Chapter 6

Comparative Sequence Analysis

Between Human and Zebrafish

6.1 Introduction

The identification of human genes and their orthologous counterparts is greatly facilitated by the generation of genomic sequences across the syntenic regions in model organisms (as discussed in previous chapters). As with genomes of other vertebrates, the gene complement of the zebrafish is also expected to show extensive similarity to that of man, thus assisting the annotation of the majority of human genes. The zebrafish genome is approximately 1.7 Gb in size and is divided into twenty-five linkage groups or chromosomes. Recent studies to place zebrafish ESTs onto linkage groups by RH mapping have shown that there is extensive synteny between the human and zebrafish genomes. Pairs of genes in the same region of the human genome are being observed in the same region of the zebrafish genome (Barbazuk, W. B.*, et al.*, 2000; Gates, M. A.*, et al.*, 1999; Postlethwait, J. H.*, et al.*, 1998). However, little is known about how the distances between these pairs of genes differs between human and zebrafish Given that the zebrafish genome is approximately half to two thirds smaller than the human genome, the distance between genes may be smaller in the zebrafish than in the human.

In addition to organisms such as the mouse, the fly and the frog, the zebrafish is one of the organisms of choice in developmental biology, as it is easy to keep, has a short generation time and produces conveniently transparent embryos (Metscher, B. D.*, et al.*, 1999). Humans and zebrafish are thought to share a common ancestor, probably a bony fish, which existed approximately 400 million years ago compared to the estimate of 70 million years since the divergence of man and mouse (O'Brien, S. J.*, et al.*, 1999). Therefore it is expected that the extent of synteny between human and

zebrafish is less than for human and mouse given the greater time available for gross chromosomal rearrangements in the respective genomes. Current estimates suggest that there are greater than 1000 homology segments between the human and zebrafish genomes (see Figure 6.1, Johnson unpublished), which compares to 200 segments seen between human and mouse (Hudson, T. J.*, et al.*, 2001). There are currently 12 separate segments showing homology to the human X chromosome on eight different zebrafish linkage groups (indicated with arrows on Figure 6.1). In this chapter the region between genes HPR6.6 and ZNF-Kaiso in human has been targeted for investigation in the zebrafish, in order to further the understanding of the syntenic relationship between the region of interest in human, mouse and zebrafish, and to identify novel orthologous genes in the zebrafish.

