5 Discussion

The subject of this thesis can be divided in three parts. The first part deals with the adaptation of IRS-PCR for physical and radiation hybrid mapping of the zebrafish (4.1). In the second part, physical mapping of zebrafish linkage group 20 using STS and IRS markers was implemented (4.2). The third part consists of radiation hybrid mapping of ESTs of biological interest (4.3). Additionally, tests were carried out to evaluate other hybridisation based, high throughput oriented mapping techniques (4.4).

5.1 IRS-based methods for mapping the zebrafish genome

Part 1 is a continuation of the work which was done on genetic mapping using IRS-PCR (Burgtorf, 1999). Genetic mapping of the zebrafish showed to be inefficient due to the lack of pure inbred strains resulting in a high level of heterozygosity in the mapping strains. IRS-PCR is particularly susceptible to errors caused by heterozygous alleles, because IRS markers are not codominant (like microsatellites), so heterozygotes and homozygotes can not be easily distinguished. To evade this problem, haploid offspring of mapping crosses can be used. Haploid embryos however are only viable up to day 3, and therefore it is not possible to obtain large amounts of DNA for big mapping projects. For that reason, a new mapping panel has been created from the rather isogenic C32 and SJD lines. From F1 females, gynogenetic diploid offspring was generated and raised to maturity (Kelly et al., 2000) (Woods et al., 2000). This panel is now used in large scale mapping of genes and ESTs, and might also be suited for IRS-PCR based mapping, but was not available to our laboratory. For a detailed discussion of IRS based genetic mapping in the zebrafish see the Ph.D. thesis of C. Burgtorf (Burgtorf, 1999). Besides dominance of the marker and heterogenic mapping strain, a third problem arises due to the high repeat content of IRS-PCR products, as revealed by sequencing. This concerns also radiation hybrid and physical mapping, but can be diminished by use of stringent hybridisation conditions and competitive reannealing of probes with unlabelled genomic DNA.

5.1.1 Identification and testing of anchor sites for IRS-PCR

For physical and radiation hybrid mapping, heterozygosity of strains is not an issue, making it better suited for IRS based techniques. In contrast, a high frequency of products per genome segment. (i.e. a high marker density) is critical to be able to order clones in large

contigs. For this reason, different repetitive elements, primer binding sites and PCR conditions have been tested.

Sequences containing the mermaid repeat and a newly discovered SINE–like *arielle* element were retrieved from databases and aligned to calculate a consensus sequence. Primers differing with respect to their binding sites and number of wobble bases were tested for their efficiency to yield IRS-PCR products. As one result of these tests, the already published primer MMA, binding to the C2 region of the DANA/mermaid element, showed the highest number of products per PAC (1 product per PAC; average PAC length ≈ 120 kb). However, under high throughput experimental conditions, not all of those products were reproducible or abundant enough to be used in hybridisation, possibly because the primer binding sites did not match the primer sequence sufficiently. Consequently, only one IRS product out of three PACs could be used. This number is similar to the rate in the mouse, which has 1 product per 360 kb using the B1 repeat as anchor (H. Himmelbauer, pers. communication). This is somehow contrary to expectation, because the zebrafish genome has approximately half the size of the mouse genome, but probably more genes (Wittbrodt et al., 1998), so one would assume, that the repeat density in the zebrafish is lower than in the mouse.

5.1.2 Size distribution of IRS-PCR products

In a multiplex, single primer PCR, there are two processes, which influence length distribution: (a) Preference for short PCR products due to premature termination of elongation. (b) Suppression of short products because single primer PCR introduces inverted terminal repeats, which cause self-annealing of smaller fragments. Taken together, these effects should suppress very short and very long products. The use of long elongation times and a high primer concentration in our IRS-PCR protocol should alleviate both effects (Shagin et al., 1999). PCR analysis of the IRS marker libraries shows, that only 5% of the cloned products are longer than 1.5 kb, and that inserts longer than 2 kb are rare. The PCR reaction was carried out with long annealing and extension times (1 min annealing and 3.5 minutes extension), so this is unlikely to be a limiting factor. The primer concentration was also very high, which is supposed to compete with self-annealing of fragments with inverted ends. Taken together with empirical tests, the PCR conditions used can be considered as optimised to deal with suppression effects and preference for small products.

