1 Introduction

The progress in sequencing the genomes of selected organisms, culminating in the complete human DNA sequence (Lander et al., 2001, Venter et al., 2001), is unquestionable a revolution in the scientific understanding of life (a status overview of sequencing projects of a number of eukaryotic organisms can be viewed under http://www.ebi.ac.uk/genomes/mot/). While genomic sequence can be used directly for answering many questions concerning number and organisation of genes and for comparison of different organisms, it will have its biggest impact if it is combined with functional data. Right now, there are only limited applications for theoretical modelling of life processes; thus analysis relies on experimental data. On a whole genome scale, data have to be collected in a systematic way, because any hypothesis driven approach has an inherent bias not suitable for the complexity of life functions. For systematic data acquisition, two complementary approaches are feasible: A gene driven and phenotype driven strategy. The gene driven analysis is based on DNA sequences, for example known genes or transcribed sequences and their protein products. Of these, the expression levels, molecular interactions, as well as the effects of gene disruptions and misexpressions are studied. A phenotype driven approach starts with a specific phenotype (mutant organism, diseased tissue) and tries to identify the process responsible for that condition.

Model organisms have a key role in systematic data acquisition for two reasons: First, because many kinds of investigations can not be performed directly in humans. Gene driven studies of human genes can only be done in a context independent way, e.g. by in vitro studies, or by expression in a different host. Phenotype driven analyses in humans are not completely systematic, because they have to rely on those phenotypes that can actually be found in human populations and are available to the researcher. In contrast to that, model organisms offer the opportunity to study the function of any gene in the context of the whole organism (e.g. by gene disruption or overexpression). Mutations can be induced randomly and systematic screens can be performed to isolate mutations in genes with specific functions (Nusslein-Volhard and Wieschaus, 1980). Secondly, model organisms are necessary for comparative evolutionary studies. Because life did not come into existence by design, but rather by a trial and error process, a real understanding of biology can only be accomplished by placing phenomena in an evolutionary framework. For example, detection of conserved sequences by comparison of different genomes is an efficient method to filter out functional regions (coding regions, regulatory elements Clark, 1999). Likewise, different species can be viewed as natural mutants from an ancestral genetic makeup, having severe (but beneficial) phenotypes.

Examples for large-scale, phenotype-driven, systematic analysis of gene functions in a vertebrate are mouse large scale random random mutagenesis screens (Hrabe de Angelis et al., 2000, Nolan et al., 2000, Justice, 2000). Complementary gene-driven strategies are followed in chemical ES cell mutagenesis, gene targeting and gene trap projects (Wiles et al., 2000, Nadeau et al., 2001), and in situ hybridisation screens in mouse (Neidhardt et al., 2000) and Xenopus (Gawantka et al., 1998).

1.1 Zebrafish as a model organism for studying vertebrate development

The Zebrafish (Danio rerio, Meyer et al., 1993), a small (3-4 cm long) tropical freshwater fish that originates from India, is well suited for a systematic cellular and genetic analysis of vertebrate embryogenesis (Streisinger et al., 1981, Driever et al., 1994; Kimmel, 1989; Kimmel and Warga, 1988). It has a short generation time (3-4 months), and mature females lay several hundred eggs at weekly intervals, which develop rapidly and synchronously outside the mother. Therefore they are inexpensive to maintain and can be bred in large numbers. Because of the optical clarity and the large size of the embryo, development of the living embryo can be observed using a standard dissecting microscope. At 12 h of development, the body axes and the overall body plan is apparent, and by 24 h all the major organ primordia are formed. Figure 1 summarises major stages of zebrafish embryonic development. Embryological techniques as known from Xenopus can also be applied in the zebrafish: Individual or groups of cells can be transplanted to new locations to test fate determination. Mutant or wild type mRNA can be injected to test the effects of ectopic expression. Individual cells can be labelled and their development followed. In contrast to that, the classical model organisms for developmental biology can be used either for sophisticated embryological manipulations (frog, chicken) or advanced genetics (fly, mouse). The zebrafish allows both approaches and therefore is a particular powerful experimental model to unravel the making of a vertebrate out of an egg.

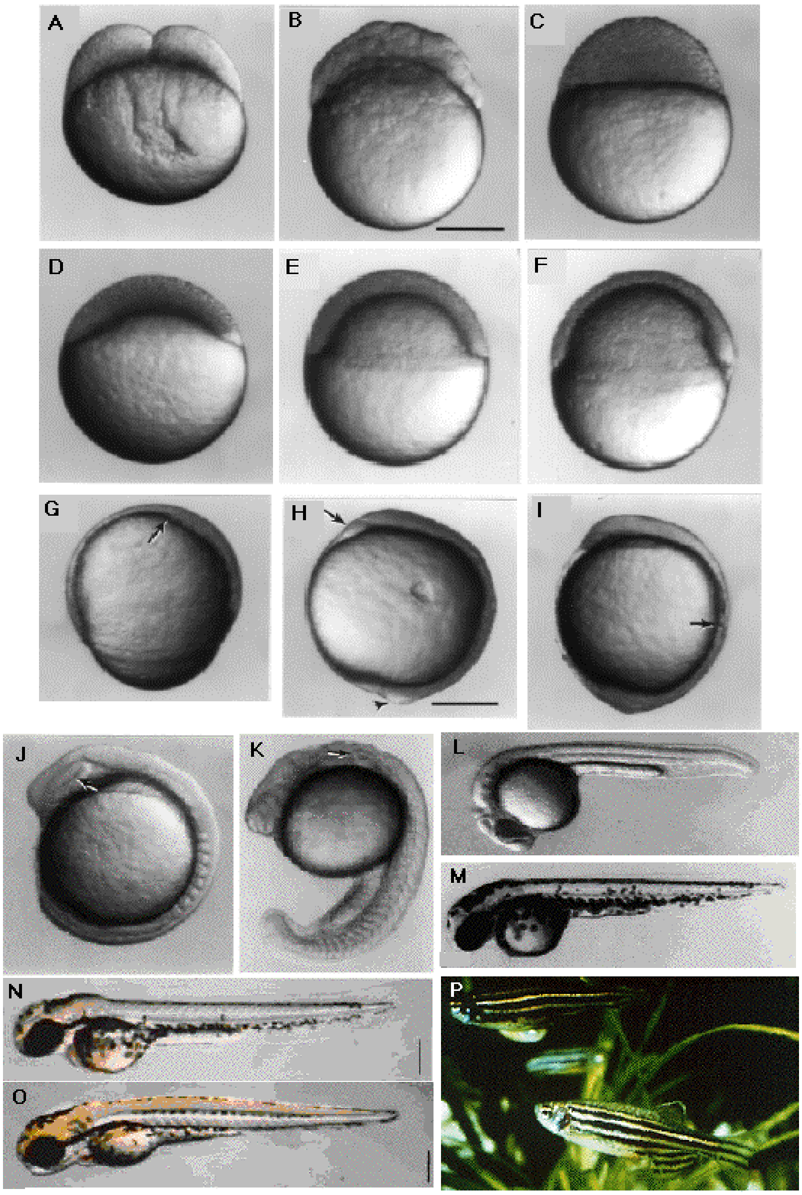

Figure 1

Zebrafish embryonic development ((Kimmel et al., 1995) adopted from (Burgtorf, 1999). A: Two-cell stage (0.75 h). B: Sixty-four cell stage (2h). C: Transition between the oblong and sphere stages (3.8h). D: Dome stage (4.3h). E: 50%-epiboly stage (5.25 h). F: Shield stage (6h). G: 70%-epiboly stage (7.7h). The arrow indicates the axial hypoblast of the prechordal plate. H: Bud stage (10h). The arrow shows the polster, and the arrowhead shows the tail bud. I: Two-somite stage (10.7h). The arrow indicates the posterior boundary of somite 2. J: Eight-somite stage (13h). The arrow indicates the optic primordium. K: Twenty-somite stage (19). The arrow indicates the otic vesicle. L: Pharyngula period (24h). M: 42h-embryo. Pigmentation by melanophores extends the whole length of the embryo. N: 48h-embryo. O: 72h-embryo. Yellow pigmentation increases owing to xanthophore development P: Adult zebrafish. Scale bars = 250 µm.

Genetics of diploid organisms like mice or zebrafish is normally carried out by breeding individual heterozygous founder animals to generate F_1 families. From those, siblings have to be crossed to drive a recessive mutation to homozygosity and reveal its phenotype. In zebrafish, however, there are tricks that circumvent this time-consuming procedure (Figure 2, Streisinger et al., 1981, Driever et al., 1994). Fertilisation of eggs with UV-inactivated sperm yields haploid embryos, which develop the basic body plan and are viable until 3 days post fertilisation. However, haploid embryos have some defects including short tail, deformed notochord and edema. Therefore, haploid embryos allow screening for mutations affecting the fundamental body structure (which can be seen in half of the haploid progeny of a heterozygous female) but are not suited for detailed investigations of more subtle changes. In fish, the second meiotic division of the egg occurs after fertilisation. Treatment with hydrostatic pressure ("early pressure") inhibits this process, resulting in a partially homozygous embryo. This provides a simple method to map mutations relative to the centromere, because the likelihood to recover an individual homozygous for a mutant allele gets lower, the further it is from the centromere (Streisinger et al., 1986, Johnson et al., 1995, Johnson et al., 1996). In analogy to tetrad analysis in fungal genetics this method is called half-tetrad analysis. Alternatively, administration of hydrostatic pressure or heat shock during the first embryonic cell cycle blocks the first mitotic cleavage and produces completely homozygous offspring, since both chromosome sets derive from a haploid one. Using this technique, it has been possible to create clonal lines of fish, which are especially useful for genetic mapping (Streisinger et al., 1981, Nechiporuk et al., 1999, Kelly et al., 2000).