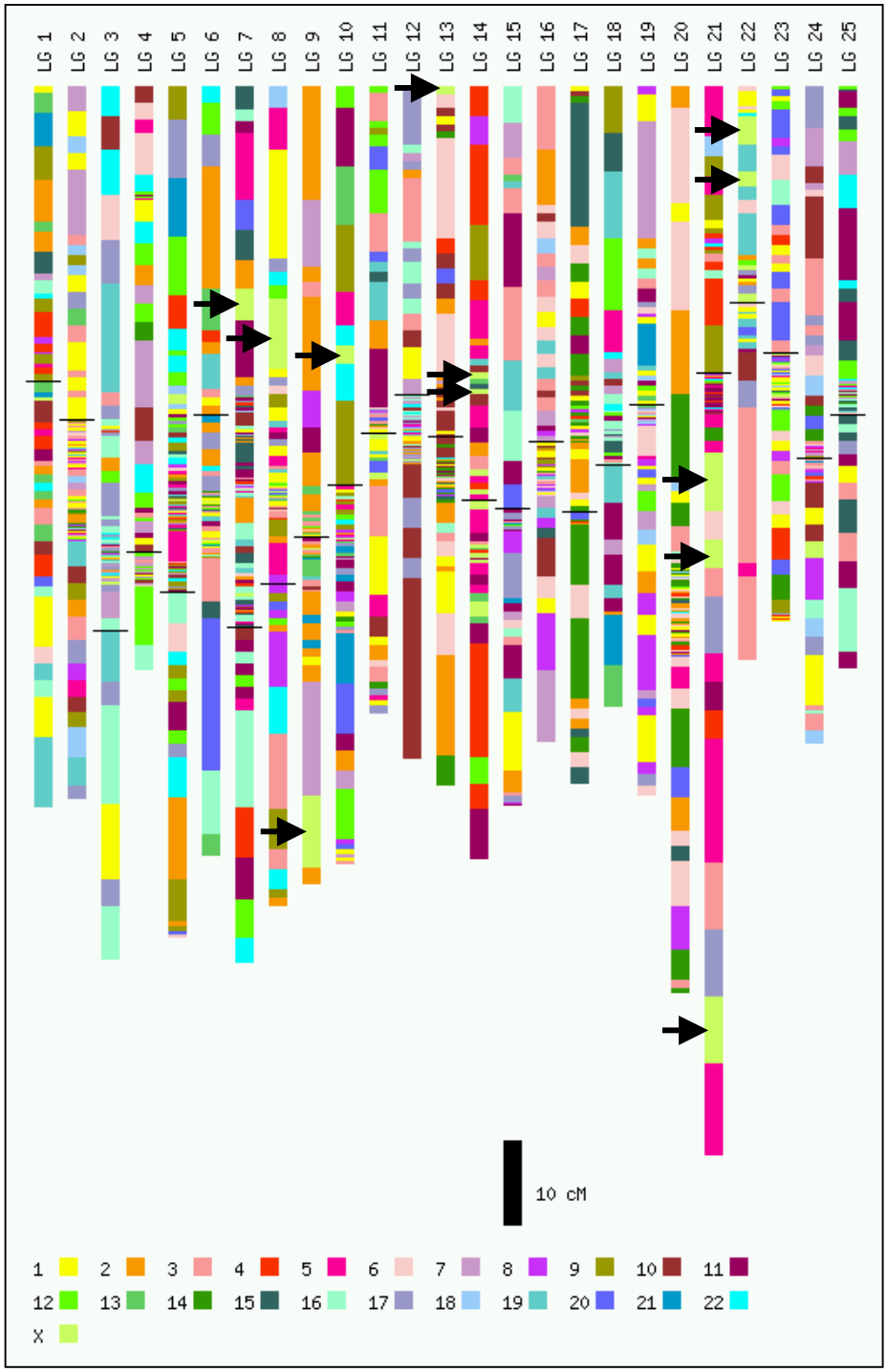


Figure 6.1: *Synteny between human and zebrafish (courtesy of Steve Johnson). The 25 linkage groups (LG) of the zebrafish genome are represented and coloured depending on the positioning of zebrafish-specific ESTs that significantly match human genes. Each colour represents a different human chromosome. The arrows indicate the position of the 11 regions showing synteny to the human X chromosome.*

RESULTS

6.2 Identification of zebrafish genomic clones

In the previous chapter, the strategy for mouse bacterial clone isolation relied upon the knowledge of the sequences of a number of orthologous pairs of genes from which STSs specific to mouse sequences could be designed and used for library screening. At the time this project began, there were no zebrafish sequences known that were orthologous to the human genes in the region between HPR6.6 and ZNF-Kaiso. Therefore a strategy for clone isolation was designed based on using human probes to isolate zebrafish clones by reduced-stringency hybridisation. Fifteen primer pairs were designed within a single exon of sixteen out of of the eighteen genes in the region (as discussed in Section 4.3.4, dJ525N14.CX.1 and bG421I3.CX.1 are 99% identical and a single STS was designed that represented both genes – thus fifteen primer pairs for sixteen genes). No primer pair was designed for two genes, dJ555N2.CX.1 and dJ525N14.CX.3 as these were identified since the clone isolation was carried out. Each STS was labelled and hybridised individually to filters of the zebrafish BAC library (RPCI-71) at 50^{o}C for 16 hours.

A series of washes of increasing stringency (see Section 2.17.3) were carried out. Initial washing was carried out at 50 ^{o}C in 6x SSC, 1% Sarkosyl for 2x 30 minutes, and stringency was increased by decreasing the amount of SSC in subsequent wash solutions (4x SSC, 2x SCC, 1x SSC, all with 1% Sarkosyl, and at 50^{o}C for 2x 30 mins). The washing continued until the number of counts present on each filter (measured using a Geiger Counter held to a single filter) dropped below 5 counts per

second. When this point was reached, it was assumed that non-specific binding of probe to the filters had been removed and any probe still bound would potentially represent a sequence-specific positive signal. An X-ray film was exposed to the filters for 36 hours at room temperature. A summary of the results is shown in Table 6.1, columns 1-4. Column 3 shows that for different probes, the filters were washed to different stringencies. For the probe derived from the human UPF3B gene the filters were washed to 4xSSC, whereas for the probes derived from ANT2, UBE2A, RPL39 and NDUFA1, the filters were washed to 1xSSC.