5.1.3 Complexity of the IRS-PCR amplicon

As an extrapolation, 1 product per PAC suggests that there are ~ 14,000 products in the haploid genome. 1 product every 3 PACs suggest ~ 5,000 products. By oligonucleotide fingerprinting of the IRS library, ca. 11,000 different products were found, among which 9345 were singletons. The library was constructed from IRS products from 5 different strains, and many products occur only in a subset of strains. Thus the number of products per haploid genome is lower. Sequencing of inter-mermaid products revealed, that many of them contain a mermaid homologous sequence only at one end. This is probably due to mispriming at the other end, given the fact, that PCR conditions were not highly selective. For that reason it is appropriate to consider it as repeat anchored PCR rather than inter-repeat PCR. IRS-PCR from genomic clones as well as analysis of the IRS marker library confirms that - despite the randomising influence of partial mispriming - fragments are amplified specifically and reproducibly. It has been estimated that there are $0.1 – 5 \times 10^5$ DANA/mermaid elements in the haploid genome. Due to the uncertainty concerning the copy number, and the fact, that probably a majority of IRS-PCR products is misprimed at one end, calculations about the number of potential mermaid products based on estimations of repeat distribution can be considered irrelevant.

5.1.4 Repeat and SNP content

Sequencing of IRS products shows that there is a high incidence of repetitive or low complexity sequence in the products (besides the mermaid ends). This can cause problems in hybridisations due to non-specific crosshybridisation, and is probably the reason for the high background particularly in the radiation hybrid hybridisations. During the progression of this work, several methods have been applied to lower background signals, including raising the hybridisation and washing temperatures and competitive reannealing of probes with unlabelled genomic DNA. This resulted in an improved signal to noise ratio, but did not totally eliminate problems with background. Additional use of unlabeled repetitive DNA (mermaid, $(CA)_n$) for competitive reannealing of the probe and blocking of the filters could further improve the outcome, but due to he nature of the probe, a residual background hybridisation cannot be totally eliminated.

Sequencing also reveals the high incidence of SNPs in orthologous IRS-PCR products from different strains (1.3%). This suggests the use of our normalised IRS marker library as a

reduced representative subset of the genome for SNP mapping. IRS markers containing SNPs can be anchored directly to the physical clone map and – although with less efficiency – to the radiation hybrid map by hybridisation. For SNP genotyping, various techniques have been developed (based on e.g. nucleotide incorporation, mass spectrometry or hybridisation to DNA arrays).

5.1.5 IRS-PCR pools for physical mapping

Pooling of clones greatly reduces the amount of spots on a filter. The increase of complexity (and accompanying risk of background hybridisation) is compensated by spotting IRS amplicons instead of whole clones. Determination of clone addresses from a pool hybridisation is sensitive against false negative results, because clone addresses have to be complete and unambiguous. To lessen this problem, we employed a redundant pooling scheme, in which each co-ordinate (plate, row, column) was represented twice. The increase in spot size owing to redundancy made it necessary to use filters of double size, thus in a way decreasing throughput. The larger number of clone addresses, which could be resolved, however compensated for this. The average number of identified clones per hybridisation was 3.3. This is far less than expected, given the 17 fold genome coverage which is represented on the filters. Several reasons for this high rate of false negatives are conceivable: 1.) IRS-PCR is inherently a multiplex PCR reaction and products are not amplified to an equal abundance independently from the background. Hence, a lot of incomplete clone addresses appear (false negatives). This effect is partially diminished by the use of a redundant pooling scheme. 2.) High background hybridisation levels result in spots falsely scored as positives. This can cause correct clone addresses to become ambiguous, since two spots from the same block of 8 plates and the same dimension (e.g. the same plate) can not be resolved. 3.) High background hybridisation levels can make entire regions of the filter unscorable. Although problems connected with PCR multiplexing persist, it is possible to influence the product length preference of the reaction.

5.1.6 Characterisation of a new interspersed repeat

A new interspersed repetitive sequence was discovered to occur in ESTs sequenced in the Washington University zebrafish EST project (M. Clark, pers. communication). For this work its structure was analysed in more detail, and its use as an anchor in IRS-PCR was

investigated. As a first step, copies of the repeat were searched in databases and aligned, to calculate a consensus sequence. This consensus sequence was investigated further for structural features. The repeat was tentatively called *arielle* and it has some features in common with composite SINEs like DANA/*mermaid*. It is around 500 bp long, and has a poly-A tail. Like DANA, it consists of four rather conserved regions interrupted by stretches of repetitive DNA. Also direct and inverted repeats at the insertion site were found. However, it does not seem to contain a conserved RNA polymerase III promoter, a diagnostic feature of SINEs. It is not recognised by a tRNA scanning program. This makes it an untypical SINE-like element and the mechanism of propagation is unclear. No homologous sequence was found in the Fugu genomic sequence available, restricting the element to the zebrafish and maybe close relatives.

