5 Discussion

The use of fish models has been very popular during the last years for research in diverse areas. Teleosts such as zebrafish (*Danio rerio*), the pufferfish (*Takifugu rubripes*), and medaka (*Oryzias latipes*) are important model systems for analyzing gene function in vertebrates due to their relatively small genome size. Of them, medaka is widely used as a very well established genetic model system because of its advantages for genetical research, for instance the availability of divergent, completely inbred strains and a genome size of 800 Mb, less than half the size of the zebrafish genome. Moreover, sexes are easily distinguished in medaka in comparison to zebrafish. *Takifugu rubripes* has also been used as a model organism due to its compact genome size (only 365 Mb) and a gene repertoire similar to human (Brenner *et al.*, 1993), but no experimental tools are available for this species (Table 2-1). These aspects make the medaka an interesting tool for genetic research.

5.1 Use of medaka strains for creating a physical map

Altogether three different medaka genomic BAC libraries were employed for construction of a medaka physical map. These libraries were generated from the inbred strains, Hd-rR (a target of the genome sequencing project) and Cab (established by Wittbrodt's group and currently in use for mutagenesis) both from the southern population, and HNI from a northern population. Both Hd-rR and Cab strains are remarkably similar, with an estimated divergence of 0.035% in coding regions, corresponding to one SNP (single nucleotide polymorphism) in 2,800 bp. The average sequence divergence between the southern and northern population is about 0.8% in the coding regions and 2.6% in the intron regions (Naruse *et al.*, 2000). Considering that the difference between genomes of human and great apes is 1-2% (Fujiyama *et al.*, 2002), the SNP frequency between these two medaka populations is high, which makes it easy to find polymorphisms for analysis of linkage. The southern populations are genetically variable, though this is not reflected by strains Cab and Hd-rR. In contrast, the northern population is genetically homogeneous, and only very few genetic variation of protein and mitochondrial sequences have been observed (Sakaizumi, 1986).

As a considerable genetic distance has been observed between southern and northern strains, employing these three BAC libraries for construction of the medaka physical map holds several advantages:

First, it provides rapid entry into regions which change rapidly during evolution, to study them in molecular details. Such region could for instance be the major histocompatibility complex (MHC) or other intervals that contain genes with immunological functions, impact on fertility or environmental adaptation. Indeed, the MHC sequence is now available both from a southern and a northern strain and an alignment between the two regions identifies several segments exceptionally diverged (Tsukamoto *et. al*, 2005).

Second, BAC maps that integrate several library resources are inherently more complete. It has often been observed that some regions will not be represented in library A, but will be found in library B. This can either be attributed to chance or, alternatively, technical issues (vector, strain) or biological effects (e.g. sequences present in region of strain A that have toxic effects on *E.coli*).

Third, once positional cloning has led to the identification of candidate genes, BAC clones containing different allelic versions of the gene could be used to create transgenic fish to see if the phenotype can be rescued.

We have initiated a project to generate a physical map of the medaka genome in BACs. In the current absence of a contiguous sequence of the medaka genome, a BAC map is an essential tool for positional cloning of genes identified in ENU mutagenesis screens. For instance, the Hd-rR and HNI BAC libraries have been used for positional cloning of several genes, including the sex-determining gene *dmrt1bY/DMY* (Kondo *et al.*, 2001; Matsuda *et al.*, 2002; Nanda *et al.*, 2002) and for the genomic structural analysis of gene complexes such as the MHC (Matsuda *et al.*, 2002).

5.2 BAC mapping

Three artificial chromosome systems have been used to hold large-insert DNAs, the yeast artificial chromosome (YAC) system, the bacterial artificial chromosome (BAC) system, and the P1-derived artificial chromosome (PAC) system. Recently, the BAC vector, which is transformed into an *E.coli* host strain, has been the most frequently used because of its low levels of chimerism and its relative stability (Asakawa *et al.*, 1997). They can also easily be separated from the *E.coli* genomic DNA, using a standard plasmid isolation protocol employing the alkaline lysis method. Although the amount of isolated DNA is low because BAC-vectors are based on the F-factor, which is a bacteria fertility plasmid and maintained at only 1-2 copies per cell. Therefore, the DNA yield in this culture is low.