Table 6.1: *Summary of Bacterial Clone Isolation*

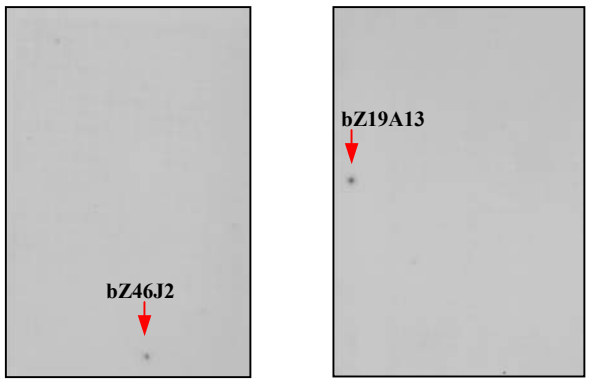
Gene	STS	Wash stringency	No. of clones identified	Clones in ctgs by fingerprinting	No. of ctgs	Sequence clone
HPR6.6	stdJ555N2.2	2xSSC	4*	-		-
dJ555N2.CX.1	-	-	-	-	-	-
dJ1139I1.CX.1	stdJ1139I1.6	2xSSC	4	4	1	bZ21D15
ANT2	stdJ404F18.4	1xSSC	2	2	1	bZ46J2
dJ876A24.CX.1	stdJ404F18.5	2xSSC	5	5	2	bZ80I7 bZ3C13
UBE2A	stdJ876A24.17	1xSSC	3	3	1	bZ46J2
dJ876A24.CX.3	stdJ876A24.16	2xSSC	2	2	0	bZ10G3 bZ20I5
SEPTIN2	stdJ876A24.11	-	-	-	-	-
RPL39	stbK38K21.3	1xSSC	1	1	0	bZ74M9
UPF3B	stdJ327A19.10	4xSSC	3	3	1	bZ79P20
ZNF183	stdJ327A19.12	-	-	-	-	-
NDUFA1	stdJ327A19.13	1xSSC	2	2	1	bZ36D5
dJ327A19.CX.3	stdJ327A19.11	2xSSC	2	2	0	bZ18K17 bZ74M9
bG421I3.CX.2	stbG421I3.4	2xSSC	7	2	1	bZ5O12
dJ525N14.CX.1 bG421I3.CX.1	stbG421I3.5	2xSSC	3	0	0	bZ30I22 bZ71M17 bZ74M9
dJ525N14.CX.3	-	-	-	-	-	-
dJ525N14.CX.4	stdJ525N14.11	2xSSC	1	1	0	bZ39A15

* no clone chosen for sequencing

Thirteen of the fifteen STSs identified a total of 33 positive clones which were assembled into eight contigs by *Hin*d III restriction digest fingerprinting (see Section 2.12.3). A summary of the contigs is given in Table 6.1, columns 5 and 6. On average, one STS identified 2.5 clones which approximately agrees with the estimate that the RPCI-71 library contains three equivalents of the zebrafish genome (RPCI-71 – see http://www.chori.org/bacpac). An example of bacterial clone contig construction using an STS designed to the human ANT2 gene is shown in Figure 6.2. The probe derived from the ANT2 gene identified two clones bZ46J2 and bZ19A13 which when fingerprinted assembled into one contig. Early evidence that two genes, ANT2 and UBE2A, were closely linked in the zebrafish genome came from the fact that the probe derived from UBE2A identified the same clones bZ46J2 and bZ19A13, along with a third clone bZ11A23 which assembled into the same contig by fingerprinting (see Figure 6.2b). This provided supporting evidence that the hybridisation method was identifying sequence-specific signals. A second example of this was seen for two STSs, designed to the genes RPL39 and dJ327A19.CX.3, both of which identified bZ74M9 among other clones (data not shown).