In IRS-PCR using *arielle* as an anchor, very few products were obtained (one product from 16 PACs, some discrete bands from whole genomic DNA). This showed, that the element is far less abundant than mermaid, and that it is not suitable for IRS-PCR unless used in combination with other primers. To further study this element it would be important to have more exact estimates about the abundance and distribution. An easy way to determine the copy number is spotting dilution series of the cloned element (as a standard) and of genomic DNA on a filter and hybridising with a repeat specific probe to compare the signal. It can also easily be tested if the element is somehow randomly distributed across the genome, or if it is associated with highly repetitive heterochromatic regions. For this, one has to hybridise total genomic DNA onto a large-insert genomic library. The strongest signals come from clones containing highly repetitive DNA. As a comparison, an arielle specific probe is used, and it is determined, if there is the same pattern as with the whole genomic probe. If this is not the case, and the signals are distributed quite evenly, one can assume that it is a rather evenly distributed interspersed element.

To investigate the mechanism of transposition is more difficult, and it might well be, that no transposition is occurring in recent times. It has been proposed that SINEs get transcribed by RNA polymerase III and reverse transcribed by the LINE encoded reverse transcriptase and integrase machinery (Okada, 1991). In vitro transcription of total genomic DNA using RNA polymerase III has therefore been suggested as a method to detect SINEs. However, the lack of a conserved RNA polymerase III promoter will leave *arielle* undetected in this kind of test. It has also been shown, that some SINE like sequences are in reality truncated 3' fragments of LINEs (Nikaido and Okada, 2000). To test this for the *arielle* repeat, a closer look at the upstream and downstream sequences has to be taken.

5.2 Physical mapping of the zebrafish linkage group 20

Part two of this work deals with physical mapping of a zebrafish chromosome. Physical maps of genomes or genomic regions are constructed to support positional cloning and sequencing. Positional cloning usually requires a genome walking step, i.e. the construction of a clone contig spanning a defined region. A global physical mapping project of a whole genome or chromosome can fulfil this task more efficiently, than if it was done for every single mutant separately. A physical map serves also as a backbone of sequencing large genomes (Green, 2001). Physical maps can be constructed by identifying clones containing genetic markers (STS content map), or by screening clones for overlaps (clone based maps). A purely clone-based map like a restriction fingerprint map or a map based on hybridisation of anonymous DNA fragments (e.g. IRS-PCR products or insert ends) consists of a path of overlapping clones. A restriction map additionally includes data about overlap length, which is important to select a minimal tiling path for sequencing. It does not require any information about sequence or chromosomal location of the contigs, although this would be necessary for the later sequence assembly step. These data can be added by screening the library with STS markers. Alternatively, following assembly of the map, BAC end sequences of contigs can be anchored to a radiation hybrid map (this is planned by the Tübingen mapping project).

In contrast to that, an STS based framework map directly anchors clone contigs to the genetic and radiation hybrid maps. Consequently, the genetic and radiation hybrid sizes of the contigs and also of the gaps are known immediately, but information about the physical size of contigs and overlaps – needed for construction of a minimal tiling path for sequencing – are not available straightaway.

5.2.1 Mapping strategy

We used an STS-based approach to construct a framework of zebrafish chromosome 20 consisting of PAC and YAC clones anchored to the genetic and radiation hybrid maps. We decided to concentrate our effort on one chromosome, because we wanted to achieve a high marker density on a defined part of the genome, with the limited resources available. The rationale behind this approach was to establish a model chromosome, which serves as a paradigm for sequencing and assembling the other 24 zebrafish chromosomes. Chromosome

20 was chosen despite its relatively large size, because a laboratory collaborating in the zebrafish EST project was especially interested in it (M. Clark and S. Johnson, pers. communication).

We hybridised STS-derived oligonucleotides against a gridded PAC library with 5.8 x genome coverage. PACs recognised by the probes were picked and used for generation of IRS-PCR probes that were hybridised against YAC and PAC pools. The STS probes were supposed to detect anchored framework PACs as seeds for contig construction. IRS hybridisation was used to find additional overlapping PACs and YACs, the latter potentially spanning larger regions. As a result we obtained an STS-tagged, chromosome 20 specific set of large insert clones spanning large portions of the chromosome.

Our approach is complementary to the restriction fingerprint map generated at the Max-Planck-Institut for Developmental Biology in Tübingen. In a collaborative effort, we are restriction fingerprinting our mapped PAC clones. This is a control to verify our contig assembly and to determine the physical sizes of contigs and overlaps. It will also integrate our STS-based map with the restriction fingerprint map.