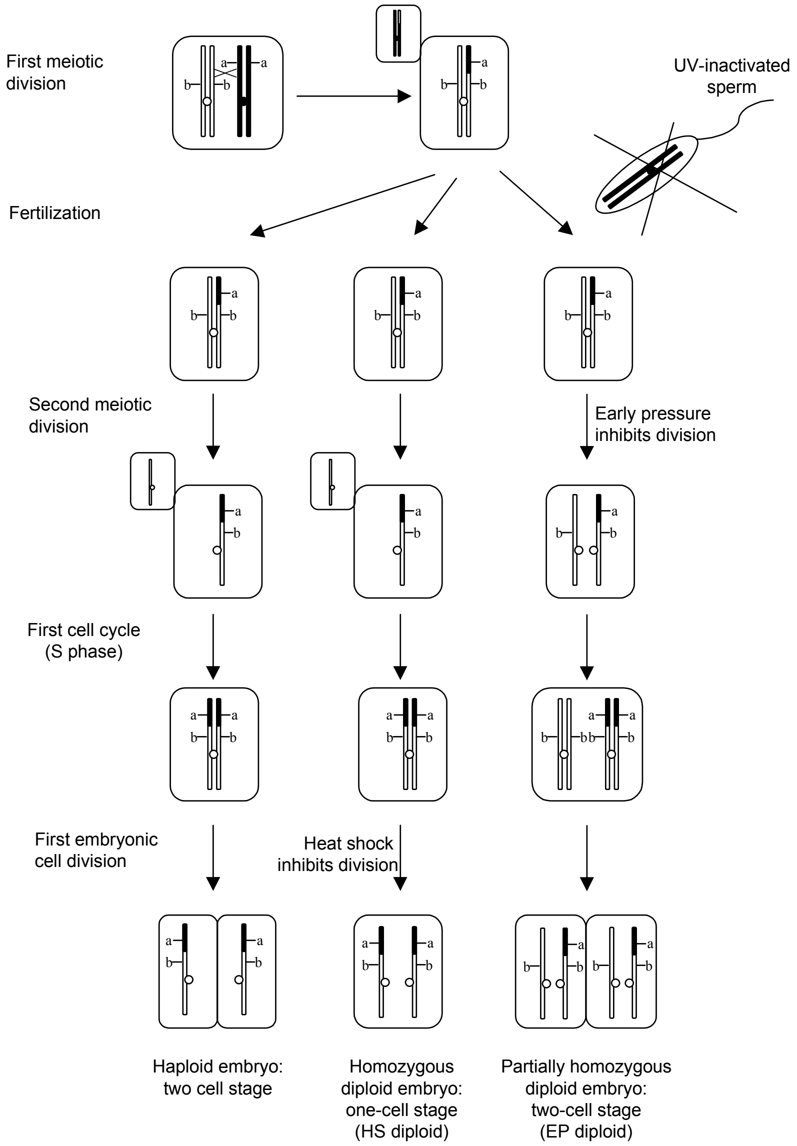
First meiotic division
UV-inactivated sperm
Fertilization
Second meiotic division
Early pressure inhibits division
First cell cycle (S phase)
First embryonic cell division
Heat shock inhibits division
a
b
Haploid embryo: two cell stage
Homozygous diploid embryo: one-cell stage (HS diploid)
Partially homozygous diploid embryo: two-cell stage (EP diploid)

Figure 2

Generation of haploid and gynogenetic diploid fish (adapted from (Driever et al., 1994)). The segregation of two mutant alleles, (a) from the paternal homologue and (b) from the maternal homologue maternal is illustrated. Maternal chromatides are shown as open rectangles, paternal chromatides by filled rectangles. UV-sterilised sperm induces development of a haploid embryo. This can be transformed to a gynogenetic diploid animal by various experimental treatments. Application of hydrostatic pressure inhibits the second meiotic division, which results in partially gynogenetic embryos. Haploid embryos can by made diploid by inhibition of the first embryonic cell division using heat shock or hydrostatic pressure. This results in completely homozygous diploid genomes.

Saturation mutagenesis and genetic screens on a large scale have the potential to discover all possible mutations in an organism. Such screens have been used to identify genes which define developmental pathways in a number of invertebrate and plant model organisms like *Drosophila melanogaster* (Nusslein-Volhard and Wieschaus, 1980, Nusslein-Volhard et al., 1984) *Caenorhabditis elegans* (Brenner, 1974, Hirsh and Vanderslice, 1976) and *Arabidopsis thaliana* (Mayer et al., 1991). Since small-scale mutagenesis screens on vertebrate development using the zebrafish have proven to be successful (Grunwald and Streisinger, 1992, Kimmel et al., 1989, Solnica-Krezel et al., 1994, Mullins et al., 1994), large-scale saturation mutagenesis screens have been set up. Such screens are initiated by exposing a male fish to the mutagenic agent ethylnitrosourea (ENU), which alkylates nucleotide bases and hence changes their base-pairing properties. Consequently, point mutations are introduced in the DNA of the spermatogonia. In order to obtain embryos being homozygous for a recessive mutation with a zygotic effect, screens are designed as two-generation breeding screens with analysis of mutant phenotypes in F_3 embryos (Figure 3).

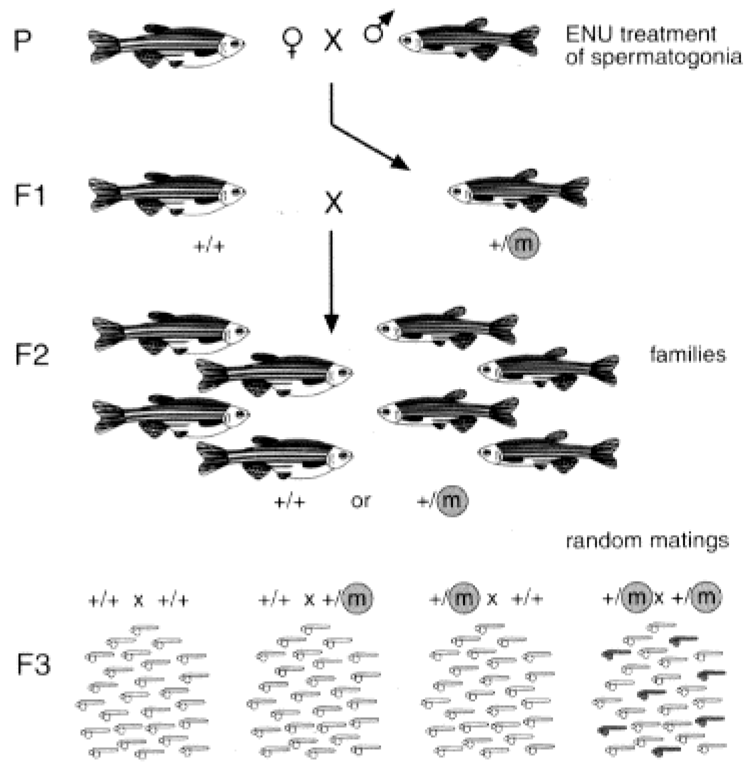

Figure 3

Outline of the three-generation breeding scheme to produce homozygous mutant embryos (from Haffter et al., 1996). Male individuals are mutagenised using ENU and mated to wild type females. In the F_1 offspring each individual is heterozygous for a different mutagenised genome. F_2 progeny are raised from single pair matings. A mutation m present in one of the F_1 parents is inherited by 50% of the F_2 fish. In 25% of random matings between F_2 siblings both parents are heterozygous and 25% of their offspring is homozygous for the mutation showing the mutant phenotype.

Mutations showing a similar phenotypic class are tested if they are different alleles of the same gene by complementation analysis. For that purpose, heterozygous F3 carrier fish are identified by sibling matings and used for complementation crosses with other mutants. If two heterozygous individuals are carriers of different alleles of the same gene, a mutant phenotype is found in their progeny. If they carry different mutated genes, their progeny is heterozygous for the wild type alleles and no mutant phenotype can be found.

The results of two large-scale screens carried out in Tübingen and Boston have been published 1996 in an issue of the journal "Development" solely dedicated to this subject (Driever et al., 1996, Haffter et al., 1996 and accompanying papers. For reviews see Eisen, 1996, Currie, 1996). Altogether 6197 genomes were screened (which essentially corresponds to the number of F2 families crossed). F3 embryos were examined for a variety of morphological features. 6647 mutant phenotypes were detected, two thirds of which fall into four major categories: general necrosis, brain necrosis, general edema or developmental retardation, and were not further taken into consideration. 1858 mutant lines that showed distinct and specific phenotypes were further characterised. At the time of publication 1225 mutations were tested for allelism and identified as 592 genes.

