Chapter Three

Positional Cloning of *doc*

3 Positional Cloning of *doc*

In this chapter the techniques used to identify the gene responsible for a mutant phenotype will be discussed. The methods required to determine a genes position and then identify the gene within that region responsible for a mutant phenotype will be covered. This introduction will also provide examples of the application of this approach to successfully identify mutant genes in the zebrafish. The results of the positional cloning of *doc* are then presented.

3.1 Introduction

There are two dominant approaches used to clone the gene responsible for a specific mutant phenotype; the candidate gene approach and positional cloning. The candidate gene approach has been used successfully on number of occasions to identify mutated zebrafish genes. This approach uses functional information, obtained from previous work, to suggest genes that may act in a way that could bring about the mutant phenotype. Two examples of this approach are the zebrafish mutants *no tail* and *floating head*, which were found to be homologues of the *Xenopus* genes *Xbra* and *Xnot* respectively (Schulte-Merker et al., 1994; Talbot et al., 1995). However, this approach is limited to well studied processes and genes, and in cases where there is no obvious candidate, such as is the case when isolating novel genes, cloning via the identification of the genes position in the genome becomes necessary (Talbot and Schier, 1999). This approach was initially developed as a means of identifying the genes responsible for human genetic diseases (Collins

et al., 1992) and in 1986 was demonstrated to be a functional approach to identifying mutations with the identification of the gene responsible for chronic granulomatous disease (Royer-Pokora et al., 1986). Since then, many inherited human diseases have had the gene responsible identified through the application of positional cloning techniques. These include Duchenne muscular dystrophy, cystic fibrosis, fragile X syndrome and breast cancer (Fu et al., 1991; Kerem et al., 1989; Miki et al., 1994; Monaco et al., 1986; Riordan et al., 1989; Rommens et al., 1989; Verkerk et al., 1991). This method has become more powerful and accessible in recent years through the availability of high quality genome sequence. The sequencing of whole genomes in human, mouse and, most importantly in terms of this thesis, zebrafish has dramatically decreased the work necessary to identify a gene based on its genetic location.

Positional cloning in zebrafish, or recombinant mapping, relies on the use of polymorphisms that occur between the mutant strain and the mapping strain, where the mapping strain is crossed with the mutant carrier to generate a mapping line from which mutant and wild type sibling embryos are collected. The identification of polymorphic markers that are closely linked to the mutation, through comparisons of wild type and mutant embryos, is a critical step in any positional cloning project. If such flanking markers can be found then the amount of work involved later in chromosome walks and candidate identification is greatly reduced. Hence, the availability of dense genetic maps is highly desirable. For zebrafish, the first genetic map was based on random amplified polymorphic DNA (RAPD) (Postlethwait et al., 1994) and identified 401 loci, which was increased to 652 by 1996 (Johnson et al., 1996). However, genetic mapping using RAPD analysis has many drawbacks and thus short sequence length polymorphism (SSLP) markers have become the preferred

marker type for genetic mapping studies (Beier, 1998). SSLPs utilise PCR primers that flank short microsatellite repeats, typically the dinucleotide cytosine-adenine (CA), that are spread throughout the genome and that vary in the number of repeats. A map based on SSLP marker was published for the zebrafish in 1996 and characterised 102 SSLPs. Since then further markers have been added and the total number now exceeds 2000 (Shimoda et al., 1999). However, one drawback of SSLP maps is that a high proportion of markers will not be polymorphic between two specific strains and so are not useful for mapping. Thus, any marker that might be useful must initially be tested to determine whether it is polymorphic between the mapping lines.

