters are harder to identify because of the high variance introduced by experimental error. The identification of small clusters would need to generate highly pure clusters by stringently set algorithmic parameters. But this would lead to an overestimation of the total number of genes due to cluster splitting and to the false assignment of clones to singletons that should be clustered and thus to a lower normalisation rate of the cDNA library [Herwig et al., 1999]. Therefore several runs of the clustering algorithm were compared to each other to identify an "optimal" solution. The quality of clustering was calculated in backhybridisation experiments (tab. 3.6), which indicated that one gene was represented on average by 2 OFP clusters. At this point the hybridisation analysis was not further refined. Clones from different OFP clusters or OFP singletons were subjected to EST sequencing and again evaluated by cluster analysis of the obtained EST sequences. It was found that one gene is still represented by 2 cDNA clones in the rearrayed and sequenced library which confirms the result of backhybridisation experiments. The result of EST cluster analysis was taken to identify a highly valuable, unique set of 10,016 gene candidates.

Sequencing work may be further reduced by comparing *in silico* obtained fingerprints of public available EST data to experimentally obtained fingerprints of OFP clusters and singletons. ESTs would therefore be checked for the occurence of oligonucleotide sequences and from that information an *in silico* fingerprint would be calculated. These *in silico* fingerprints would then be compared to the consensus fingerprints of all OFP clusters and singletons. In case similar fingerprints were found, these fingerprints may be obtained from the same gene, on one hand an EST from that gene and on the other a fingerprinted cDNA clone from that gene. This cDNA clone than does not to be sequenced again therefore saving sequencing ressources. This method was implemented by the bioinformatics group of R. Herwig, MPI for Molecular Genetics, but did not work sufficiently. There were too little agreements between *in silico* data and experimentally obtained data, which does not reflect the real world and therefore this method was not used.

4.2.3 OFP analysis as a method for transcriptome analysis

Like all methods which are based on mRNA analysis, OFP allows to detect the full set of active genes in whole tissues and also to identify yet unknown genes. Another advantage is that it takes into account the whole sequence of a cDNA clone and provides therfore more information than tagging approaches like EST approaches. This method reduces the redundancy of a cDNA library

before sequencing and therefore reduces sequencing costs. During normalisation and subtraction of libraries by OFP, library quality is not affected like in other commonly used normalisation and subtraction methods, which tend to select shorter, truncated and internally primed inserts, whereas OFP selects for the clone with longest fingerprint, most likely one of the longest cDNA clones. There were also other methods developed to avoid problems caused by normalisation like selecting for full-length cDNA clones [Carninci et al., 2000].

One disadvantage of the OFP approach lies in high error rate results from hybridisations of short oligonucleotide probes, where many clones are hybridised to many probes under the same hybridisation conditions. This approach is also slower than the EST tag method and also cost- and material-consuming but maybe highly efficient if enough resources are available.

This project showed good results in normalisation of the provided cDNA libraries. The redundancy of the four unnormalised libraries was successfully reduced almost 10-fold, reminding that 100,660 cDNA clones were successfully subjected to OFP analysis and 10,016 unique sequences were left after EST clustering. Considering the 100 superprevalent transcripts, which were represented by OFP cluster of sizes from 122 to 1855 clones, the occurrence of these genes was reduced up to 216-fold (OFP15 cluster, probably representing a gene for Keratin, contains 433 clones and its corresponding EST cluster, CL3450Contig1, contains 2 ESTs; see table 3.8). Noting that the OFP analysis provided 18,378 potentially different transcripts, but which were grouped into 10,016 unique sequences, it may be estimated that only 54% of obtained OFP clusters comprise different genes. This result is comparable to the OFP analysis of zebrafish cDNA libraries (68% gene diversity after OFP analysis; [Clark et al., 2001]), but better results were already obtained for sugar beet (89%; [Herwig et al., 2002]) and sea urchin (92%; [Poustka et al., 2003]). Within this project good results were obtained in normalisation and subtraction, but work was less efficient in clustering unknown sequences by OFP compared to other projects. This may be largely caused by the low amplification rates of bacteria containing cDNA clones used, which affected the quality of all hybridisations and therefore the OFP clustering result. Further only a little number of oligonucleotide probes was successfully hybridised to all libraries, 124 oligonucleotides compared to 246 in sugar beet, which may be caused by the different composition of the different libraries, e.g. different mRNA sets are present.

4.3 Analysis of the Medaka transcriptome

4.3.1 EST sequencing

In the course of this project 5' ESTs were obtained from a subset of cDNA clones chosen according to their oligonucleotide fingerprint. To produce a minimal set of sequences, only one representative of each OFP cluster or OFP singletons was chosen for sequencing. Sequences were obtained by an improved approach were cDNA inserts were amplified before by a rolling circle amplification method (2.2.2). Sequence analyses of EST data may be improved by 3'sequencing, but the production of

3'ESTs was not successfully established.

4.3.2 Sequence quality

Like in all EST experiments results largely depend on a good quality of ESTs, which is not an easy task because they are single-pass produced. Therefore low base quality signals were called with PHRED [Ewing et al., 1998, Ewing and Green, 1998] and these bases were together with vector sequences trimmed and not subjected to further analysis. Also sequences shorter than 60 bp were excluded from subsequent experiments like EST clustering. During first round of OFP analysis 6909 high-quality 5' EST sequences with an average length of 654 bp after trimming were obtained. From second round analysis 11,486 high-quality reads were produced, with an average size of 511 bp.

Because of their high coding probability, ESTs should not include too many repetitive elements. Such fragments were identified by RepeatMasker. In 4.5% (449 of 10016 sequences, stretches of low complexity were identified and only 3.5% (349 sequences) contained microsatellites, meaning repeats of 1 to 6 nucleotides. Of all sequences 9% were largely covered with repeats. The proportion of small sequences ($<$ 200 bp) is quite low with 3%.

Possible genomic contamination of RNA samples was analysed by blasting randomly selected public full-length mRNAs against the medaka genome. This was compared to the BLAST analysis of randomly selected EST singletons and EST contigs (tab. 3.13). In cases where public data matched the Medaka genome, always additional public evidence or at least some genscan predicted information was found. EST contigs show similar results as for public full-length coding sequences, but EST singletons maybe show some contamination as the amount of hits to public evidences is clearly reduced (tab. 3.13). Therefore within EST singletons there may be a higher amount of sequences containing genomic DNA but no transcript information.

4.3.3 EST clustering

To cope with the fragmented nature of ESTs these were clustered into bigger contigs. EST sequences were subjected to two rounds of clustering. At first by the application of TGICL all ESTs were automatically grouped and consensus sequences were calculated. These consensus and singleton sequences were during a second round again grouped according to BLAST results and new consensus sequences were calculated. This resulted in 10,016 unique sequences.

EST sequencing and clustering was only done from the 5' end, which may be improved by additional 3' sequencing. [Carninci et al., 2003] stated that 5'-end clustering should only be undertaken with caution, because of alternative promotors and multiple transcriptional starts of the same gene in different tissues.

The EST clustering approach may be improved by obtaining longer sequence reads, which was only partially realised resulting in an average insert length of 565 bp. In case of too short sequence

data there is a high risk that 5' ESTs or EST cluster (and if available 3' ESTs) will not overlap which will in turn also overestimate the number of genes.

4.3.4 EST annotation

The annotation of ESTs was improved by clustering them firstly into longer contigs. Low quality data had already been removed before the clustering and did not influence this task. Annotation was done automatically by BLAST searches. These results were visualised if available by gene ontology means (4.3.5). This makes it easy to filter interesting genes and to decide for further experiments. The annotation results were not further confirmed, but the employment of *in situ* hybridisations within the MEPD project may provide further hints about the function of a protein.

There may arise disadvantages for gene discovery in merely sequencing 5' ends due to large 5' untranslated regions [Carninci et al., 2003]. In many cases no homology is detected at either DNA or protein level between new ESTs and previously described sequenced genes or cDNA sequences because of low similarities between different 5' UTRs of the respective genes and the therefore small portion of coding sequence included within the EST, which reduces further the probability to detect homology to database sequences. To estimate the influence of 5' UTRs firstly the average length of 5' UTRs was searched for. Unfortunately there are only some full-length mRNAs available for Medaka and only one mRNA has an entry in UTRdb [Mignone et al., 2005] with an 5' UTR length of 154 bp. Therefore results for zebrafish (500 mRNAs randomly collected with an average 5' UTR length of 179bp) and fugu (34 mRNAs available in UTRdb with an average 5' UTR length of 230bp) were taken as an estimate. The average sequence length of the unique 10,016 sequences was 625 bp (4.2). Besides the small average 5' UTR length, 3% of zebrafish 5' UTRs are larger than the average length of the unique data set. Therefore around 300 sequences of 10,016 are maybe made up of 5' UTR sequences only.

The classification of ESTs is also problematic because certain motifs have to be present to decide for a certain function but such motifs may lie outside of deduced protein sequence from the respective ESTs (example endoplasmic retention signal present at the C-terminus of proteins).

4.3.5 GO annotation

A nice way to visualise sequence annotation is gene ontology presentation of sequence data. In analysing data this way one has to keep in mind that sequences may be categorised with two functions, e.g one protein can have catalytic function and for this it will maybe bind a protein. Such a protein will be put into at least two groups, "binding, protein binding" and "catalytic activity". This way it is not possible to really dissect annotated sequences into distinct groups. Therefore all numbers have to be treated with caution. But on the other hand it is possible to compare different categories between different species, for instance one can raise questions like, is in Arabidopsis a higher percentage of sequences annotated as binding to nucleotides than in fish.

To 5590 EST clusters or singletons (56% of unique sequences) GO functions were assigned. Within the most interesting categories (concerning this project: regulation of embryonic development, embryonic development, embryonic eye morphogenesis, eye morphogenesis, neurogenesis, glial cell differentiation, neuron differentiation) 488 ESTs or EST clusters were identified. The most functions were categorised into neurogenesis (292 ESTs/EST cluster compared to 38 in gastrulation group). This is remarkable as OFP fingerprints were mainly taken from cDNA clones from a gastrula library. Clones from neurogenesis library were less involved compared to the other three libraries as to their low library quality (in respect to increased amount of vector-only clones and decreased length of amplified clone inserts).

A closer look was given to EST sequences showing hits to proteins involved in eye development. This way homology to key proteins responsible for eye development were detected, including PAX6 and SIX1, which are homebox-containing transcription factors functioning during sensory organ development throughout the animal kingdom [Gehring, 2005]. Also Eyes absent was found, which is as one component of the retinal determination gene network essential for eye fate specification in metazoans [Silver et al., 2003]. Also surprisingly many GTP proteins were categorized into eye development, but also for these literature evidence was found, like Rap1 (Ras-related Rap GTPase which is highly conserved across diverse species) being responsible for regulation of morphogenesis in the Drosophila eye disk [Asha et al., 1999]. For more annotation results see appendix D. These results clearly show the efficiency in filtering important genes by means of applying oligonucleotide fingerprinting to cDNA libraries and using *in silico* GO annotation on sequences.

4.3.6 Non-annotated ESTs

For around half of the unique sequence data set no annotation by gene onotology was obtained. This is largely caused by the little amount of proteins that was GO annotated until the day of analysis (see 2.3). Further BLAST analyses provided further evidence for protein functions of the unique data set. Finally 19% (1861) of sequences were left without similarity to any known sequence, even though this part of unannotated ESTs or EST contigs were not of lower quality (tab. 4.2) or did not contain more repetitive motifs (tab. 3.12). It was observed that the amount of annotated data has grown over time as the amount of public available transcriptome data increases very fast. This indicates that the amount of non-annotated ESTs may still decrease in the future. Lately evidence about non-coding RNAs was collected, which may be polyadenylated and therefore synthesised to cDNA, but which do not code for proteins. Therefore non-coding RNAs may increase the amount of sequence data not showing similarities to any other known proteins.

4.3.7 Estimating gene numbers

To get a good estimation for obtained gene numbers it is crucial to start with full length cDNA clones [Carninci et al., 2003], but which this project was not provided with, as the average insert

	sequences	average contig length	sequences < 200bp	
total 10016 unique sequences	10016	625bp	331	3%
non annotated	1861	510bp	155	8%

Table 4.2: Analysis of the length of ESTs or EST contigs, which were not annotated by any means, showed only little differences to the same analysis of the total set of ESTs.

size was measured with 1021 bp (tab. 4.1), which is smaller than the size of an average coding sequence (1311 bp for worm, 1497 bp for fly, 1340 bp for human; [HGS Consortium, 2001]). Non-full-length clones from low-quality libraries would artificially increase the apparent number of clusters, because of the high risk of splitting full coding sequences from their partial sequence(s).

The number of full transcript length clones among the sequenced cDNAs was estimated with 46% (tab. 4.3). Based on the fact, that poly-dT priming was used for production of the cDNA libraries, this estimation is based on protein searches of the sequences where 39.4% of the 5'-ESTs show stringent protein hits (BLASTX e-value < 1.0e-10) matching the start methionine of the non-redundant protein database sequence, which indicates a clone including the first transcribed exon, and additional 6.2% ESTs match the first 2 to 4 amino acids of query proteins. Because of non-overlapping EST sequences the actual number of genes represented in the rearrayed library could still be lower.

To some extent the amount of non-overlapping ESTs may be estimated from clustering experiments, where project data was clustered together with all publicly available EST data from ricefish (Spring 2005, see 2.3.1). This approach resulted in clustering of 1297 of unique sequences into 605 new clusters together with public data. These clusters were visually analysed for their composition. From the unique data set 761 ESTs did not overlap in the calculated 354 consensus sequence (these are clearly cDNA fragments belonging to one gene), 106 do overlap between 50-200 bp in 52 contigs (this small overlap makes it difficult to cluster these in the case where no further data is available, so these sequences may also belong to one gene) and 430 sequences do overlap by more than 200bp in 199 contigs, but were not grouped during our clustering approach (this may indicate, that they belong to alternative transcripts of one gene or are members of the same gene family). Taking 761+106 (867) sequences into account, the amount of EST fragments belonging to the same gene is estimated with around 8.7% of the obtained unique sequence set. Therefore the data set of 10,016 unique sequences may represent 9,144 genes (in 2005).

4.3.8 The Medaka transcriptome

Within the cooperative project on molecular and genetic analysis of lens-retina interactions in vertebrates four cDNA libraries were subjected to oligonucleotide fingerprinting to obtain results on the Medaka transcriptome. This should provide a rough picture about genes active in certain stages or different tissues.

experiment	OFP1		OFP2		total	
seqs included	6909		10229		17138	
nrprot hits	4616	100%	6300	100%	10916	
1st AA included	1886	40,90%	2389	37,90%	4275	39,40%
2nd to 4th AA included	260	5,60%	419	6,70%	679	6,20%
						45,60%

Table 4.3: To estimate the amount of full-length cDNA clones, high quality traces were blasted and the amount of hits, which included the start methionine were counted.

4.3.8.1 A Medaka gene catalogue

Analyses on the ricefish transcriptome were started with three cDNA libraries (gastrulation, neurula and organogenesis) covering three embryonic stages, which are besides other things important for eye development. The eye development is initiated at late gastrula stage and at the end of gastrulation the eye field is largely determined. During neurula stage were the neural tube is built, also the evagination of optic vesicles happens. Finally during the stage of organogenesis the organs are developed. Additionally one library was used including cDNA inserts synthesised from one adult organ, the ovary.

In concentrating on four cDNA libraries analyses of the transcriptome are already restricted to a certain subset of genes. Additionally during the course of oligonucleotide fingeprinting genes and gene families will also be missed because of various reasons, e.g. because of problems during production of hybridisation filters or during hybridisation work (see possible disadvantages of OFP in 4.2.3). Therefore the picture obtained during this thesis may be quite incomplete. But still interesting findings were taken during all stages of analyses, were examples of differentially expressed genes or alternatively spliced transcripts were identified.

By the means of gene ontology analysis genes were identified responsible for all stages of development, even that one stage, the gastrula stage, was largely overrepresented in this analysis. 488 candidate genes were identified with functions in regulation of embryonic development, embryonic development, embryonic eye morphogenesis, eye morphogenesis, neurogenesis, glial cell differentiation and neuron differentiation. Many of the key proteins responsible for eye development were identified (see 4.3.5). This way the unique sequence data will also provide a source of highly valuable cDNA data in other aspects of research.

4.3.8.2 Differentially expressed genes

237 candidates for differentially expressed genes were identified which show putative functions in *cell signaling and signal transduction* (Cellular retinoic acid-binding protein, Pleiotrophic factor-beta-2 precursor), *cell cycle* (Cyclin B2, RAN protein, beta 1 integrin isoform A), *cell adhesion and cytoskeleton* (keratin, Zona pellucida glycoproteins, Fast muscle troponin I, Claudin-like protein ZF-A89), *transcription* (Small nuclear ribonucleoprotein D1, CCAAT-binding transcription factor

subunit, PP2A inhibitor), *translation and protein metabolism* (Ubiquitin, Choriolysin H 1, Elongation factor 1-alpha, ribosomal proteins), *immune response* (thymosin beta b, Px19-like protein) and *general metabolism* (Cytochrome, Aldolase A fructose-bisphosphate, Apolipoprotein E).

It has to be noted that bigger clusters are more likely to be called differentially expressed. Therefore these results have to be treated with caution. But still the R-test was the most appropriate test for comparing expression of cDNA clones in more than two libraries [Stekel et al., 2000].