The advantage of using a chemical mutagen inducing point mutations is twofold: First, mutation frequency is high and mutations are randomly distributed over the chromosomes, whereas mobile DNA elements are biased towards certain chromatin structures. Secondly, allelic series of a gene can be generated, ranging from very mild alleles caused by an amino acid exchange which only weakly hampers the activity of the gene, to 5' nonsense mutations which totally abolish gene expression. This facilitates the study of gene function. The disadvantage of this approach lies in the fact, that it is not possible to directly isolate the mutation on a molecular level and clone the gene. For this purpose, more indirect methods like positional cloning and candidate gene approaches have to be applied (as discussed in 1.4).

To circumvent the problems associated with cloning of point-mutated genes, alternative mutagenesis methods have been used. Most importantly, insertional mutagenesis using retroviral vectors has been applied in a modified two-generation breeding scheme (Gaiano et al., 1996, Amsterdam et al., 1999). The overall efficiency of mutagenesis is lower compared to ENU treatment, and it is not known if the integration sites are randomly distributed over the genome. It is possible, that certain structures in the genome are resistant to retroviral insertion, and as a result, some genes can not be mutagenised. Nevertheless, the ease with which a mutant gene can be cloned by using the inserted retrovirus as a tag compensates for these disadvantages. Currently, a large-scale insertional mutagenesis screen underway at the Massachusetts Institute of Technology aims to isolate 600-1200 embryonic lethal mutations by mid of 2001.

The widespread use of the P-element transposon in *Drosophila melanogaster* for mutating genes and transferring foreign DNA to the genome attracted interest in transposons as molecular biological research tools in zebrafish. Transposons of the Tc1 family of Caenorhabditis elegans have been found in zebrafish and evidence for active transposition has been collected (Lam et al., 1996, Gottgens et al., 1999). Heterologous transposable elements from *C. elegans* and *D. melanogaster* integrate in the zebrafish germline (Raz et al., 1998, Fadool et al., 1998) and are capable of being mobilised upon introduction of transposase protein *in trans*. A remarkable experiment was done in the group of P. Hacket (Ivics et al., 1997): A consensus sequence of a transposase gene of the salmonid subfamily of TC1-like transposons was engineered by eliminating the inactivating mutations in the sequence of a silent transposon. The transposase gene provided *in trans* mediates transposition in a cut-and-paste mechanism. However, the suitability of transposon technology for mutagenesis screens in the zebrafish as still to be proven.

Ionising radiation can induce a variety of chromosomal changes in *Drosophila*, including point mutations, deletions ranging in size from a few base pairs to megabases, inversions, and translocations. Gamma rays have been used to induce chromosomal deletions in the zebrafish, resulting in visible mutant phenotypes (Fritz et al., 1996). Deletion sites were detected by scanning the genome using multiplex PCR. Only a minority of the deletions characterised turned out to be small enough to disrupt the function of a single gene. Hence, the method is rather crude and chemical mutagenesis may be better suited to target the function of single genes

1.2 Mapping and sequencing of complex genomes

One of the major obstacles in the analysis of vertebrate genomes is their large size and their highly repetitive nature. The haploid DNA content of a zebrafish cell is approximately 1.8 pg, equivalent to ca 1.7 x 10^9 bp. The zebrafish genome is therefore ca. 10 times bigger than the Drosophila genome (Hinegardner, 1968). It has about half the size of the human genome but more genes, because in the teleosts, duplications of genes or possibly of the whole genome have occurred (Wittbrodt et al., 1998, Robinson-Rechavi et al., 2001). This suggests, that the zebrafish genome has a higher gene density than the mammalian genome, and therefore a possibly less repeat sequences. The zebrafish genome is organised in 25 chromosomes (Endo and Ingalls, 1968).

1.2.1 Genetic linkage mapping

Genetic linkage is based on the fact that two or more loci (genes or traits) are located on the same chromosome, and therefore physically linked. This becomes obvious, when transmission of the two loci marked by specific alleles tend to be inherited together more often than not, violating Mendel's law of independent assortment. Linkage of genes can be broken up by meiotic crossing over, resulting in recombinant or nonparental chromosomes, with the frequency of recombination of two loci depending on the distance between them. T.H. Morgan and A. Sturtevant recognised that this correlation could be used to position genes according to their chromosomal location, creating the first genetic map in *Drosophila* (Sturtevant, 1914). The unit of genetic distance between two loci is usually the centiMorgan (cM), which represents 1% recombination. Classically, linkage maps have been constructed in well-studied organisms like bacteria, yeast and the fruit fly, where a lot of visible mutations were available as genetic markers. With the establishment of recombinant DNA techniques, molecular marker systems based on DNA polymorphisms have been developed. These have the advantage, that (dependent on the method used) high marker density is achievable, polymorphism rate (and thus information content of a locus) is high, and in most systems alleles are codominant, i.e. both alleles are equally detectable in heterozygotes. Moreover, DNA markers are applicable in any organism. Restriction fragment length polymorphisms (RFLP, Botstein et al., 1980) was the first DNA marker system used for successful linkage of a human disease gene (Huntington disease, Gusella et al., 1983), eventually resulting in the cloning of the gene (HDCRG, 1993). Subsequently, PCR based techniques became widely used, most prominently length polymorphisms of microsatellite markers (e.g. CA repeats, Weber and May, 1989), also called simple sequence repeats (SSR) or simple sequence length polymorphisms (SSLP). Using microsatellites, a comprehensive genetic linkage map of the human genome was constructed. It consisted of more than 5000 markers with an average spacing of 1.6 cM (Dib et al., 1996).

Concurrent with the isolation of zebrafish mutants, projects to construct genetic maps of this organism have been initiated. The first map published was based on randomly amplified polymorphic DNA (RAPD, Williams et al., 1990, Johnson et al., 1994, Postlethwait et al., 1994). This uses single decamer primers in a PCR, which amplify several bands from genomic DNA. Mutations at the primer binding sites or insertions and deletions between primer binding sites can create recessive, or less frequently, codominant genetic variants. To circumvent the disadvantages linked to recessive alleles, haploid embryos were generated and used for genotyping. RAPD-based mapping does not require any sequence information or extensive primer synthesis, and is thus a quick and cheap method for building a genetic map in organisms where haploid stages occur. However, because microsatellite based markers are codominant, highly polymorphic, transferable between different strains and yield more reliable genotyping results, they have essentially replaced RAPD markers for linkage mapping. In the group of Mark Fishman at the Massachusetts General Hospital, an ongoing mapping project has been started to place SSLP markers on the zebrafish genome (Knapik et al., 1996, Knapik et al., 1998, Shimoda et al., 1999). To date, more than 3100 markers have been placed on the map, providing an average resolution of 1.2 cM (see http://zebrafish.mgh.harvard.edu/)

To directly position genes on the map that are normally not highly polymorphic between strains, PCR primers are designed from 3' untranslated regions (3'UTRs), which are less conserved than coding sequences and therefore more likely to contain a polymorphic site. PCR products are denatured to separate the complementary strands and single stranded DNA fragments are run on a gel under nondenaturing conditions. Due to alternative secondary structures, polymorphic DNA sequences can be distinguished on the gel. Single strand conformation polymorphism (SSCP) based mapping has been performed on haploid mapping panels (Postlethwait et al., 1994, Fornzler et al., 1998, Gates et al., 1999). Subsequently, a gynogenetic homozygous diploid panel was created using heat shock (Figure 2), which could be grown to adulthood, providing a much larger amount of DNA compared to haploid embryos. This panel is used in an ongoing SSCP based mapping project Stanford University (Kelly et al., 2000, Woods et al., 2000). To date more than 1500 genes and expressed sequence tags (ESTs) have been mapped on this panel.

Genetic maps are the entry point for identifying mutated genes in a positional cloning approach, and are therefore crucial for every kind of molecular genetic analysis. However, genetic mapping has its inherent limitations. First, sites of recombinatorial hot spots and regions of low recombination frequency cause genetic maps to be linearly distorted compared to the actual physical distance of markers. Generally, telomeric regions have a higher recombination frequency than centromeric regions. Thus, some segments of a chromosome might not be resolved using genetic mapping. Genetic marker systems have their inherent limitations: Transcripts and genes (sometimes called type I markers, Miller, 1997) have only a low degree of polymorphism, but are conserved across different species, allowing the construction of synteny maps. Anonymous DNA markers (like microsatellites) on the other hand, are highly polymorphic sequences which are valuable in construction of a genetic map, but are rarely conserved between species and not useful in interspecies comparisons.