The positional cloning approach has been used to successfully identify many zebrafish genes, including novel genes (Zhang et al., 1998), genes that provide functional models for human disease (Parsons et al., 2002a), genes that highlight the conserved nature of development (Kikuchi et al., 2001) and, most relevantly to this thesis, genes involved in development of the notochord (Coutinho et al., 2004; Parsons et al., 2002b) (**Table 3.1**). Positional cloning is a labour intensive process and utilises a multitude of genetic tools to identify the gene responsible. Since the publication of the large-scale ENU mutagenesis zebrafish screens in 1996 such tools have become a source of much development in zebrafish, steadily increasing the ease with which mutations can be mapped. Currently, there is a relatively comprehensive SSLP map, containing thousands of markers spread throughout the genome (Postlethwait et al., 1994; Shimoda et al., 1999), which has been supported recently by genome sequencing. The generation of large scale genome sequence has allowed markers to be more accurately positioned and also offers up identified insertion/deletion (INDEL) fragments and BAC end fragments as alternate markers

for use in fine mapping. Radiation hybrid (RH) maps, large-insert genomic libraries and the large number of published expressed sequence tags (ESTs) (reviewed in (Beier, 1998)), also aid in positional cloning. Further to these tools, zebrafish genome sequence, which is currently estimated at more than 1.5Gb of sequence, allows for the rapid and easy identification of candidate genes and gene sequence that, when coupled with the ability to easily test gene function through 'knock-down' with anti-sense morpholino oligonucleotides (MOs), enables the rapid identification and testing of candidate genes once genomic location has been identified. Such techniques have succeeded in making positional cloning a straightforward means to characterise a genes identity from the identification of mutant phenotypes (Fishman, 1999; Talbot and Schier, 1999).

Mutant	Protein	Reference
casanova	Sox related protein	(Kikuchi et al., 2001)
gridlock	Hairy-related bHLH	(Zhong et al., 2000)
heart and soul	PKC lambda	(Horne-Badovinac et al., 2001)
miles apart	Sphingosine-1-phosphate receptor	(Kupperman et al., 2000)
one-eyed pinhead	EGF-related protein	(Zhang et al., 1998)
pickwick	titin	(Xu et al., 2002)
ogon	Sizzled	(Yabe et al., 2003)
grumpy	laminin β1	(Parsons et al., 2002b)
sleepy	laminin γ1	(Parsons et al., 2002b)
bashful	laminin α1	(Pollard, 2002)
sneezy	COPα	(Coutinho et al., 2004)

Table 3.1 Examples of Zebrafish genes identified by positional cloning.

one-eyed pinhead was the first zebrafish gene identified by positional cloning. Positional cloning has been used to identify four of the seven zebrafish dwarf mutants.

Positional cloning is performed in a series of steps where the mutation is defined to smaller and smaller genomic regions. The primary step in positionally cloning any gene is to define the mutation to a specific linkage group. There are several methods that can be used in this step, including gynogenic half-tetrad diploids (Johnson et al., 1995) and bulk segregant analysis (BSA), which involves the identification of polymorphic markers that are linked to a mutation by studying the segregation of markers between pools of WT and mutant embryo DNA (Beier, 1998). Once linkage to a specific linkage group has been confirmed, genetic mapping is performed to find linked markers on either side of the mutation, with markers within that region then tested to define to mutation to regions of decreasing size. Once known genetic markers are exhausted, the genomic sequence can be used to design new markers, which can be tested for polymorphisms between the two mapping strains.

The zebrafish is an excellent system in which to perform meiotic mapping studies, as large numbers of meioses can be analysed with relative ease. This then increases the probability of identifying useful crossovers and enables the mutation to be mapped with greater resolution to a genomic region. By crossing two genetically distinct strains of zebrafish fish, one of which contains the mutation and the other of wild-type phenotype, hybrid fish are produced that can then be interbred to produce embryos suitable for mapping. By using these embryos, genetic markers can be used to perform meiotic mapping, where the frequency of recombination between the marker and the genetic locus is measured through the use of markers that vary between the two strains. Approximately one in three known SSLP markers vary between any two zebrafish strains. The recombination events of these polymorphism markers between the mapping strains are then used identify a closely linked marker,