An important stage in the analysis of complex genomes is usually the production of a physical map, which is a diagram of the genome depicting the physical locations of various landmarks along the DNA. In other words, physical maps are maps of cloned genomic DNA.
The conceptual utility of a physical map is that it affords a single framework for organizing and integrating diverse types of genetic information, including the position of chromosome bands, chromosome break points, mutant genes, transcribed regions, and DNA sequences. The practical utility of a physical map is that it contains the information needed for positional cloning.

Different approaches have been suggested for BAC map construction. BAC fingerprinting is implemented by restriction digestion of BAC DNA, followed by electrophoretic separation of restriction fragments and band size estimation. If two BAC clones overlap, they will have restriction fragments in common, and thus will be placed into the same contig. BAC maps by fingerprinting have been constructed for multiple species, including man, mouse, rat and zebrafish (McPherson *et al.*, 2001; Gregory *et al.*, 2002 and Krzywinski *et al.*, 2004; http://www.sanger.ac.uk/Projects/D_rerio/WebFPC/zebra/small.shtml), to name a few. There are several major drawbacks of the fingerprinting approach. First, clone contigs that are produced will initially be to a large extent anonymous. Therefore, in a second step, markers with known location in the genome will need to be typed on the clones, to provide anchoring points. Second, fingerprinting will only be successful if BAC resources generated from genotypes are used that do not differ to drastically. Polymorphisms that alter the size of restriction fragments will complicate contig construction and local clone order. Finally, small stretches of overlap between clones will remain undetected by fingerprinting, to prevent an unacceptably high background of falsely assembled contigs. On the positive side, contigs assembled by fingerprinting allow the very accurate size determination of target regions, because the final product of a BAC fingerprinting project essentially is the restriction map of a genome.

Marker-content based approaches do not suffer from the drawbacks described above. Markers can be selected on the basis that contig location is known early on, BAC libraries from very different genotypes can be used and small overlaps between clones are readily detected. Technically, marker content mapping is conducted by PCR on pooled BAC DNA templates, prepared according to pooling schemes that reduce the amplification effort. Alternatively, hybridization assays of labeled PCR products or oligonucleotide probes against arrayed BAC

pools or BAC colonies are performed. STS content mapping will result in a BAC map that is useful already at an early project stage, because of the advantage to have markers with known features (sequence, location) on the map. When final maps prepared by fingerprinting or STS mapping, respectively, are compared, both resources will be equally useful. By locating markers on fingerprinted clones, or by providing integration points via BAC end sequencing, BACs can readily be selected for further experimentation.

5.2.1 BAC-based sequencing in the era of whole genome shotgun

The whole genome shotgun (WGS) approach has gained wide spread popularity for genome sequencing, because the generation of a large set of sequence data can easily be achieved (Venter *et al.*, 1998). However, the assembly of WGS sequences into the context of large sequence scaffolds is not trivial (Batzoglou *et al.*, 2002). As a compromise, a combination of WGS and BAC clone-based sequencing has been suggested. In practice, low-coverage sequencing of BACs is accompanied by parallel WGS read production. Subsequently, skimmed BACs are used as baits to identify matching sequence reads in the WGS data set, to obtain high sequence coverage of BAC inserts. This strategy has been successfully implemented to obtain a high-quality draft sequence of the rat genome (Rat genome sequencing project consortium, 2004). Low coverage sequencing from ordered medaka BACs, clone by clone (hence dubbed the CBC approach), will therefore significantly improve genome sequence quality, because assembly problems are essentially reduced to BAC-sized genome intervals. The map presented here will be of central importance to the CBC-based sequencing of the medaka genome. Our BAC map is to a large extent based on gene-derived markers, and encompasses a large proportion of markers currently employed for whole genome scans. Thus, genetic map assignments provide immediate access to underlying clones and contigs, simplifying molecular access to candidate gene regions and their characterization.