Bulked segregant analysis is a fast and economic method to link a mutant phenotype to a marker of known chromosomal position. For this, crosses of carrier and wildtype individuals are set up, and pools of mutant progeny and their unaffected siblings are collected (Figure 4, Michelmore et al., 1991, Churchill et al., 1993). Due to recombination, mutant offspring have an identical genotype in the region of the mutated gene, but random genotypes at loci unlinked to that region. By analysing the pools using codominant markers (e.g. SSLPs), it can be determined if there is recombination between a marker and the gene of interest. This method has been used to place mutations found in the Tübingen screen on the genetic map (Rauch et al., 1997) http://www.eb.tuebingen.mpg.de/abt.3/research_interests/geisler_lab/gen_mapping.html

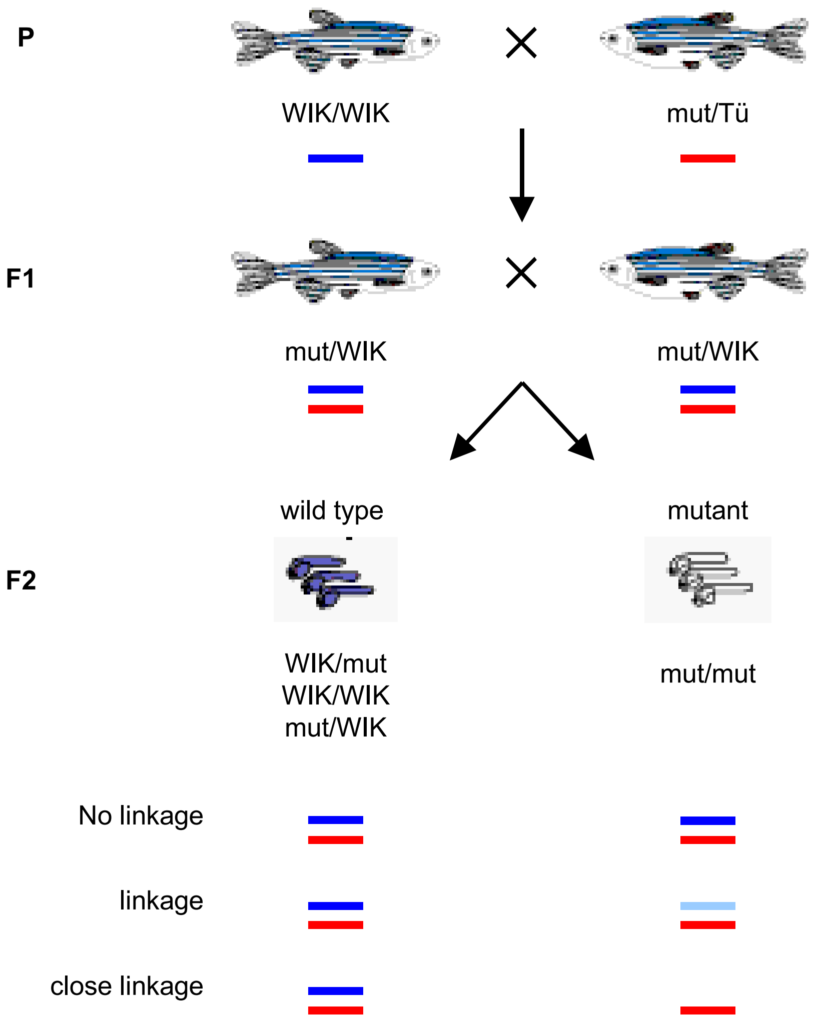

Figure 4

Bulked segregant analysis of zebrafish mutants

1.2.2 Radiation hybrid mapping

Radiation hybrid (RH) mapping is a method of ordering sequences according to their chromosomal position analogous to conventional genetic mapping, in which the background is a different species. To generate a radiation hybrid mapping panel, a donor cell line (e.g. human fibroblast) is exposed to γ-rays, which fragments their DNA in a dose dependent manner. Irradiated cells are fused to recipient cells (usually a rodent cell line). The hybrid cells retain fragments of the donor genome, usually in the order of 15-50%. A 96-well microtitreplate containing DNA of different hybrid cell lines and control DNA is then screened for sequence markers, usually by PCR.

The existence of the marker is scored and the result (the radiation hybrid vector) is compared by a computer program to the radiation hybrid vectors of other markers. Identical or similar radiation hybrid vectors of two markers indicate linkage. Analogous to genetic mapping, the measure for genomic distances between two markers is the breakpoint frequency occurring between the markers (1 centiRay (cR) = 1% of breakage occurrence for a specific radiation dose). Radiation hybrid panels were initially created from single chromosomes (Goss and Harris, 1975, Cox et al., 1990), later radiation hybrid panels of whole genomes were generated (Walter et al., 1994, McCarthy, 1996, McCarthy et al., 1997). Compared to genetic linkage mapping, the radiation hybrid based technique has the advantage, that sites with a low level of polymorphisms between different strains, e. g. genes, can easily be mapped using PCR. Additionally, radiation break frequency is solely dependent on radiation dose, unbiased towards genomic regions, and can be very high (20 times as many breaks as there would be crossovers in a similar number of mice, Schalkwyk et al., 1998) resulting in a higher resolution. Thus, radiation hybrid mapping is the method of choice to quickly position a large number of transcribed sequences in a genome. In human, an international consortium has mapped to date more than 30,000 genes using this method (Deloukas et al., 1998). The map can be viewed and search at the URL http://www.ncbi.nlm.nih.gov/genemap99/. However, not every genomic region might be propagated in radiation hybrids, resulting in gaps in the map. Likewise, genetic and RH maps cannot be easily superimposed, because the respective regions might be distorted or inverted relative to each other.

Panel	Retention	Genome size	Physical size	Resolution	Reference

T51	18.4%	27729 cR	61 kb/cR	350 kb	Geisler et al., 1999
LN54	22%	13286 cR	128kb/cR	500 kb	Hukriede et al., 1999 Yi Zhou (pers. comm.)

Table 1

Comparison of the zebrafish radiation hybrid panels available

Two radiation hybrid panels have been generated from the Zebrafish (Kwok et al., 1998, Geisler et al., 1999, Hukriede et al., 1999). At the time of June 2001, the LN54 has now ca. 7400 markers mapped on it (SSLP markers, ESTs and genes), while 7800 markers are positioned on the T51 panel. Although algorithms have been developed to integrate the maps from different radiation hybrid panels from one organism (Agarwala et al., 2000), this has not been done on the zebrafish panels.

1.2.3 Physical mapping

Two basic concepts of physical genome maps can be distinguished: On the one hand, there are clone-based maps, which are basically sets of overlapping clones (contigs, Staden, 1980). Otherwise, a physical map can be conceived as a collection of genetic landmarks, ordered according to their physical location on the genome. Such landmarks can be for example recognition sites for restriction endonucleases (in the case of complex genomes these have to be rare cutters, e.g. *Not*I). In 1989, short DNA sequences, called sequence tagged sites (STS), which are identifiable by PCR, were proposed as a “common language for physical mapping” (Olson et al., 1989) and became subsequently the most widely used marker system.

Clone overlaps (contigs) can be identified by PCR assays (STS-content mapping), by hybridisation based techniques, or by restriction fingerprinting. Ideally, a contig covers an entire chromosome without gaps. The chromosomal location of a contig can be determined by anchoring it to genetic (if the STS is a polymorphic marker) or radiation hybrid maps. PCR-based STS-content mapping is a highly robust and straightforward technique, and it is not dependent on the physical availability of a certain DNA marker, facilitating merging of data from different laboratories. It can be applied to all kinds of cloning systems and mapping panels. On the other hand it demands primer design for each marker studied, thus increasing costs of a project. PCR has to be done on every clone of a library, followed by detection of products, a task, which gets impracticable, if the number of markers and the number of clones are large. To obviate this problem, clones can be collected in pools reducing the number of PCR reactions to be done. Pools are set up according to a sophisticated pooling scheme, which allows the identification of a positive clone by deconvoluting the results of plate- row- and column pools (Green and Olson, 1990, Bruno et al., 1995, Hunter, 1997). A pooling scheme, which was used during this work, is delineated in Figure 24. Large-scale, STS-based maps of the human (Chumakov et al., 1992, Hudson et al., 1995) and the mouse (Nusbaum et al., 1999) have been constructed, integrating genetic, RH- and physical maps. These maps serve as valuable resources for sequencing the genomes.

As an alternative to PCR, genomic libraries spotted on high-density gridded filters are screened by hybridisation, which allows screening of a high genome coverage by one single hybridisation. The most commonly used probes used for hybridisation are subcloned genomic fragments, end sequences of genomic clones or inter-repeat sequences. (Mozo et al., 1999, Hildmann et al., 1999). Reduction of the number of hybridisation targets by pooling can cause problems due to the high complexity of the target and the presence of repetitive sequences. Instead, hybridisation of pooled short oligonucleotide probes, followed by deconvolution of results and attribution of positive clones to a single probe has been shown to be feasible, reducing the numbers of hybridisations (Cai et al., 1998, Klysik et al., 1999, Asakawa and Shimizu, 1998, Han et al., 2000). Hybridisation-based physical maps have been constructed from the genome of Schizosaccharomyces pombe (Hoheisel et al., 1993) and of human chromosomes 21 (Hattori et al., 2000) and X (Roest Crollius et al., 1996).

