Chapter 3

Positional cloning of *schnecke*

Chapter 3: Positional cloning of *schnecke*

3.1 Summary

In this chapter I have described the positional cloning of the muscle mutant, *schnecke* (*sne*). Using simple sequence length polymorphism (SSLP) and insertion-deletion (indel) markers the location of the *sne* mutant locus was initially defined to a 1 centimorgan (cM) interval on chromosome 8. Subsequent sequencing of one of the candidate genes within this region revealed a point mutation at a splice site of *capzα1*. Unexpectedly, the splice site mutation induces the transcription of three mis-spliced transcripts in the *sne* mutant.

3.2 Introduction

3.2.1 ENU mutagenesis screens

Forward mutagenesis screens using the mutagen *N*-ethyl-*N*-nitrosurea (ENU) were first performed in *Drosophila* (Nusslein-Volhard, 1994; Nusslein-Volhard and Wieschaus, 1980) and many genes involved in embryonic patterning were successfully identified such as *wingless* (*wg*), *decapentaplegic* (*dpp*) and *hedgehog* (*hh*). In vertebrates, zebrafish have proved to be an ideal model organism for use in mutagenesis screens to identify genes important in development. Zebrafish have a short generation time (2-4 months) and hundreds of progeny are produced from each mating, therefore many mutant phenotypes can be scored and mutant lines can be established quickly. Additionally, very early developmental phenotypes can be detected as the embryos are transparent and develop externally from the first cell division.

In the mid 1990s two landmark zebrafish ENU mutagenesis screens were performed and almost 2000 mutations affecting approximately 600 genes were identified (Driever et al., 1996; Haffter et al., 1996). Over the past decade the study of these developmental mutants has dramatically assisted in extending our knowledge of vertebrate development, however, many mutants still remain uncharacterized. The *sne* mutant is one of more than 50 muscle mutants generated in the Tübingen screen (Granato et al., 1996; Haffter et al., 1996). In this screen single base mutations were induced into the premeiotic germ cells of Tübingen male zebrafish by incubating the male in a solution of ENU for 1 hour. The treatment was repeated up to six times at weekly intervals. The male was then outcrossed with a Tübingen longfin (TL) wild type female zebrafish, resulting in heterozygote F1 progeny. Sibling F1s were then incrossed in single pair matings, so that half of the subsequent F2 progeny were heterozygous for a specific mutation carried by either of the F1 parents. The F2 progeny were then intercrossed and the embryos analyzed for a phenotype (Fig. 3.1). If both F2 parents were heterozygous for a specific recessive mutation then a quarter of their progeny would have a mutant phenotype that could be scored. In this type of screen only mutations in genes that have unique and partially non-redundant functions and that produce a scorable phenotype will be detected. Once a mutant line has been established the mutation is identified by positional cloning.

3.2.2 Positional cloning

Positional cloning is currently the most common method used to identify mutated loci in ENU generated mutants and can be divided into three main steps. Firstly, the location of the mutation is mapped to a genomic interval on a particular chromosome via linkage analysis. Secondly, candidate genes within the chromosomal region are selected, and finally the mutation within a candidate gene is identified by sequencing.