5.3 Source of probes utilized for map construction

To construct a physical map of the medaka genome based on BAC clones, probes derived from different sources were utilized. At the beginning of this project, we generated a set of random BAC end-fragments without any information about the position of these clones in the map. Use of BAC end-fragments has some gains and drawbacks. For instance, markers derived from opposite ends of a given BAC clone are separated by the length of the insert and may also

cover the regions without genes. Furthermore, the chance of not detecting overlap between contigs is minimal. Another side to use the BAC end-fragments is that the repeat content of genome must be known. Otherwise, many probes will hybridize to many targets in the genome. During production of these probes (September 2001-June 2002), the sequence of the whole genome of *Takifugu rubripes* was published, which had great influence on genomic research in medaka and other fish systems. We also had received medaka cDNA clones from the lab of Professor Hiroshi Mitani, University Tokyo, Japan (see below). The insert sequences of cDNA clones could be used as anchors for our map. Because creating probes using the synteny between medaka and fugu and also amplification of inserts of cDNA clones was easier and faster than the production of BAC end-fragments, this work was discontinued, and the marker-content approach was used for further work. Altogether, 187 end-fragment sequences from Cab and Hd-rR BAC clones were used to complete the medaka map.

For the construction of a medaka BAC map by STS-content mapping, we took advantage of two other available resources, first, medaka cDNA sequences deposited in GenBank and second, the draft sequence of the pufferfish *Takifugu rubripes* genome.

The largest group of probes was obtained by comparison of published medaka EST sequences (NCBI) with the draft sequence of the pufferfish *Takifugu rubripes* genome. Phylogenetic analysis places medaka in close relationship to the pufferfish, with an assumed divergence time of 60-80 Myr (Figure 2-1), which is less than the evolutionary distance between human and mouse. As genes and gene order remained largely conserved during evolution, it was suggested that the synteny between medaka and fugu should be high, even though genome sizes differ twofold between the two species. Our assumption is supported by an analysis of sequence data generated from BAC clones located on medaka linkage group (LG) 22, revealing conservation of synteny, despite the expected size discrepancies that distinguish the two genomes (Sasaki *et al.*, 2004). Therefore, fugu genome scaffolds could be used as reference points to create the medaka physical map. In the fugu assembly, scaffolds are sequentially ordered with respect to their size in base pairs. The large number of scaffolds does not adequately reflect the true quality of the sequence assembly. For instance, scaffolds 1–300 together contain 30% and the first 1500 scaffolds encompass 70% of the total assembly. At the same time, a large number of EST sequences from medaka cDNA clones had been generated and deposited in the public domain.

5.4 Map construction

103,144 public medaka EST sequences were downloaded from the NCBI website. Large-scale medaka EST analysis and gene mapping are essential for positional cloning of genes responsible for mutants and the genomewide comparison of linkage relationship among vertebrate species. To produce a representative non-redundant medaka EST set a clustering procedure and Cap3 software (Poustka *et al.*, 2003; Altschul *et al.*, 1990) were used, and the 103,144 ESTs were merged into 21,121 sequence clusters of which 9,379 were true clusters containing more than one EST, while 11,742 remained singletons. The number of 21,121 clusters will certainly still be inflated with respect to the number of different genes represented in the data set for several reasons. First, we adopted a conservative approach in cluster generation such that true sequence overlap was necessary for a decision whether two ESTs belonged to the same cluster or not. However, many cDNA clones had been sequenced from both ends and sequences did not overlap, and therefore were assigned to different clusters. Second, cap3 will split pre-clusters that contain an internal region of non-homology, observed, for instance, in the case of alternative splicing of transcripts. Third, the 3'-end of a gene may vary due to the usage of alternative polyadenylation sites. Since the bulk of cDNA libraries had been constructed from oligo (dT) primed mRNA, the variation in the length of 3'UTR (3' untranslated region) sequences could have a profound impact on clustering statistics. The last assumption is substantiated by the observation that only 38% of clusters matched a known protein in nrprot at a stringency cutoff of 1.0e-10.