A total of eight clones were identified for sequencing from the eight contigs constructed by fingerprinting (see Table 6.1, column 7). A further six clones were identified for sequencing (shown in red in Table 6.1), from cases in which the STS either identified only one clone, or identified two or three clones that did not show any significant overlap by fingerprinting. For instance, the STS designed to both dJ525N14.CX.1 and bG421I1.CX.1 identified three clones that showed no significant overlap by fingerprinting and all three clones were selected for sequencing.

(a



(b

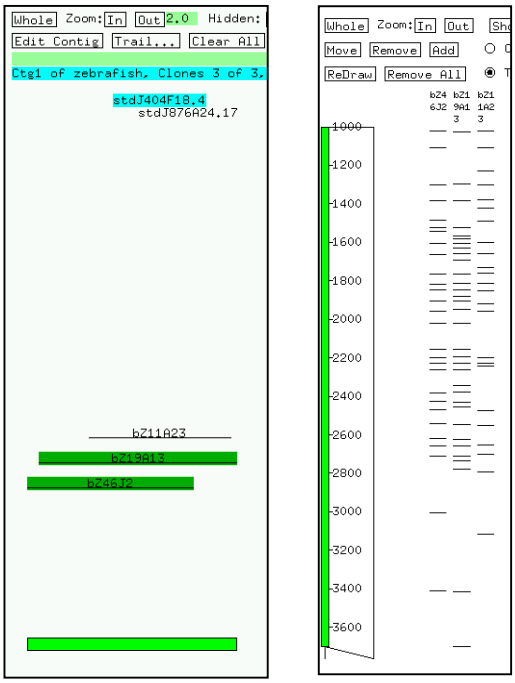


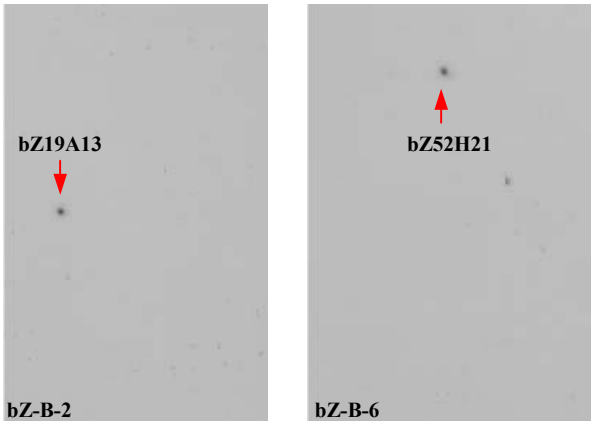
Figure 6.2: *BAC isolation by reduced stringency hybridisation. (a) An example of positive clones when an STS designed within an exon of the human ANT2 gene was hybridised to the zebrafish BAC library. The clones bZ46J2 and bZ19A13 were identified and (b) were assembled into a single contig by Hind III fingerprinting. Both clones along with bZ11A23 (shown un-highlighted) were identified by an STS designed to UBE2A. The fingerprints of all three clones are also given.*

6.3 Evaluation of strategy for the identification of orthologous genes

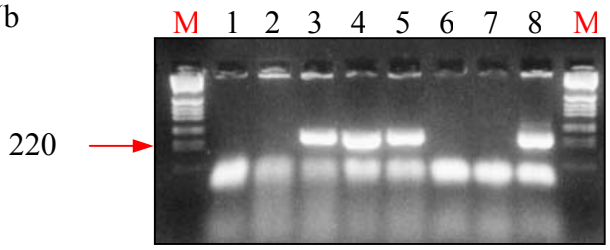
Analysis of the genomic sequence of fourteen BAC clones by BLAST revealed that only two contained potential orthologous genes. bZ46J2, detected with STSs from UBE2A and ANT2, and bZ74M9, detected with STSs from RPL39 and dJ327A19.CX.3. The remaining 12 BACs did not contain any sequences orthologous to human sequence, and appeared to be false positives identified during the hybridisation procedure. Even though stbG421I3.5 identified bZ74M9, no orthologous sequence was present for this STS. Therefore, it appears that bZ74M9 was identified as a false positive for stbG421I3.5. There was no obvious difference between the signal intensity of real positives versus the false positives. However, analysis of the washing stringency showed that the filters containing the real positive clones were washed to a higher stringency (1xSSC at 50^{o}C) when compared to the false positive clones (greater than or equal to 2xSSC at 50^{o}C).