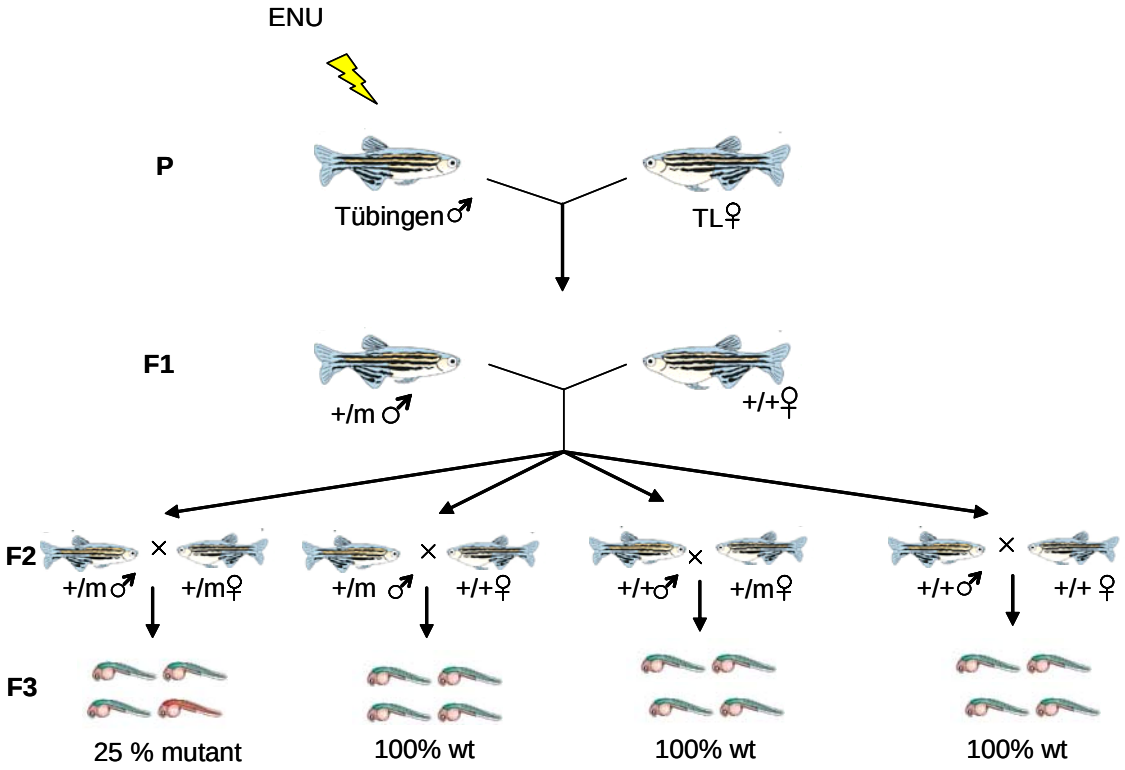

Fig. 3.1. Generation of a zebrafish ENU mutant library. The ENU mutagenized founder male is outcrossed to a wild type female of a different strain (P). The subsequent progeny (F1) are intercrossed in single pair matings and produce an F2 generation where half the offspring are heterozygous for a particular mutation. In a quarter of the F2 intercrosses both the male and female will be heterozygotes, therefore 25% of their progeny will be homozygous for a mutation which can be phenotyped. This figure was adapted from Lieschke and Currie, 2007.

A number of PCR based mapping technologies such as RAPDs (random amplified polymorphic DNAs), AFLPS (amplified fragment length polymorphisms) and SSLPs have been utilized to create genetic maps that enable the locus of a mutation to be assigned to a particular region on a chromosome by linkage (Johnson et al., 1994; Knapik et al., 1998; Postlethwait et al., 1994; Vos et al., 1995). The premise behind all these different mapping techniques is to detect polymorphisms between strains that can be linked to the mutant locus. SSLP mapping is by far the simplest positional cloning technique compared to RAPD or AFLP mapping. The amplification of SSLP markers is generally consistent, and as fewer amplified products are generated it is much easier to score the linkage of a particular marker. Moreover, as the SSLP markers tend to be co-dominant, heterozygous and homozygous genotypes can be distinguished, thus diploid embryos can be used for mapping. For these reasons, SSLP markers are now routinely used to identify mutations in ENU mutagenized zebrafish.

SSLPs (also known as microsatellites) were originally identified in the zebrafish genome by cloning and sequencing of hundreds of CA repeat regions from genomic DNA (Goff et al., 1992; Knapik et al., 1996; Knapik et al., 1998). These regions are highly polymorphic between strains. Therefore, PCR amplification of the microsatellites enables linkage to be established between a marker and the mutant locus. Linked SSLP markers that are closer to the mutation will have a lower recombination rate than markers further away from the mutant locus. By determining the recombination frequency of each linked SSLP, the distance (in cM) of the marker from the mutation can be estimated (see Fig 3.2 as an example). The recombination frequency at a particular SSLP marker is determined by identifying the recombinations that have occurred at this locus in many individual mutant embryos. The distance in cM is calculated by dividing the number of recombinations observed by the total number of meioses (total number of mutant embryos scored x 2) then multiplying this ratio by a 100. The accuracy of this distance