To avoid oversampling of gene-dense regions, these 21,121 sequence clusters were aligned against the fugu genome draft. The comparison of medaka ESTs with fugu is available at (http://www.molgen.mpg.de/~hennig/medak/fugu-v3-update-12-2003/SCAFF-HITS-table.html). 11,254 of clusters (53%) successfully matched a fugu scaffold. Altogether, 3,397 fugu scaffolds were hit by at least one medaka cluster. These 3397 scaffolds encompass 249 Mb of the fugu sequence, corresponding to 75% of the total assembly. The "matched" fraction of medaka clusters is vastly enriched in protein coding entities. 62% of "matched" clusters contained protein homology to an nrprot entry at a stringency threshold of $1.0e^{-10}$. Conversely, only 10% of "non-matched" EST clusters found a match in nrprot. The fugu genome sequence is not entirely complete, for instance, 4.1% (13.7 Mb) of scaffold sequence is composed of unspecified nucleotides, bridging known gaps within the scaffolds. Protein-coding genes represented in the medaka EST set that do not match a fugu scaffold could therefore be localized within gaps in the present sequence assembly. The high proportion of medaka EST

sequences without protein homology in the 'non-matched' EST fraction, however, is highly indicative for an enrichment of sequences that are less well conserved during evolution, e.g. extended 3'UTR sequences.

For many of medaka genes it is the sequences that are publicly available, but not the clones, as physical entities. We therefore devised a strategy that allowed us to fully exploit the wealth of existing medaka cDNA sequence data, by designing 35mer oligonucleotide probes. Our approach maximizes the flexibility to convert sequences to markers on the map, because the strategy is of course not restricted to probe design based on cDNA data. In fact, any STS or genetically mapped marker or BAC end-fragment, for which sequence information exists, can be converted into a marker for the mapping project. In addition, 35mer probes are free of vector sequences and can be designed such that known repeat sequences are avoided. For the current project, we designed 2363 35mer probes based on the alignment of medaka EST clusters against the fugu genome sequence. Since the exon structures of genes remain conserved during evolution, this strategy prevents the inadvertent design of 35mer probes that are disrupted by intronic sequences in genomic DNA. As expected, the density of matched medaka EST clusters varied greatly between fugu scaffolds. On average, we observed one hit per 24 kb of fugu genomic sequence. As some clusters hit more than one location, 9586 cluster hits, corresponding to 8588 different clusters, are contained within the 1500 largest fugu scaffolds. These scaffolds range in size between 1145 kb (scaffold 1) and 52 kb (scaffold 1499). We aimed at selecting medaka clusters for probe design that matched 100 kb apart from each other in the fugu genome. Assuming the fugu genome as half the size of medaka and knowing that BAC clones in the Hd-rR library exceed 200 kb average size, we anticipated that our probe design strategy would be most economical in terms of translating the fugu genome sequence into medaka BAC contigs. To design these probes, data from the website (http://www.molgen.mpg.de/~hennig/medak/fugu-v3-update-12-2003/SCAFF-HITS-table.html) was used.