A third, clone based approach is restriction fingerprinting (Olson et al., 1989, Coulson et al., 1986). Large insert clones are cut by a restriction enzyme and run on a gel, resulting in a pattern of small fragments characteristic of each clone (i.e. a fingerprint of each clone). Overlapping clones are identified by comparison of restriction fingerprints. Software tools have been developed for reading and handling restriction fingerprint data and for assembling the clone overlaps (Staden, 1980, Sulston et al., 1988), which makes large-scale projects possible. An advantage of this technique is that the range of overlap between two clones can be easily determined. Hence, clones can be selected that cover a genomic region with a minimum of overlaps, establishing a minimal tiling path, and reducing the amount of clones to be sequenced. However, contigs generated by restriction fingerprinting can not be anchored to chromosomal locations directly. To achieve this, clone ends have to be sequenced and mapped to radiation hybrid maps, or STS content maps of the same library have to be generated. Clone contigs generated by restriction fingerprinting have been the basis for genomic sequencing of Caenorhabditis elegans (1998), Arabidopsis thaliana

(Marra et al., 1999) and human (McPherson et al., 2001) and also contributed to the sequencing of the Drosophila melanogaster genome (Adams et al., 2000, Hoskins et al., 2000). In the zebrafish, a physical map is constructed by restriction fragment fingerprinting of 100,000 BAC clones (10x genome coverage). This is done in a collaboration of G.-J. Rauch and C. Nüsslein-Volhard (Max-Planck-Institute for Developmental Biology) and R. Plasterk (Netherlands Institute for Developmental Biology; Utrecht, the Netherlands).

During the progression of various mapping projects, the theoretical background of physical map construction has been worked out. This resulted in algorithms and software tools for manipulation of experimental data and for ordering of clones to contigs, as well as for dealing with clone chimerism and experimental noise (Lander and Waterman, 1988, Mott et al., 1993, Grigoriev, 1993, Grigoriev et al., 1994, Roach, 1995).

Table 2 shows the most important vectors used to create large-insert libraries. The choice of the vector depends on the purpose the library is made for. YACs can carry inserts lager than 1 Mb, but are difficult to isolate and are often chimeric. Conversely, PACs and BACs are restricted in their size to up to 300 kb, but can easily be isolated by conventional plasmid-minipreps. Thus, YACs are suited to construct a map covering a large region without using too many probes, while bacterial cloning systems are preferred for fine mapping and construction of sequence-ready maps.

	YAC	BAC	PAC
Host cells	*Saccharomyces cerevisiae* AB1380, J57D	*Escherichia coli* DH10B	*Escherichia coli* DH10B
Transformation	Spheroblast transformation	Electroporation	Electroporation
DNA topology of recombinants	Linear	Circular, supercoiled	Circular, supercoiled
Maximum insert size	>1Mb	~300 kb	~300 kb
Selection for recombinants	Ade2 *sup*F red-white colour selection	LacZ blue-white colour selection	SacIIB selective growth
Selection for vector	Media lacking tryptophan and uracil	Chloramphenicol	Kanamycin
Enzyme for partial digestions	*Eco*RI	*Hin*dIII	*Mbo*I or *Sau*3A
Stability	Varies from clone to clone, but can be very unstable	Very stable	Very stable
Degree of chimerism	Varies from library to library but can be higher than 50% of clones in a library	Very low	Very low

Purification of intact insert	Relatively difficult	Easy	Easy
Direct sequencing of insert	Difficult	Relatively easy	Relatively easy
Reference	(Burke et al., 1987)	(Shizuya et al., 1992)	(Ioannou et al., 1994)

Table 2

Comparison of different large-insert cloning systems (adapted from (Amemiya et al., 1999)). Derivatives of these vectors exists which have altered features.

For the work presented here, three genomic libraries were used. Clones, pools and filters of these libraries are distributed by the Resource Centre of the German Genome Project (RZPD http://www.rzpd.de/).

Library name	Library type	Number of clones	Average insert size	Genome coverage	Reference
MGH_y932	YAC	19,000	470 kb	5.2x	(Zhong et al., 1998)
HACHy914	YAC	34,560	240 kb	4.8x	(Amemiya et al., 1999)
BUSMP706	PAC	104,064	115 kb	7x	(Amemiya and Zon, 1999)

Table 3

Large-insert genomic libraries available at the Resource Centre of the German Genome Project (RZPD http://www.rzpd.de/), which were used during this study.

1.2.4 Sequencing strategies for complex genomes

So far, two strategies have been applied for the sequencing of large eukaryotic genomes: Clone-by-clone shotgun sequencing and whole genome shotgun sequencing (Green, 2001). Clone-by-clone shotgun sequencing (also called hierarchical shotgun sequencing) is the original and so far the most commonly used strategy for sequencing large eukaryotic genomes. It has been applied for sequencing yeast (Mewes et al., 1997), *C. elegans* (1998), *Arabidopsis thaliana* (2000) and by the publicly funded Human Genome Project (Lander et al., 2001). First, a physical map based on large-insert clones is constructed of the target genome, whereupon individual clones that span the region of interest are selected and subjected to shotgun sequencing. In the finishing phase, gaps are closed and misassemblies are resolved using the appropriate techniques. The big advantage of this method lies in its modularity. Problems in the assembly and finishing step can be dealt with on the level of the single clone, and therefore, no long range disturbing factors from other regions of the genome (repeats, segmental duplications etc.) can interfere with the process. Moreover, problematic regions can be isolated and targeted in the finishing phase. These benefits have to be paid for with the initial effort of constructing a high quality, clone based physical map containing a minimal number of gaps, before sequencing can be started. In practise it has been shown that, once the maps are available, hierarchical shotgun sequencing is a robust technique which yields clean data, as was exemplified with the sequences by first human chromosomes being finished - 21 (Hattori et al., 2000) and 22 (Dunham et al., 1999).

As an alternative to the clone based approach, a whole genome shotgun (WGS) strategy has been proposed (Weber and Myers, 1997), and became the subject of a controversial debate (Green, 1997). In WGS, the entire genome of an organism is fragmented in small sizes and subcloned in suitable plasmid vectors of different size classes. Sequence reads are generated from both insert ends and assembled without the use of a clone-mapping information. The feasibility of this method in bacteria has been proven with the first complete bacterial genome sequenced (*Haemophilus influenzae* Fleischmann et al., 1995), and it is now routinely used in sequencing archea and eubacteria (http://www.tigr.org/tdb/mdb/mdbcomplete.html). Its application to large, repeat rich eukaryotic genomes, however, faces enormous problems. Most importantly, repetitive sequences and segmental duplications can cause links between remote sequences, resulting in forks and loops in the assembled sequence. Finishing without a backbone of mapped clones might be impossible to achieve at reasonable costs. To deal with some of these problems, libraries of several size classes of subclones, as well as BAC ends are used. Read pairs (“mates”) of clones with long inserts are used to span gaps and to form scaffolds of linked sequences. In the final stage, available STS based maps are used to order and orient the scaffolds on the chromosomes. Whole genome shotgun sequencing was first applied for the Drosophila genome, albeit only as a component of a hybrid strategy which also included information from clone maps (Adams et al., 2000). The sequence of the human genome produced by the private company Celera using WGS also took advantage of data from the publicly funded Human Genome Project, which it was competing with (Venter et al., 2001). The advantages of whole-genome-shotgun sequencing include the ability to initiate the sequencing phase without an existing map. Thus a large amount of sequencing data is produced early on, which can be used for homology search and for polymorphism (SNP) detection.

Besides the above approaches, additional schemes to sequence complex genomes have been proposed and modelled: Random sequencing of BAC clones (Wendl et al., 2001); random sequencing of BAC clones combined with BAC end sequencing to select clones with a minimal overlap with an already sequenced clone (Venter et al., 1996); a parking strategy, i.e. an iterative sampling-without-replacement scheme, where BACs which do not overlap with BACs already sequenced, are selected randomly for sequencing (Roach et al., 2000).

To combine the advantages of the clone-by-clone shotgun and the whole genome shotgun strategies, zebrafish, mouse and rat will be sequenced using a hybrid strategy. The zebrafish sequencing project at the Sanger Centre http://www.sanger.ac.uk/Projects/D_rerio/ is now starting with a low coverage whole genome shotgun and random BAC sequencing (Rogers J., Coulson A., pers. communication). At a later stage, when the restriction fragment fingerprint map will be finished, it will be switched to the mapped clones.

1.3 High throughput characterisation of the zebrafish transcriptome by gene catalogues and whole mount in situ screens

A large scale zebrafish EST project is carried out in our group in collaboration with Steve Johnson, Washington University School of Medicine, St. Louis, MO, USA (Clark *et al.* in press). This involves the creation of the libraries, picking, normalisation by oligonucleotide fingerprinting, and single-pass sequencing. By clustering the sequences a "unigene set" i.e. a nonredudant catalogue of transcripts or genes is obtained. Similar project carried out in our group target sea urchin and amphioxus, both basic representatives of the deuterostome and chordate lineage, respectively (Panopoulou *et al.*, manuscript in preparation, Poustka et al., 1999).

A whole mount in situ hybridisation (WMISH) screen of zebrafish embryos was set up for high throughput functional characterisation of the cloned transcripts. This screen follows two different approaches: The main part is based on systematic, unbiased use of normalised cDNA libraries, with the final goal to characterise the expression patterns of all zebrafish transcripts. From the 1536 clones selected so far, 20 % of the clones chosen showed an expression that was spatially or temporally restricted. The rest showed ubiquitous or undetectable expression (Musa *et al.*, manuscript in preparation). A restricted expression pattern indicates a specific function of the gene examined in embryonic development. It also suggests its potential use as a cell- and tissue-specific marker.