4.3.8.3 Alternatively spliced transcripts

Alternative splicing is an important mechanism of modulating gene expression and function. The splicing mechanism during transcription is a carefully controlled process, where 98% of intron sequences have a consensus splice site of 5'GT ... AG 3'. There is little high-throughput experimental data (such as mass spectrometry) available surveying alternative splicing throughout the proteome, but mainly alternative splicing events are searched by EST analysis. The approach used during this project involved EST cluster anaylsis and alignment of EST clusters to the Medaka genome draft sequence, to utilise also information about intronic part of the genome. Only EST clusters were taken into account which showed some sequence similarities but were not clustered during the EST clustering approach. There may still be alternative spliced transcripts within some EST clusters. Therefore these experiments do not give the complete picture about alternative splicing in this project's data set and no estimate about the amount of alternative splicing in Medaka is done. ASAP2 [Kim et al., 2007], a database collecting alternative splicing events in several species, noticed that out of the total set of human multi-exon genes (22,220) 53% were detected to contain alternative splicing. For other species they identified 53% (mouse), 24% (rat), 13% (zebrafish) and 3% (fugu) of alternative splicing rate, but noted that the detection of alternative splicing grows still rapidly as a function of increased EST and mRNA counts, therefore these are only preliminary results. One problem in using EST data arises from genomic contamination of EST data due to mis-priming of the polyA-stretches present in the genomic DNA. This contaminated EST will turn out to be unspliced, therefore cannot be distinguished from retained introns. Further problems of EST data may be incomplete mRNA processing, or library construction artifacts such as chimeric sequences.

By the analysis of Medaka EST clusters 122 alternative splicing cases were found in 101 gene candidates: 49 exon skipping (40% of 122 cases), 7 retained introns (6%), 34 competing 5' donor site (28%), 32 (26%) 3' acceptor site. [Kim et al., 2005] gives classification results of alternative splicing types in human, mouse and rat: 62%, 55%, 44% exon-skipping, 14%, 11%, 2% intron retention, 49%, 45%, 30% 5' donor splice site variation, 51%, 45%, 33% 3' acceptor splice site variation. The rate of intron retention in Medaka lies within the range of the results found in this publication, therefore there is only little contamination with unspliced sequence data. For all other classification types little less events were found than in rat, but because of the small number of candidate genes involved these numbers have to be treated with caution.

These experiments provide evidence for alternatively spliced transcripts within this project's data set. All candidates always need to be checked experimentally, e.g. by RT-PCR. But to give a hint about the values of these experiments it was checked if some of the candidates are reported as alternatively spliced in other organisms. There are several databases providing accesss to alternatively spliced proteins. ASAP2 [Kim et al., 2007] was chosen, where unigene clusters were used to calculate isoform sequences (different protein-coding sequences from a single gene) from these clusters aligned to the respective genome for several different species. Isoform calculation [Xing et al., 2004] is a problematic task because there are many sources of variation producing multiple isoforms: alternative splicing, alternative initiation, polyadenylation, intron retention, nonsense-mediated decay, RNA editing. To distinguish isoforms that give rise to a real protein product from those that probably reflect EST artifacts, calculated isoforms were additionally filtered. They must contain an complete ORF, the translation of each transcript must exceed 50 amino acids and the translation of each minor isoform must match the major isoform by at least 50% (this approach may also miss isoforms of real biological interest). Therefore there may be genes where some transcripts show alternative splicing but no isoforms were calculated.

Also orthologous genes were provided in ASAP2, which were found by multigenome alignments [Kim et al., 2007] constructed for the UCSC genome browser. If exons or introns share at least one of their splice junctions in multigenome alignments, they are annotated as orthologous exons or introns. This was largely used in finding evidence for alternative splicing in orthologous to this project's candidate genes.

Candidate sequences were blasted using NCBI BLAST against the non-redundant nucleotide database (March 2009). The best hit (only hits with e-values smaller than 1e-20 were taken into account) in zebrafish, fugu, human, mouse or rat was taken, because these are available in ASAP2, and the unigene identifier to which the hit sequence belongs to was taken. These unigene IDs were searched in ASAP2 for alternative splicing events. In using this approach the problem of highly dynamic nature of unigene clusters become visible, where lots of these clusters retired after the calculation of ASAP2 in January 2006. Therefore ASAP2 was also searched for the protein name if it is known and this way the old unigene IDs were identified. If an old unigene cluster did not change completely, e.g. only some sequences dropped out or were added, therefore the unigene ID was changed, the result for the old unigene cluster was taken as result for the new unigene cluster. The different events of alternative splicing (exon skipping, 5' competing donor site, 3' competing acceptor site) were noted if provided in ASAP2 and compared to the events noticed in candidates. From all candidate genes for 47 data was found in ASAP2. Of these 47 almost all showed evidence of alternative splicing (see table 4.4 for some examples and supplemental material for all data), 12 candidate genes with evidence found in zf or fugu and 28 candidates with evidence found in other species (human, mouse, rat).

cand.	EST/EST contig	annotation	best hit vs. Danio, Fugu, Human, Mouse, Rat (AC ID, Unigene ID, e-value); NCBI BLAST March 2009	reference in ASAP2; DB of Jan06
competing 5' splice sites				
C11	CL2778Contig1	similar to UP|Q767K9 Protein phosphatase 1 regulatory subunit 10	BC165667; Dr.80488; Danio rerio phosphatase 1, regulatory subunit 10 (ppp1r10); 4e-99	2 isoforms for Dr.6364 (retired; now Dr.80488); 4 isoforms for Mm.29385 (5' donor)
	McF0032E14	similar to UP|Q767K9 Protein phosphatase 1 regulatory subunit 10	BC165667; Dr.80488; Danio rerio phosphatase 1, regulatory subunit 10 (ppp1r10); 3e-88	
C13	McF0031L08	similar to SP|P45973|CBX5_HUMAN Chromobox protein homolog 5 (Heterochromatin protein 1 homolog alpha)(HP1 alpha) (Antigen p25)	BC164295; Dr.104817; Danio rerio chromobox homolog 5 (cbx5); 1e-59	NA for Dr.104817; 4 isoforms for Hs.349283; 1 isoform in Rn.101856
	CL1448Contig1	similar to SP|P45973|CBX5_HUMAN Chromobox protein homolog 5 (Heterochromatin protein 1 homolog alpha)(HP1 alpha) (Antigen p25)	BC164295; Dr.104817; Danio rerio chromobox homolog 5 (cbx5); 5e-84	

continued on next page

cand.	EST/EST contig	annotation	best hit vs. Danio, Fugu, Human, Mouse, Rat (AC ID, Unigene ID, e-value); NCBI BLAST March 2009	reference in ASAP2; DB of Jan06
C15	CL1307Contig1	similar to SP|Q02374|NIGM_BOVIN NADH-ubiquinone oxidoreductase AGGG subunit, mitochondrial precursor (Complex I-AGGG) (CI-AGGG)	BC063026; Hs.655788; Homo sapiens NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 2 (NDUFB2); 6e-32	4 isoforms for Hs.567309 (retired; now mainly Hs.655788; exon skipping); 1 isoform for Rn.18013
	McF0043O22	similar to SP|Q02374|NIGM_BOVIN NADH-ubiquinone oxidoreductase AGGG subunit mitochondrial precursor (Complex I-AGGG) (CI-AGGG)	BC063026; Hs.655788; Homo sapiens NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 2 (NDUFB2); 7e-31	
competing 3' splice sites				
C31	McF0030A06	similar to (Q7SYW3) Xenopus laevis protein	BC155085; Dr.114812; Danio rerio carbonic anhydrase IV a (ca4a); 1e-28	3 isoforms for Hs.89485 (5' donor, exon skipping); 2 isoforms for Rn.51389; 3 isoforms for Mm.1641 (5' donor)
	McF0021C18	similar to (Q7SYW3) Xenopus laevis protein	BC155085; Dr.114812; Danio rerio carbonic anhydrase IV a (ca4a); 7e-38	

continued on next page

cand.	EST/EST contig	annotation	best hit vs. Danio, Fugu, Human, Mouse, Rat (AC ID, Unigene ID, e-value); NCBI BLAST March 2009	reference in ASAP2; DB of Jan06
C32	McF0050N12	similar to UniRef100_Q80ZW8 RIKEN cDNA 1700018O18	NM_032793; Hs.655177; Homo sapiens major facilitator superfamily domain containing 2 (MFSD2), transcript variant 2; 7e-43	1 isoform Mm.331842; 1 isoform for Dr.34052 (retired; now Dr.319); 2 isoforms for orthologous Dr.37117 (retired; now Dr.86869); 2 isoforms for Hs.75668 (mutually exclusive exon; exon skipping)
	McF0002M14	similar to UniRef100_Q96F59 Hypothetical protein FLJ14490	NM_032793; Hs.655177; Homo sapiens major facilitator superfamily domain containing 2 (MFSD2), transcript variant 2; 2e-49	
C37	CL47Contig1	similar to GB|AAH44522.1|27882547| BC044522 wu:fb39h07 protein	NM_001536; Hs.20521; Homo sapiens protein arginine methyltransferase 1 (PRMT1), transcript variant 1; 0.0	4 isoforms for Hs.20521 (5' donor); 2 isoforms for orthologous Mm.27545 (3' acceptor; exon skipping); 1 isoform for Dr.879 (retired; now mainly Dr.110639)
	CL47Contig2	similar to GB|AAH44522.1|27882547| BC044522 wu:fb39h07 protein	BC044522; Dr.110639;Danio rerio protein arginine methyltransferase 1; 4e-179	

continued on next page

cand.	EST/EST contig	annotation	best hit vs. Danio, Fugu, Human, Mouse, Rat (AC ID, Unigene ID, e-value); NCBI BLAST March 2009	reference in ASAP2; DB of Jan06
exon skipping				
C53	McF0018J09	similar to (Q9BTT2) Homo sapiens Hypothetical protein (Fragment)	NM_001045231.1; Dr.75502; novel IBR domain containing protein; 3e-62	3 isoforms for Mm.245537 (1 exon-skipping, early polyA); 1 isoform for Rn.60559; 1 isoform for Dr.5019 (retired, now Dr.75502); 1 isoform for Dr.45833 (retired, now Dr.75502)
	McF0045E12	similar to (Q7Z4H3) Homo sapiens NS5ATP2	NM_027168.2; Mm.245537; Mus musculus HD domain containing 2, Hddc2; 9e-19	
C60	McF0038N20	similar to UP|Q8BV13 Mus musculus 16 days embryo head cDNA RIKEN full-length enriched library clone:C130003E18 product:COP9 (constitutive photomorphogenic) homolog subunit 7b (Arabidopsis thaliana); full insert sequence	BC059697.1; Dr.11003; Danio rerio COP9 constitutive photomorphogenic homolog subunit 7A (cops7a), partial cds; 3e-125	0 isoforms in Dr.11003 (5' donor); 3 isoforms for Mm.1444
	CL2413Contig1	similar to UP|Q8BV13 Mus musculus 16 days embryo head cDNA RIKEN full-length enriched library clone:C130003E18 product:COP9 (constitutive photomorphogenic) homolog subunit 7b (Arabidopsis thaliana); full insert sequence	BC059697.1; Dr.11003; Danio rerio COP9 constitutive photomorphogenic homolog subunit 7A (cops7a), partial cds; 4e-123	

continued on next page

cand.	EST/EST contig	annotation	best hit vs. Danio, Fugu, Human, Mouse, Rat (AC ID, Unigene ID, e-value); NCBI BLAST March 2009	reference in ASAP2; DB of Jan06
C81	CL318Contig1	similar to AAP13460 Breast cancer-associated protein SGA-1M	BC076296; Dr.109068, Nedd4 family interacting protein 1, like (ndfip1l); Danio rerio mRNA, complete cds; 2e-124	2 isoforms for Hs.9788 (5' donor); 3 isoforms for Mm.102496; 1 isoform for Dr.1866
	CL318Contig2	similar to AAP13460 Breast cancer-associated protein SGA-1M	BC076296; Dr.109068, Nedd4 family interacting protein 1, like (ndfip1l); Danio rerio mRNA, complete cds; 1e-94	
different events				
C83	McF0041N07	similar to GB|AAH52118.1|30354417| BC052118 microfibrillar-associated protein 1	BC052118; Dr.77030; Danio rerio microfibrillar-associated protein 1 (mfap1); 4e-111	1 isoform in Dr.10072 (retired, split into Dr.122557 and Dr.77030); 1 isoform for Hs.61418
	CL3487Contig1	similar to GB|AAH52118.1|30354417| BC052118 microfibrillar-associated protein 1	BC052118; Dr.77030; Danio rerio microfibrillar-associated protein 1 (mfap1); 4e-175	
C84	McF0002D22	similar to (Q9QXV3) Inhibitor of growth protein 1	AY624105; Rn.145491 (Ing1); Rattus norvegicus p24ING1c variant 2 mRNA; 4e-36	NA for Rn; 6 isoforms for Mm.25709
	McF0018K06	similar to GB|AAG12175.1|10039549| HSP33ING2 p33ING1	BC166594; Rn.221994 (Ing1); Rattus norvegicus inhibitor of growth family, member 1; 4e-42	

continued on next page

cand.	EST/EST contig	annotation	best hit vs. Danio, Fugu, Human, Mouse, Rat (AC ID, Unigene ID, e-value); NCBI BLAST March 2009	reference in ASAP2; DB of Jan06
C86	CL2783Contig1	similar to GB|AAH48047.1|28856130| BC048047 acidic (leucine-rich) nuclear phosphoprotein 32 family member A	BC048047; Dr.20261; Danio rerio acidic (leucine-rich) nuclear phosphoprotein 32 family, member A; 3e-177	2 isoforms for Dr.20261 (exon skipping); 1 isoform for Hs.458747
	McF0015J02	similar to GB|AAH48047.1|28856130| BC048047 acidic (leucine-rich) nuclear phosphoprotein 32 family member A	BC048047; Dr.20261; Danio rerio acidic (leucine-rich) nuclear phosphoprotein 32 family, member A; 7e-160	

Table 4.4: Annotation of alternative splicing candidates with references in ASAP2.

4.3.9 Contribution to Medaka resources

The provided unique sequence data is an important step towards the medaka gene catalogue. This project's data was already used in other projects on the evaluation of gene or protein functions, for example cooperation partners (Bourrat, Gif-sur-Yvette, France) were provided with EST clones, to support their work on cell proliferation and morphogenesis of the Medaka brain.

An important step towards elucidation of gene function is the location of mRNAs by *in situ* hybridisations which is done within the MEPD project [Henrich et al., 2003] by high-throuhgput means. The unique sequence data of rearrayed libraries produced during this project was used to reduce the scale of this project.

Experiments on differential expression of genes also need unique sequence data. The OFP clustering results presented only little amount of data in respect to this point. More evidence on differential expression may be provided by microarray analyses (oligonucleotide arrays).

Unique sequence data of this project was also used to help in physical mapping of the Medaka genome [Zadeh Khorasani et al., 2004]. Within that project unique sequence data (obtained from publicly available Medaka EST data and this project's data) were aligned to the fugu genome to identify gene-specific markers. Oligonucleotide markers were then designed against non-overlapping cDNA clones from different Fugu scaffolds. These marker were used for construction of a physical map for Medaka.

Despite the available draft sequence of the Medaka genome EST collections are necessary for the

analysis of that genome in helping open reading frame predictions or as test sequences in training programs for intron-exon predictions. Gene expression analyses largely depend on cDNA data, as experiments with proteins are more complicated in handling, especially in large scale projects.

Unique sequence data is also needed to find a Medaka gene number and especially useful for comparative transcriptome analysis, e.g. for questions what makes medaka a medakafish and not a fugu or a zebrafish. For fugu, zebrafish and tetraodon protein sets were calculated from their genome draft sequences, but these are still to a large extent only *in silico* predicted proteins. Data on medaka is mostly confined on EST sequence data. Simple comparison of sets by BLAST analysis is not applicable in including more than two protein sets in this analysis. Protein sets have to be clustered according to their all vs. all BLAST results. This is a very difficult task especially in respect to fish because of their large amount of duplicated genes, which makes it very hard to distinguish paralogous from orthologous genes.

4.3.10 Outlook - How to obtain all Medaka genes?

How do we fish all Medaka genes? To identify all genes by using EST analysis is a tiresome approach, including many new cDNA libraries of different stages or time frames, also the normalisation of these libraries during or after their production, and afterwards massive EST collection (as described for example in [Carninci et al., 2003]), which will always include much redundant work as sequencing of high prevalent transcripts will be not completely avoided.

Another tagging approach has been applied to the Medaka genome [Kasahara et al., 2007], where more than one million 5' SAGE (serial analysis of gene expression) tags corresponding to transcription start sites where obtained to predict 20,141 non-redundant gene structures from the draft Medaka genome sequence. They used cDNA libraries from oligo-capped mRNA preparations to obtain 20mer SAGE tags which contain the transcription start site and therefore simplifies the identification of the entire expressed gene.

Tiling arrays, as a subtype of microarray chips, can produce an unbiased look at gene expression, in case of using whole-genome arrays where partially-overlapping or non-overlapping probes cover the whole genomic sequence. One application of tiling arrays is transcriptome mapping. Because of the use of whole-genome arrays it is possible to identify noncoding RNAs, natural antisense transcripts, RNA that is not polyadenylated, very short RNAs and also RNAs with extensive secondary structure [Mockler and Ecker, 2004], which are usually missed during the EST analysis approach. The detection of rare RNA molecules or the finding of genes that are only active in response to signals or specific to a time frame is easier compared to EST analysis because of the ultra-sensitive detection of hybridisation signals and the use of total RNA of several different tissues or stages is by far easier than to create all these different cDNA libraries for subsequent EST analysis. As for EST analysis hybridisation signals and corresponding gene structures obtained by transcriptome mapping remain predictions that must be confirmed by RT-PCR amplification, cloning of full-length cDNAs, and sequencing. Tiling arrays are also used for the discovery and characteri-

sation of alternatively spliced transcripts [Clark et al., 2002] using overlapping probes tiled across splice junctions. For relatively rare splicing events, still a large number of distinct tissues or cell populations must be surveyed like in transcriptome mapping. Compared to EST analysis, arrays cannot distinguish whether two splicing events observed in one sample occur in the same or distinct transcripts, if the two isoforms are expressed at similar levels. All candidates must be validated by further experimentation such as sequencing of RT-PCR results [Mockler and Ecker, 2004]. The use of tiling arrays is a huge improvement for transcriptome analysis, but there is one big disadvantage of tiling arrays: the price of arrays makes it impractical to actually use genome wide tiling arrays for larger genomes like mammalian (e.g. a tiling array design representing human chromosomes 21 and 22 and interrogating around 35 Mb of nonrepetitive sequence with probes spaced about every 35 bases on average required already 3 chips [Kapranov and et al., 2002]).