5.5 Medaka LG22 project

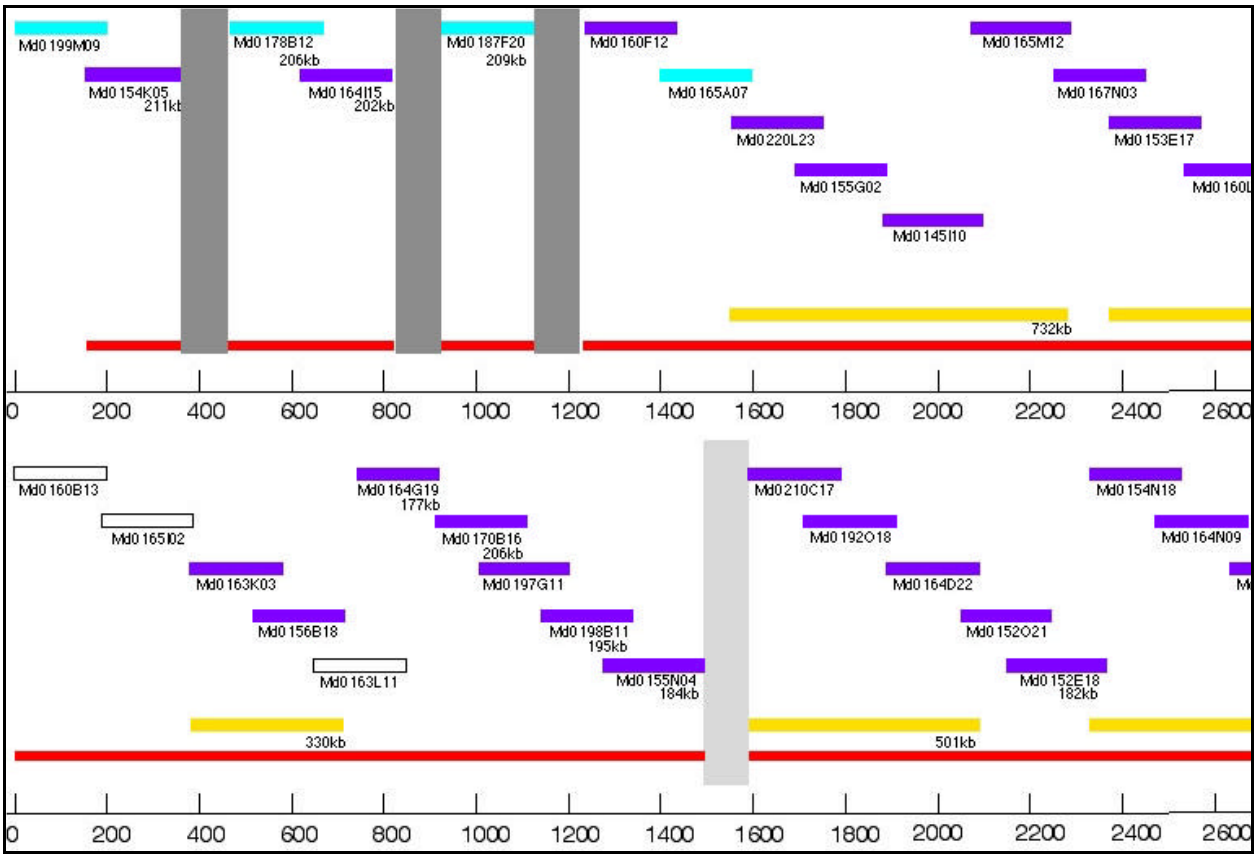
The aim of the LG22 project is to generate the sequence of the entire chromosome based on BAC map. As a model chromosome, it will provide important information about the characteristics of the medaka genome, including GC contents, repeat elements, gene density, promoters, and average intron size.