5.2.2 Mapping results

As described in 4.2, 344 STS were used, including pooled probes. 3196 clones were hit by these probes, 2147 of them with hybridisation strength ≥ 2. 284 IRS probes were used, hitting 561 YAC and PAC clones. Altogether, 3416 PACs were hit. These PACs were picked and rearrayed as a linkage group 20 specific sublibrary. Currently they undergo restriction fingerprinting. For the contig assembly, we only used single probes and hybridisation results with signal strength ≥ 2.

A subset of the above hybridisation data were used for contig construction with the program wprobeorder (Mott et al., 1993). These included hybridisations with signal strengths ≥2 in the PAC colony hybridisations, as well as pool hybridisations resulting in complete and unambiguous clone addresses. These criteria are met by 388 probes, hitting 2007 clones, representing a marker density of 1 per 205 kb. The contig assembly resulted in 249 contigs, of which 19% were anchored by more than one probe. To test, if our hybridisation results match the theoretical predictions, we calculated the predicted number and size of the contigs using a program written by A. Grigoriev (Grigoriev, 1993). The algorithm works under the assumption, that the probes are independent from each other and distributed randomly across the chromosome. This assumption holds for the STS probes, but not for the

IRS probes, because they are derived from PACs hit by STS probes. For that reason, only hybridisations of STS probes on PAC filters are used for a comparison of the theoretically expected and the empirically determined contig data. 216 STS probes hit 1476 PAC clones with signal strength ≥ 2 (6.8 clones per probe). This is higher than the genome coverage of the clones on the filters (5.8). There are several reasons for this discrepancy: (1) EST derived probes hit clones representing gene-rich regions, which might be overrepresented in the library as opposed to repetitive, heterochromatic structures. (2) Even at high signal strength, there are false positive results, probably because some probe sequences have more than one copy in the genome. Eliminating all hybridisations hitting more than a specific number of PACs can reduce that problem. In our empirical results, the clones assemble into 155 contigs, 18% of which are anchored by more than one probe. In our prediction, there are 167 contigs with an average length of 212 kb. 23% of the contigs are anchored by 2 or more probes. 40% of the genome is covered by contigs. This shows that our empirical results are consistent with theoretical predictions. The small deviation of the number of contigs is probably caused by the marker localisation, which is not completely random. Some chromosomal regions have a higher marker density, and as a result of that, there are less but longer contigs (anchored by more probes). The genome coverage predicted should be seen as the lower limit of the real value, because IRS based mapping data are not included in the dataset underlying the simulation. Coverage of the chromosome and the average contig length are higher, when the IRS hybridisation data are integrated, particularly due to YAC clones, which have an average insert size of 480 kb. However, as outlined in 4.2.2, hybridisation of IRS probes onto pool filters was not an efficient process (probes could be generated from 1 out of 3 PACs, 60% of probes gave unambiguous results). Under these circumstances, an alternative walking strategy, based on hybridisation of PAC insert ends, seems to be more promising. The advantage of this method lies in the fact, that in principle two probes with a distance defined by the insert length can be generated from each clone. But in contrast to IRS probes, PAC ends have to be hybridised on large PAC colony filters, thereby decreasing throughput. Currently the use of PAC end probes generated by inverse PCR is tested (P. Nierle, diploma thesis, in preparation).

During the time of this writing, the physical mapping project is still in progress. Since the last contig assembly (described above), hybridisation work has been continued. Small (7x11 cm) colony filters were created from the LG 20 specific sublibrary. 100 STS probes, which had been hybridised as pools against the whole PAC library before, were hybridised now as single probes on the small filters. Additionally 309 IRS probes were hybridised (P. Nierle,

diploma thesis, in preparation). The ongoing mapping effort will further increase the marker density, contig sizes and chromosome coverage. All mapping data, including marker sequences, mapped clones etc. will be accessible at our webpage (http://www.molgen.mpg.de/~ag_zebrafish/, in preparation).

5.2.3 Significance of the map

Integration of STS and restriction fingerprinting data of our map of chromosome 20 with the restriction fingerprint data of the Tübingen map will make the map of this chromosome the so far best anchored and most complete of all. Sequencing of this chromosome is facilitated and one can expect that there will be a high quality assembly of this chromosome early during the process of sequencing the whole genome. This will make chromosome 20 a model for other zebrafish chromosomes. Such a model is useful for studying chromosome structure and function in many ways.