or, ideally, two closely linked markers that flank the mutation on either side. This involves genotyping many phenotypically mutant fish, in an attempt to identify recombination events that assist in the determination of the distance and order of markers. Thus, a candidate interval is defined, which is reduced through further identification of new markers and continued mapping. Once all known polymorphic markers are exhausted, further markers can be designed from genomic sequence to further resolve the genomic region. By comparing INDELs between the two mapping strains, polymorphic markers can be identified that can be further used to refine the mapping region. Once the mutation has been defined to a suitable region sequence either from the genome sequencing project or, in cases where this sequence is uncertain, from self generated sequence from BACs can be analysed to identify candidate genes. Once candidates are identified there are three well-established techniques that can be used to test and prioritise them: the attempted rescue of the mutant phenotype, through over-expression of cRNA or DNA (Yan et al., 1998), though this is not always straight forward, due to complications of gene mis-expression; the expression of the candidates can be tested through *in situ* hybridisation, to examine if the gene is expressed in a manner befitting the phenotype; and the function of a gene can be examined through targeted gene knock-down with MOs in an attempt to 'phenocopy' the mutant (Nasevicius and Ekker, 2000). Once a suitable candidate is identified, the mutation itself can be characterised through cloning and sequencing of wild type and mutant cDNA.

In recent years, the zebrafish has become a far more versatile system for studying developmental biology through the use of the MO gene knock down technique, which provides a quick, easy and relatively cheap method of removing a gene to study its function. MOs are short oligonucleotides that possess a morpholino

ring, rather than a ribose sugar and phosphoamidite rather than phosphodiesters. This abnormal backbone means that MOs are more resistant to degradation than conventional oligonucleotides and are nevertheless capable of hybridising with endogenous nucleic acids. Mos have proved to be particularly effective in the study of developmental biology in a number of organisms and has rapidly become a popular and powerful tool (Heasman, 2002; Heasman et al., 2000). The phenotypes of *no tail*, *chordino* and *one-eyed pinhead* have all been successfully phenocopied using MOs (Nasevicius and Ekker, 2000). Moreover, all of the identified zebrafish notochord mutants have had their respective phenotypes copied using MOs, demonstrating that these phenotypes are easily replicated with this technology (Coutinho et al., 2004; Parsons et al., 2002b). MOs are generally designed to prevent translation through hybridisation to the endogenous mRNA at either the start of translation, which knocks down both maternal and zygotic transcripts, or at a splice sites in the un-edited RNA (Draper et al., 2001), which knocks down only the zygotic transcript. MOs are stable (operating via steric blocking), cheap, and extremely effective. Production of a given mutant phenotype after injection of a MO against a particular candidate gene provides strong evidence that it represents the mutated gene.

3.2 Initial Mapping of *doc*

Initial mapping of *doc* involved the linking of the mutation to a specific chromosome. Pools of 48 mutant and 48 wild type sibling (which would include homozygous wild type as well as heterozygous embryos) were used to test linkage of

doc to two sets of SSLP markers spaced throughout the genome. Each panel consisted of 192 markers. Markers were amplified from both mutant and wild type pools, then products run on 3% agarose to examine linkage of polymorphic markers. Using this method, non-mendelian inheritance of the marker D3 on plate H2-2 was noted. This lack of recombination between the mutant locus and the marker Z9484 linked *doc* to LG18. (**Figure 3.1**)