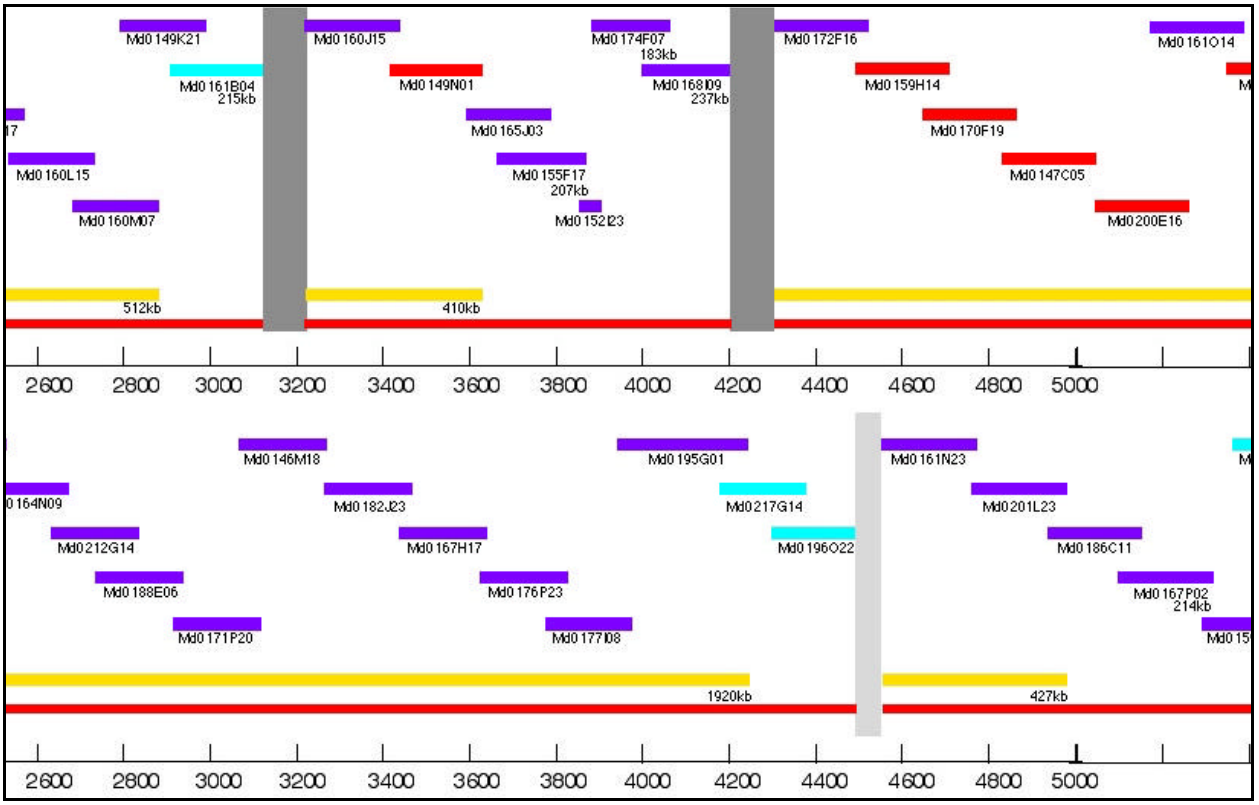
LG22 was chosen as a first medaka chromosome for complete sequencing because it is one of the smallest chromosomes (19.2 Mb) in medaka genome and it is relatively rich in genes as compared to other medaka chromosomes (Naruse *et al.*, 2004). Furthermore, it was shown that several orthologs of known genes on medaka LG22 are located on tetraodon chromosome 10, zebrafish chromosome 17 and several human chromosomes (Mitani, unpublished data). The sequence of medaka LG22 is the first whole chromosome sequence from non-mammalian vertebrate.

The medaka chromosome LG22 BAC map consisted of 127 BAC clones with only 5 clone gaps, which determined the genomic DNA sequence of 18,803,338 bp. The overall GC content of LG22 is 40.9%; almost the same as in the human (41%) (International Human Genome Sequencing Consortium, 2001) and mouse genomes (42%) (Mouse Genome Sequencing Consortium, 2002), but about 5% lower than the both pufferfish (*Takifugu rubripes* and *Tetraodon nigroviridis*) genomes (45%) (Aparico *et al*., 2002; Jaillon *et al*., 2004). Average GC content of all the exons is 51.6%; about 10% higher than the overall GC content of LG22 (40.9%) (Sasaki *et al.*, submitted). Figure 5-1 shows a current map of medaka LG22.

Furthermore, analysis of the repetitive elements with RepeatMasker2 (http://repeatmasker.org) indicated that these repeat sequence elements occupy 26.7% (5,028,599bp) of the LG22 genomic sequence (Sasaki *et al.*, submitted).

Moreover, on the medaka chromosome LG22 633 protein-coding genes, 18 RNA genes that apparently produce novel transcripts without coding sequence, and 7 pseudogenes were identified. The average number of exons in 633 protein-coding genes is 10.0, and the average size of exons is 183 bp. The average number of coding exon per gene is 9.6; not much different from fugu (8.8), tetraodon (6.9), mouse (8.8) and human (9.0). The overall gene density of medaka LG22 is 33.7 genes/Mb; 4 times higher than the human genome (approximately 8 genes/Mb) (International Human Genome Sequencing Consortium, 2001; Mouse Genome Sequencing Consortium, 2002).