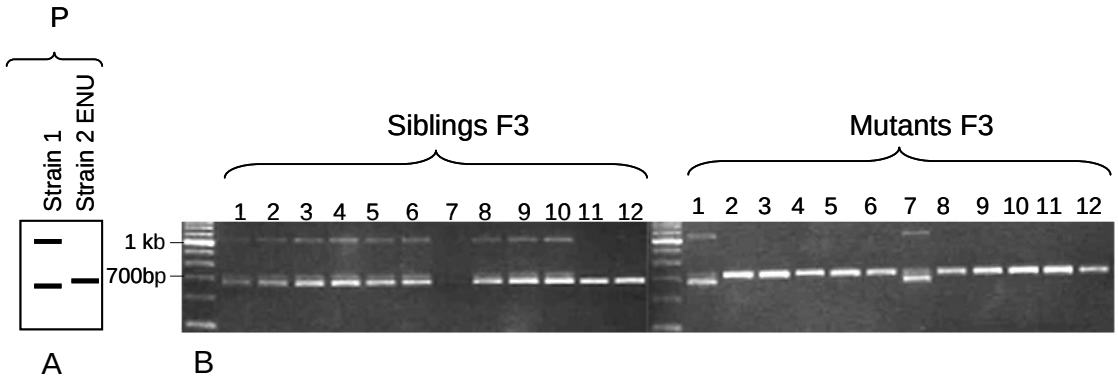

Fig. 3.2. Example of how the recombination frequency of a linked SSLP polymorphic marker is determined by PCR. A) The SSLP marker produces 3 products at 680bp, 710bp, and 1050bp. The 680bp and 1050bp product are linked to parental strain 1 and the 710bp product is specific for strain 2 (which has been mutagenized). B) PCR products generated from individual F3 sibling and mutant embryos using the linked SSLP marker. Most of the F3 siblings amplified all 3 products (1-6 and 8-10) or only the 680bp product (11 and 12) (PCR 7 failed). In the mutant embryos the 710bp product (specific to strain 2) was amplified in most individuals. However, due to recombinations in two of the mutants (1 and 7), products from the unmutagenized strain (strain 1) were also amplified. The number of mutants which have a banding pattern corresponding to the unmutagenized strain can therefore be used to determine the recombination frequency and subsequently the distance of the marker from the mutant locus.

measurement improves as greater numbers of mutant embryos are scored for recombination events.

There are several methods available to pinpoint the affected gene once the genomic interval that contains the mutation has been established. Usually the candidate gene approach is taken, whereby likely candidate genes within the defined region are selected and sequenced. Antisense morpholino oligonucleotides (MOs) can also be designed to knockdown the candidate gene, to determine whether the mutant phenotype can be copied. Unfortunately, this approach is limited by the number of genes that have been fully sequenced in the region and the size of the interval.

3.3 Positional cloning of *sne*

The mutation in the *sne* mutant was originally roughly mapped to linkage group 8 using a subset of SSLP markers derived from the Massachusetts General Hospital (MGH) marker map (Knapik et al., 1998; Shimoda et al., 1999) http://zebrafish.mgh.harvard.edu/). 384 SSLP markers from the G4 and H2 marker set (Geisler 2007) were tested for linkage, and the *sne* locus was mapped to a 10.7 cM region between markers Z21483 and Z21115 on chromosome 8 by Dr. E. Busch-Nentwich (Fig. 3.3). Recently a mutant mapping screen was published that also roughly mapped the *sne* locus between SSLP markers Z14312 and Z21115 (Geisler et al., 2007).

The region containing the *sne* locus was subsequently refined by initially testing whether 24 of the markers within this region were polymorphic by PCR, on pooled mutant and sibling genomic DNA. Seven markers were found to be polymorphic (Fig. 3.3) and the recombination frequency was determined at these loci by scoring individual mutant embryos. A discrepancy between the position of three of the markers (Z21315, Z10456 and Z60737) on the MGH map