In order to determine whether increasing the washing stringency could increase the sequence-specificity of detection by hybridisation, multiple filters containing DNA from both the false positive clones and the two real positives were generated. A pooled probe of nine STSs representing ten genes (two STSs that gave real positives and seven STSs that gave false positives) was hybridised to the filters for 16 hours at 50^{o}C. The filters were then washed in steps with increasing stringency, from 4xSSC to 0.1xSSC, at 50^{o}C. After each wash step, one of the filters was removed and stored in 2xSCC at room temperature. The results after exposing the filters to X-ray film for 36 hours at room temperature (the same exposure time as was originally used for the clone isolation) are shown in Figure 6.3.

(a

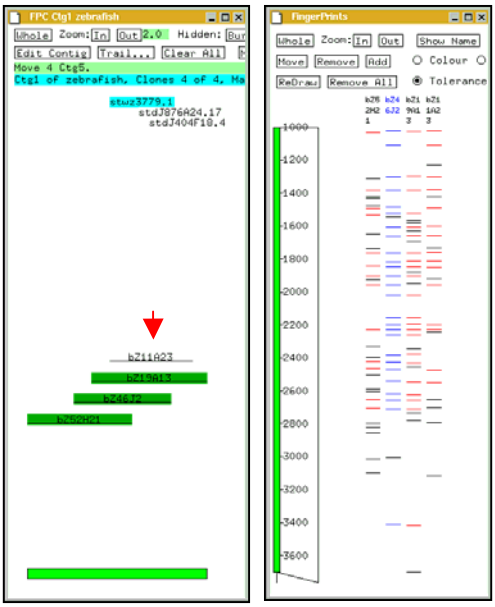


(b



1. bZ67m17
2. bZ74M9
3. bZ46J2
4. bZ52H21

(c



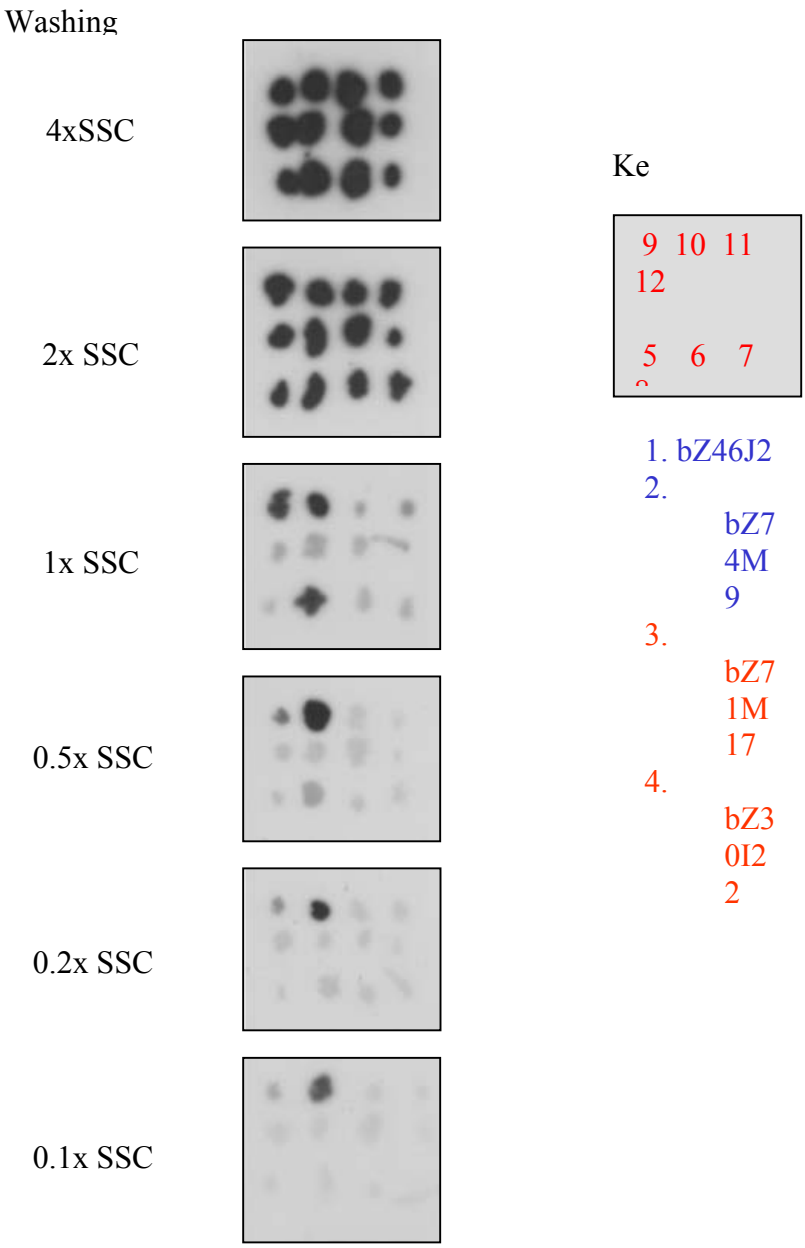
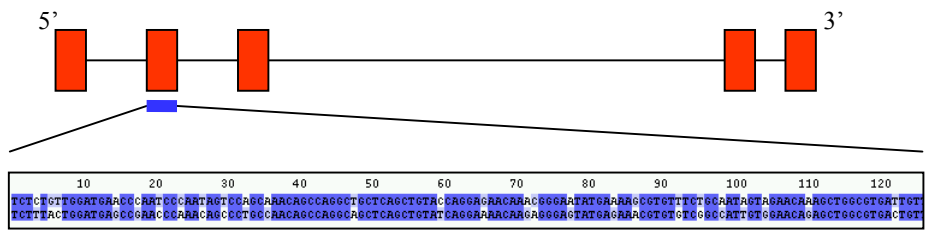


Figure 6.3: *Evaluation of false positives. A pooled probe containing nine STSs was hybridised to six copies of a filter with two true positives (1 and 2) and ten false positives (3-12). The filters were washed with increasing stringency