Md0199M09
Md0178B12
206kb
Md0187F20
209kb
Md0160F12
Md0165M12
Md0154K05
211kb
Md0164I15
202kb
Md0165A07
Md0167N03
Md0220L23
Md0153E17
Md0155G02
Md0145I10
732kb
0
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
2600
Md0160B13
Md0164G19
177kb
Md0210C17
Md0154N18
Md0165I02
Md0170B16
206kb
Md0192O18
Md0164N09
Md0163K03
Md0197G11
Md0164D22
Md0156B18
Md0198B11
195kb
Md0152O21
Md0163L11
Md0155N04
184kb
Md0152E18
182kb
330kb
501kb
0
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
2600


Md0149K21
Md0160J15
Md0174F07
183kb
Md0172F16
Md0161O14
Md0161B04
215kb
Md0149N01
Md0168I09
237kb
Md0159H14
Md0165J03
Md0170F19
Md0160L15
Md0155F17
207kb
Md0147C05
Md0160M07
Md0152I23
Md0200E16
512kb
410kb
2600
2800
3000
3200
3400
3600
3800
4000
4200
4400
4600
4800
5000
Md0146M18
Md0195G01
Md0161N23
Md0164N09
Md0182J23
Md0217G14
Md0201L23
Md0212G14
Md0167H17
Md0196O22
Md0186C11
Md0188E06
Md0176P23
Md0167P02
214kb
Md0171P20
Md0177I08
1920kb
427kb
2600
2800
3000
3200
3400
3600
3800
4000
4200
4400
4600
4800
5000