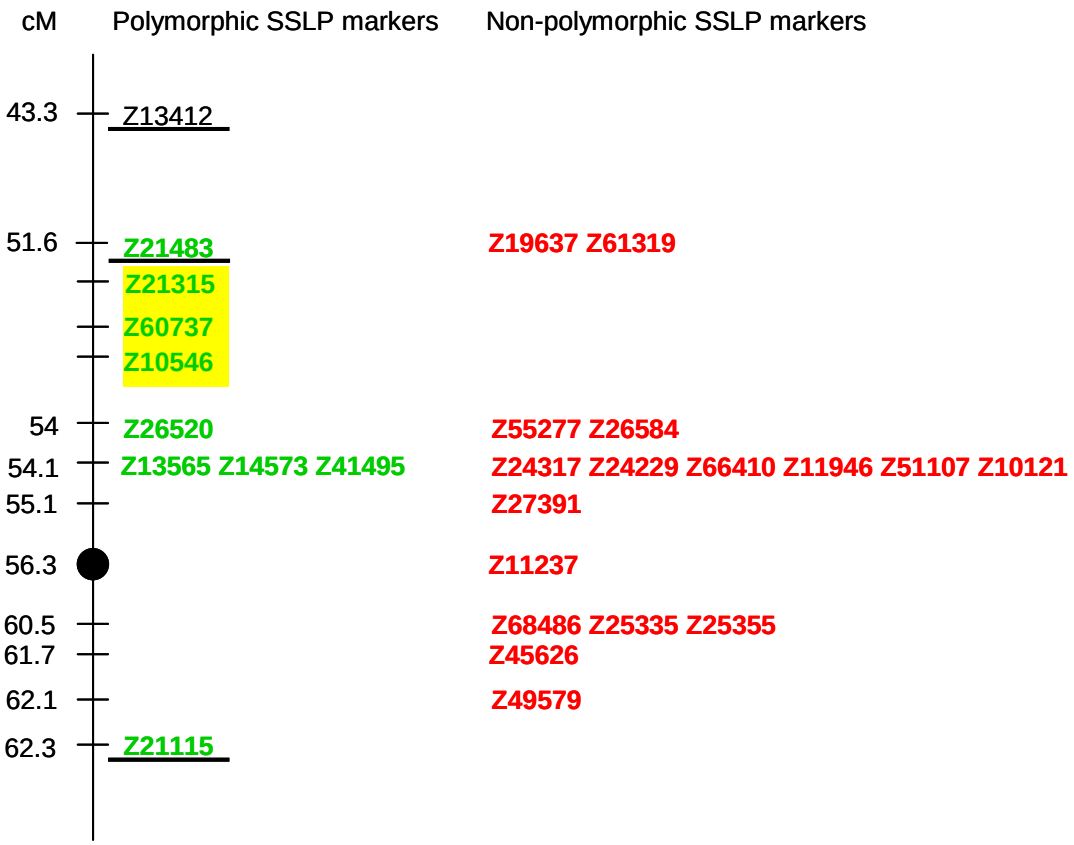

Fig. 3.3. SSLP marker map of chromosome 8 between 43.3 and 62.3cM. Markers identified as non-polymorphic are in red, polymorphic markers are in green and were tested on individual mutant embryos to determine recombination frequencies. The linked markers determined by E. Busch-Nentwich and Geisler et al., 2007 are underlined. Markers that were in a position on the MGH map that were inconsistent with the recombination frequencies I obtained have been re-positioned and are highlighted in yellow. On the published MGH map Z21315 and Z10456 are located at 54.1cM and Z60737 is at 56.5cM. The black dot indicates the centromeric region of the chromosome.

compared to the recombination frequencies I calculated was observed, therefore they have been repositioned on the map shown in Fig. 3.3.

Due to the relatively low number of polymorphic SSLP markers in the region containing the *sne* mutation, indels were also used as markers to more accurately pinpoint the *sne* mutant locus. Indels are polymorphic insertions or deletions that were identified from the initial sequencing of the zebrafish genome, when DNA from ~ 1000 individual Tübingen zebrafish were pooled and used as the template for whole genome shotgun sequencing. They are usually found in repeat regions, are greater than 4bp and have been mapped to the zebrafish genome assembly (shown on a DAS (distributed annotation system) track in Zv5).

Prior to selection of the indel markers, the existing SSLP markers that encompassed the *sne* locus had to be mapped directly to the zebrafish Ensembl genome assembly. Unfortunately, when this was performed the assembly (Zv5) was still rudimentary. Many contigs were not joined and a number of markers mapped to more than one chromosome, reducing the efficiency of finding closely linked markers. Nevertheless, the markers that were identified ultimately proved to be useful in determining the site of the mutation.