Most hybridisation probes are derived from transcripts, and the PAC sequences will cover the whole genomic region of the genes. By isolating and sequencing the full length cDNA or maybe just by comparison to the NCBI zebrafish unigene set, it will be possible to study the genomic structure of genes. This is very useful as a test and training set of gene prediction, which is still a difficult operation particularly in the zebrafish, because the software is usually optimised for the human genome. Sequenced PACs will also contain upstream regulatory elements of the genes. These can be identified using comparisons with the Fugu, Tetraodon, human and rodent genomes. Evolutionary comparison of control elements of developmentally relevant genes between these clades will be of special interest to explain changes in body plans. The increasing amount of expression data of the developing zebrafish embryo will be used also for the definition of promoter function. Genes that have identical expression patterns (synexpression groups, Niehrs and Pollet, 1999) are likely to posses homologous promoter elements. These can be found by aligning the upstream regions, using profile searches and hidden Markov models. Subsequently, their functions can be tested by injection of GFP reporter constructs.

Not only the structure of single genes, but also the more global chromosome architecture is best studied on a single model chromosome, before the complete genomic sequence is finished. This concerns the distribution of genes and repeats, which is an important factor for assembly. Particularly structures containing highly repetitive DNA e.g. centromeres (sometimes called the “black holes” of the genome) are difficult to assemble and analyse. A

high-quality map can very much support this task. However, due to the EST probes we used, gene-rich regions will probably be overrepresented in our map at the expense of gene-poor, repetitive regions.

In summary, the STS-based physical map we generated is the first of this kind of a zebrafish chromosome. Our data support the assembly of the restriction fingerprint map and the sequence of this chromosome. They provide a control for empirical optimisation of parameters for the assembly processes. They help to establish chromosome 20 as a model chromosome to study genome structure and function.

5.3 Radiation hybrid mapping

5.3.1 Mapping of ESTs with specific expression patterns during embryogenesis

In this part of the work, chromosomal locations of zebrafish ESTs were determined by radiation hybrid mapping. These ESTs were selected because they showed interesting expression patterns in a systematic whole mount in situ hybridisation (WMISH) screen, or because they had a high sequence homology with human disease genes. The goal was to find candidate genes for mutations and to further characterise genes and chromosomes. Many zebrafish mutations are mapped coarsely to chromosomal regions but are not cloned yet. By generating information about expression patterns and chromosomal locations of ESTs, the gap between EST sequences (structural data) and mutant phenotypes (functional data) is bridged, providing potential candidate genes. We are therefore in the process of radiation hybrid mapping transcripts showing interesting specific expression patterns (Musa et al., in preparation see Table 11).

5.3.2 ESTs homologous to human disease genes

Model organisms are excellent tools to study the pathophysiology of diseases. The identification, expression analysis and mapping of human disease gene homologues in the zebrafish is supposed to further establish this organism as a model for the study of human diseases. In the OMIM morbid map, 1792 genetic diseases of the human are listed (ftp://ncbi.nlm.nih.gov/repository/OMIM/morbidmap). To study the functions of these genes, their action in the context of the whole organism has to be investigated experimentally, an enterprise that cannot be undertaken in the human. The zebrafish is especially suited for the study of the function of disease genes in embryonic development,

because the role of known disease genes can be tested experimentally, and new genes can be found by screening of mutant phenotypes that resemble disease conditions (Dooley and Zon, 2000).

In our group, zebrafish homologues of human disease genes were identified by BLAST searches, and expression patterns of those genes were analysed in a WMISH screen (see 1.2.4). However, sequence similarity alone is not sufficient to prove, that the genes are true orthologues. So far, only a fraction of all zebrafish genes is represented in the EST set yet. As a consequence, a paralogous gene can have the highest similarity to a human disease gene, if the true orthologue is not present. The evolution of the teleostei further complicates the issue, because there is evidence, that fish have generally more genes than mammals. It is still not clear if this is caused by a whole genome duplication, or by frequent independent duplications (Amores et al., 1998 Robinson-Rechavi et al., 2001).

There are different hypotheses concerning what might happen to a duplicated gene in evolution, when both copies stay fixed in the genome: (1) Both genes keep their function in a redundant manner. (2) One gene keeps the old function, while the other evolves into picking up a new function. (3) Both genes change their functions in a way, that the function of the ancestral gene is split between the two. This makes a direct functional comparison of fish and human genes problematic, although both copies would be true orthologues of the human gene. To further explore the degree of functional and evolutionary relationships of the human-zebrafish homologues, we investigated the expression patterns (R. Zeller, doctoral thesis, in preparation) and chromosomal locations (this work) of the transcripts in the zebrafish. By determining the expression patterns of zebrafish genes, evidence was obtained in many cases, that the gene has a functional analogy to the human disease gene. By mapping the gene, we were able to detect syntenic relationships between zebrafish and human homologues.