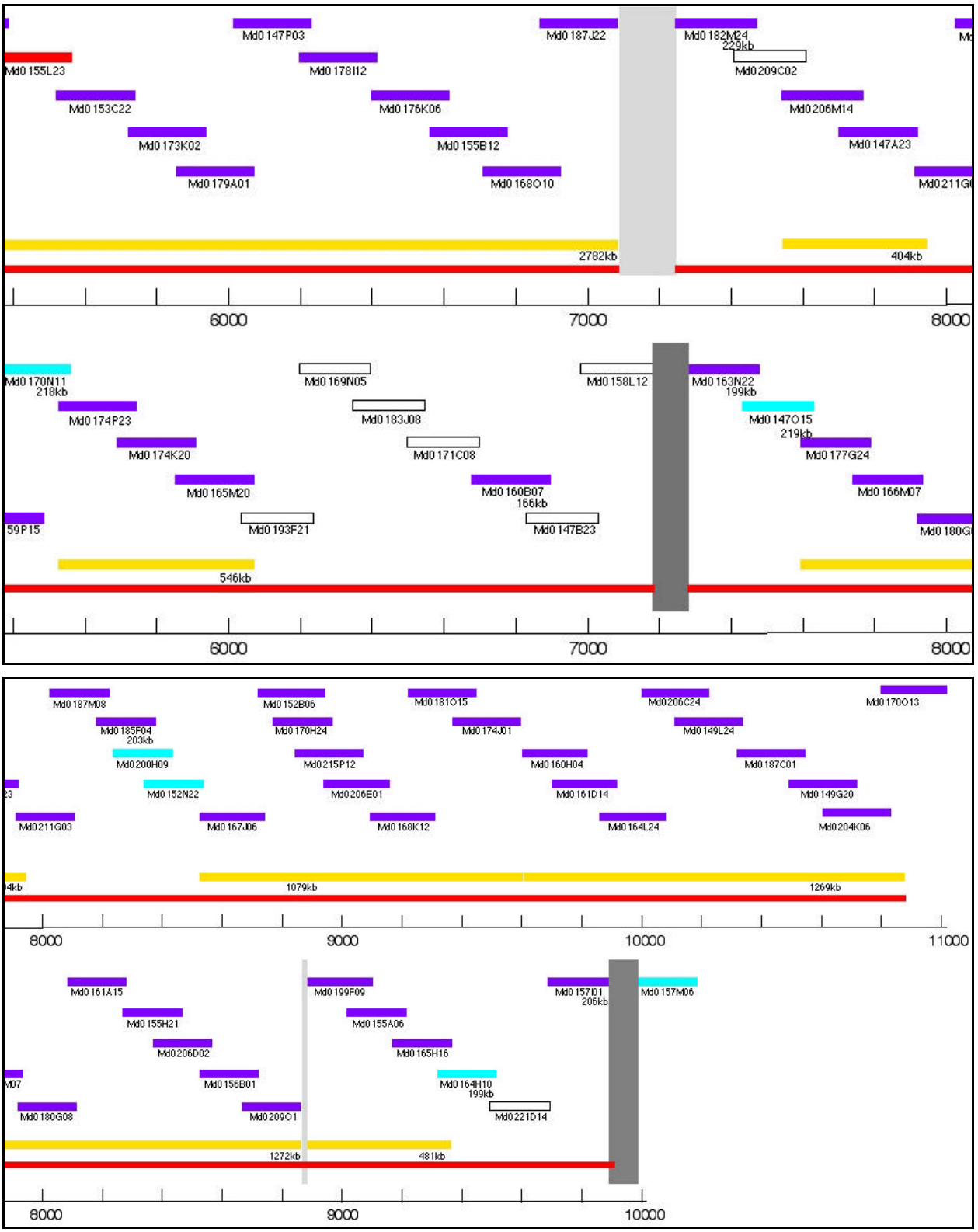


Figure 5-1: The medaka LG22 map (Sasaki, *et al.* submitted).

Sequence comparison between medaka LG22 and tetraodon chromosome 10 suggested that they might have been derived from a common ancestral chromosome (Naruse *at al.*, 2004).

This comparison revealed only a few incidences of intra-chromosomal shuffling (~1%) and a quite rare inter-chromosomal rearrangement (less than 5%).

Of 633 protein-coding medaka genes, 588 genes have obvious human orthologs as counterparts. Among these orthologous genes, 248 genes (42.2%) are located on human chromosome 14, indicating that the majority of these genes must have been located on a single chromosome of the common ancestor of medaka and human. Most of the remaining 232 genes are located on four other human chromosomes 1, 2, 6 and 15. All together, on these five human chromosomes represent 81.6% (480/588) of all human genes that are orthologous to LG22. Furthermore, analysis under the new concept "conserved gene cluster" (GCG) revealed that 37.5% of the pair-wise gene relationship on the ancient chromosome of common ancestor in human and medaka were conserved. CGC terms that the order and direction of two or more genes is conserved among species without interruption by any other genes. It was then estimated that approximately 20,000 breaks of pair-wise gene relationship would have occurred between the human and medaka lineage (Sasaki *et al.*, submitted).

5.6 Outlook

This work presents the first generation physical map of medaka in BAC clones containing 2,534 gene-derived markers and 187 end-fragments. As the next step, we would like to resolve ambiguities present in the current map, filling the gaps within the map segments and therefore to enlarge the map segments to complete the map. Thus, we designed another 1,152 35mer oligonucleotides from medaka ESTs aligned against fugu scaffold and hybridized them against BAC filters. Although the current map is not complete and corresponds to 74% of the medaka genome, this map has been widely used by other groups, working on the positional cloning of mutations affecting the formation of heart and thymus, as well as showing defects in eye and CNS development (Table 5-1).

The clone by clone (CBC) sequencing of the medaka genome has been initiated in the framework of the medaka genome initiative, and our map will be used as a backbone. As the medaka LG22 is one of the smallest medaka chromosomes, the CBC sequencing currently focuses on LG22 as a medaka model chromosome. In conclusion, the data presented here will have a profound impact on the efficient use of medaka as a model system, to promote the translation of genomical information into biological knowledge.

Positional cloning of ENU-induced medaka mutations		
mutant	**phenotype**	**cooperation**
koepke	tailless	Carl/Wittbrodt
fukuwarai	head malformation	Furutani-Seiki/Kondoh
34-7A	no thymus	Furutani-Seiki/Kondoh
21-30C	no thymus	Furutani-Seiki/Kondoh
sakura	heart defect	Furutani-Seiki/Kondoh
sunglasses	eye defect	Martinez/Wittbrodt
ojoplano	eye defect	Martinez/Wittbrodt
Chinese Ink	eye defect	Martinez/Wittbrodt

Table 5-1: Current utilization of BAC map data for positional cloning projects.