MPSS (massively parallel signature sequencing) captures data by counting virtually all mRNA molecules in a tissue or cell sample by generating a 17-base sequence (signature sequence) for each mRNA at a specific site upstream from its poly(A) tail, obtained by a ligation-based method from cDNAs cloned on the surfaces of microbeads [Brenner et al., 2000]. To measure the level of expression of any given gene, the total number of signatures for that gene's mRNA is counted. The assignment of a signature sequence or tag to a specific gene on the genome is much less ambiguos with MPSS than with SAGE (theoretically signature lengths of 14 nucleotides are 80 per cent unique, while the 17-nucleotide signature lengths generated with MPSS are approximately 95 per cent unique on the human genome) [Reinartz et al., 2002]. This method provides very deep transcriptome analyses of individual tissues or cell types, it samples the transcriptome deeply enough to detect transcripts expressed at levels as low as three copies per cell [Lin et al., 2005]. Starting with one million mRNA molecules from a particular cell or tissue sample, one million beads will be produced, each containing 100,000 cloned copies of cDNA from each mRNA molecule. The time, effort and cost to generate data from one million mRNAs in a sample with MPSS is a small fraction of that required to sequence the same number of clones using conventional technologies.

High costs in sequencing maybe circumvented by new sequencing technologies, like the 454 system [Margulies et al., 2005] or the Solexa system [Bennett, 2004]. These methods improve costs and throughput of sequencing work by far, but these technologies are accompanied by a substantial reduction in read-length to 36-120 bases for the Solexa system and to approximately 500 bases for the 454 system, which is no problem in genome re-sequencing, where new sequences just need to be long enough to align to the reference genome [Bentley, 2006]. The 454 system has already been used for expression analysis as described in [Torres et al., 2009], where 454 sequencing of 3' cDNA fragments generated by nebulization was done producing longer reads compared to SAGE or MPSS, which makes this approach a better method if the underlying genome is not available or incomplete.

Application of these new techniques onto the Medaka transcriptome will permit to obtain a complete picture of the *Oryzias latipes* gene catalogue.

Bibliography

[Abrahams et al., 2002] Abrahams, B., Mak, G., Berry, M., Palmquist, D., Saionz, J., Tay, A., Tan, Y., Brenner, S., Simpson, E. and Venkatesh, B. (2002). Novel vertebrate genes and putative regulatory elements identified at kidney disease and NR2E1/fierce loci. Genomics 80, 45–53.

[Adams et al., 1992] Adams, M., Dubnick, M., Kerlavage, A., Moreno, R., Kelley, J., Utterback, T., Nagle, J., Fields, C. and Venter, J. (1992). Sequence identification of 2,375 human brain genes. Nature 355, 632–634.

[Adams et al., 1991] Adams, M., Kelley, J., Gocayne, J., Dubnick, M., Polymeropoulos, M., Xiao, H., Merril, C., Wu, A., Olde, B. and Moreno, R. e. a. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252, 1651–1656.

[Aida, 1921] Aida, T. (1921). On the inheritance of colour in a freshwater fish Aplocheilus latipes Temmninck and Schlegel, with special reference to sex-linked inheritance. Genetics 6, 554–573.

[Altschmied et al., 2002] Altschmied, J., Delfgaauw, J., Wilde, B., Duschl, J., Bouneau, L., Volff, J.-N. and Schartl, M. (2002). Subfunctionalization of duplicate mitf genes associated with differential degeneration of alternative exons in fish. Genetics 161, 259–267.

[Altschul et al., 1990] Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. (1990). Basic local alignment search tool. J Mol Biol 215, 403–410.

[Amores et al., 1998] Amores, A., Force, A., Yan, Y., Joly, L., Amemiya, C., Fritz, A., Ho, R., Langeland, J., Prince, V. and Wang, Y. (1998). Zebrafish hox clusters and vertebrate genome evolution. Science 282, 1711–1714.

[Anderson and Seilhamer, 1997] Anderson, L. and Seilhamer, J. (1997). A comparison of selected mRNA and protein abundances in human liver. Electrophoresis 18, 533–537.

[Aparicio et al., 2002] Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Gelpke, M., Roach, J., Oh, T., Ho, I., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S., Clark, M., Edwards, Y., Doggett, N., Zharkikh, A., Tavtigian, S., Pruss, D., Barnstead, M., Evans, C., Baden, H., Powell, J., Glusman, G., Rowen, L., Hood, L., Tan, Y., Elgar, G., Hawkins,

T., Venkatesh, B., Rokhsar, D. and Brenner, S. (2002). Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297, 1301–1310.

[Aparicio et al., 1997] Aparicio, S., Hawker, K., Cottage, A., Mikawa, Y., Zuo, L., Venkatesh, B., Chen, E., Krumlauf, R. and Brenner, S. (1997). Hox clusters: evidence for continuing evolution of vertebrate Hox complexes. Nat Genet 16, 79–83.

[Aparicio et al., 1995] Aparicio, S., Morrison, A. and Goud, A. (1995). Detecting conserved regulatory elements with the model genome of the japanese puffer fish, $f$ugu rubripes. Proc Natl Acad Sci 92, 1684–1688.

[Arai et al., 1988] Arai, R., Suzuki, A. and Akai, Y. (1988). The karyotype and DNA value of a cypriniform algae eater, Gyrinocheilus aymonieri. Jpn J Ichthyol 34, 515–517.

[Asha et al., 1999] Asha, H., de Ruiter, N., Wang, M. and Hariharan, I. (1999). The rap1 gtpase functions as a regulator of morphogenesis in vivo. EMBO J. 18, 605–615.

[Ashburner and Lewis, 2002] Ashburner, M. and Lewis, S. (2002). On ontologies for biologists: the gene ontology - untangling the web. Novartis Found Symp 247, 66–80.

[Audic and Claverie, 1997] Audic, S. and Claverie, J. (1997). The significance of digital gene expression profiles. Genome Res 7, 986–995.

[Bairoch et al., 2005] Bairoch, A., Apweiler, R., Wu, C., Barker, W., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M., Natale, D., O'Donovan, C., Redaschi, N. and Yeh, L. (2005). The universal protein resource (uniprot). Nucleic Acids Res 33, D154–159.

[Bennett, 2004] Bennett, S. (2004). Solexa ltd. Pharmacogenomics 5, 433–438.

[Benson et al., 2002] Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., Rapp, B. and Wheeler, D. (2002). Genbank. Nucleic Acids Res 30, 17–20.

[Bentley, 2006] Bentley, D. (2006). Whole-genome re-sequencing. Current Opinion in Genetics & Development 16, 545–552.

[Boguski et al., 1993] Boguski, M., Lowe, T. and Tolstoshev, C. (1993). dbest–database for "expressed sequence tags". Nat Genet 4, 332–333.

[Bonaldo et al., 1996] Bonaldo, M., Lennon, G. and Soares, M. (1996). Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res 6, 791–806.

[Bouck et al., 1999] Bouck, J., Yu, W., Gibbs, R. and Worley, K. (1999). Comparison of gene indexing databases. Trends Genet 15, 159–162.

[Breitbart et al., 1987] Breitbart, R., Andreadis, A. and Nadal-Ginard, B. (1987). Alternative splicing: A ubiquitous mechanism for the generation of multiple protein isoforms from single genes. Annu Rev Biochem 56, 467–495.

[Brenner et al., 1993] Brenner, S., Elgar, G., Sandford, R., Macrae, A., Venkatesh, B. and Aparicio, S. (1993). Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. Nature 366, 265–268.

[Brenner et al., 2000] Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. and et al. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nature Biotechnology 18, 630–634.

[Brentani et al., 2003] Brentani, H., et al. Cancer Genome Project, H., Consortium, C. G. A. P. A. and Consortium, H. C. G. P. S. (2003). The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. Proc Natl Acad Sci 100, 13418–13423.

[Candal et al., 2005] Candal, E., Ngyen, V., Joly, J. and Bourrat, F. (2005). Expression domains suggest cell-cycle independent roles of growth-arrest molecules in the adult brain of the medaka, Oryzias latipes. Brain Res Bull 66, 426–430.

[Carninci et al., 2000] Carninci, P., Shibata, Y., Hayatsu, N., Sugahara, Y., Shibata, K., Itoh, M., Konno, H., Okazaki, Y., Muramatsu, M. and Hayashizaki, Y. (2000). Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. Genome Res 10, 1617–1630.

[Carninci et al., 2003] Carninci, P., Waki, K., Shiraki, T., Konno, H., Shibata, K., Itoh, M., Aizawa, K., Arakawa, T., Ishii, Y., Sasaki, D., Bono, H., Kondo, S., Sugahara, Y., Saito, R., Osato, N., Fukuda, S., Sato, K., Watahiki, A., Hirozane-Kishikawa, T., Nakamura, M., Shibata, Y., Yasunishi, A., Kikuchi, N., Yoshiki, A., Kusakabe, m., Gustincich, S., Beisel, K., Pavan, W., Aidinis, V., Nakagawara, A., Held, W., Iwata, H., Kono, T., Nakauchi, H., Lyons, P., Wells, C., Hume, D., Fagiolini, M., Hensch, T. K., Brinkmeier, M., Camper, S., Hirota, J., Mombaerts, P., Muramatsu, M., Okazaki, Y., Kawai, J. and Hayashizaki, Y. (2003). Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. Genome Res 13, 1273–1289.

[Chen et al., 2004] Chen, W.-J., Orti, G. and Meyer, A. (2004). Novel evolutionary relationship among four fish model systems. TRENDS in Genetics 20, 424–431.

[Cheng and Chen, 1999] Cheng, C. and Chen, L. (1999). Evolution of an antifreeze glycoprotein. Nature 401, 443–444.

[Chervitz et al., 1998] Chervitz, S., Aravind, L. and Sherlock, G. (1998). Comparison of the complete protein sets of worm and yeast: orthology and divergence. Science 282, 2022–2028.

[Chiu, 2004] Chiu, C. (2004). Bichir hoxa cluster sequence reveals surprising trends in ray-finned fish genomic evolution. Genome Res 14, 11–17.

[Christoffels et al., 2004] Christoffels, A., Koh, E., Chia, J., Brenner, S., Aparicio, S. and Venkatesh, B. (2004). Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Mol Biol Evol. 21, 1146–1151.

[Clark et al., 2001] Clark, M., Hennig, S., Herwig, R., Clifton, S., Marra, M., Lehrach, H., Johnson, S. and the WU-GSC EST Group (2001). An oligonucleotide fingerprint normalized and expressed sequence tag characterized zebrafish cDNA library. GenomeResearch 11, 1594–1602.

[Clark et al., 2002] Clark, T., Sugnet, C. and Ares Jr., M. (2002). Genomewide analysis of mrna processing in yeast using splicing-specific microarrays. Science 296, 907–910.

[Craig et al., 1990] Craig, A., Nizetic, D., Hoheisel, J., Zehetner, G. and Lehrach, H. (1990). Ordering of cosmid clones covering the herpes simplex virus type i (hsv-i) genome: a test case for fingerprinting by hybridisation. Nucleic Acids Res 18, 2653–2660.

[Cresko et al., 2003] Cresko, W., Yan, Y., Baltrus, D., Amores, A., Singer, A., Rodriguez-Mari, A. and Postlethwait, J. (2003). Genome duplication, subfunction partitioning, and lineage divergence: Sox9 in stickleback and zebrafish. Dev Dyn. 2003 Nov;228(3):480-9. 228, 480–489.

[Deyts et al., 2005] Deyts, C., Candal, E., Joly, J. and Bourrat, F. (2005). An automated in situ hybridization screen in the medaka to identify unknown neural genes. Dev Dyn 234, 698–708.

[Driever et al., 1996] Driever, W., Solnica-Krezel, L., Schier, A., Neuhauss, S., Malicki, J., Stemple, D., Stainier, D., Zwartkruis, F., Abdelilah, S., Rangini, Z., Belak, J. and Boggs, C. (1996). A genetic screen for mutations affecting embryogenesis in zebrafish. Development 123, 37–46.

[Drmanac et al., 1990] Drmanac, R., Strezoska, Z., Labat, I., Drmanac, S. and Crkvenjakov, R. (1990). Reliable hybridization of oligonucleotides as short as six nucleotides. DNA and Cell Biology 9, 527–534.

[Drmanac et al., 1996] Drmanac, S., Stavropoulos, N., Labat, I., Vonau, J., Hauser, B., Soares, M. and Drmanac, R. (1996). Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. Genomics 37, 29–40.

[Eickhoff et al., 2000] Eickhoff, H., Schuchhardt, J., Ivanov, I., Meier-Ewert, S., O'Brien, J., Malik, A., Tandon, N., Wolski, E.-W., Rohlfs, E., Nyarsik, L., Reinhardt, R., Nietfeld, W. and Lehrach, H. (2000). Tissue gene expression analysis using arrayed normalized cDNA libraries. GenomeResearch 10, 1230–1240.

[Ekker et al., 2007] Ekker, S., Stemple, D., Clark, M., Chien, C.-B., Rasooly, R. and Javois, L. (2007). Zebrafish genome project: Bringing new biology to the vertebrate genome field. Zebrafish 4, 239–251.

[Ewing and Green, 1998] Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using phred. ii. Error probabilities. Genome Res 8, 186–194.

[Ewing et al., 1998] Ewing, B., Hillier, L., Wendl, M. and Green, P. (1998). Base-calling of automated sequencer traces using phred. i. Accuracy assessment. Genome Res 8, 175–85.

[Feinberg and Vogelstein, 1983] Feinberg, A. and Vogelstein, B. (1983). A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal Biochem 132, 6–13.

[Fischer et al., 2000] Fischer, C., Ozouf-Costaz, C., Roest Crollius, H., Dasilva, C., Jaillon, O., Bouneau, L., Bonillo, C., Weissenbach, J. and Bernot, A. (2000). Karyotype and chromosome location of characteristic tandem repeats in the pufferfish Tetraodon nigroviridis. Cytogenet Cell Genet 88, 50–55.

[Fishman et al., 1997] Fishman, M., Stainier, D., Breitbart, R. and Westerfield, M. (1997). Zebrafish: genetic and embryological methods in a transparent vertebrate embryo. Methods Cell Biol 52, 67–82.

[Force et al., 1999] Force, A., Lynch, M., Pickett, F., Amores, A., Yan, Y. and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151, 1531–1545.

[Froese and Pauly, 2005] Froese, R. and Pauly, D. (2005). Fishbase. www.fishbase.org; version 03/2005.

[Fuchs et al., 2002] Fuchs, T., Malecova, B., Linhart, C., Sharan, R., Khen, M., Herwig, R., Shmulevich, D., Elkon, R., Seinfath, M., O'Brien, J., Radelof, U., Lehrach, H., Lancet, D. and Shamir, R. (2002). DEFOG: A practical scheme for deciphering families of genes. Genomics 80, 295–302.

[Fukamachi et al., 2001] Fukamachi, S., Shimada, A. and Shima, A. (2001). Mutations in the gene encoding B, a novel transporter protein, reduce melanin content in medaka. Nat Genet 28, 381–385.

[Furutani-Seiki et al., 2004] Furutani-Seiki, M., Sasado, T., Morinaga, C., Suwa, H. and Niwa, K. (2004). A systematic genome-wide screen for mutations affecting organogenesis in medaka, oryzias latipes. Mech Dev 121, 647–658.

[Furutani-Seiki and Wittbrodt, 2004] Furutani-Seiki, M. and Wittbrodt, J. (2004). Medaka and zebrafish, an evolutionary twin study. Mech Dev. 121, 629–637.

[Gehring, 2005] Gehring, W. (2005). New perspectives on eye development and the evolution of eyes and photoreceptors. J. Hered. 96, 171–184.

[GO Consortium, 2000] GO Consortium, T. (2000). Gene ontology: tool for the unification of biology. Nature Genetics 25, 25–29.

[GO Consortium, 2004] GO Consortium, T. (2004). The gene ontology (GO) database and informatics resource. Nucleic Acids Research 32, D258–D261.

[Goodrich, 1926] Goodrich, H. (1926). The development of mendelian characters in aplocheilus latipes. PNAS USA 12, 649–652.

[Grabher et al., 2003] Grabher, C., Henrich, T., Sasado, T., Arenz, A., Wittbrodt, J. and Furutani-Seiki, M. (2003). Transposon-mediated enhancer trapping in medaka. Gene 322, 57–66.

[Grabher and Wittbrodt, 2004] Grabher, C. and Wittbrodt, J. (2004). Efficient activation of gene expression using a heat-shock inducible Gal4/Vp16-UAS system in medaka. BMC Biotechnol 4, 26.

[Groth et al., 2004] Groth, D., Lehrach, H. and Hennig, S. (2004). Goblet: a platform for gene ontology annotation of anonymous sequence data. Nucleic Acids Res. 32, W313–W317. Web Server issue.

[Grützner et al., 1999] Grützner, F., Lutjens, G., Rovira, C., Barnes, D., Ropers, H. and Haaf, T. (1999). Classical and molecular cytogenetics of the pufferfish Tetraodon nigroviridis. Chromosome Res 7, 655–662.

[Guerasimova et al., 2001] Guerasimova, A., Nyarsik, L., Girnus, L., Steinfath, M., Wruck, W., Griffiths, H., Herwig, R., Wierling, C., O'Brien, J., Eickhoff, H., Lehrach, H. and Radelof, U. (2001). New tools for oligonucleotide fingerprinting. BioTechniques 31, 490–495.

[Gygi et al., 1999] Gygi, S., Rochon, Y., Franza, B. and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19, 1720–1730.

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

[Hawkins, 2000] Hawkins, M. (2000). Identification of a third distinct estrogen eceptor and reclassification of estrogen receptors in teleosts. Proc Natl Acad Sci U.S.A. 97, 10751–10756.

[Hawkins et al., 2001] Hawkins, W., Clark, M. and Shima, A. (2001). Four resource centers for fishes: specifies, stocks and services. Mar Biotechnol 3, S239–S248.

[He et al., 2003] He, C., Chen, L., Simmons, M., Li, P., Kim, S. and Liu, Z. (2003). Putative SNP discovery in interspecific hybrids of catfish by comparative EST analysis. Anim Genet 34, 445–448.