5.3.3 Mapping and determination of conserved synteny

Conserved synteny is an important component in identifying candidate genes (see 1.4, Karlstrom et al., 1999). In our case we used it as a test, if our zebrafish gene is an orthologue or a paralogue of the human disease gene. If a zebrafish gene maps to a chromosomal region that corresponds to the location of the human homologue, it is likely that they are true orthologues. We determined map positions by radiation hybrid mapping or by using published data (see 4.3). We found that a large portion (at least 50 %) of the

markers could be mapped to published conserved synteny or homology blocks. These are probably true orthologues (although there is a certain error probability, due to randomising effects in gene translocations).

Map positions outside the known regions of conserved synteny might indicate paralogues (genes that have been split by duplication before the split between actinopterygii and sarcopterygii 450 million years ago), which means that they are not the closest relatives (Fitch, 2000). This is a matter of concern, because it is known that in the evolution of vertebrates, duplications of genes or whole genomes have occurred. It is also possible, that zebrafish-human gene pairs are orthologues, but map to homology blocks that have not been detected yet, because the zebrafish-human comparison has only been done for a minority of the genes.

To determine the relationship of zebrafish-human homologues with a greater certainty, a careful phylogenetic analysis has to be done including all available zebrafish and human paralogues, and using a suitable outgroup, e.g. Drosophila. Because of the growing number of zebrafish genes and ESTs, this is supposed to provide an increasingly higher resolution of the phylogenetic relationships of orthologues and paralogues.

A similar search for human disease gene homologues has been done in Drosophila (Reiter et al., 2001). It is also based on a BLAST search of human disease genes against the sequence of the model organism. However, in contrast to the situation in the zebrafish, there is a complete Drosophila genome sequence. Thus, mapping of the Drosophila genes is not necessary and finding the disease gene orthologues is easier. The results of the BLAST searches have been put in a database accessible via the world wide web (http://homophila.sdsc.edu/). No whole mount in situ expression screen has been performed, but for many of the genes, published functional data are available.

5.4 Conclusion and outlook

Over the past 20 years, the zebrafish has been established as an attractive model organism for the study of vertebrate development. Milestones of this evolution were among others the introduction of the fish as a genetic system and the generation of clones (Streisinger et al., 1981), the realisation of large-scale mutagenesis screens (Haffter et al., 1996, Driever et al., 1996) and the first positional cloning of a mutant gene (Zhang et al., 1998). This has been accompanied by the development of resources like genetic and radiation hybrid maps, cDNA and genomic libraries, strain collections, databases etc. A giant leap forward

concerning the usefulness of the zebrafish will be the availability of the sequenced genome (presumably in 2003).

Model organisms are useful in basic research as well as in applied biomedical investigations. They are particularly advantageous, because they allow to manipulate and study components of life functions in the context of the whole organism (*in organismo*), which is only to a very limited degree possible in the human. This enables systematic gene-driven as well as phenotype-driven approaches, e.g. mutant screens, in situ hybridisation screens, systematic gene targeting etc.

So far, zebrafish research has had its biggest impact through the availability of mutations affecting embryogenesis. Other fields of biomedical research, like identification of quantitative trait loci (QTL) for common diseases, or the genetic control of social behaviour, are probably best studied in mammals, particularly mice and rats, but also in monkeys and apes.

After the zebrafish genome sequence will be made available, the next step will be to actually identify genes in the sequence. In the human this is not a trivial undertaking, because ESTs represent only fragments of genes, and gene prediction software tends to come up with erroneous results. This task will probably be simpler in the zebrafish, because by comparison with the human and mouse sequences as well as with other fish sequences (fugu, tetraodon) it will be possible to differentiate between well-conserved exons and less conserved introns. To support genome annotations, it also requires a large collection of full-length cDNA sequences. Expression analysis in combination with structural comparisons will reveal the organisation and function of transcriptional control elements. It is possible with DNA chip technology, to obtain expression profiles of all zebrafish tissues, all developmental stages and of all mutants. This will reveal modules of co-ordinate transcription (synexpression groups, Niehrs and Pollet, 1999) and will further give insight into the interaction of molecules and pathways. Since the transcriptome is only the (albeit easily accessible) intermediate between the genotype and the phenotype, the interaction of proteins and metabolites is the subject, which will eventually explain, “how life works”. Proteomics will get a big boost by the availability of the genome sequence, because the identification of peptides by mass spectrometry will be facilitated. As in expression analysis, systematic study of the proteome of all tissues, developmental stages and mutants is desirable.

Large-scale mutagenesis screens are still running, with increasingly specialised focus on single organs and pathways. Even when saturation is reached, i.e., when there is a mutant

for each mutable gene, the search will still go on for different alleles with different phenotypes. Cloning of mutant genes will be highly facilitated, but still be a bottleneck. Microsatellite-based mapping of mutants to genomic regions is a time consuming step, even when bulked segregant analysis is used. So there is a great demand to simplify this process. The solution of this problem would probably be to map single nucleotide polymorphisms (SNPs) by using DNA chips. Initial steps in this direction have been undertaken by the group of William Talbot at Stanford University (H. Stickney, pers. communication).