To ensure that any polymorphic indel size differences could be detected on agarose gels, primers were designed to amplify indels that were either repetitive sequences or were greater than 20bp in length (see appendix, Table 1). Out of 79 indel markers identified, seven were found to be polymorphic between the Tübingen and TL strains (Fig. 3.4). Six indel markers were used to assist in the mapping process along with seven SSLP markers. PCR amplification of these markers was performed on a maximum of 741 mutant embryos and the recombination frequency

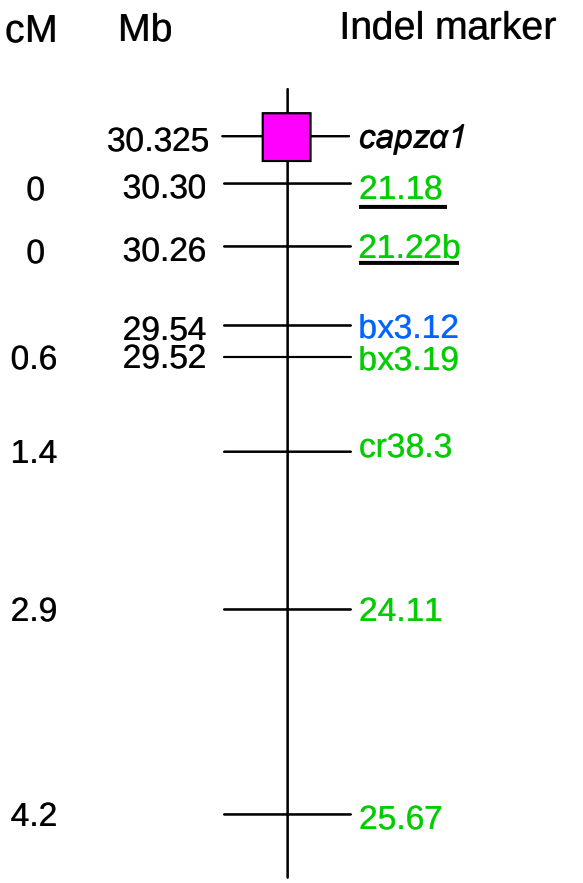

Fig. 3.4. Map of indel markers on chromosome 8. Indel markers were mapped according to their location established from the Ensembl genome assembly (Zv6). All markers except bx3.12 (labelled in blue) were tested on individual mutant embryos. No recombination events were identified at markers 21.18 and 21.22b (underlined). I was unable to determine the location of cr38.3, 24.11 and 25.67 due to discrepancies between the assembly and my analysis.

determined (Table 3.1). No recombination events were detected at indel markers 21.22b and 21.18 in all the mutant embryos tested. The two nearest SSLP markers that flanked the indel markers (Z13565 and Z26520) spanned a region of 1 cM (Z14573 was omitted as not all mutant embryos had been tested for recombination events at this marker). Mapping of Z13565 and Z26520 onto the Ensembl genome assembly revealed that this 1 cM interval corresponded to 0.7Mb.

Having mapped the *sne* mutation to as small an interval as possible, a candidate gene approach was taken to identify which of the three genes present in this region carried the causative mutation: *wnt2bb*, *capzα1*, and a hypothetical gene (orthologous to *cortactin-binding protein 2 N-terminal-like protein* (*cttnbp2*) in mouse and human) (Fig. 3.5). The most likely candidate for the *sne* locus appeared to be *capzα1*. This gene encodes an important muscle component and both the closest indel markers (21.22b and 21.18) were positioned within the intronic region. For this reason, *capzα1* was sequenced first.

3.4 The sne locus is *capzα1*

The coding region of *capzα1* was initially cloned from RT-PCR products derived from pooled mutant and sibling cDNA. The cDNA was amplified by two overlapping sets of primers and covered the 5′ and 3′ regions of the gene (Fig. 3.6, see appendix, Table 2 for primer sequence). Surprisingly, from *sne* mutant pooled cDNA, two PCR products were amplified from the 3′ region of *capzα1*. This indicated that a mutation in this region may be producing aberrant splice transcripts (Fig. 3.7). The PCR was repeated using cDNA from individual wild-type sibling and mutant 5 day old embryos (Fig. 3.8A). Three PCR products were amplified from

Table 3.1. Table of polymorphic SSLP and indel markers used to define the region of the *sne* locus. The distance of each marker locus from the mutation locus was calculated by dividing the number of recombinations that have occurred by the number of meioses and the multiplying this fraction by100.