[Hedges, 2002] Hedges, S. (2002). The origin and evolution of model organisms. Nat Rev Genet 3, 838–849.

[Hennig et al., 2003] Hennig, S., Groth, D. and Lehrach, H. (2003). Automated gene ontology annotation for anonymous sequence data. Nucleic Acids Res 31, 3712–3715.

[Henrich et al., 2003] Henrich, T., Ramialison, M., Quiring, R., Wittbrodt, B., Furutani-Seiki, M., Wittbrodt, J., Kondoh, H. and Expression Pattern Database, M. (2003). Mepd: a medaka gene expression pattern database. Nucleic Acids Res 31, 72–74.

[Herwig et al., 1999] Herwig, R., Poustka, A., Muller, C., Bull, C., Lehrach, H. and O'Brien, J. (1999). Large-scale clustering of cDNA-fingerprinting data. Genome Res 9, 1093–1105.

[Herwig et al., 2000] Herwig, R., Schmitt, A., M., S., O'Brien, J., Seidel, H., Meier-Ewert, S., Lehrach, H. and Radelof, U. (2000). Information theoretical probe selection for hybridisation experiments. Bioinformatics 16, 890–898.

[Herwig et al., 2002] Herwig, R., Schulz, B., Weisshaar, B., Hennig, S., Steinfath, M., Drungowski, M., Stahl, D., Wruck, W., Menze, A., O'Brien, J., Lehrach, H. and Radelof, U. (2002). Construction of a 'unigene' cDNA clone set by oligonucleotide fingerprinting allows access to 25000 potential sugar beet genes. The Plant J 32, 845–857.

[HGS Consortium, 2001] HGS Consortium, I. (2001). Initial sequencing and analysis of the human genome. Nature 409, 860–921.

[HGS Consortium, 2004] HGS Consortium, I. (2004). Finishing the euchromatic sequence of the human genome. Nature 431, 931–945.

[Hild et al., 2003] Hild, M., Beckmann, B., Haas, S., Koch, B., Solovyev, V., Busold, C., Fellenberg, K., Boutros, M., Vingron, M., Sauer, F., Hoheisel, J. and Paro, R. (2003). An integrated gene annotation and transcriptional profiling approach towards the full gene content of the Drosphila genome. Genome Biol 5, R3. epub.

[Hinegardner, 1976] Hinegardner, R. (1976). The cellular DNA content of sharks, rays and some other fishes. Comp Biochem Physiol 55, 367–370.

[Hinegardner and Rosen, 1972] Hinegardner, R. and Rosen, D. (1972). Cellular DNA content and the evolution of teleostean fishes. Amer Naturalist 106, 621–644.

[Hisaoka and Battle, 1958] Hisaoka, K. and Battle, H. (1958). The normal developmental stages of the zebrafish, Brachydanio rerio (Hamilton-Buchanan). J Morphol 102, 311–328.

[Hisaoka and Firlit, 1960] Hisaoka, K. and Firlit, C. (1960). Further studies on the embryonic development of the zebrafish, Brachydanio rerio (Hamilton-Buchanan). J Morphol 107, 205–225.

[Hoegg et al., 2004] Hoegg, S., Brinkmann, H., Taylor, J. and Meyer, A. (2004). Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. J Mol Evol 59, 190–203.

[Hogenesch et al., 2001] Hogenesch, J., Ching, K., Batalov, S., Su, A., Walker, J., Zhou, Y., Kay, S., Schultz, P. and Cooke, M. (2001). A comparison of the Celera and Ensembl predicted gene sets revelas little overlap in novel genes. Cell 106, 413–415.

[Hoheisel et al., 1991] Hoheisel, J., Lennon, G., Zehetner, G. and H., L. (1991). Use of high coverage reference libraries of drosophila melanogaster for relational data analysis. a step towards mapping and sequencing of the genome. J Mol Biol 220, 903–914.

[Hoheisel et al., 1994] Hoheisel, J., Ross, M., Zehetner, G. and Lehrach, H. (1994). Relational genome analysis using reference libraries and hybridisation fingerprinting. J Biotechnol 35, 121–134.

[Huang and Madan, 1999] Huang, X. and Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Res 9, 868–877.

[Hubank and Schatz, 1994] Hubank, M. and Schatz, D. (1994). Identifying differences in mRNA expression by representational difference analysis of cDNA. Nucleic Acids Research 22, 5640–5648.

[Hubbard et al., 2005] Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X. M., Gilbert, J., Hammond, M., Herrero, J., Hotz, H., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Kokocinsci, F., London, D., Longden, I., McVicker, G., Melsopp, C., Meidl, P., Potter, S., Proctor, G., Rae, M., Rios, D., Schuster, M., Searle, S., Severin, J., Slater, G. Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C. and Birney, E. (2005). Ensembl 2005. Nucleic Acids Res. 33, D447–D453.

[Hyodo-Taguchi, 1996] Hyodo-Taguchi, Y. (1996). Inbred strains of the medaka, oryzias latipes. The Fish Biol. J. Medaka 8, 11–14.

[Hyodo-Taguchi and Egami, 1985] Hyodo-Taguchi, Y. and Egami, N. (1985). Establishment of inbred strains of the medaka oryzias latipes and the usefulness of the strains for biomedical research. Zool.Sci. 2, 305–316.

[Iriki, 1932] Iriki, S. (1932). Studies on the chromosomes of pisces. on the chromosomes of Aplocheilus latipes. Sci Rep Tokyo Bunrika Daigaku, Sec. B 1, 127–131.

[Ishida, 1944a] Ishida, J. (1944a). Hatching enzyme in the fresh-water fish, oryzias latipes. Annot Zool Japon 22, 137–154.

[Ishida, 1944b] Ishida, J. (1944b). Further studies on the hatching enzyme in the fresh-water fish, oryzias latipes. Annot Zool Japon 22, 155–164.

[Ishikawa, 2000] Ishikawa, Y. (2000). Medakafish as a model system for vertebrate developmental genetics. Bioessays 22, 487–495.

[Iwamatsu, 2004] Iwamatsu, T. (2004). Stages of normal development in the medaka oryzias latipes. Mech Dev. 121, 605–618.

[Jaillon et al., 2004] Jaillon, O., Aury, J., Brunet, F. and Petit, J. (2004). Genome duplication in the teleost fish tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431, 946–957.

[Johnson et al., 2005] Johnson, J., Edwards, S., Shoemaker, D. and Schadt, E. (2005). Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. Trends Genet 21, 93–102.

[Johnson et al., 1996] Johnson, S., Gates, M., Johnson, M., Talbot, W., Horne, S., Baik, K., Rude, S., Wong, J. and Postlethwait, J. (1996). Centromere-linkage analysis and consolidation of the zebrafish genetic map. Genetics 142, 1277–1288.

[Kamito, 1928] Kamito, A. (1928). Early development of the japanese killifish (oryzias latipes), with notes on its habitats. J.Coll.Agric.Tokyo Univ. 10, 21–38.

[Kapranov and et al., 2002] Kapranov, P. and et al. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. Science 296, 916–919.

[Kasahara et al., 2007] 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., Hashimoto, S., Yang, J., Lee, Y., Matsushima, K., Sugano, S., Sakaizumi, M., Narita, T., Ohishi, K., Haga, S., Ohta, F., Nomoto, H., Nogata, K., Morishita, T., Endo, T., Shin-I, T., Takeda, H., Morishita, S. and Kohara, Y. (2007). The medaka draft genome and insights into vertebrate genome evolution. Nature 447, 714–719.

[Katayama, 1937] Katayama, M. (1937). On the spermatogenesis of the teleost, oryzias latipes (t. and s.). Bull Japan Soc Sci Fish 5, 277–278. (In Japanese).

[Khan et al., 1992] Khan, A., Wilcox, A., Polymeropoulos, M., Hopkins, J., Stevens, T., Robinson, M., Orpana, A. and Sikela, J. (1992). Single pass sequencing and physical and genetic mapping of human brain cDNAs. Nature Genet 2, 180–185.

[Kim et al., 2007] Kim, N., Alekseyenko, A., Roy, M. and Lee, C. (2007). The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. Nucleic Acids Research 35, D93–D98.

[Kim et al., 2005] Kim, N., Shin, S. and Lee, S. (2005). ECgene: Genome-based EST clustering and gene modeling for alternative splicing. Genome Research 215, 566–576.

[Kimmel et al., 1995] Kimmel, C., Ballard, W., Kimmel, S., Ullmann, B. and Schilling, T. (1995). Stages of embryonic development of the zebrafish. Dev Dyn 203, 253–310.

[Kimmel et al., 1990] Kimmel, C., Warga, R. and Schilling, T. (1990). Origin and organization of the zebrafish fate map. Development 108, 581–594.

[Kimura et al., 2004] Kimura, T., Jindo, T., Narita, T., Naruse, K., Kobayashi, D., Shin-I, T., Kitagawa, T., Sakaguchi, T., Mitani, H., Shima, A., Kohara, Y. and Takeda, H. (2004). Large-scale isolation of ESTs from medaka embryos and its application to medaka developmental genetics. Mech Dev 121, 915–932.

[Kimura et al., 2005] Kimura, T., Yoshida, K., Shimada, A., Jindo, T., Sakaizumi, M., Mitani, H., Naruse, K., Takeda, H., Inoko, H., Tamiya, G. and Shinya, M. (2005). Genetic linkage map of medaka with polymerase chain reaction length polymorphisms. Gene 363, 24–31.

[Kingsley et al., 2004] Kingsley, D. M., Zhu, B., Osoegawa, K., de Jong, P. J., Schein, J., Marra, M., Peichel, C., Amemiya, C., Schluter, D., Balabhadra, S., Friedlander, B.and Cha, Y. M., Dickson, M., Grimwood, J., Schmutz, J., Talbot, W. S. and Myers, R. (2004). New genomic tools for molecular studies of evolutionary change in threespine sticklebacks. Behaviour 141, 1331–81344.

[Kondo et al., 2002] Kondo, M., Froschauer, A., Kitano, A., Nanda, I., Hornung, U., Volff, J., Asakawa, S., Mitani, H., Naruse, K., Tanaka, M., Schmid, M., Shimizu, N., Schartl, M. and Shima, A. (2002). Molecular cloning and characterization of DMRT genes from the medaka oryzias latipes and the platyfish xiphophorus maculatus. Gene 295, 213–222.

[Kondo et al., 2001] Kondo, S., Kuwahara, Y., Kondo, M., Naruse, K., Mitani, H., Wakamatsu, Y., Ozato, K., Asakawa, S., Shimizu, N. and Shima, A. (2001). The medaka rs-3 locus required for scale development encodes ectodysplasin-a receptor. Curr Biol 11, 1202–1206.

[Kruglyak, 1997] Kruglyak, L. (1997). The use of a genetic map of biallelic markers in linkage studies. Nat Genet 17, 21–24.

[Kubota et al., 1992] Kubota, Y., Shimada, A. and Shima, A. (1992). Detection of gamma-ray-induced DNA damages in malformed dominant lethal embryos of the Japanese medaka (oryzias latipes) using AP-PCR fingerprinting. Mutat Res. 1992 Dec;283(4):263-70. 283, 263–270.

[Kubota et al., 1995] Kubota, Y., Shimada, A. and Shima, A. (1995). DNA alterations detected in the progeny of paternally irradiated Japanese medaka fish (oryzias latipes). Proc Natl Acad Sci U S A. 92, 330–334.

[Kumar and Hedges, 1998] Kumar, S. and Hedges, S. (1998). A molecular timescale for vertebrate evolution. Nature 392, 917–920.

[Laale, 1977] Laale, H. (1977). The biology and use of zebrafish, Brachydanio rerio in fisheries research. A literature review. J Fish Biol 10, 121–173.

[Lamatsch et al., 2000] Lamatsch, D., Steinlein, C., Schmid, M. and Schartl, M. (2000). Noninvasive determination of genome size and ploidy level in fishes by flow cytometry: detection of triploid Poecilia formosa. Cytometry 39, 91–95.

[Lehnert et al., 2001] Lehnert, V., Holzwarth, J., Ott, M., Thompson, A., Demmak, S. and Foernzler, D. (2001). A semi-automated system for analysis and storage of SNPs. Hum Mutat 17, 243–254.

[Li, 1980] Li, W. (1980). Rate of gene silencing at duplicate loci: A theoretical study and interpretation of data from tetraploid fishes. Genetics 95, 237–258.

[Li et al., 2002] Li, Y., Hill, J., Yue, G., Chen, F. and Orban, L. (2002). Extensive search does not identify genomic sex marker in tetraodon nigroviridis. J Fish Biol 61, 1314–1317.

[Liang et al., 2000] Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. and Quackenbush, J. (2000). An optimized protocol for analysis of EST sequences. Nucleic Acids Res 28, 3657–3665.

[Liang and Pardee, 1995] Liang, P. and Pardee, A. (1995). Recent advances in differential display. Curr Opin Immunol 7, 274–280.

[Liang and Pardee, 1998] Liang, P. and Pardee, A. (1998). Differential display. A general protocol. Mol Biotechnology 10, 261–267.

[Lin et al., 2005] Lin, B., White, J., Lu, W., Xie, T. and Utleg, A. (2005). Evidence for the presence of disease-perturbed networks in prostate cancer cells by genomic and proteomic analyses: a systems approach to disease. Cancer Res 65, 3081–3091.

[Lockhart et al., 1996] Lockhart, D., Dong, H., Byrne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14, 1675–1680.

[Loosli et al., 2000] Loosli, F., Koster, R., Carl, M., Kuhnlein, R., Henrich, T. and Mucke, M. (2000). A genetic screen for mutations affecting embryonic development in medaka fish (oryzias latipes). Mech Dev 97, 133–139.

[Loosli et al., 2003] Loosli, F., Staub, W., Finger-Baier, K., Ober, E., Verkade, H., Wittbrodt, J. and Baier, H. (2003). Loss of eyes in zebrafish caused by mutation of chokh/rx3. EMBO Rep 4, 894–899.

[Loosli et al., 2001] Loosli, F., Winkler, S., Burgtorf, C., Wurmbach, E., Ansorge, W., Henrich, T., Grabher, C., Arendt, D., Carl, M., Krone, A., Grzebisz, E. and Wittbrodt, J. (2001). Medaka eyeless is the key factor linking retinal determination and eye growth. Development 128, 4035–4044.

[Lynch and Conery, 2000] Lynch, M. and Conery, J. (2000). The evolutionary fate and consequences of duplicate genes. Science 290, 1151–1155.

[Lynch and Force, 2000] Lynch, M. and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. Genetics 154, 459–473.

[Lynch et al., 2001] Lynch, M., O'Hely, M., Walsh, B. and Force, A. (2001). The probability of preservation of a newly arisen gene duplicate. Genetics 159, 1789–1804.

[MacQueen, 1967] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, (LeCam, L. and Neyman, J., eds), vol. 1, pp. 281–297, University of California Press, Los Angeles, CA.

[Maier et al., 1994] Maier, E., Meier-Ewert, S., Ahmadi, A., Curtis, J. and Lehrach, H. (1994). Application of robotic technology to automated sequence fingerprint analysis by oligonucleotide hybridisation. J Biotechnol 35, 191–203.

[Maissey, 1996] Maissey, J. (1996). Discovering fossil fishes. Henry Holt and Company, New York.

[Malde et al., 2003] Malde, K., Coward, E. and Jonassen, I. (2003). Fast sequence clustering using a suffix array algorithm. Bioinformatics 19, 1221–1226.

[Margulies et al., 2005] Margulies, M., Egholm, M., Altman, W., Attiya, S., Bader, J., Bemben, L., Berka, J., Braverman, M., Chen, Y., Chen, Z. and et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380.

[Marquart et al., 1999] Marquart, J., Alexief-Damianof, C., Preiss, A. and Maier, D. (1999). Rapid divergence in the course of drosophila evolution reveals structural important domains of the Notch antagonist Hairless. Dev Genes Evol 209, 155–164.

[Marth et al., 1999] Marth, G., Korf, I., Yandell, M., Yeh, R., Gu, Z., Zakeri, H., Stitziel, N., Hillier, L., Kwok, P. and Gish, W. (1999). A general approach to single-nucleotide polymorphism discovery. Nat Genet 23, 452–456.

[Matsuda et al., 2001] Matsuda, M., Kawato, N., Asakawa, S., Shimizu, N., Nagahama, Y., Hamaguchi, S., Sakaizumi, M. and Hori, H. (2001). Construction of a BAC library derived from the inbred Hd-rR strain of the teleost fish, oryzias latipes. Genes Genet Syst. 2001 Feb;76(1):61-3 76, 61–63.

[McClintock, 2002] McClintock, J. (2002). Knockdown of duplicated zebrafish hoxb1 genes reveals distinct roles in hindbrain patterning and a novel mechanism of duplicate gene retention. Development 129, 2339–2354.

[McCune et al., 2004] McCune, A., Houle, D., McMillan, K., Annable, R. and Kondrashov, A. (2004). Two classes of deleterious recessive alleles in a natural population of zebrafish, Danio rerio. Proc Biol Sci 271, 2025–2033.

[McKinnon et al., 2004] McKinnon, J., Mori, S., Blackman, B., David, L., Kingsley, D., Jamieson, L., Chou, J. and Schluter, D. (2004). Evidence for ecology's role in speciation. Nature 429, 294–298.

[Megy et al., 2002] Megy, K., Audic, S. and Claverie, J. (2002). Heart-specific genes revealed by expressed sequence tag (EST) sampling. Genome Biol 3, Research0074.1–Research0074.11. epub.

[Meier-Ewert et al., 1998] Meier-Ewert, S., Lange, J., Gerst, H., Herwig, R., Schmitt, A., Freund, J., Elge, T., Mott, R., Herrmann, B. and Lehrach, H. (1998). Comparative gene expression profiling by oligonucleotide fingerprinting. Nucleic Acids Res 26, 2216–2223.

[Meyer, 1990] Meyer, A. (1990). Morphometrics and allometry of the trophically polymorphic cichlid fishes, cichlasoma citrinellum: alternative adaptations and ontogenetic changes in shape. J Zool 221, 237–260.