SNPs can be found by sequencing a reduced representative subset of the genome, e.g. IRS-PCR products (4.1.9). SNPs will also detected as a side product of the zebrafish genomic sequence, because the Tuebingen strain used to construct the BAC library is not totally homozygous, and therefore SNPs will be found in the regions, where clones overlap.

The heterozygosity of the zebrafish strains also poses a problem to SNP mapping, because for hybridisation on DNA chips, the mapping strains should be homozygous. For that reason, C32 and SJD lines are preferred for SNP mapping experiments. However, most mutations have been created and maintained in the rather heterozygous AB or Tübingen strains. It would take several generations to breed these mutations to isolation on a SJD or C32 background. This is why it is highly advisable to use inbred, ideally clonal lines when setting up mutagenesis screens.

The accumulation of data resulting from the systematic unravelling of gene functions has to be accompanied by further improvements of computational tools to store and connect these data. For example tools have to be developed which support data mining in different databases, that are not compatible. Work in this direction includes the development of the extensible markup language (XML), gene ontologies and the semantic web (http://www.w3.org/; http://www.geneontology.org/). With this background, it will be possible, to simulate life processes on the computer, thereby generating hypotheses that can be tested experimentally. Finally these progresses will identify a large number of potential drug target, considerably more than the ca. 500 known today. This will be the basis of a new age of drug design and personalised medicine.

6 Summary

6.1 Abstract

The zebrafish has been shown to be a powerful system for the genetic, molecular and embryological study of vertebrate development. In spite of a growing number of available genomic resources (genetic maps, radiation hybrid maps, genomic libraries, cDNA libraries) the cloning of genes disrupted in mutants is still difficult and time-consuming. The availability of physical maps of the genome anchored to the genetic and radiation hybrid maps can facilitate the identification and cloning of mutated genes and is crucial as a framework for sequencing the genome. Gene catalogues containing expression and mapping data of transcripts also help to identify candidate genes for mutations and hence have the potential to enable the cloning of the affected genes.

The subject of this thesis is the establishment and application of physical and radiation hybrid mapping techniques in the zebrafish genome. The work can be divided in three parts: In the first part, interspersed repetitive sequence (IRS)-PCR was adapted and optimised for the zebrafish system, especially for physical and radiation hybrid mapping. Different repetitive elements, primer annealing sites and PCR conditions were tested. A new zebrafish repetitive element was identified and characterised. Libraries of IRS-PCR products derived from 5 different zebrafish strains were constructed and 27,000 clones were picked. The libraries were normalised and clustered by oligonucleotide fingerprinting, suggesting 11,000 different IRS-PCR products in the library. This method also identified IRS products with +/- polymorphism among different zebrafish strains. IRS-PCR products were further characterised by sequencing. The average frequency of single nucleotide polymorphisms (SNPs) between IRS-PCR products of different strains tested was found to be $p = 0.013$. Thus IRS-PCR products can serve as a reduced representation subset of the zebrafish genome for SNP mapping and detection. Radiation hybrid mapping of IRS-PCR products was established.

Part two describes the construction of a physical framework map of zebrafish linkage group 20. A mapping procedure was set up by hybridising STS-derived oligonucleotides on a PAC library, and by hybridising IRS-products to YAC and PAC pools. So far, 344 STS derived oligonucleotide probes and 284 IRS probes have been used. A total of 3416 PACs were identified as positive, and are currently being restriction fingerprinted for integration in the Tuebingen restriction fingerprinting map and for selection of tiling paths for sequencing. A subset of the hybridisation results was used for contig construction. This

consisted of 388 probes, hitting 2007 clones, representing a marker density of 1 per 205 kb. The contig assembly resulted in 249 contigs, of which 19% were anchored by more than one probe. These data agree with theoretical predictions. Thus, an STS-based framework map of zebrafish chromosome 20 has been constructed, which is anchored to genetic and radiation hybrid maps, and integrated in the restriction fingerprint map. This is the first map of its kind of a zebrafish chromosome.

In the third part of this work zebrafish ESTs with homologies to human disease genes or with localised expression patterns were mapped on the radiation hybrid map. So far we mapped > 120 clones. The aim of this work is to identify candidate genes for zebrafish mutations and to refine the map of human:zebrafish syntenic relations; synteny data itself are used to determine if sequences homologous between zebrafish and human are true orthologues.