The second approach aims at the characterisation of zebrafish orthologues of human disease genes. The goal of this project is to evaluate the zebrafish as a model for the study of human diseases. For this purpose, zebrafish orthologues of human genes involved in diseases are identified, used for in situ hybridisation and RH mapped. Expression analysis gives evidence, if the gene might have an analogous function in the zebrafish. Mapping and synteny comparison, together with phylogenetic analysis, shows if the gene is a true orthologue. For a discussion of the terms analogy, homology, orthology and paralogy see Fitch, 2000.

A set of 922 human genes involved in human diseases was selected from the GeneCards database (http://www.rzpd.de/cards/index.html). These genes were used as queries in a BLAST database search against zebrafish ESTs. Sequences homologous with an expectation value of $p \leq 10^{10}$ were further selected for a similarity of ≥ 70% (amino acid level). By this procedure, 288 clones were selected and picked. To account for mistakes in linking clones and sequences (which happens to a certain extend in the large-scale sequencing performed at Washington University), the picked clones were resequenced before they were used for in situ hybridisation. Ca. 26% of clones showed localised expression during the first 24 hours of development. This number is higher than in the systematic, random in situ screen, because the clones were selected on the basis of highly specific phenotypes (in humans). The fact, that not more clones are localised may be due to the high number of metabolic enzymes included in the clone set, which are usually ubiquitously expressed (R. Zeller, doctoral thesis, in preparation). During this Ph.D. work, a procedure was set up to determine the chromosomal location of clones used in whole mount in situ screens (see 4.3).

1.4 Cloning of genes identified in mutants

The goal of mutagenesis screens is to define the specific functions of mutated genes. The next step towards understanding the gene activity is to clone the gene. If mutated genes carry an easily identifiable tag (i.e. by insertion of a retrovirus), this is a rather trivial task. Chemically induced mutations, however, can not be found directly. Instead, they have to be identified by testing of candidate genes, or by positional cloning. There are several criteria to select genes as good candidates Candidate genes are selected using several criteria: The mutant resembles a mutant from a different species, where the mutated gene is already known (*kreisler*/*valentino* Moens et al., 1998). The expression pattern of a known transcript coincides with the tissue affected by the mutation (*casanova* Dickmeis et al., 2001 Kikuchi et al., 2001). The mutant gene is positioned in a chromosomal region syntenic to a well-characterised region in a different species (*you-too,* Karlstrom et al., 1999).

To find out, if a candidate gene is the cause of the mutant phenotype, several criteria have to be met: (1) A disruptive mutation in the gene or its regulatory region is linked to the mutant phenotype. (2) The wild type gene, when injected, can rescue the wild type phenotype, while the mutant gene can not. (3) A disruption of the gene function, e.g. by injection of an antisense construct (morpholino, (Ekker, 2000)), phenocopies the mutation. However, tests 2 and 3 work only for recessive, loss-of-function mutations and can not exclude the possibility, that a factor acting downstream of the mutant gene is tested. Therefore, test 1 is necessary.

Positional cloning, in contrast, is an unbiased method not requiring any *a priori* hypotheses about the nature of the gene. The procedure includes mapping of mutations to a chromosomal region, construction of contigs of large insert clones covering that region (chromosomal walking), and identification of genes in that region. These genes are tested in an equivalent manner as in the candidate gene approach, including fine mapping, expression analysis, mutation detection etc. A tutorial of positional cloning in the zebrafish can be found in http://134.174.23.167/zonrhmapper/positionCloningGuide.htm. Because positional cloning is tedious, and in some cases does not give a result, it is usually exerted, when all candidate genes have shown to be wrong. That is why there is only a limited amount of genes identified by a pure positional cloning strategy in the zebrafish, the first of which being *one-eyed pinhead* (*oep* Zhang et al., 1998).

All gene identification approaches rely on the availability of genomic resources like large insert libraries, dense genetic maps, gene catalogues and physical maps. While genetic radiation hybrid maps and EST projects have progressed very well in the past few years, construction of physical maps has only recently been started (this work, Rauch et al., pers. communication).

1.5 Genome mapping by interspersed repetitive sequence (IRS)-PCR

Interspersed repetitive sequences (IRS) can be used as anchorpoints for PCR amplification of single copy DNA probes located between a pair of the repeat element (Figure 5). Initially, this technique has been applied to isolate human specific sequences propagated in a different background, e.g. human/rodent somatic cell hybrids, and the human *Alu* repeat was used as an anchor (Nelson et al., 1989). Repetitive elements homologous to the Alu repeat are also found in the rodent genome, however, sequence divergence between the two phyla is high enough to reduce cross hybridisation. IRS-PCR can also be used to isolate DNA fragments from genomic clones, which are otherwise difficult to isolate, e.g. YACs. The method was also applied to the mouse, using primers that bind to the B1 repeat (Cox et al., 1991). Due to variant repeat distribution in different mouse strains, a subset of IRS products shows absence/presence polymorphism, and can therefore be mapped genetically in mapping crosses (Figure 5, McCarthy et al., 1995, Elango et al., 1996). IRS mapping has also been used to determine the genetic position of the mouse HIP1 gene by using a polymorphic fragment which is located on the same BAC (Himmelbauer et al., 1998). Accordingly, non-polymorphic IRS products can be mapped on a genome by hybridising it to IRS-PCR products of a radiation hybrid panel (Himmelbauer et al., 2000). In physical mapping, IRS-PCR products are used to detect clone overlaps by hybridisation against genomic libraries (Hunter et al., 1996, Hunter, 1997, Roest Crollius et al., 1996). Hybridisation of IRS probes derived from chromosome specific somatic cell hybrids is used to detect a chromosome specific subset of a genomic library (Liu et al., 1995). If a library is to be screened with IRS-PCR fragments, the complexity of the library and the number of spots on the filter can be reduced by pooling the clones, applying IRS-PCR to the pools and spotting the IRS amplicons on filters. By using the same probe for hybridisation on clone pools, radiation hybrids and genetic mapping panels, an integrated physical, radiation hybrid and genetic map can be constructed.

The benefit of IRS-PCR-based mapping lies in the fact, that it allows generating markers without prior sequencing and primer design. If advantage is taken from pooling, a high genome coverage can be screened in one hybridisation. Because a single researcher is capable of screening hundreds of clones per week, this method is well suited for high-throughput applications, as was demonstrated by the recent completion of a physical mapping project of the mouse, involving hybridisation of 15,000 IRS-markers on a YAC library (Himmelbauer, pers. communication). Conversely, in STS-based mapping, markers have to be identified by sequencing and unique primers have to be designed for each marker. Genotyping by PCR requires gel electrophoresis of each reaction.

For IRS-based mapping it is critical, that the repetitive element chosen as PCR anchor is highly abundant, evenly distributed over the whole genome, and sufficiently conserved to provide a primer binding site. In zebrafish, the DANA/mermaid element has been suggested to fulfil these criteria (Shimoda et al., 1996, Izsvak et al., 1996, Burgtorf, 1999, this work).

In contrast to microsatellites, IRS markers are not codominant and therefore it is not possible to distinguish between a heterozygous and a homozygous individual with a dominant allele (i.e. presence of the product). Pseudo-homozygous alleles in the P generation can cause mapping errors because they do not segregate in the predicted Mendelian fashion. As a consequence, founder individuals of a genetic cross have to be pure inbred strains homozygous for the vast majority of markers. Because zebrafish mapping strains have been shown not to be purely homozygous, this makes genetic mapping using dominant markers difficult (Burgtorf, 1999). A possible solution for this is to use gynogenic diploids (which are completely homozygous) as P generation or haploid offspring as a mapping panel.

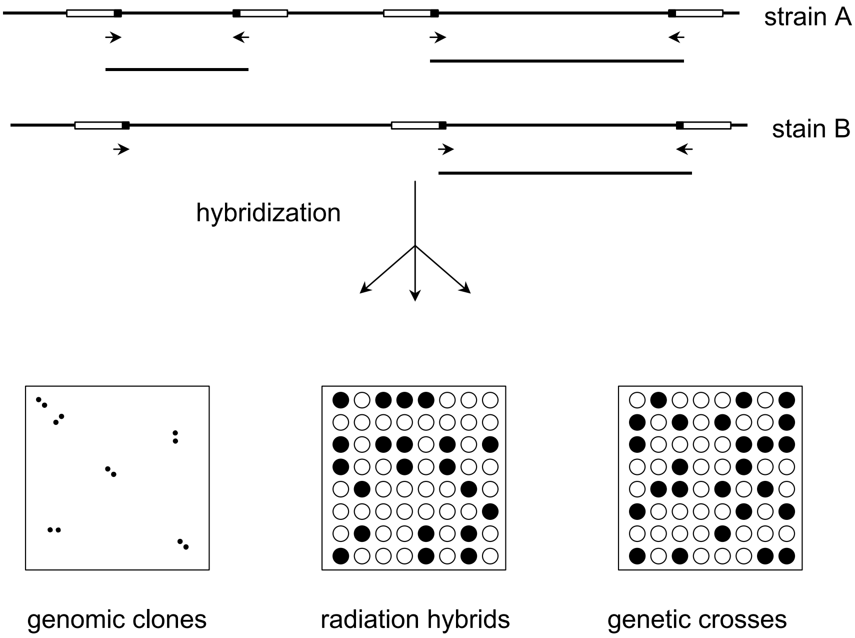

Figure 5

Principle of IRS-PCR based marker generation and hybridisation. An oligonucleotide primer binding to an interspersed repetitive element is used to amplify DNA fragments placed between two repeats in opposite orientation. Performing IRS-PCR on the whole genome and cloning of the products generates a library of IRS markers. IRS-PCR on large insert clones yields distinct bands, which can be isolated from a gel. Both kinds of IRS markers are hybridised to IRS-PCR products amplified from clone pools, radiation hybrids or genetic crosses. Absence/presence polymorphisms in different strains due to differences in repeat distribution can be used for genetic mapping.