Markers	*Z21115*	*25.67*	*24.11*	*cr38.3*	*bx3.19*	*Z13565*	*Z14573*	*21.22b*	*21.18*	*Z26520*	*Z10546*	*Z60737*	*Z21315*
Mutant embryos with recombinant markers	47	12	32	3	6	7	0	0	0	7	3	7	11
Total number of mutant fish tested	185	142	545	104	474	738	193	687	715	741	285	473	283
Distance from mutation in cM	12.7	4.2	2.9	1.4	0.6	0.5	0.0	0.0	0.0	0.5	0.5	0.7	1.9

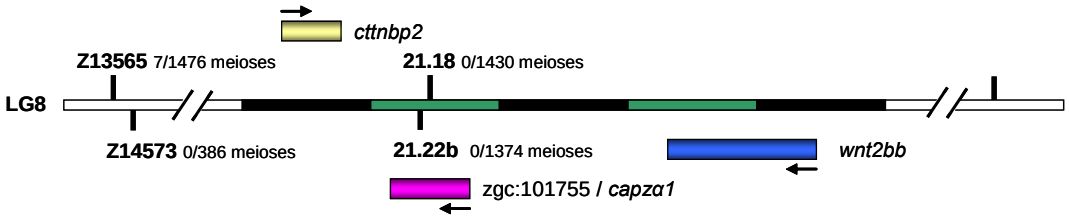

Fig. 3.5. Schematic diagram of the genomic region on chromosome 8 containing the three candidate genes for the *sne* locus. Each green or black bar represents 0.02Mb. Arrows indicate the orientation of the genes with respect to transcription.

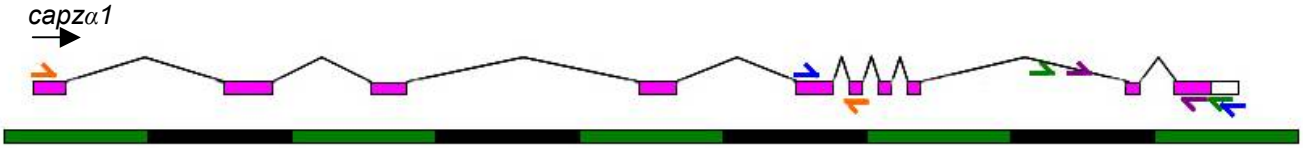

Fig. 3.6. Coding region of *capzα1* exons illustrating the position of the primer binding sites used to amplify this gene. Primers that amplified *capzα1* cDNA are represented by orange arrows (primer pair 2) and blue arrows (primer pair 4). Nested primers that amplified genomic DNA for sequencing are represented in green (fact3) and purple (fact2).The 3′ untranslated region is illustrated in white. Each green or black bar represents 2kb.

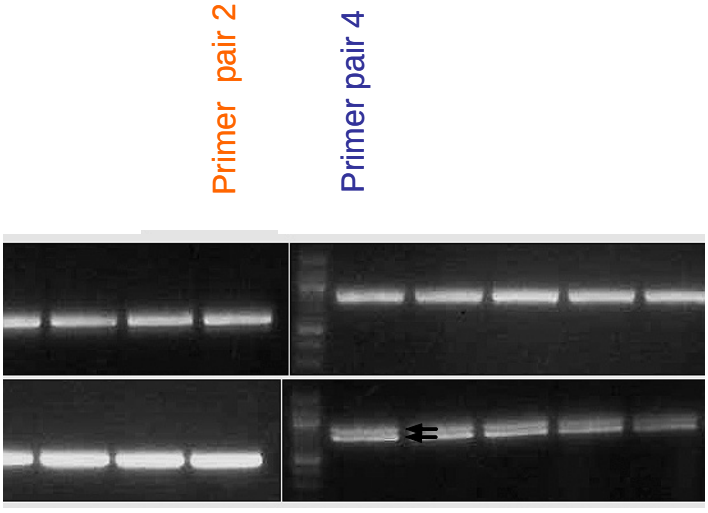

Fig. 3.7. Gel of RT-PCR products from pooled *sne* sibling and mutant cDNA, using primers that amplified the 5′ and 3′ prime regions of *capzα1*. Arrows indicate the two PCR products amplified from *sne* mutant cDNA. See Fig. 3.6 for position of the primers and the appendix, Table 2 for primer sequence.

mutant embryos, but not wild-type siblings. All three products were cloned into TOPO blunt II® vectors and sequenced. This revealed that three aberrant splice transcripts were expressed in the mutant (Fig 3.8B). In the 750bp product exon 9 had been skipped. In the 900bp and 950bp product, 24bp and 46bp of intron 9 (respectively) were retained in the mRNA. *In silico* translation of these products indicated that the 950bp product would produce a frame shift, which would result in a premature stop codon 19 amino acids into the translation of the intronic region and the first part of exon 10. Translation of the 750bp and 900bp product would result in a 21 amino acid deletion (loss of exon 9) and an 8 amino acid insertion from residue 240 (between exon 9 and 10) respectively (see appendix, Fig. 2. for the predicted protein sequence of all aberrant splice transcripts). Therefore, both these aberrant splice transcripts would still produce in-frame protein products and exon 10 would be translated.