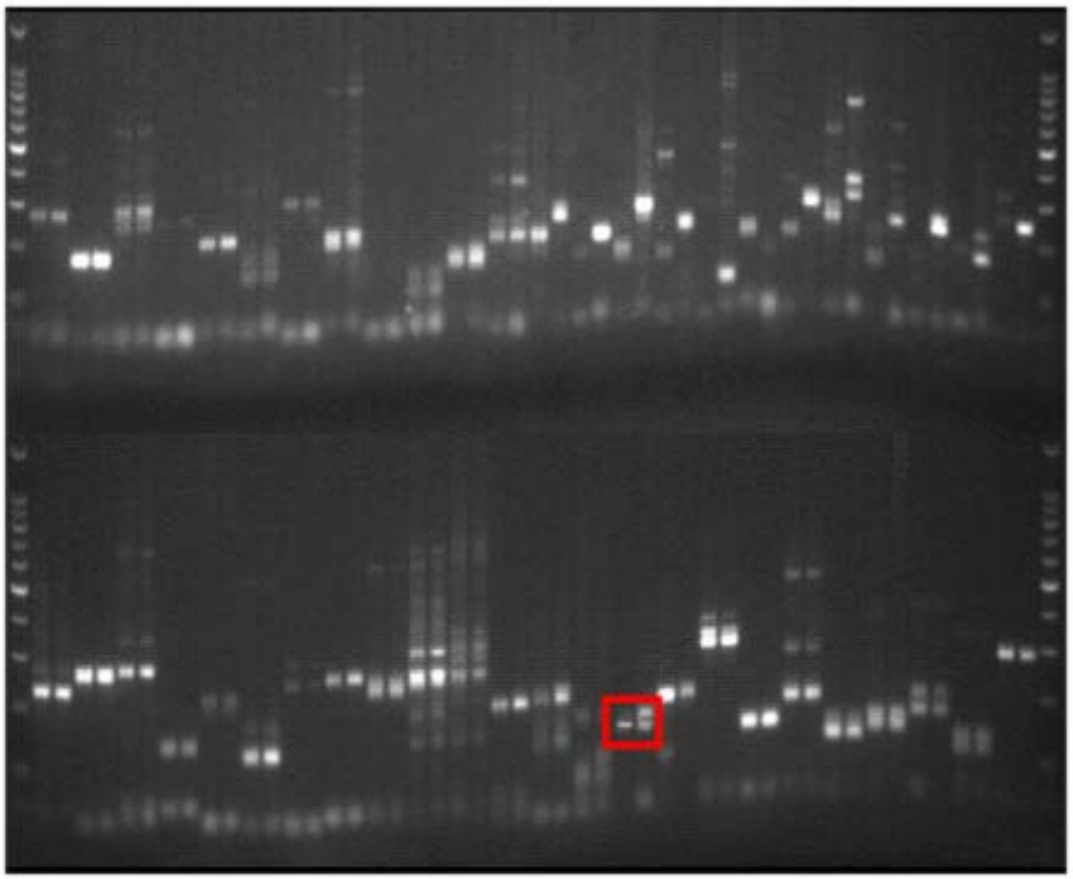

Figure 3.1 Likage group mapping of *doc.*

Mapping panel H2-2, marker a1-12 and b1-12 (top row) and c1-12 and d1-12 (bottom row). PCR products of mutant pool (left) and wild type sibling pool (right) run in tandem. Bordered in red is marker d3 (Z9484), mutant pool shows one marker product where the wild type sibling pool shows both polymorphic products. Products run on 3% agarose, 100bp ladder (promega #G210A) run on the left and right of the gel.

3.3 Confirmation of Linkage and Further mapping

3.3.1 Confirmation of Linkage to LG18

To confirm linkage of *doc* to linkage group 18, three polymorphic markers, Z9484, Z7654 and Z7417 were identified using the MGH zebrafish microsatellite map (Shimoda et al., 1999) (http://zebrafish.mgh.harvard.edu/). Products were amplified from 48 single *doc* mutant fish and 48 single wild type siblings and run on 3% agarose to examine recombination events. Linkage was demonstrated to all three markers, with the 48 mutant embryos showing four recombination events between *doc* and marker Z9484, 10 recombination events between *doc* and marker Z7654 and four recombination events between *doc* and marker Z7417 demonstrating a distance of 4.1cM, 10.4cM and 4.1cM from the markers respectively.

3.3.2 Defining a Region for *doc*

By comparing the occurrence of recombination in specific individuals, markers can be placed on either side of the mutant locus, as recombination occurs only once on a single linkage groups during meiosis in zebrafish (Streisinger et al., 1986). All four recombinant embryos between *doc* and marker Z9484 were unique, whereas the four recombinants between Z7417 and *doc* were also recombinant for Z7654. Thus, the markers Z7417 and Z7654 flank one side of *doc* with Z9484 flanking the other side. The two closest markers to *doc* on either side were identified as Z9484 and Z7417. Using publicly available genome sequence

(http://www.ensembl.org/Danio_rerio/), this placed *doc* in a region of ~12Mbases on linkage group 18.