1.6 Repetitive elements in the zebrafish genome

In humans, 45% of the genome is made of various kinds of interspersed repetitive elements (Lander et al., 2001). There is no exact determination of the repeat fraction in the zebrafish genome, but the number might be similar. There are different ways to classify repeat elements, (based on their complexity, mobility, mode of mobilisation, arrangements etc.). Some repeat elements play obvious roles in chromosomal dynamics during cell cycle and replication, such as centromeric and telomeric repeats (not discussed here), while others are probably parasitic, using the genome as a vector for transmission. In zebrafish, a variety of different classes of repeats have been discovered (as summarised in Table 4 (Izsvak et al., 1997, Ivics et al., 1999): Microsatellites, in particular CA-repeats are abundant, so that length polymorphisms are used for genetic mapping (Knapik et al., 1996, see 1.2.1). Several different satellite elements (tandemly arrayed sequence repeats) with distinct consensus sequences and differences in abundance have been found (He et al., 1992, Ekker et al., 1992). Transposons of the Tc1/mariner family (reviewed in Plasterk, 1996) have also been discovered in the zebrafish (Izsvak et al., 1995, Lam et al., 1996, Gottgens et al., 1999). Another class of mobile elements is the so-called miniature inverted-repeat transposable element (MITE). It consists of a palindromic DNA sequence 80-500 bp in length and is found in a wide range of organism. However, the mechanism of proliferation has not been established yet. In zebrafish an element of this class, called Angel, has been reported, with an estimated abundance of 10^3-10^4 copies per haploid genome (Izsvak et al., 1999). Retroelements transpose via an RNA intermediate and a reverse transcription step. Some retroelements have long terminal repeats (LTRs), a fact that puts them in close relationship to retroviruses. Non-LTR containing retroelements are characterised by the absence of repeated termini, the existence of an A-rich 3'-tail, and target–site duplications flanking their insertion points (reviewed in Weiner et al., 1986 and Okada, 1991). They can be divided in LINEs (long interspersed repetitive element) and SINEs (short interspersed repetitive elements). Full-length mammalian LINEs are 6-7 kb long and contain two open reading frames, one of which is homologous to the *pol* (reverse transcriptase) gene of retroviruses. SINEs are short (100 - 400 bp), high copy number transposable elements, which propagate non-autonomously via an RNA intermediate. SINEs posses a compound structure: The 5' part harbours an internal RNA polymerase III promoter and is derived from a tRNA sequence, with the exception of the primate Alu family and the rodent B1 family, which are derived from the signal recognition particle component 7SL. The 3' portion is thought to be derived from a LINE-like sequence, consistent with the assumption, that SINEs transpose using the LINE-encoded reverse transcription machinery.

Zebrafish contains a family of SINE like repeats, that was discovered independently by two different research groups and given different names (mermaid, Shimoda et al., 1996, and DANA Izsvak et al., 1996), along with different data about structure, distribution etc.

Sequence analysis however shows that it is a single repeat family (see 4.1.1). Estimates on the abundance of DANA/mermaid in the zebrafish genome differ by one order of magnitude: According to Izsvak et al. there are 4-5 x 10^5 copies per haploid genome, equivalent to 1 copy every 3.8 kb, or 10% of the genome. Results from Shimoda et al. suggest an abundance of 1.2 x 10^4 copies, or one copy every 140 kb. Using hybridisations to dilution series of genomic zebrafish DNA and a cloned mermaid element during the course of our project, it was estimated, that there are about 50,000 copies of the mermaid element, equivalent to 1 repeat per 37 kb (Burgtorf, 1999).

Class	Name	Structure	Properties	Copies/haploid Genome	% of genome	Reference
Satellite DNA		A+T-rich	Type 1a		5-8	a, b
			Type 1b		0.2	b
			Type 1c		>0.5	c
		G+C-rich	Type 2a		1	a
			Type 2b		>0.5	c
Retroelements						
SINEs	Mermaid	Interspersed	tRNA-related, found in vertebrates	$1.2x10^4$	1	d
	DANA	Interspersed	tRNA-related Danio specific	$4-5x10^5$	10	e
LINEs	Not described				n.d.	
DNA transposons						
Tc1/mariner like	Tdr1				0.07	f, g
	Tdr2				0.1	h, i
Miniature inverted repeat transposable elements (MITEs)	Angel				0.1	j

Table 4

Repetitive elements in the zebrafish genome adapted from Ivics et al., 1999 and Burgtorf, 1999.
a: He et al., 1992; b: Ekker et al., 1992; c: Izsvak et al., 1997; d: Shimoda et al., 1996; e: Izsvak et al., 1996; f: Radice et al., 1994; g: Izsvak et al., 1995; h: Ivics et al., 1997; i: Lam et al., 1996; j: Izsvak et al., 1999.

The DANA/mermaid repeat has been tested for its usability as an anchor for IRS PCR (Burgtorf, 1999, this work). It is interspersed in the genome, highly abundant, and inter-mermaid PCR provides a large number of different PCR products, which can be successfully used for hybridisation against mapping panels and genomic libraries. Genetic mapping, however, has been shown to be inefficient, due to the dominant nature of IRS markers in combination with the high degree of heterozygosity in the available mapping strains. One of the goals of this work was therefore, to evaluate and establish IRS-PCR based techniques for physical and radiation hybrid mapping, where these problems do not occur.

1.7 Physical mapping of zebrafish chromosome 20

One of the goals of this work is to construct a physical framework map of zebrafish chromosome 20 by using a combined STS and IRS-PCR strategy. Chromosome 20 was chosen on the basis of interest expressed by collaborators of our group, searching for mutations on this chromosome (M. Clark and S. Johnson, pers. communication). This project has a twofold purpose: (1) To construct a physical framework map to facilitate positional cloning of mutations on this chromosome and to provide the data to the community. (2) To integrate the mapping data with other maps, e.g. restriction fragment fingerprinting maps, as a basis for eventually sequencing the chromosome, thereby establishing a model chromosome, which would work as a control in the assembly of the whole-genome fingerprint map and sequence. As discussed in 1.2.3, restriction fingerprint maps are *per se* not anchored to the genetic map, and assembly of the map requires determination of optimal stringency parameters. This is facilitated, if anchor points are available. Therefore, our STS based mapping project of a chromosome is complementary to the Tübingen restriction map and should provide a valuable control to empirically determine optimal parameters for the assembly process. The same applies for the sequencing of the zebrafish genome, initiated by the Sanger Centre. Clones anchored to chromosomal locations will be templates for sequencing and will aid in the assembly of sequencing data. Once the chromosome is sequenced, it will serve as a model chromosome to study chromosome structure, repeat distribution, gene structure etc. in the zebrafish. For a detailed discussion of the potential benefits of our project see paragraph 5.2.

The mapping procedure starts with the selection of LG20-specific genes, ESTs and STS. These are used to design oligonucleotide probes for hybridisation on PAC filters. IRS-PCR probes generated from positive PACs are used for hybridisation on IRS-PCR pool filters of PAC and YAC libraries. Finally, all chromosome 20 specific PACs are restriction fingerprinted (For a more detailed description of the mapping process see 4.2).