[Mignone et al., 2005] Mignone, F., Grillo, G., Licciulli, F., Iacono, M., Liuni, S., Kersey, P., Duarte, J., Saccone, C. and Pesole, G. (2005). Utrdb and utrsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mrnas. Nucleic Acids Research 33, D141–D146.

[Mirkin, 1996] Mirkin, B. (1996). Mathematical classification and clustering. Kluwer Academic Publishing, Dordrecht, The Netherlands.

[Mockler and Ecker, 2004] Mockler, T. and Ecker, J. (2004). Applications of dna tiling arrays for whole-genome analysis. Genomics 85, 1–15.

[Modrek and Lee, 2002] Modrek, B. and Lee, C. (2002). A genomic view of alternative splicing. Nature Genetics 30, 13–19.

[Montpetit et al., 2003] Montpetit, A., Wilson, M., Chevrette, M., Koop, B. F. and Sinnett, D. (2003). Analysis of the conservation of synteny between fugu and human chromosome 12. BMC Genomics 4, 30–37.

[Mott, 1997] Mott, R. (1997). EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. CABIOS 13, 477–478.

[Nadeau and Sankoff, 1997] Nadeau, J. and Sankoff, D. (1997). Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics 147, 1259–1266.

[Naruse et al., 2004] Naruse, K., Hori, H., Shimizu, N., Kohara, Y. and Takeda, H. (2004). Medaka genomics: a bridge between mutant phenotype and gene function. Mech Dev 121, 619–628.

[Naruse et al., 1994] Naruse, K., Sakaizumi, M. and Shima, A. (1994). Medaka as a model organism for research in experimental biology. The Fish Biology Journal Medaka 6, 47–52.

[Nobrega and Pennacchio, 2004] Nobrega, M. and Pennacchio, L. (2004). Comparative genomic analysis as a tool for biological discovery. J Physiol 554, 31–39.

[Ohno, 1970] Ohno, S. (1970). Evolution by gene duplication. Springer-Verlag, New York.

[Okazaki et al., 2002] Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R. and Suzuki, H. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420, 563–573.

[Ozato et al., 1986] Ozato, K., Kondoh, H., Inohara, H., Iwamatsu, T., Wakamatsu, Y. and Okada, T. (1986). Production of transgenic fish: introduction and expression of chicken delta-crystallin gene in medaka embryos. Cell Differ. 19, 237–244.

[Panopoulou et al., 2003] Panopoulou, G., Hennig, S., Groth, D., Krause, A., Poustka, A., Herwig, R., Vingron, M. and Lehrach, H. (2003). New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. Genome Res 13, 1056–1066.

[Pertea et al., 2003] Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J. and Quackenbush, J. (2003). TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19, 651–652.

[Phillips and Rab, 2001] Phillips, R. and Rab, P. (2001). Chromosome evolution in the Salmonidae (Pisces): an update. Biol Rev Camb Philos Soc 76, 1–25.

[Pontius et al., 2003] Pontius, J., Wagner, L. and Schuler, G. (2003). Unigene: a unified view of the transcriptome. In The NCBI Handbook. Bethesda (MD): National Center for Biotechnology Information.

[Postlethwait et al., 1994] Postlethwait, J., Johnson, S., Midson, C., Talbot, W., Gates, M., Ballinger, E., Africa, D., Andrews, R., Carl, T., Eisen, J. and et al. (1994). A genetic linkage map for the zebrafish. Science 264, 699–703.

[Postlethwait and Talbot, 1997] Postlethwait, J. and Talbot, W. (1997). Zebrafish genomics: from mutants to genes. Trends Genet 13, 183–190.

[Postlethwait et al., 1998] Postlethwait, J., Yan, Y.-L., Gates, M. and Horne, S. (1998). Vertebrate genome evolution and the zebrafish gene map. Nature genetics 18, 345–349.

[Poustka, 2000] Poustka, A. (2000). Sea Urchin OFP. PhD thesis, Universitaet Salzburg.

[Poustka et al., 2003] Poustka, A., Groth, D., Hennig, S., Thamm, S., Cameron, A., Beck, A., Reinhardt, R., Herwig, R., Panopoulou, G. and Lehrach, H. (2003). Generation, annotation, evolutionary analysis, and database integration of 20,000 unique sea urchin EST clusters. Genome Res 13, 2736–2746.

[Poustka et al., 1999] Poustka, A., Herwig, R., Krause, A., Hennig, S., Meier-Ewert, S. and Lehrach, H. (1999). Toward the gene catalogue of sea urchin development: The construction and analysis of an unfertilized egg cDNA library highly normalized by oligonucleotide fingerprinting. Genomics 59, 122–133.

[Quackenbush et al., 2001] Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R. and White, J. (2001). The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res 29, 159–164.

[Quiring et al., 2004] Quiring, R., Wittbrodt, B., Henrich, T., Ramialison, M., Burgtorf, C., Lehrach, H. and Wittbrodt, J. (2004). Large-scale expression screening by automated whole-mount in situ hybridization. Mech Dev 121, 971–976.

[Radelof et al., 1998] Radelof, U., Hennig, S., Seranski, P., Steinfath, M., Ramser, J., Reinhardt, R., Poustka, A., Francis, F. and Lehrach, H. (1998). Preselection of shotgun clones by oligonucleotide fingerprinting: an efficient and high throughput strategy to reduce redundancy in large-scale sequencing projects. NAR 26, 5358–5364.

[Rasooly et al., 2003] Rasooly, R., Henken, D., Freeman, N., Tompkins, L., Badman, D., Briggs, J., Hewitt, A. and of Health Trans-NIH Zebrafish Coordinating Committee., N. I. (2003). Genetic and genomic tools for zebrafish research: the nih zebrafish initiative. Dev Dyn 228, 490–496.

[Reese et al., 2000] Reese, M., Hartzell, G., Harris, N., Ohler, U., Abril, J. and Lewis, S. (2000). Genome annotation assessment in drosophila melanogaster. Genome Res 10, 483–501.

[Reinartz et al., 2002] Reinartz, J., Bruyns, E., Lin, J.-Z., Burcham, T. and Brenner, S. e. a. (2002). Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. Brief Funct Genomic Proteomic 1, 95–104.

[Rice et al., 2000] Rice, P., Longden, I. and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16, 276–277.

[Roberts and Smith, 2002] Roberts, G. and Smith, C. (2002). Alternative splicing: combinatorial output from the genome. Current Opinion in Chemical Biology 6, 375–383.

[Robinson-Rechavi et al., 2001a] Robinson-Rechavi, M., Marchand, O., Escriva, H., Bardet, P.-L., Zelus, D., Hughes, S. and Laudet, V. (2001a). Euteleost fish genomes are characterized by expansion of gene families. Genome Research 11, 781–788.

[Robinson-Rechavi et al., 2001b] Robinson-Rechavi, M., Marchand, O., Escriva, H. and Laudet, V. (2001b). An ancestral whole-genome duplication may not have been responsible for the abundance of duplicated fish genes. Current Biology 11, R458–R459.

[Sakaizumi and Jeon, 1987] Sakaizumi, M. and Jeon, S. (1987). Two divergent groups in the wild population of medaka oryzias latipes (pisces: oryziatidae) in korea. Korean J. Limnol. 20, 13–20.

[Sakaizumi et al., 1983] Sakaizumi, M., Moriwaki, K. and Egami, N. (1983). Allozymic variation and regional differentiation in wild populations of the fish oryzias latipes. Copeia 2, 311–318.

[Salzburger and Meyer, 2004] Salzburger, W. and Meyer, A. (2004). The species flocks of east african cichlid fishes: recent advances in molecular phylogenetics and population genetics. Naturwissenscahften 91, 277–290.

[Schartl, 1995] Schartl, M. (1995). Platyfish and swordtails: a genetic system for the analysis of molecular mechanisms in tumor formation. Trends Genet 11, 185–189.

[Schartl, 2004] Schartl, M. (2004). A comparative view on sex determination in medaka. Mech Dev 121, 639–645.

[Schena et al., 1995] Schena, M., Shalon, D., Davis, R. and Brown, P. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270, 467–470.

[Schmid et al., 2003] Schmid, K., Sorensen, T., Stracke, R., Torjek, O., Altmann, T., Mitchell-Olds, T. and Weisshaar, B. (2003). Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in Arabidopsis thaliana. Genome Res 13, 1250–1257.

[Schulze-Kremer, 2002] Schulze-Kremer, S. (2002). Ontologies for molecular biology and bioinformatics. In Silico Biol 2, 179–193.

[Schwartz et al., 2000] Schwartz, S., Zhang, Z., Frazer, K., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R. and Miller, W. (2000). PipMaker - A web server for aligning two genomic DNA sequences. Genome Research 10, 577–586.

[Shannon, 1948] Shannon, C. (1948). A mathematical theory of communication. Bell Syst Tech J 27, 379–423.

[Shima et al., 2003] Shima, A., Himmelbauer, H., Mitani, H., Furutani-Seiki, M., Wittbrodt, J. and Schartl, M. (2003). Fish genomes flying. Symposium on Medaka genomics. EMBO Rep 4, 121–125.

[Shima and Mitani, 2004] Shima, A. and Mitani, H. (2004). Medaka as a research organism: past, present and future. Mech Dev. 121, 599–604.

[Shimada and Shima, 1998] Shimada, A. and Shima, A. (1998). Combination of genomic DNA fingerprinting into the medaka specific-locus test system for studying environmental germ-line mutagenesis. Mutat Res. 399, 149–165.

[Silver et al., 2003] Silver, S., Davies, E., Doyon, L. and Rebay, I. (2003). Functional dissection of Eyes absent reveals new modes of regulation within the retinal determination gene network. Mol. Cell. Biol. 23, 5989–5999.

[Smith et al., 2002] Smith, S., Snell, P., Gruetzner, F., Bench, A., Haaf, T., Metcalfe, J., Green, A. and Elgar, G. (2002). Analyses of the extent of shared synteny and conserved gene orders between the genome of Fugu rubripes and human 20q. Genome Res 12, 776–784.

[Soares et al., 1994] Soares, M., Bonaldo, M., Jelene, P., Su, L., Lawton, L. and Efstatiadis, A. (1994). Construction and characterization of a normalized cDNA library. Proc Natl Acad Sci 91, 9228–9232.

[Staden et al., 2000] Staden, R., Beal, K. and Bonfield, J. (2000). The Staden package, 1998. Methods Mol Biol 132, 115–130.

[Stahl, 1985] Stahl, F. (1985). George Streisinger (December 27, 1927-August 11, 1984). Genetics 109, 1–2.

[Steinfath et al., 2001] Steinfath, M., Wruck, W., Seidel, H., Lehrach, H., Radelof, U. and O'Brien, J. (2001). Automated image analysis for array hybridization experiments. Bioinformatics 17, 634–641.

[Stekel et al., 2000] Stekel, D., Git, Y. and Falciani, F. (2000). The comparison of gene expression from multiple cdna libraries. Genome Res 10, 2055–2061.

[Stevens et al., 2000] Stevens, R., Goble, C. and Bechhofer, S. (2000). Ontology-based knowledge representation for bioinformatics. Brief Bioinform 1, 398–414.

[Stiassny and Meyer, 1999] Stiassny, M. and Meyer, A. (1999). Cichlids of the rift lakes. The extraordinary diversity of cichlid fishes challenges entrenched ideas of how quickly new species can arise. Sci Am 280, 44–49.

[Streisinger et al., 1986] Streisinger, G., Singer, F., Walker, C., Knauber, D. and Dower, N. (1986). Segregation analyses and gene-centromere distances in zebrafish. Genetics. 112, 311–319.

[Streisinger et al., 1981] Streisinger, G., Walker, C., Dower, N., Knauber, D. and Singer, F. (1981). Production of clones of homozygous diploid zebrafish (Brachydanio rerio). Nature 291, 293–296.

[Syvänen, 2001] Syvänen, A.-C. (2001). Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat Rev Geneticcs 2, 930–942.

[Takehana et al., 2003] Takehana, Y., Nagai, N., Matsuda, M., Tsuchiya, K. and Sakaizumi, M. (2003). Geographic variation and diversity of the cytochrome b gene in japanese wild populations of medaka, Oryzias latipes. Zoolog Sci 20, 1279–1291.

[Takehana et al., 2004] Takehana, Y., Uchiyama, S., Matsuda, M., Jeon, S. and Sakaizumi, M. (2004). Geographic variation and diversity of the cytochrome b gene in wild populations of medaka (oryzias latipes) from korea and china. Zoolog Sci. 21, 483–491.

[Taylor et al., 2001] Taylor, J., Van de Peer, Y. and Meyer, A. (2001). Revisiting recent challenges to the ancient fish-specific genome duplication hypothesis. Current Biology 11, R1005–R1007.

[Temminck and Schlegel, 1846] Temminck, J. and Schlegel, H. (1846). In Fauna Japonica, (von Siebold, ed.), pp. 224–225. A. Arnz et Socios, Leiden. ,plate 103, fig. V.

[Torres et al., 2009] Torres, T., Metta, M., Ottenwälder, B. and Schlötterer, C. (2009). Gene expression profiling by massively parallel sequencing. Genome Research 18, 172–177.

[Toyama, 1916] Toyama, K. (1916). On some examples of Mendelian characters. Rep Jap Breed Soc 1, 1–19. in Japanese.

[Trower et al., 1996] Trower, M., Orton, S., Purvis, I., Sanseau, P., Riley, J., Christodoulou, C., Burt, D., See, C., Elgar, G., Sherrington, R., Rogaev, E., George-Hyslop, P., Brenner, S. and Dykes, C. (1996). Conservation of synteny between the genome of the pufferfish (Fugu rubripes) and the region on human chromosome 14 (14q24.3) associated with familial Alzheimer disease (ad3 locus). Proc Natl Acad Sci U.S.A. 20, 1366–1369.

[Van de Peer, 2004] Van de Peer, Y. (2004). Tetradon genome confirms Takifugu findings: most fish are ancient polyploids. Genome Biol 5, 250.

[Van de Peer et al., 2002] Van de Peer, Y., Frickey, T., Taylor, J. and Meyer, A. (2002). Dealing with saturation at the amino acid level: a case study based on anciently duplicated zebrafish genes. Gene 295, 205–211.

[Van de Peer et al., 2003] Van de Peer, Y., Taylor, J. and Meyer, A. (2003). Are all fishes ancient polyploids? J of Structural and Functional Genomics 2, 65–73.

[Vandepoele et al., 2004] Vandepoele, K.and De Vos, W., Taylor, J., Meyer, A. and Van de Peer, Y. (2004). Major events in the genome evolution of vertebrates: Paranome, age and size differ considerably between ray-finned fishes and land vertebrates. PNAS 101, 1638–1643.

[Velculescu et al., 1999] Velculescu, V., Madden, S., Zhang, L., Lash, A., Yu, J., Rago, C., Lal, A., Wang, C., Beaudry, G., Ciriello, K., Cook, B., Dufault, M., Ferguson, A., Gao, Y., He, T., Hermeking, H., Hiraldo, S., Hwang, P., Lopez, M., Luderer, H., Mathews, B., Petroziello, J., Polyak, K., Zawel, L. and Kinzler, K. (1999). Analysis of human transcriptomes. Nat Genet 23, 387–388.

[Velculescu et al., 1995] Velculescu, V., Zhang, L., Vogelstein, B. and Kinzler, K. (1995). Serial analysis of gene expression. Science 270, 484–487.

[Venkatesh, 2003] Venkatesh, B. (2003). Evolution and diversity of fish genomes. Curr Opin Genet Dev 13, 588–592.

[Venter, 1993] Venter, J. (1993). Identification of new human receptor and transporter genes by high throughput cDNA (EST) sequencing. J Pharm Pharmacol 45, 355–360.

[Verheyen et al., 2003] Verheyen, E., Salzburger, W., Snoeks, J. and Meyer, A. (2003). Origin of the superflock of cichlid fishes from Lake Victoria, East Africa. Science 300, 325–329.

[Vingron and Hoheisel, 1999] Vingron, M. and Hoheisel, J. (1999). Computational aspects of expression data. J Mol Med 77, 3–7.

[Volff et al., 2003] Volff, J.-N., Bouneau, L., Ozouf-Costaz, C. and Fischer, C. (2003). Diversity of retransposable elements in compact pufferfish genomes. TRENDS in Genetics 19, 674–678.

[Wada et al., 1995] Wada, H., Naruse, K., Shimada, A. and Shima, A. (1995). Genetic linkage map of a fish, the Japanese medaka Oryzias latipes. Mol Mar Biol Biotechnol. 4, 269–74. Erratum in: Mol Mar Biol Biotechnol 1996 Sep;5(3):239.

[Wilcox et al., 1991] Wilcox, A., Khan, A., Hopkins, J. and Sikela, J. (1991). Use of 3' untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: Implications for an expression map of the genome. Nucleic Acids Res 19, 1837–1843.

[Wilming et al., 2008] Wilming, L., Gilbert, J., Howe, K., Trevanion, S., Hubbard, T. and Harrow, J. (2008). The vertebrate genome annotation (Vega) database. Nucleic Acids Res 36, D753–D760.

[Winkler et al., 1994] Winkler, C., Wittbrodt, J., Lammers, R., Ullrich, A. and Schartl, M. (1994). Ligand-dependent tumor induction in medakafish embryos by a xmrk receptor tyrosine kinase transgene. Oncogene 9, 1517–1525.

[Wittbrodt et al., 1998] Wittbrodt, J., Meyer, A. and Schartl, M. (1998). More genes in fish? BioEssays 20, 511–515.

[Wittbrodt et al., 2002] Wittbrodt, J., Shima, A. and Schartl, M. (2002). Medaka - a model organism from the far East. Nat Rev Genet 3, 53–64.

[Wright et al., 2003] Wright, D., Rimmer, L., Pritchard, V., Krause, J. and Butlin, R. (2003). Inter and intra-population variation in shoaling and boldness in the zebrfish (Danio rerio). Naturwissenschaften 90, 374–377.

[Wruck et al., 2002] Wruck, W., Griffiths, H., Steinfath, M., Lehrach, H., Radelof, U. and O'Brien, J. (2002). Xdigitise: visualization of hybridization experiments. Bioinformatics 18, 757–760.