3.3.3 Fine mapping of *doc*

Using publicly available genome sequence, BAC end sequence and INDELS (insertion/deletion sequences) were identified and primers designed to amplify ~200bp regions to test for polymorphisms. Initially, 20 BAC ends and 18 INDEL markers were tested, spaced approximately evenly across the region. One of the BAC end markers and two of the INDEL markers were polymorphic between the mapping strains. Testing of the BAC end marker with the 48 mutant embryos identified that one of the embryos demonstrating recombination between Z9484 and *doc* was also recombinant. Thus, this marker, at roughly the 11.1Mb position on LG18 is positioned between Z9484 and *doc*. Mapping of the INDEL markers enabled finer positioning of *doc*. Both of the polymorphic markers demonstrated recombination within one the mutant embryos that demonstrated recombination for Z7417, demonstrating that they are positioned between *doc* and Z7417. The closest of these markers to *doc* was located at approximately the 11.5Mb position on LG18. Thus, *doc* is situated between the two identified polymorphic markers in a region of ~0.5Mbases between 11.1 and 11.5 Mb on LG18.

3.4 Identifying Candidates

Using SSLP and INDEL markers, the *doc* locus was defined to a region of ~0.5Mbases on linkage group 18. Using publicly available genomic sequence, this region was examined for identified or predicted genes that might play a role in notochord development. Three such genes were identified: Cadherin-13, Syntaxin-8 and a novel multiple WD40 domain protein.

3.5 Discussion

Through the use SSLP markers covering the entire zebrafish genome, linkage of *doc* to linkage group 18 was established, since the panel marker Z9484 showed no mendelian inheritance in pooled mutant samples. This linkage was confirmed through the further use of SSLP markers, including Z9484 and known linked markers from the genetic map. These markers also assisted in defining the *doc* locus to a specific region, since the recombinants for Z9484 and the markers Z7417 and Z7654 were discrete individuals. Thus, since recombination occurs only once on each linkage group during meiosis in zebrafish, the markers were considered to be situated on either side of the mutation, as no mutant embryo demonstrated recombination for both Z9484 and Z7417 and/or Z7654. Since the frequency of recombination between *doc* and Z7417 was less than the frequency of recombination between *doc* and Z7654, *doc* was characterised as lying between Z9484 and Z7417.

The genetic map provided no polymorphic markers closer to *doc* than Z9484 and Z7417, which provided a genomic distance of ~12Mb in which the mutation was situated. Such a region was far too large to attempt to find candidate genes from available sequence and so genomic sequence was instead used to try to provide further markers for recombinant mapping. INDELs are commonly identified

polymorphisms in the zebrafish genome, resulting from the insertion or deletion of small fragments of non-coding genome between strains. These INDELs were used to design and identify polymorphic markers for use in further mapping. Using such markers, mapping of *doc* was further refined and the mutation was mapped to a region of ~0.5Mbases. Within this region, three possible candidate gene were identified from the genomic sequence (see **Figure 1.1** for an overview of mapping).

3.6 Summary

- *doc* is located on linkage group 18.
- *doc* lies between the markers Z9484 and Z7417, in a defined region of ~0.5Mbases.
- Within this region, three gene were identified as likely candidates.

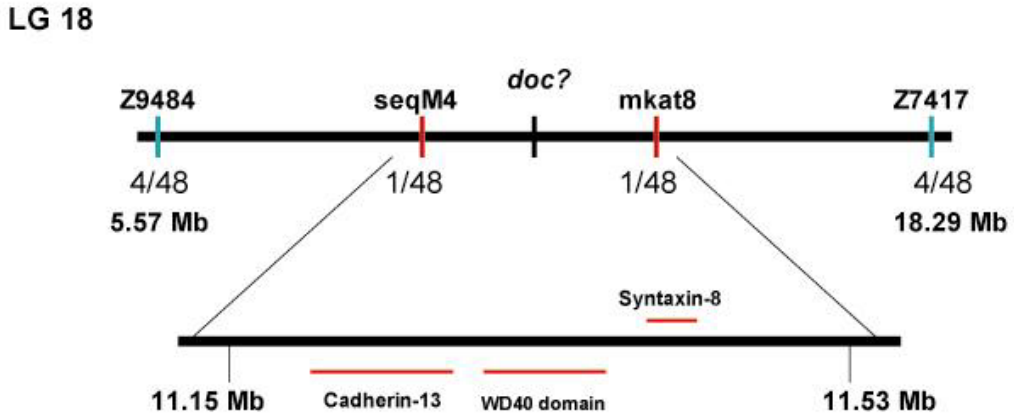

Figure 3.2 Overview of mapping of *doc*