6.2 Zusammenfassung

Der Zebrafisch ist ein wichtiger Modellorganismus für genetische, molekulare und embryologische Untersuchungen der Embryonalentwicklung. Obwohl immer mehr genomische Ressourcen zur Verfügung stehen (genetische Karten, Bestrahlungshybridkarten, genomische Bibliotheken, cDNA Bibliotheken), ist die Klonierung von Genen, die durch einer Mutante entdeckt wurden, schwierig und zeitaufwendig. Eine physikalische Karte des Genoms, die mit genetischen und Bestrahlungshybridkarten verankert ist, kann die Identifizierung und Klonierung mutierter Gene erleichtern und dient als Grundlage für die Sequenzierung des Genoms. Genkataloge mit Expressions- und Kartierungsdaten sind entscheidend für die Auswahl von Kandidatengenen und ermöglichen daher die Identifizierung mutierter Gene.

Das Thema dieser Arbeit ist die Entwicklung und Anwendung der physikalischen Kartierung und Bestrahlungshybridkartierung im Zebrafisch. Die Arbeit gliedert sich in drei Teile: Im ersten Teil wird beschrieben, wie die auf eingestreute repetitive Elemente beruhende Polymerase-Kettenreaktion (IRS-PCR) an das Zebrafischgenom angepaßt und optimiert wurde, insbesondere für die physikalische und Bestrahlungshybridkartierung. Dazu wurden verschiedene repetitive Elemente, Primerbindestellen und PCR Bedingungen getestet. Ein neuer Zebrafischrepeat wurde gefunden und charakterisiert. Eine Markerbank aus IRS-PCR Produkten von fünf Zebrafischstämmen wurde konstruiert, und 27.000 Klone wurden gepickt. Diese Markerbanken wurden durch Oligunukleitid-Fingerprinting normalisiert, wodurch etwa 11.000 verschiedene Sequenzen identifiziert werden konnten. Dadurch wurden auch Produkte mit +/- Polymorphismus in den verschieden Stämmen identifiziert. IRS-PCR Produkte wurden durch Sequenzierung näher charakterisiert. Die Durchschnittsfrequenz von SNPs (single nucleotide polymorphisms) beträgt $p = 0.013$. Aus diesem Grund ist unsere normalisierte IRS-PCR Markerbank ein geeignetes repräsentatives Subset des Genoms für die SNP-Detektion. Die Kartierung von IRS-PCR Produkten mittels der Hybridisierung auf Bestrahlungshybride wurde ebenfalls entwickelt.

Im zweiten Teil wurde eine physikalische Rahmenkarte der Kopplungsgruppe 20 des Zebrafischs erstellt. Ein Kartierungsverfahren wurde entwickelt, in dessen Verlauf kartierte Oligonukleotide auf eine PAC-Bibliothek und IRS-Marker auf YAC- und PAC-Pools hybridisiert wurden. Bis zum jetzigen Zeitpunkt wurden 344 Oligonukleotid- und 284 IRS Sonden verwendet. Insgesamt wurden 3416 PACs getroffen, die zur Zeit durch Restriktions-fingerprinting analysiert werden. Dadurch werden sie in die Tübinger

Restriktionskarte integriert und ein Klonpfad für die genomische Sequenzierung wird konstruiert. Ein Teil der Hybridisierungsergebnisse wurde für die Konstruktion von Contigs verwendet. Dies umfaßte 388 Sonden, die 2007 Klone trafen, und damit eine Markerdichte von 1/205 kb repräsentieren. Das Contigassembly besteht aus 249 Contigs, von denen 19% durch mehr als eine Sonde verankert sind. Diese Daten stimmen mit der theoretischen Erwartung überein. Auf diese Weise wurde eine physikalische STS-Rahmenkarte des Chromosoms 20 des Zebrafischs erstellt, die mit den genetischen und Bestrahlungshybridkarten verankert und in die Restriktionsfingerprintkarte integriert ist. Dies ist die erste derartige Karte eines Zebrafischchromosoms.

Im dritten Teil dieser Arbeit wurden Zebrafisch ESTs, die humanen Krankheitsgenen sequenzhomolog sind oder spezifische Expressionsmuster zeigen, auf der Bestahlungshybridkarte positioniert. Bis jetzt wurden > 120 Klone auf diese Weise kartiert. Ziel dieses Teils der Arbeit ist die Identifizierung von Kandidatengenen für Zebrafischmutationen und die verfeinerte Kartierung der Syntänieverhältnisse zwischen Mensch und Zebrafisch. Die Syntäniedaten dienen zur Bestimmung, ob es sich bei homologen Sequenzen von Mensch und Zebrafish um Orthologe oder Paraloge handelt.