[Xing et al., 2004] Xing, Y., Resch, A. and Lee, C. (2004). The multiassembly problem: reconstructing multiple transcript isoforms from EST fragment mixtures. Genome Research 14, 426–441.

[Yamamoto, 1958] Yamamoto, T. (1958). Artificial induction of functional sex-reversal in genotypic females of the medaka (`oryzias latipes`). J Exp Zoo 137, 227–264.

[Yamamoto, 1975] Yamamoto, T. (1975). Medaka (Killifish) - Biology and Strains. Keigaku, Tokyo.

[Yamanoue et al., 2006] Yamanoue, Y., Miya, M., Inoue, J., Matsuura, K. and Nishida, M. (2006). The mitochondrial genome of spotted green pufferfish `tetraodon nigroviridis` (Teleostei: Tetraodontiformes) and divergence time estimation among model organisms in fishes. Genes Genet Syst 81, 29–39.

[Yanai et al., 2004] Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., Lancet, D. and Shmueli, O. (2004). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics [Epub ahead of print].

[Zadeh Khorasani et al., 2004] Zadeh Khorasani, M., Hennig, S., Imre, G., Asakawa, S., Palczewski, S., Berger, A., Hori, H., Naruse, K., Mitani, H., Shima, A., Lehrach, H., Wittbrodt, J., Kondoh, H., Shimizu, N. and Himmelbauer, H. (2004). A first generation physical map of the medaka genome in BACs essential for positional cloning and clone-by-clone based genomic sequencing. Mech Dev 121, 903–913.

[Zhang et al., 2000] Zhang, Z., Schwartz, S., Wagner, L. and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. J Comput Biol 7, 203–214.

Appendix A

Oligonucleotide sequences used for OFP analysis

Oligonucleotide sequences successfully hybridised to medaka cDNA inserts, either during first round (OFP1) and/or second round (OFP2) of experiments, are summarised in the following tables.

A.1 Standard oligonucleotides

name	oligonucleotide sequence	OFP1/OFP2 successfully hybridised	zf analysis
o005	NGGAATGGAN	1/2	
o008	NCCCTCATCN	1/2	
o009	NTGATGATGN	1/2	
o010	NGGAGTGGAN	1/2	
o012	NCCAGCCTGN	2	
o013	NCAGGCTGGN	2	
o015	NCAGCCTCCN		
o016	NAGCCTGGGN	2	
o019	NAGGCTGAGN	2	
o034	NGCCACCTGN	2	
o035	NAAGGAAAAN	2	
o037	NCCTCCCTGN	2	
o038	NAGGACCTGN	2	
o039	NAGAAGAGAN	2	
o041	NCATCCTGGN(m)	2	
o048	NCTGGGTCCN		
o049	NGATGAGAAN	2	
o050	NAAAGAGAAN(m)	1/2	

continued on next page

name	oligonucleotide sequence	OFP1/OFP2 successfully hybridised	zf analysis
o052	NTGCTGGCCN		
o053	NTGGCAGTGN	2	
o055	NCACCTGGAN(m)	1	
o058	NGAAGACAGN	2	
o059	NAGCCAGAAN	2	
o060	NGCAGAAGCN(m)		
o061	NCTTCTTTTN		
o062	NCTGGGCTGN	2	
o063	NTGGAGAGAN		
o064	NTGGGGCAGN	2	
o067	NACCCCCTGN		
o068	NTTTGCAGAN		
o069	NTGGGAGAGN		
o070	NGGACACCTN		
o073	NGGAGACCCN		
o074	NGACCTGCTN		
o075	NCTGCTGCTN	1/2	zf
o076	NGGAGCTGGN	1/2	zf
o077	NCAGCCTGGN	1	zf
o078	NTGAAGAAGN	1/2	zf
o080	NAGGAGGAGN	1/2	zf
o082	NGCTGCTGGN	1/2	zf
o083	NAGGAGAAGN(m)	1/2	zf
o084	NCCTGGAGCN	1/2	zf
o085	NCTGGAAGAN	1	zf
o086	NCCCAGCCCN	1/2	zf
o087	NCTCCTGCTN	1/2	zf
o088	NAGGAGCAGN(m)	1/2	zf
o090	NCTGGAGGAN	1/2	zf
o094	NGGCTGGGGN	1/2	zf
o097	NGCTGCAGCN	1	zf
o098	NCTCCTGGAN(m)		
o100	NAGCAGCTGN	1/2	zf
o101	NTCCTCCTGN	1/2	zf
o102	NGGGGCTGGN	1/2	zf
o104	NCCTGGCCAN	1/2	zf
o105	NTGCTGGAGN	1/2	zf
o106	NGCTGCTGCN	1/2	zf
o107	NCCTGGGCTN	1/2	zf
o108	NAGCAGGAGN(m)	1	zf
o109	NTGGAGGAGN	1/2	zf

continued on next page

name	oligonucleotide sequence	OFP1/OFP2 successfully hybridised	zf analysis
o110	NTGGAGAAGN	1/2	zf
o111	NCCTGGGCAN		zf
o112	NGCTGGGGCN	1/2	zf
o114	NCCTCAGCCN	1/2	zf
o115	NGAAGGAGGN	1/2	zf
o116	NGCTCCTGGN	1/2	zf
o119	NCTCCAGCCN	1	zf
o120	NTGCTGGTGN	1/2	zf
o121	NTGGAGCAGN	1	zf
o122	NTGGCCCTGN	1/2	zf
o123	NGGAGGAAGN	1/2	zf
o125	NAGCTGGAGN(m)	1/2	zf
o128	NCAGCCCCAN	2	zf
o130	NGGAAGGAGN	1	zf
o133	NTTCCTGGAN	1	zf
o134	NGAGGAGAAN(m)	1	zf
o135	NCCAGGAGGN	1/2	zf
o136	NCTGGAGCAN	1/2	zf
o137	NGAGCTGGGN	1/2	zf
o138	NGGAGCAGCN	1/2	zf
o140	NAGCTGCTGN(m)	1	zf
o142	NCCTGGCTGN	1/2	zf
o143	NGCTGGGCCN	1/2	zf
o144	NGCCCTGGGN	1	zf
o145	NCCTGGAAGN	1/2	zf
o146	NGAGGAGGAN(m)	1/2	zf
o147	NAAGGAGAAN	1/2	zf
o148	NCAGCCCTGN(m)	1	zf
o149	NCCTGCTGCN(m)	1/2	zf
o150	NGGAGGTGGN	1	zf
o151	NAGGAAGAGN	1/2	zf
o152	NCCTCCTGCN	1/2	zf
o155	NTCCTGGAGN	1/2	zf
o156	NTGGAGCTGN	1/2	zf
o158	NGGAGAAGAN	1	zf
o159	NTGCTGCTGN	1/2	zf
o160	NCTGCAGCCN	1	zf
o162	NAGGAGGAAN	1/2	zf
o163	NGGCCCTGGN	1/2	zf
o164	NGCTGGTGGN	1/2	zf
o165	NTGCAGCTGN	1	zf

continued on next page

name	oligonucleotide sequence	OFP1/OFP2 successfully hybridised	zf analysis
o166	NCCCTGCTGN	1/2	zf
o168	NGAGGAAGAN(m)	1	zf
o177	NGGCCAAGGN	1	zf
o180	NCTGGGGCCN	1/2	zf
o181	NCCCTGCCCN	1	zf
o183	NCAGCCTGAN(m)	1/2	zf
o185	NCTCACCATN	1	zf
o187	NGAAGAGGAN	1/2	zf
o189	NCCATCTCCN	1	zf
o192	NTGAAGAAAN		
o196	NCCTGGTCAN	1	zf
o199	NCTTCCTGGN	1	zf
o201	NGTACCAGCN		zf
o203	NCCCTCCAGN	1	zf
o204	NAGTGGCTGN	1/2	zf
o205	NCTGAGCTGN	1/2	zf
o206	NAGCCCAAGN	1/2	zf
o207	NTTTCTTCCN	1/2	zf
o209	NGCCATGGAN	1/2	zf
o211	NTTCATCTAN	1	zf
o212	NCAGCCACCN	1/2	zf
o214	NTGTTATTTN		zf
o215	NTCACTGTGN	1	zf
o219	NAGGGAGTGN	1	zf
o220	NCATCACCAN(m)	1/2	
o224	NTGGGGGAGN	2	
o227	NTCTCTCCCN	1	zf
o228	NAGCTCACCN	1	zf
o229	NCCAAGGTGN	1	zf
o231	NTTGTTTTCN	1	zf
o232	NTGCTGTGTN	1/2	zf
o233	NCCAGAACCN(m)	1	zf
o234	NAATGAGGAN		
o235	NCGTCTCCTN	1/2	zf
o236	NTGCTCCTGN	1	zf
o237	NGTGGTGGTN	1/2	zf
o241	NTTCTGGAAN	1/2	zf
o247	NTTTCAGAAN	1	zf
o249	NCTACTGGGN		zf
o250	NTTCCTGCAN		zf
o251	NTTGCCTTTN		zf

continued on next page

name	oligonucleotide sequence	OFP1/OFP2 successfully hybridised	zf analysis
o252	NCTCCCACAN	2	zf
o253	NAGCTCACTN		zf
o254	NTGGGATGGN	2	zf
o255	NAGAAGCCCN		zf
o256	NGCTGGGTGN	2	zf
o258	NCCTTTGCTN	2	zf
o259	NCCCTGTCCN	2	zf
o260	NGAGGCGGAN	2	zf
o261	NGAAGCAGAN		zf
o262	NTTTCTCTGN		zf
o265	NATGAGCAGN	2	zf
o267	NGCCAGGACN	2	zf
o268	NCATGGCCCN		zf
o291	NCCTCCTCCN	2	
o293	NCTGCAGGAN(m)	1/2	
o314	NAGAAAGAAN(m)	1	
o318	NAGAAAACAN(m)	1	
o322	NCAAAGAAAN(m)		
o343	NAGCTCAGCN(m)	1	
o345	NAAATGGAAN(m)		
o407	NTCTCCTCAN(m)	1	
o416	NAAAACACAN(m)	1/2	
o422	NAAACAGAAN		
o438	NTGGAGGAAN(m)	1/2	
o440	NCAGAGCAGN(m)	1/2	
o469	NGATGAAGAN(m)	1	
o470	NAAACATCTN(m)	1	
o641	NTCCCTGCAN(m)	1	
o660	NCAGAGATGN(m)	1	
o702	NCTCCTTCAN(m)	1/2	
o703	NCAGCTTCAN(m)	1/2	
o706	NCTCCACCAN(m)	2	
o711	NAGCTGAAGN(m)	1/2	
o717	NCACCTTCAN(m)	1/2	
o720	NCAGCTCCAN(m)	1/2	
o756	NTCCCTTCAN(m)	1	
o759	NGAAGAACCN(m)	1/2	
o769	NCAACAACAN(m)	1/2	
o780	NCCTCTGCCN(m)	1/2	
o783	NCTGGAAAAN(m)	1/2	
o786	NCAGAAATGN(m)	1	

continued on next page

name	oligonucleotide sequence	OFP1/OFP2 successfully hybridised	zf analysis
o797	NACAGAAAGN(m)	1	
oz043	NCAGCAGAGN(m)	2	

Table A.1: Oligonucleotides taken from the standard set. m - these sequences were also calculated as being specific for medaka EST sequences. zf - these oligonucleotides were successfully hybridised in the zebrafish OFP analysis [Clark et al., 2001].

A.2 Calculated oligonucleotides

name	oligonucleotide sequence	OFP1/OFP2 successfully hybridised
m001	NCTGCAGCAN	2
m004	NCTGCTGTCN	1/2
m005	NAAAAACAGN	1/2
m006	NAGGAAAAAN	1/2
m007	NN	
m008	NCAGAAAAAN	1
m009	NCTGCAGAAN	1/2
m010	NACAAAACAN	1
m011	NACAGGAAGN	
m012	NAAAGCAGCN	1/2
m013	NCATCATCAN	1/2
m014	NAAAGCAAAN	1
m015	NCAGAGGAAN	1
m016	NAGAGAAAAN	1
m017	NGAACAAAAN	1
m018	NACACAAAAN	1
m020	NCCAGCTGAN	1/2
m021	NAAAATGTCN	1
m022	NAAACAAAGN	1
m024	NCCTCCACCN	1
m025	NGATGCTGAN	1/2
m029	NATGGAGAAN	1
m030	NCAGAGCCAN	1
m031	NGACAAAAAN	1
m035	NCAAAAGAAN	1
m041	NCACAGCCAN	1
m083	NCCAGCCTCN	1/2
m135	NGAGGAGCCN	1

continued on next page

name	oligonucleotide sequence	OFP1/OFP2 successfully hybridised
m153	NCTGCAGCAN	1
m154	NGCAGCAGAN	1/2
m155	NCCAGCAGGN	1/2
m156	NCTGTTTGAN	2
m157	NAATGTTTGN	1
m158	NTCCAAAAAN	1
m159	NCATTTGAAN	1
m160	NCTTCAACAN	1
m161	NAGCTTCTGN	1
m162	NCCTTCAGAN	1
m230	NCAGCTTCAN	2
m231	NCTTCTTCAN	2

Table A.2: Oligonucleotides specifically calculated for medaka sequence data.

Appendix B

cDNA libraries in pCS2 vector

B.1 Vector design

cDNA libraries were provided by cooperation partners in a modified pCS2 vector. The fragment of CMV promoter was deleted from CS2+ by digestion with SalI and HindIII. This 3 kb fragment was made blunt by Klenow enzyme and self-ligated to make short CS2+ (psCS2+). Then, an annealed oligonucleotide (5'-tcgaggcgcgccgttaaacccggccattatggcctgcagcatgcggccgcctcggccct-3'and 5'-ctagagggccgaggcggccgcatgctgcaggccataatggcccgggtttaaacggcgcgcc-3') with SfiIA-PstI-SphI-SfiIB site was inserted at the XhoI and XbaI site of psCS2+ vector to create psCS2+(SfiIA-B).

B.2 Cloning cDNA inserts into pCS2

Production of cDNA libraries was done in the group of M. Furutani-Seiki. Total RNA was extracted from of each pool by the ToTALLY RNA extraction kit (Ambion, TX, USA) producing about 6mg of total RNA/sample. Poly(A)+ RNA was purified from the total RNA by the Poly(A) + quick mRNA kit (Stratagene, TX, USA). The mRNA was converted to cDNA with directional SfiI sites (SfiIA site at 5' end and SfiIB site at 3' end), using the SMART cDNA kit (Clontech, CA, USA). Synthesized cDNA was ligated to psCS2+(SfiIA-B) vector and transformed by electroporation using ElectroTen-Blue Electroporation Competent Cells (Stratagene, TX, USA).

Appendix C

Differentially expressed OFP cluster

This table continues to show the annotation of differentially expressed OFP clusters, started in table 3.15.

OFP cluster	OFP cluster size	R-value	stage overexpressed	stage underexpressed	EST cluster	Annotation
translation and protein metabolism						
OFP7	1855	711.768	neurula, ovary	gastrula	CL1Contig6	similar to SP|P02383|RS26_HUMAN 40S ribosomal protein S26
OFP20	381	198.081	neurula, ovary	gastrula	CL17Contig1	similar to SP|P25111|RS25_HUMAN 40S ribosomal protein S25
OFP147	89	143.852	organo		CL3304Contig1	similar to UniRef100_Q7ZYX1 Fts protein
OFP181	77	83.483	organo		CL1532Contig1	similar to UP|Q9IBE7 Alveolin
OFP189	88	74.132	ovary		CL850Contig1	similar to SP|P46782|RS5_HUMAN 40S ribosomal protein S5
OFP136	98	62.748	gastrula		CL2782Contig1	similar to UP|Q90YR5 40S ribosomal protein S9

continued on next page

OFP cluster	OFP cluster size	R-value	stage overexpressed	stage underexpressed	EST cluster	Annotation
OFP384	49	61.043	ovary		CL398Contig1	similar to SP|P31580|HCE1_ORYLA High choriolytic enzyme 1 precursor (Hatching enzymezinc-protease HCE 1 subunit) (Choriolysin H 1)
OFP14	491	60.467	gastrula, organo	ovary	CL4Contig17	similar to UP|Q9PT09 Ubiquitin
OFP23	314	59.829	gastrula		CL20Contig1	similar to SP|Q9YIC0|EF1A_ORYLA Elongation factor 1-alpha (EF-1-alpha)
OFP28	298	48.747	ovary	gastrula	CL32Contig1	similar to UP|Q90YX0 Ribosomal protein P1
OFP56	212	45.655	gastrula		CL22Contig1	similar to GB|AAH53240.1|31418910| BC053240 nucleophosmin 1
OFP214	74	44.630	ovary		CL1290Contig1	similar to PIR|S52084 ribosomal protein L22 cytosolic
OFP142	95	42.645	gastrula		CL2512Contig1	similar to UP|AAP20208 Ribosomal protein L37
OFP393	36	39.237	ovary		CL1343Contig1	similar to UP|Q8JH38 Muscle-type creatine kinase CKM2
OFP472	43	37.942	organo		CL422Contig1	similar to UP|Q6YI49 Ubiquitin C-terminal hydrolase L1
OFP320	73	37.122	gastrula		CL126Contig1	similar to UP|Q9IA73 Ribosomal protein large P2
OFP1858	26	34.920	neurula		CL52Contig1	similar to UP |AAQ97821 Nit protein
OFP51	179	30.820	neurula		CL86Contig1	similar to GB|AAF64459.1|7595809| AF240376 ribosomal protein L18