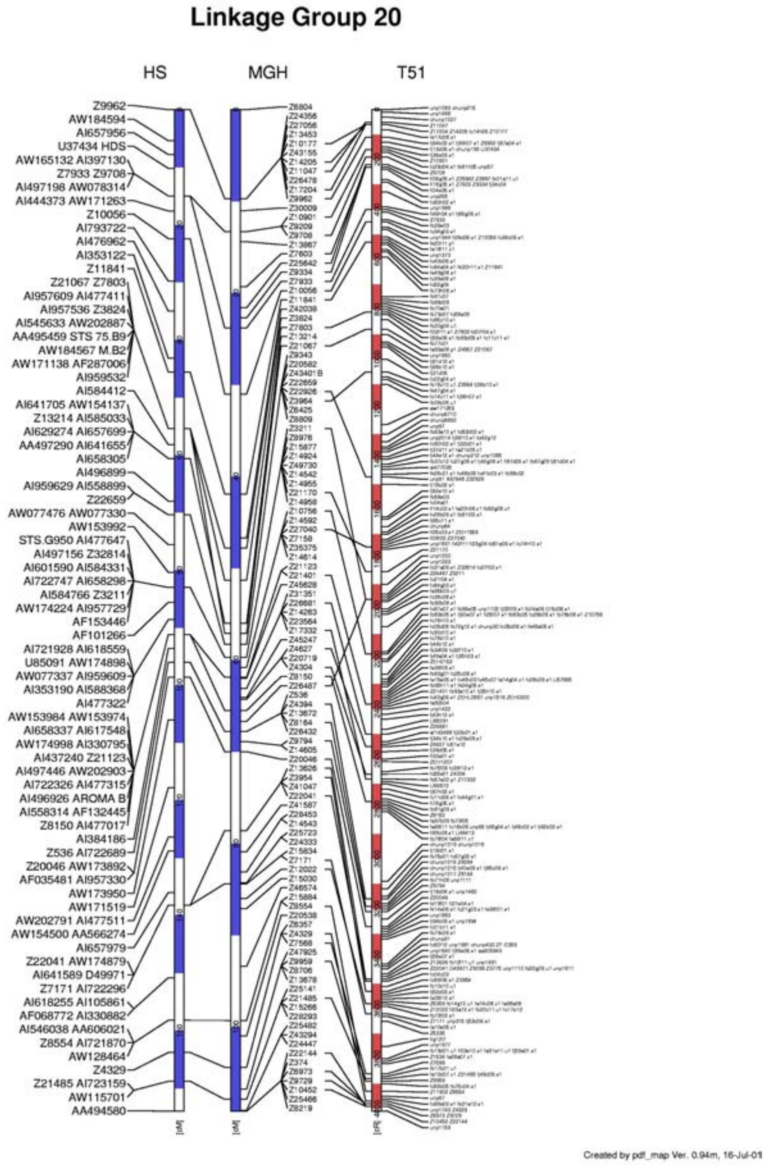

Figure 6

Alignment of genetic and radiation hybrid maps of linkage group 20. Two genetic maps (the MGH microsatellite map, Shimoda et al., 1999, and the heat shock panel map Woods et al., 2000) and one radiation hybrid map (T51, Geisler et al., 1999) are aligned. Microsatellite framework markers, which the MGH map shares with one of the other maps, are connected by lines. Note the distortions and inversions particularly between the genetic an radiation hybrid maps. The illustration was produced using a PHP program written by Thomas Kreitler.

In Table 5, mapping data of linkage group 20 are shown from the MGH and heat shock genetic panel, and from the T51 radiation hybrid panel. The T51 map still contains gaps, therefore chromosome size can only be estimated using this panel. Based on the MGH map, it can be stated, that LG20 has a genetic size of 109.2 cM, thus being somewhat larger than the average (91.8 cM). It comprises ca. 5% of the genome.

	MGH	HS	T51
Size	109.2 cM	174 cM	-
Calculated physical size	82 Mbp	98 Mbp	-
Markers mapped	110	125	238

Table 5

Mapping data of the zebrafish linkage group 20. Shown are data from the MGH and heat shock genetic panels, and from the T51 radiation hybrid panel. No size estimates are given for the radiation hybrid map, because it contains gaps.

The maps shown in Table 5 have markers in common, which allows integration of maps. Particularly, microsatellite markers from the MGH map are used to anchor the heat shock and radiation hybrid maps. Different ESTs might be derived from the same transcript, and might even contain largely overlapping sequences. Therefore, overlapping sequences have to be identified by sequence clustering to create a set of non-redundant markers.

The three maps combined have 383 markers with sequences stored in Genbank (not including the unpublished ones). From these sequences, 15 sequences are annotated as cloned genes with complete coding sequences (defined as VRT, i.e. “other vertebrate sequences” in the Genbank entry); 275 sequences are annotated as expressed sequence tags; 93 are annotated as sequence tagged sites (microsatellite sequences). The latter number is less than the 123 microsatellite markers on the map. This is probably due to the fact, that not all marker sequences have been submitted to Genbank. To give an overview of what is known about genes on chromosome 20, Table 6 shows the 15 markers, which are annotated in Genbank as cloned genes.

Accession numbers	Genbank definitions
AF035481	Danio rerio connexin 43 (Cx43) mRNA, complete cds.
AF068772	Danio rerio heat shock protein hsp90beta mRNA, complete cds.
AF101266	Danio rerio DNA binding protein (sox25) mRNA, complete cds.
AF132445	Danio rerio signalling molecule lefty2 (lft2) mRNA, complete cds.
AF143488	Danio rerio acetyl-CoA acetyltransferase 2 (ACAT2) mRNA, complete
AF153446	Danio rerio kit receptor tyrosine kinase mRNA, complete cds.
AF287006	Danio rerio T-box brain 1 mRNA, partial cds.
D49971	Danio rerio mRNA for bone morphogenetic protein, complete cds.
U37434	Danio rerio L-isoaspartate (D-aspartate) O-methyltransferase (PCMT)
U49413	Danio rerio Zg09 gene, partial cds.
U57965	Danio rerio ribonucleotide reductase protein R2 class I mRNA,
U66872	Danio rerio enhancer of rudimentary homolog mRNA, complete cds.
U85091	Danio rerio transcriptional regulator Sox-11B (sox11B) mRNA,
X67648	B.rerio ZN-CAD mRNA.
Z32814	B.rerio mRNA for platelet-derived growth factor receptor alpha.

Table 6

Cloned genes (Genbank annotations) mapped on linkage group 20. The list was created by filtering all Genbank entries of chromosome 20 markers containing “VRT” (other vertebrate genes) in the definition field.

Some of the genes listed in Table 6 have been studied for their roles in zebrafish embryonic development. The mutant *swirl* for example is disrupted in the bone morphogenetic protein 2b (*bmp2b*) gene on chromosome 20 (Nguyen et al., 1998). It has a dorsalized phenotype, suggesting a role of *bmp2b* in dorsoventral axis specification. The phenotype of the *sparse* mutant, showing defects in melanocyte migration and survival, is due to a mutation in the zebrafish *c-kit* orthologue (Parichy et al., 1999). *Lefty2* a member of the TGF-β protein family, is thought to be involved in establishing bilateral symmetry (Bisgrove et al., 1999).

An interesting aspect of the zebrafish chromosome structure is its syntenic relationship with other organisms, especially human. Since the split between the actinoptrygian and sarcopterygian phyla the chromosomes of the common ancestor of fish and humans have been rearranged by interchromosomal translocations and intrachromosomal inversions. Nevertheless regions of conserved synteny are maintained, the size of which is dependent of the rate at which rearrangements are fixed. Conserved synteny means that two or more genes are linked in two different species. Homology blocks are uninterrupted segments containing two or more contiguous genes or ESTs with conserved map order between two genomes. Such comparisons help to understand the evolutionary history of the different phyla. They are also needed to define the relationship between zebrafish genes and mutations with those of humans. For zebrafish chromosome 20, conserved synteny is detectable primarily with human chromosomes 6 and 14, but also with 2, 4 and 20 (Barbazuk et al., 2000) (Woods et al., 2000). A large number of genes shows conserved syntenies between zebrafish and humans, but gene orders are usually inverted and transposed. This suggests that intrachromosomal rearrangements have been fixed more frequently than translocations (Postlethwait et al., 2000). Uninterrupted homology blocks on zebrafish chromosome 20 can be up to (estimated) 6 Mbp large.

2 Objective

The goal of this work is to generate data and resources that facilitate the genetic study of the zebrafish as a model organism for vertebrate embryonic development and human diseases. In spite of a growing number of available genomic resources the cloning of genes disrupted in zebrafish mutants is still often difficult and time-consuming. The anticipated availability of the sequence of the zebrafish genome will accelerate this drastically. Both positional cloning of genes as well as genomic sequencing is dependent on physical maps, anchored to genetic and radiation hybrid maps. However, PCR-based physical mapping of large libraries is laborious, mainly due to the large number of samples, which have to be visually inspected on gels.

One specific goal of this work is therefore to establish, optimise and apply IRS-PCR based methods to physical and radiation hybrid mapping of the zebrafish. This is carried out by testing of different anchor repeats and PCR conditions; additionally, by generation and characterisation of an IRS marker library and IRS-pool filters for hybridisation. The suitability of IRS-PCR products as a reduced representative subset of the genome for SNP detection and mapping is also tested.

A second goal is to generate a physical framework map of a zebrafish chromosome, anchored to the already existing genetic and radiation hybrid maps by using a combined IRS-PCR and STS-content based approach. By restriction fingerprinting of mapped clones, these data will be integrated in the Tuebingen restriction fingerprint map. Thus a first STS-based physical framework map of a zebrafish chromosome is generated. This will support the positional cloning of mutant genes genetically mapped on this chromosome. It will also establish a model chromosome for the empirical optimisation of parameters for assembly processes in map construction and sequencing, and for the study of gene and chromosome structure in this species.

A third goal of this work is to further establish the zebrafish as a model organism for the study of human diseases and to provide candidate genes for mutations. This is carried out by the mapping of ESTs with homologies to human disease genes and of ESTs showing localised expression patterns in an in situ hybridisation screen. Synteny comparisons between human and zebrafish help to determine if a zebrafish homologue of a human disease gene is a true orthologue. By adding mapping information to gene catalogues containing expression data, potential candidate genes for mutations are provided.