The aberrant splice transcripts detected in the mutant indicated that there was a point mutation at the donor splice site of exon 9. This was confirmed by sequencing of the exon/intron boundary using nested primers that flanked exon 9 and 10 (Fig. 3.6). A single G-A base pair change was found at the exon 9 donor splice site (Fig 3.8C). RNA splicing depends on the recognition of pairs of splice junctions that flank each intron. The generic consensus at the 5′ donor splice site is GU, and at the 3′ acceptor splice site is AG. As the mutation in *capzα1* disrupts the 5′ donor splice site consensus sequence, it is highly likely that the splice site becomes unrecognizable to the splicing machinery, thus alternative donor splice sites are used. Indeed, two potential generic donor splice sites were detected 23bp and 47bp into intron 9 and correlate with the aberrant splice transcripts detected in the mutant (Fig. 3.9).

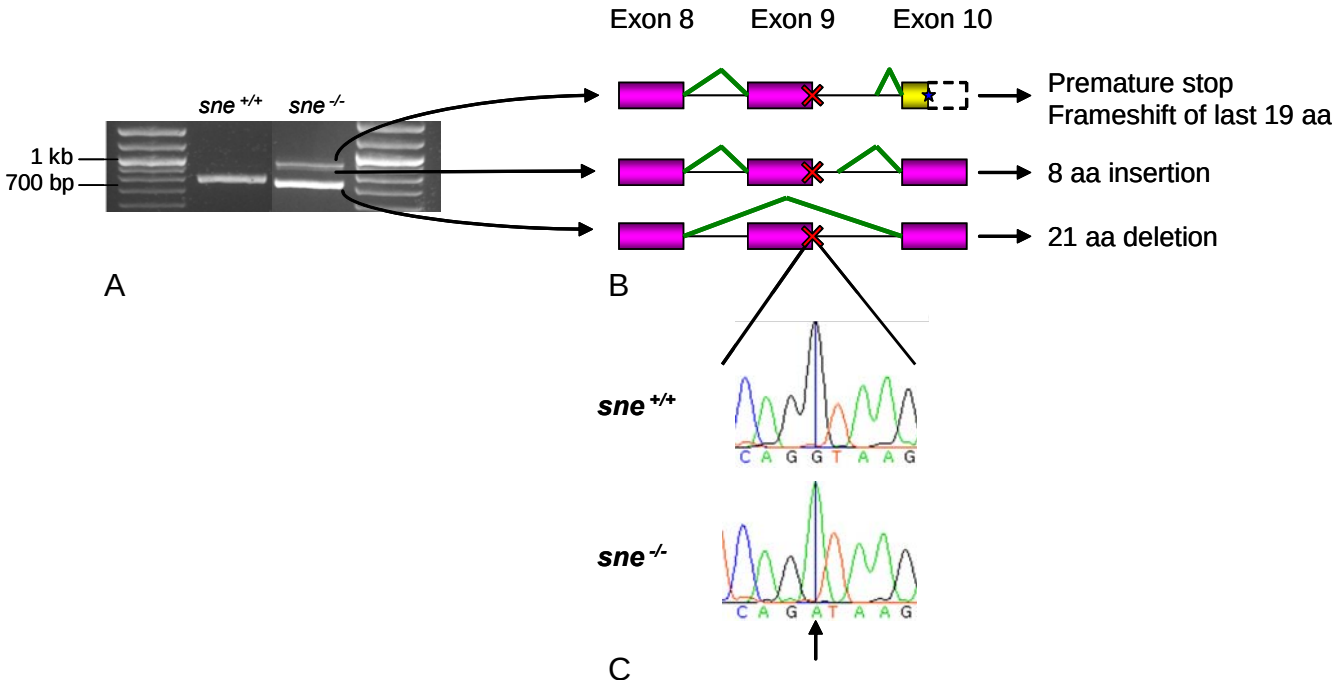

Fig. 3.8. Identification of the mutation in *capzα1* of the *sne* mutant. A) Gel of RT-PCR products from *capzα1* cDNA of individual *sne* mutant and wild-type sibling embryos. The three RT-PCR products that were amplified from *sne* mutant cDNA were cloned into TOPO Blunt II® vectors and sequenced. B) A schematic diagram illustrating the exons and introns at the 3′ end of *capzα1* that are aberrantly spliced in the *sne* mutant. The red crosses indicate the site of the mutation, the partially filled yellow exon indicates that it was translated out of frame and the blue star indicates a stop codon. C) Sequence traces of the exon 9 donor splice site revealed the base change in the mutant from a G to A (arrow).

Fig. 3.9. Position of alternative donor splice sites in intron 9 of *capzα1*. Two alternative donor splice sites were identified (underlined in blue) at positions that correspond to the aberrant splice transcripts expressed in the *sne* mutant (dashed line represents intronic sequences that are transcribed in the *sne* mutant). The normal splice site is highlighted in green and matches the vertebrate donor splice site consensus (5′-AGGUAAGU-3′). In both alternative splice sites the general donor splice site consensus is present (GU).

3.5 Discussion

The ENU muscle motility mutant *sne* was positionally cloned to a 1cM region using SSLP and indel markers. SSLPs are advantageous in mapping because they are co-dominant markers i.e. both alleles are equally detectable in heterozygotes, they are also abundant and widely distributed throughout the genome. Moreover, they can be assayed relatively easily by PCR and, coupled with the generation of the MGH SSLP marker map, it is possible to perform high throughput positional cloning and roughly map mutants derived from ENU screens with relatively ease (Geisler et al., 2007). The efficiency of SSLP mapping is limited, however, by a number of