continued on next page

OFP cluster	OFP cluster size	R-value	stage overexpressed	stage underexpressed	EST cluster	Annotation
OFP17	393	30.405	ovary	organo	CL135Contig1	similar to SP|Q9YGF2|RS6_ONCMY 40S ribosomal protein S6
OFP133	118	29.504	gastrula		CL65Contig1	similar to GB|AAH44698.1|28278283| BC044698 MGC54031 protein
OFP91	136	29.371	gastrula		CL168Contig1	similar to GB|AAH49061.1|29124464| BC049061 ribosomal protein L12
OFP59	169	27.493	gastrula		CL115Contig1	similar to UP|Q8HLX0 ATPase subunit 6
OFP577	34	26.096	organo		CL742Contig1	similar to PIR|E70358 HupE hydrogenase related function
OFP40	192	26.046	neurula		CL786Contig1	similar to UP|Q90YU7 Ribosomal protein L21
OFP101	126	26.003	gastrula		CL57Contig1	similar to SP|P02401|RLA2_RAT 60S acidic ribosomal protein P2
OFP564	39	25.277	neurula		CL93Contig1	similar to UP|O42448 (O42448) Id2 protein
OFP44	207	24.174	ovary		CL101Contig1	similar to PIR|I84474 RagA (ras-related alternatively spliced GTPase A)
OFP302	58	21.438	gastrula		CL257Contig1	similar to UP|AAH34898 Ube2n protein (Fragment)
OFP65	176	21.113	gastrula		CL74Contig1	similar to GB| AAH49478.1|29612581| BC049478 ik:tdsubc_1f2 protein

continued on next page

OFP cluster	OFP cluster size	R-value	stage overexpressed	stage underexpressed	EST cluster	Annotation
OFP83	131	20.834	gastrula		CL1029Contig1	similar to GB|AAK95200.11529404|AF402826 40S ribosomal protein S17
OFP140	171	20.820	ovary		CL180Contig1	similar to SP|P14118|RL19_HUMAN 60S ribosomal protein L19
OFP536	122	20.458	gastrula		CL1Contig7	similar to UP|AAP20215 40S ribosomal protein S24
OFP239	99	18.134	gastrula		CL66Contig1	similar to UP|Q17000 Polyubiquitin
OFP90	122	18.114	gastrula		CL483Contig1	similar to GB|AAK95157.1|15293929|AF401585 ribosomal protein L30
general metabolism						
OFP5	1239	1532.857	neurula, ovary	gastrula	CL77Contig1	similar to UP|Q8AYQ8 Alpha-type globin
OFP16	517	588.951	neurula, ovary		CL131Contig1	similar to UP|Q8AYQ6 Beta-type globin
OFP97	124	105.669	ovary		CL1959Contig1	similar to UP|Q8JIM9 Adult beta-type globin
OFP782	65	68.838	ovary		CL41Contig1	similar to UP|Q8AYQ6 Beta-type globin
OFP210	71	67.943	ovary		CL1237Contig1	similar to UP|Q8AYQ7 Alpha-type globin
OFP321	43	42.073	ovary		McF15H22	similar to UP|Q8JIN1 Embryonic beta-type globin
OFP8	903	1193.173	ovary	gastrula, neurula	CL153Contig1	similar to UP|Q8UUS2 Parvalbumin
OFP57	178	151.054	ovary	gastrula	CL2908Contig1	similar to UP|Q8UUS2 Parvalbumin
OFP19	387	285.459	organo	gastrula	CL31Contig1	similar to UP|O57691 Fatty acid binding protein H6-isoform
OFP13	594	251.731	gastrula	ovary, organo	CL9Contig1	similar to UP|Q9PT73 Apolipoprotein E

continued on next page

OFP cluster	OFP cluster size	R-value	stage overexpressed	stage underexpressed	EST cluster	Annotation
OFP75	143	143.136	ovary		McF07K04	similar to PIR|A49184 fatty acid-binding protein
OFP22	299	227.238	neurula, ovary	gastrula	CL450Contig1	similar to UP|Q98TG0 14kDa apolipoprotein
OFP24	318	247.015	organo	gastrula	CL98Contig1	similar to UP|Q9W7D5 Quinone reductase
OFP32	323	139.157	gastrula		CL8Contig1	similar to UP|CYB_SALSA (Q35925) Cytochrome b
OFP18	278	105.120	gastrula		CL1Contig1	similar to UP|Q704S4 Beta-galactosamide alpha-2,6-sialyltransferase
OFP128	108	128.957	ovary		CL906Contig1	similar to UP|Q90X19 Muscle-specific creatine kinase
OFP48	194	82.702	neurula, ovary	gastrula	CL254Contig1	similar to UP|Q9I901 Retinol-binding protein
OFP184	69	73.620	ovary		McF20D13	similar to UP|TRFE_ORYLA (P79819) Serotransferrin precursor
OFP35	278	69.772	gastrula		CL100Contig1	similar to UP|CAD19164 Sperm protein 8
OFP242	70	63.269	organo		CL359Contig1	similar to GB|AAH55971.1|33416776|BC055971 MGC68838 protein
OFP123	103	55.280	gastrula		CL26Contig1	similar to GB|AAH45302.1|28279507|BC045302 selenophosphate synthetase

continued on next page

OFP cluster	OFP cluster size	R-value	stage overexpressed	stage underexpressed	EST cluster	Annotation
OFP138	128	53.808	gastrula		CL21Contig1	similar to SP| P29218| MYOP_HUMAN Inositol-1(or 4)-monophosphatase (IMPase) (IMP)(Inositol monophosphatase) (Lithium-sensitive myo-inositolmonophosphatase A1)
OFP470	47	47.856	ovary		CL1172Contig1	similar to SP|P12115|KAD_CYPCA Adenylate kinase (ATP-AMP transphosphorylase)
OFP46	199	46.508	gastrula		CL537Contig1	similar to UP|O57691 Fatty acid binding protein H6-isoform
OFP573	27	43.338	ovary		CL2520Contig1	similar to AAQ94593 Aldolase A fructose-bisphosphate
OFP42	205	41.451	neurula, ovary		CL63Contig1	similar to GB|AAH48051.1|28856132| BC048051 phosphoribosylaminoimidazole carboxylase phosphoribosylaminoimidazole succinocarboxamide synthetase
OFP105	130	40.954	gastrula		CL18Contig1	similar to UP|NU1M_BRARE (Q9MIZ0) NADH-ubiquinone oxidoreductase chain 1
OFP148	96	34.612	ovary		CL242Contig1	similar to SP|O42204|ITMB_CHICK Integral membrane protein 2B (Transmembrane protein E3-16)

continued on next page

OFP cluster	OFP cluster size	R-value	stage overexpressed	stage underexpressed	EST cluster	Annotation
OFP322	46	34.589	ovary		CL53Contig3	similar to UP|Q9DEB7 Tropomyosin
OFP316	92	30.203	ovary		CL154Contig1	similar to UP|Q9PVL0 Cytochrome c oxidase subunit VIa precursor
OFP121	103	29.757	ovary		CL1926Contig1	similar to GB|AAO86704.1|29648601|AY216590 phospholipid hydroperoxide glutathione peroxidase A
OFP556	35	29.647	ovary		CL858Contig1	similar to UP|Q9IB34 Myosin light chain 2
OFP440	34	29.483	ovary		McF17I04	similar to SP|P82160|MLE3_MUGCA Myosin light chain 3 skeletal muscle isoform (A2 catalytic) (Alkali)(LC-3) (LC3)
OFP788	20	27.466	organo		CL3456Contig1	similar to UP|Q90XY6 Omega class glutathione-S-transferase
OFP377	54	21.094	gastrula		CL84Contig1	similar to UP|NU1M_BRARE (Q9MIZ0) NADH-ubiquinone oxidoreductase chain 1
OFP73	149	20.789	gastrula		CL402Contig1	similar to SP|P00026|CYC_CYPCA Cytochrome c iso-1/iso-2
OFP94	136	18.009	gastrula		CL48Contig1	similar to GB|AAH00088.1|12652679|BC000088 glutathione S-transferase M3
hypothetical						
OFP3	1434	2139.775	organo	gastrula	CL2Contig12	similar to UniRef100_Q9PTQ5 Hypothetical protein

continued on next page

OFP cluster	OFP cluster size	R-value	stage overexpressed	stage underexpressed	EST cluster	Annotation
OFP4	1254	1873.202	organo	gastrula	CL2Contig1	similar to UniRef100_Q9W7D3 Hypothetical protein
OFP54	164	237.213	organo		CL2Contig2	similar to UniRef100_Q9W7D0 Hypothetical protein
OFP120	118	172.201	organo		CL2Contig9	similar to UniRef100_Q9W7D4 Hypothetical protein
OFP77	122	166.971	organo		CL2349Contig1	similar to UPI0000360821 UniRef100 entry
OFP185	80	114.264	organo		CL896Contig1	similar to UPI000032EB82 UniRef100 entry
OFP68	159	78.496	gastrula		CL954Contig1	similar to UP|Q9BYD5 mRNA related with psoriasis
OFP590	33	55.399	organo		CL2Contig16	similar to UniRef100_Q9PTQ5 Hypothetical protein
OFP328	52	42.467	organo		CL748Contig1	similar to UniRef100_Q9W7C8 Hypothetical protein
OFP895	28	41.652	organo		CL186Contig1	similar to UniRef100_Q9W7C7 Hypothetical protein
OFP975	29	39.329	organo		CL1446Contig1	similar to UniRef100_Q9X9L8 Hypothetical protein
OFP862	22	28.153	organo		CL1016Contig1	similar to UniRef100_UPI0000360616
OFP166	157	25.365	gastrula		CL11Contig1	similar to UP|Q7VCS4 Predicted protein
no classification						
OFP9	824	1154.931	organo	gastrula	CL15Contig3	similar to UP|Q7ZTS2 Cldni protein
OFP12	663	835.966	ovary	gastrula	CL6Contig3	similar to UP|Q31555 Gamma-crystallin M2-1

continued on next page

OFP cluster	OFP cluster size	R-value	stage overexpressed	stage underexpressed	EST cluster	Annotation
OFP208	70	82.701	ovary		McF03H18	similar to PIR|S45015 gamma-crystallin M2
OFP124	111	141.547	ovary		CL1244Contig1	similar to UP|Q90WT1 Crystallin B1 protein
OFP309	74	108.089	ovary		CL3446Contig1	similar to UP|Q31555 Gamma-crystallin M2-1
OFP877	78	83.867	ovary		CL6Contig1	similar to UP|Q31555 Gamma-crystallin M2-1
OFP187	77	95.114	ovary		CL2151Contig1	similar to UP|Q8VHL5 GammaN-crystallin
OFP666	58	65.918	ovary		CL6Contig2	similar to UP|Q31555 Gamma-crystallin M2-1
OFP78	145	230.788	organo		CL334Contig1	similar to GP|12659138|gb mage-d3
OFP82	233	134.990	gastrula		CL132Contig1	similar to UP|AAS40882 Peptidase M23/M37 family
OFP537	42	58.315	organo		CL798Contig1	similar to UP|O57150 H88
OFP736	31	52.041	organo		CL2Contig4	similar to UP|Q99738 Pinin
OFP365	37	45.286	organo		McF12B23	similar to UP|Q7Z4E2 MSTP096
OFP202	70	40.544	gastrula		CL28Contig1/ Contig2	similar to PIR|T39903 serine-rich protein
OFP98	17	28.539	organo		McF51A11	similar to UP|Q9BIU9 Flagelliform silk protein (Fragment)
OFP224	71	27.581	ovary		CL951Contig1	similar to UP|AAH59742 MGC75717 protein
OFP627	45	25.838	gastrula		CL13Contig1	similar to UP|Q8C962 Odd Oz/ten-m homolog 1 (Fragment)
OFP485	40	25.611	gastrula		CL245Contig1	similar to UP|Q7U280 ISONIAZID INDUCTIBLE gene protein INIB
OFP576	54	24.841	neurula		CL28Contig1	similar to PIR|T39903 serine-rich protein

continued on next page

OFP cluster	OFP cluster size	R-value	stage overexpressed	stage underexpressed	EST cluster	Annotation
no annotation						
OFP100	129	199.194	organo		CL687Contig1	NA
OFP260	55	77.573	organo		CL2866Contig1	NA
OFP213	69	44.180	gastrula		McF23I03	NA
OFP287	61	29.865	gastrula		CL24Contig1	NA
OFP504	71	25.003	gastrula		CL35Contig1	NA
OFP509	28	47.005	organo		McF54L10	NA
no traces						
OFP10	796	660.119	neurula, ovary	gastrula		no trace
OFP25	286	460.349	organo			no trace
OFP87	132	221.594	organo			no trace
OFP108	113	189.698	organo			no trace
OFP27	276	176.719	gastrula			no trace
OFP89	134	156.590	ovary			no trace
OFP139	94	151.225	organo			no trace
OFP186	76	116.739	ovary			no trace
OFP212	69	115.833	organo			no trace
OFP205	70	112.359	ovary			no trace
OFP226	64	96.041	organo			no trace
OFP279	53	85.072	ovary			no trace
OFP285	52	83.467	ovary			no trace
OFP276	53	82.974	organo			no trace
OFP169	82	79.528	ovary			no trace
OFP311	46	77.222	organo			no trace
OFP99	113	75.378	neurula, ovary			no trace
OFP312	46	73.836	ovary			no trace
OFP338	41	68.828	organo			no trace
OFP88	134	66.188	gastrula			no trace
OFP290	51	65.874	neurula			no trace
OFP310	46	65.756	ovary			no trace
OFP380	39	65.471	organo			no trace
OFP346	42	64.030	organo			no trace
OFP359	41	61.182	ovary			no trace
OFP385	38	60.995	ovary			no trace
OFP366	40	56.412	ovary			no trace
OFP291	50	55.499	organo			no trace
OFP432	34	54.574	ovary			no trace
OFP471	32	53.720	organo			no trace

continued on next page

OFP cluster	OFP cluster size	R-value	stage overexpressed	stage underexpressed	EST cluster	Annotation
OFP286	48	53.591	organo			no trace
OFP397	35	51.712	ovary			no trace
OFP481	31	50.460	neurula			no trace
OFP21	313	50.154	gastrula			no trace
OFP484	31	49.759	ovary			no trace
OFP520	29	48.684	organo			no trace
OFP422	35	48.297	organo			no trace
OFP383	38	48.281	ovary			no trace
OFP501	30	48.154	ovary			no trace
OFP546	28	47.005	organo			no trace
OFP360	39	45.397	neurula			no trace
OFP182	77	45.003	gastrula			no trace
OFP513	30	44.939	organo			no trace
OFP201	70	44.820	gastrula			no trace
OFP492	30	43.843	ovary			no trace
OFP247	55	43.777	neurula			no trace
OFP570	27	43.338	ovary			no trace
OFP635	25	41.969	organo			no trace
OFP594	26	41.733	ovary			no trace
OFP203	70	40.617	gastrula			no trace
OFP469	32	40.244	ovary			no trace
OFP562	27	40.011	organo			no trace
OFP530	28	39.665	ovary			no trace
OFP388	34	39.364	organo			no trace
OFP510	30	37.527	ovary			no trace
OFP567	27	36.356	ovary			no trace
OFP727	21	33.708	ovary			no trace
OFP281	53	30.795	gastrula			no trace
OFP864	18	30.217	organo			no trace
OFP904	18	30.217	organo			no trace
OFP857	18	28.892	ovary			no trace
OFP307	36	28.857	ovary			no trace
OFP340	43	27.532	gastrula			no trace
OFP1607	11	27.066	neurula			no trace
OFP1036	16	26.860	organo			no trace
OFP354	41	26.252	gastrula			no trace
OFP223	60	23.109	gastrula			no trace

Table C.1: Differentially expressed OFP cluster.

Appendix D

Genes important for embryonic development

Based on gene ontology analysis ESTs and EST contigs were detected which show similarities to genes important for embryonic development. For each category only few examples are shown, all annotations are found within supplemental material in file Annotation\GO_embryodevo.xls.

EST/EST contig	Similarity to ...
regulation of embryonic development	
CL1027Contig1	GB|AAH44082.1|28422594|BC044082 LOC398558 protein (*Xenopus laevis*)
CL139Contig1	GB|AAH44082.1|28422594|BC044082 LOC398558 protein (*Xenopus laevis*); UP|Q6DRE3 (Q6DRE3) NOP5/NOP58
embryonic development	
embryonic cleavage	
CL1681Contig1	UP|AAH00338 (AAH00338) PSMD7 protein (Fragment); PIR|S65491|S65491 26S proteasome regulatory chain 12 - human (*Homo sapiens*)
CL1951Contig1	SP|Q90336|M14A_CYPCA Mitogen-activated protein kinase 14a (Mitogen-activated protein kinase p38a) (MAP kinase p38a) (cp38a). (*Cyprinus carpio*); UP|M14B_BRARE (Q9DGE1) Mitogen-activated protein kinase 14b (Mitogen-activated protein kinase p38b) (MAP kinase p38b) (zp38b); UP|Q6P3M1 (Q6P3M1) Mapk14b protein
CL691Contig1	UP|Q9WTX5 (Q9WTX5) SCF complex protein SKP1 (Transcription elongation factor B (SIII) polypeptide 1 (15 kDa) -like); GB|AAD16036.1|4322377|AF083214 SCF complex protein Skp1 (*Mus musculus*); UP|Q6PBY8 (Q6PBY8) S-phase kinase-associated protein 1A
embryonic development	
CL1450Contig1	UP|Q8AYB5 (Q8AYB5) Bone morphogenetic protein 4 (Fragment); UP|O57574 (O57574) Bone genetic protein 4 (Bone morphogenetic protein 4)

continued on next page

EST/EST contig	Similarity to ...
gastrulation	
CL1431Contig1	UP|Q9PW31 (Q9PW31) Rac GTPase; SP|P15154|RAC1_HUMAN Ras-related C3 botulinum toxin substrate 1 (p21-Rac1) (Ras-likeprotein TC25). (*Homo sapiens; Canis familiaris; Bos taurus; Mus musculus*); UP|O93466 (O93466) GTPase cRac1B; UP|Q6LC82 (Q6LC82) GTPase cRac1A
CL1676Contig1	GB|AAH48035.1|28856238|BC048035 cell division cycle 42 homolog (*Danio rerio*)
CL1905Contig1	UP|Q9PVF7 (Q9PVF7) Cell-adhesion protein plakoglobin; UP|CTNB_MOUSE (Q02248) Beta-catenin
embryonic morphogenesis	
embryonic eye morphogenesis	
CL1450Contig1	UP|Q8AYB5 (Q8AYB5) Bone morphogenetic protein 4 (Fragment); UP|O57574 (O57574) Bone genetic protein 4 (Bone morphogenetic protein 4)
McF0001MGR-1B01bd1	UP|PAX5_MOUSE (Q02650) Paired box protein Pax-5 (B-cell specific transcription factor) (BSAP); UP|O57685 (O57685) Paired box protein; UP|O57676 (O57676) Paired box protein
McF0039C13-MGRbd1	UP|EYA4_MOUSE (Q9Z191) Eyes absent homolog 4; UP|Q9W6E8 (Q9W6E8) Eyes absent homolog 1
embryonic pattern specification	
embryonic axis specification	
CL1861Contig1	UP|HS9A_BRARE (Q90474) Heat shock protein HSP 90-alpha; PIR|A32298|HHCH90 heat shock protein 90 - chicken (*Gallus gallus*); UP|Q9W6K6 (Q9W6K6) Heat shock protein hsp90 beta
CL2098Contig1	SP|P50550|UBCI_HUMAN Ubiquitin-like protein SUMO-1 conjugating enzyme (SUMO-1-protein ligase) (Ubiquitin carrier protein) (Ubiquitin-conjugatingenzyme UbcE2A) (P18). (*Xenopus laevis; Gallus gallus; Homo sapiens; Mus musculus; Rattus norvegicus*)
CL2370Contig1	UP|Q7T3L3 (Q7T3L3) Chaperone protein GP96; UP|AAQ95586 (AAQ95586) HSP-90

Table D.1: ESTs or EST contigs showing similarity to genes with function(s) during general embryonic development.

EST/EST contig	Similarity to ...
embryonic morphogenesis	
embryonic eye morphogenesis	
CL1450Contig1	UP|Q8AYB5 (Q8AYB5) Bone morphogenetic protein 4 (Fragment); UP|O57574 (O57574) Bone genetic protein 4 (Bone morphogenetic protein 4)
McF0001MGR-1B01bd1	UP|PAX5_MOUSE (Q02650) Paired box protein Pax-5 (B-cell specific transcription factor) (BSAP); UP|O57685 (O57685) Paired box protein; UP|O57676 (O57676) Paired box protein

continued on next page

EST/EST contig	Similarity to ...
McF0039C13-MGRbd1	UP|EYA4_MOUSE (Q9Z191) Eyes absent homolog 4; UP|Q9W6E8 (Q9W6E8) Eyes absent homolog 1
organogenesis	
eye morphogenesis	
CL1431Contig1	UP|Q9PW31 (Q9PW31) Rac GTPase; SP|P15154|RAC1_HUMAN Ras-related C3 botulinum toxin substrate 1 (p21-Rac1) (Ras-like protein TC25). (*Homo sapiens; Canis familiaris; Bos taurus; Mus musculus*); UP|O93466 (O93466) GTPase cRac1B; UP|Q6LC82 (Q6LC82) GTPase cRac1A
CL1450Contig1	UP|Q8AYB5 (Q8AYB5) Bone morphogenetic protein 4 (Fragment); UP|O57574 (O57574) Bone genetic protein 4 (Bone morphogenetic protein 4)
CL1644Contig1	UP|Q6NX50 (Q6NX50) CDC10 protein (Fragment); UP|AAH25987 (AAH25987) CDC10 protein (Fragment); UP|Q6Q137 (Q6Q137) Cell division cycle 10; SP|Q16181|SEP7_HUMAN Septin 7 (CDC10 protein homolog). (*Homo sapiens*)
neurogenesis	
CL1002Contig1	SP|O88567|DL2A_MOUSE Dynein light chain 2A cytoplasmic (Dynein-associated proteinKM23) (Bithoraxoid-like protein) (BLP). (*Mus musculus; Rattus norvegicus*)
CL1034Contig1	UP|O93431 (O93431) Ephrin A-L1; UP|CSP6_HUMAN (Q9NVC6) Cofactor required for Sp1 transcriptional activation subunit 6 (Transcriptional coactivator CRSP77) (Vitamin D3 receptor-interacting protein complex 80 kDa component) (DRIP80) (Thyroid hormone receptor-associated protein complex 80 kD
CL1074Contig1	SP|P38408|GB14_BOVIN Guanine nucleotide-binding protein alpha-14 subunit (GL1). (*Bos taurus*); GB|AAF59930.1|7329187|AF234260 heterotrimeric guanine nucleotide-binding protein alpha q subunit (*Rattus norvegicus*)

Table D.2: ESTs or EST contigs with similarity to genes responsible for morphogenesis during embryonic development.

EST/EST contig	Similarity to ...
glial cell differentiation	
McF0009F13-MGRbd1	GB|AAH45014.1|27924279|BC045014 rab1-prov protein (*Xenopus laevis*); UP|Q8QFR9 (Q8QFR9) Basic fibroblast growth factor
McF0019F13-MGRbd1	UP|Q8QFR9 (Q8QFR9) Basic fibroblast growth factor; UP|Q8NG24 (Q8NG24) GTP-binding protein Rab25
neuron differentiation	
CL1132Contig1	UP|AAQ72492 (AAQ72492) 14-3-3E2 protein; UP|AAQ72491 (AAQ72491) 14-3-3E1 protein; UP|Q6PUV9 (Q6PUV9) 14-3-3 protein
CL1626Contig1	GB|AAC34258.1|3523107|AF022224 Bcl-2-binding protein (*Homo sapiens*)
CL1724Contig1	UP|Q7T3G1 (Q7T3G1) GATA-binding protein 2; UP|GAT3_BRARE (Q91428) Transcription factor GATA-3 (GATA binding factor-3)

Table D.3: ESTs or EST contigs with similarity to genes responsible for neuronal cell differentiation during embryonic development.

Appendix E

Alternative splice events - exon-intron boundaries

All sequences of candidates for alternative splicing were aligned against the Medaka genome by est2genome and intron-exon borders were calculated. The first and last nucleotide of each exon are depicted for three examples. All other exon borders are found in supplemental material in file AlternativeSplicing\AlternSplices_Exons.xsl. Different alternative splicing events were noted: case 1 - retained intron, case 2 - competing 5' splice sites, case 3 - competing 3' splice sites, case 4/5 - cassette exons or mutual exclusive exons.

C53(revers) - scaff1673 (24000-30000)		
	McF0018J09	McF0045E12
	367 630	439 630
	738 876	738 876
case 4	1546 1614	
	1833 1935	1833 1935
	2043 2164	2043 2164
case 4		2429 2481
	2933 2986	2933 3010

C60 - scaff4355				
	McF0038N20	CL2413Contig1	BJ711472	BJ726420
	929 969	901 969	901 969	
case 4	2299 2352			
	6550 6672	6550 6672	6550 6672	
	8055 8130	8055 8130	8055 8130	
	8892 8980	8892 8980	8892 8980	
	11296 11500	11296 11500	11296 11498	11462 11498
			11606 11711	11606 11711
			16164 16216	16164 16333
				17422 17871

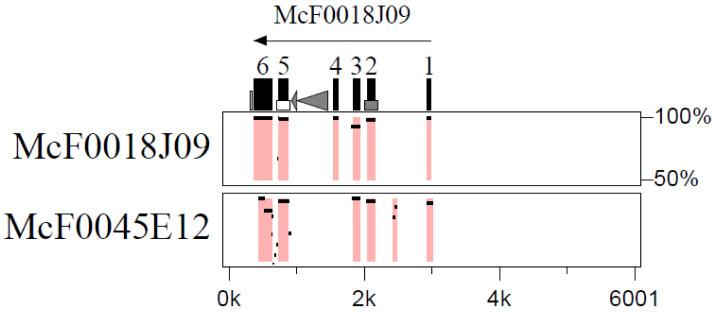
C81 - scaff650				
	CL318Contig1	CL318Contig2	BJ529575	BJ541889
	134 279		107 279	
case 4/5		885 981		
	1398 1482	1398 1482	1398 1482	
	1693 1823	1693 1823	1693 1823	1746 1823
	3161 3248	3161 3248	3161 3248	3161 3248
	3371 3495	3371 3495	3371 3495	3371 3495
	3736 3802	3736 3782	3736 3802	3736 3802
	5831 6106		5831 5861	5831 6127

Appendix F

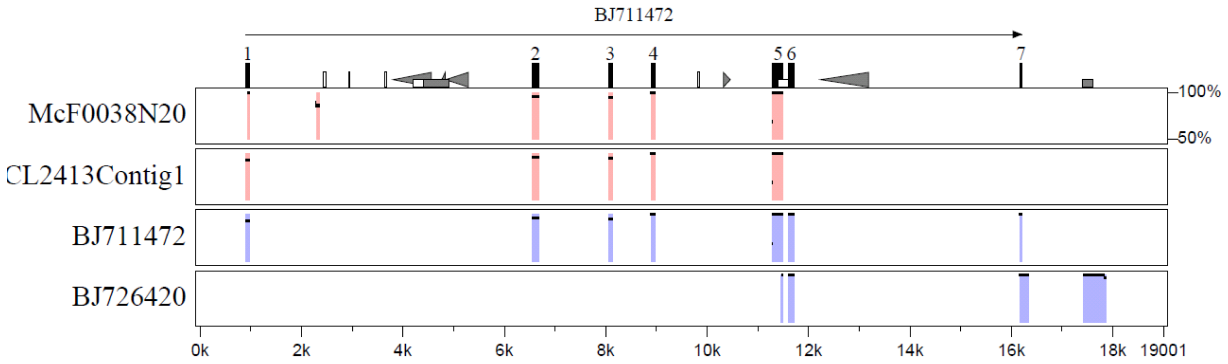
Alternative splice events - PIP plots

PIP plots of alternative splicing events showing alignment of all known sequences to the longest available sequence were calculated for all candidates (4.3.8.3). Different alternative splicing events were noted: case 1 - retained intron, case 2 - competing 5' splice sites, case 3 - competing 3' splice sites, case 4/5 - cassette exons or mutual exclusive exons. Below are three examples shown. All PIP plots are found as supplemental material on CD (in file AlternativeSplicing\AlternativeSplices_PIPplots.pdf).

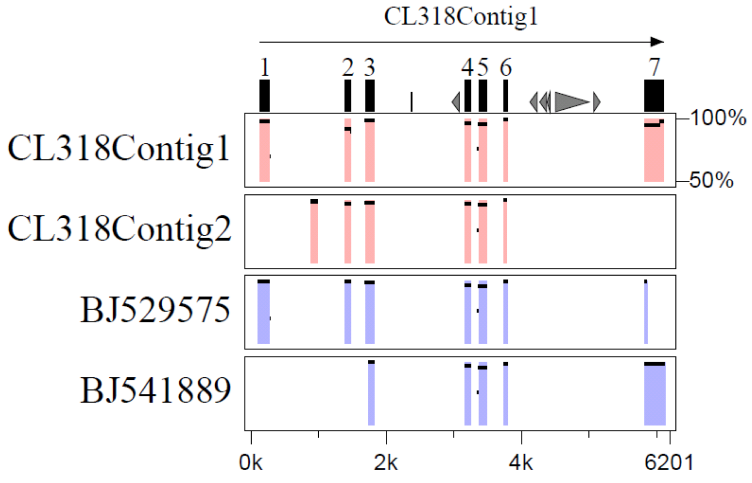
C53 - scaffold 1673: case 4-5.



C60 - scaffold 4355: case 4-5.



C81 - scaffold 650: case 4-5.



Appendix G

Supplemental material on CD

directory	file	description of content
Alternative Splicing	AlternativeSplices_PIPplots.ppt	PIP plots of all candidate genes
	AlternSplices_Exons.xls	Exons of est2genome output of all candidate genes
	AlternSplices_ASAP2.xls	Annotation of candidates for alternative splicing events with references in ASAP2
Annotation	GO_embryodevo.xls	annotation of interesting proteins from different GO categories
Clustering \plates1to20	tgiclSingletons_plates1to20.singletons.txt	result of tgicl EST clustering of the first 20 plates (first OFP run); EST singletons
	tgiclACE_plates1to20.txt	result of tgicl EST clustering of the first 20 plates (first OFP run); composition of EST contigs (CO - EST contig name; RD - EST belonging to contig)
Clustering \plates21to55	tgiclSingletons_plates21to55.txt	result of tgicl EST clustering of plates 21 to 55 (second OFP run); EST singletons
	tgiclACE_plates21to55.txt	result of tgicl EST clustering of plates 21 to 55 (second OFP run); composition of EST contigs

continued on next page

directory	file	description of content
Clustering \plates1to55	tgiclSingletons_ plates1to55.txt	result of tgicl EST clustering of all plates (both OFP runs); EST singletons
	tgiclACE_ plates1to55.txt	result of tgicl EST clustering of all plates (both OFP runs); composition of EST contigs
	tgiclSingletons_ plates1to55_ 10016_mod.txt	result of tgicl EST clustering of all plates (both OFP runs) modified to obtain 10,016 unique sequences; EST singletons
	tgiclACE_ plates1to55_ 10016_mod.txt	result of tgicl EST clustering of all plates (both OFP runs) modified to obtain 10,016 unique sequences; composition of EST contigs
Clustering \plates1to55_public	AllData_nr.fasta	result of tgicl EST clustering of all sequence data to all publicly available data; sequences of calculated EST cluster contigs and EST singletons in FASTA format
	AllData_Medaka. singletons_McF.txt	result of tgicl EST clustering of all sequence data to all publicly available data; EST singletons of this project not clustered with public data
	AllData_Medaka. singletons_CL.txt	result of tgicl EST clustering of all sequence data to all publicly available data; EST clusters of this project not clustered with public data
	ACE_605.txt	result of tgicl EST clustering of all sequence data to all publicly available data; this project's data which was clustered into 605 clusters together with publicly available sequences; result of contig analysis (clusters are grouped into three categories according to the amount of overlapping sequence; below this detailed results are found)
Fingerprint	medaka.r1.res.txt	fingerprinting result of first OFP run; first number after # is number of OFP cluster
	medaka.r2.res.txt	fingerprinting result of second OFP run; first number after # is number of OFP cluster
	newOFP_oldOFP.txt	fingerprinting result of second OFP run was changed according to EST clustering result; all modified OFP clusters or singletons are listed with their new OFP cluster (first column - new OFP cluster; second column - singleton clone or old OFP cluster now grouped into OFP cluster of first column

continued on next page

directory	file	description of content
PerlScripts	Ace.pl	see tab. 2.5
	Calc_Divindex.pl	see tab. 2.5
	Calc_Rearray.pl	see tab. 2.5
	Cluster_ESTs.pl	see tab. 2.5
	Edit_acefiles.pl	see tab. 2.5
	Edit_CloneAC.pl	see tab. 2.5
	Run_CAP3.pl	see tab. 2.5
PerlScripts\web	MedakaDB_new.cgi	cgi script, which manages database handling and GUI
	AbundanceProfile.html	OFP cluster abudance profiles; see output in figure 3.17
	CloneInfo.html	starting page; see fig. 3.15 for an example output
	MedakaDB_help.html	help page
	OFP_project.html	OFP project web page
Repeats	Repeats.xls	ESTs or EST contigs are listed containing repeats, classified according to kind of repeats (low complexity, microsatellites, TEs and others, RNAs and unknown)
SQLiteDB	sqlite3.exe	executes sqlite3 on the command line
	fifth.sqlite	SQLite database file of this project's data
UTRs	Align_mRNA.txt	genome alignment of known mRNAs to the Medaka genome in comparison to the alignment of EST contigs and EST singletons of this project, which were not annotated successfully

Publications

Manuscripts

Zadeh Khorasani, M. and Hennig, S. and Imre, G. and Asakawa, S. and Palczewski, S. and Berger, A. and Hori, H. and Naruse, K. and Mitani, H. and Shima, A. and Lehrach, H. and Wittbrodt, J. and Kondoh, H. and Shimizu, N. and Himmelbauer, H. (2004) A first generation physical map of the medaka genome in BACs essential for positional cloning and clone-by-clone based genomic sequencing. Mech Dev, 121, 903-913.

Henrich, T. and Ramialison, M. and Wittbrodt, B. and Assouline, B. and Bourrat, F. and Berger, A. and Himmelbauer, H. and Sasaki, T. and Shimizu, N. and Westerfield, M. and Kondoh, H. and Wittbrodt, J. (2005) MEPD: a resource for medaka gene expression patterns. Bioinformatics, 21 (14), 3195-3197.

Conference abstracts

Himmelbauer, H., Berger, A., Zadeh Khorasani, M., Hennig, S., Sasaki, T., Asakawa, S., Imre, G., Palczewski, S., Herwig, R., Hori, H., Mitani, H., Shima, A., Wittbrodt, J., Kondoh, H., Lehrach, H. and Shimizu, N. Transcriptome analysis and a first-generation physical map in BACs for the medaka. HUGO HGM2004, Berlin, 4th-7th April, 2004.

Acknowledgements

Firstly, my thanks go to PD Dr. H. Himmelbauer for his valuable supervision of this project and jointly with Prof. Dr. H. Lehrach, head of department of Vertebrate Genomics, for giving me the opportunity to carry out this project at the Max Planck Institute for Molecular Genetics.

Many thanks go also to Prof. Dr. G. Korge and to Prof. Dr. R. Mutzel from the Freie University of Berlin for their support and the supervision of this project.

I want to thank all people within the MPI, who contributed to this work. I got valuable support from Jana Illiger (Michal Janitz group) concerning the oligofingerprinting laboratory methods and computational support was kindly provided by the group of Ralf Herwig. I really appreciated the helpful advice in computational methods from Dr. Detlev Groth, especially in gene ontology analysis.

During this project I had the opportunity to spent one month in Prof. Shimizu's Lab at Keio Medical School in Tokyo were I was kindly supervised by Dr. Takashi Sasaki and Sabine Ishigawa.

I would also like to acknowledge all my colleagues in and around the Heinz Himmelbauer group: Juliane Dohm, Claudia Gösele, Boris Greber, Gabi Imre, Tobias Nolden, Stephanie Palczewski, Ruben Rosenkranz, Yinyan Sun, Helena Tandara, Maryam Zadeh Khorasani, Heike Zimdahl.

I would also like to thank Peter for his untiring support and love. I am glad you are always there for me!