factors. Firstly, SSLP markers have to be polymorphic between the two strains used in the ENU screen. There is still a significant amount of allele sharing between strains (Knapik et al., 1998) so not all SSLP markers will be polymorphic. Secondly, two closely positioned markers that flank the mutation locus are ideally required. Thirdly, these markers need to be accurately positioned onto the genetic map i.e. the zebrafish genome assembly.

Within the large region that the *sne* locus was previously mapped to on linkage group 8 (51.6-62.3cM), nine SSLP markers were polymorphic (37.5%). Out of the nine linked markers, six were genetically mapped to the genome assembly and correctly corresponded to the positions that I had obtained from the recombination data. As it was very difficult to accurately define the region containing the mutation using only SSLP markers, indel markers were also selected to assist in postional cloning of the *sne* mutant. Although less than 10% of the indel markers were polymorphic, they were crucial to placing the SSLP markers on to the genome assembly and in determining the likely candidate gene to contain the mutation. The low polymorphic differences of the indel markers I tested may be due to the fact that the indels in Ensembl were derived from

differences between individual Tübingen fish, which is one of the strains that was used to generate the *sne* mutant.

A mutation at the exon 9 donor splice site of the primary candidate gene, *capzα1*, was identified by sequencing. RT-PCR of *capzα1* cDNA products from *sne* mutants indicated that the mutation induces mis-splicing of the *capzα1* transcript. The aberrant splicing produces three *capzα1* transcripts: 1) exon 9 is completely spliced out of the transcript, 2) 24bp of intron 9 is included in the transcript, 3) 46bp of intron 9 is included in the transcript. This finding indicates that the mutation disrupts the exon 9 donor splice site causing the use of alternative donor splice sites within intron 9, resulting in the production of three different transcripts. *In silico* analysis of two of the transcripts predicted partial translation of the intron, however, only one of these transcripts encoded a premature stop. It remains to be determined whether all the aberrantly spliced transcripts are equally abundant, and whether any of the transcripts are translated. Quantitative RT-PCR or Northern blotting may be useful in determining the ratio of expression levels between the mis-spliced *capzα1* transcripts. If the three transcripts are translated they may be detectable by Western blotting, however, due to the small differences in the predicted sizes and isoelectric points of the protein products, (exon 9 deletion: 30.3kDa, p I5.5, 24bp insert: 33.6kDa, pI 5.4 and 46bp insert: 29.5 kDa, pI 4.2) it may be difficult to separate them, even on a two dimensional gel.