from 4xSSC at 50°C to 0.1xSSC at 50°C. The signal strength remains even for all clones until at 1xSSC only the two true positives, bZ46J2 and bZ74M9, and one false positive, bZ36D5 remain. As the washing stringency continues to increase, the signal of the remaining positives reduces.*

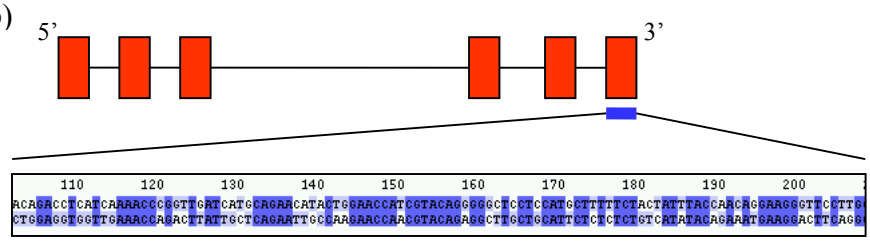
It can be seen that the signals for the real positive clones are still present after washing at 1xSSC, whereas the signal for the false positive clones has all but been removed. One of the clones identified as being a false positive, bZ36D5 was still showing a significant signal even after washing at 1xSSC. Comparison of the sequence of the STS that was designed to the gene NDUFA1, with bZ36D5 by BLAST (Altschul, S. F.*, et al.*, 1990) showed that there was a region of 30 bp that was 75% identical between the two sequences. This would be sufficient to account for the apparent sequence-specific signal observed at a wash stringency of 1x SSC (Eric Green, personal communication). Further increasing the stringency of washing by the use of 0.5x SSC shows that the signal from bZ36D5 is removed, but that the signal from the true positive bZ46J2 is also significantly reduced. These results show that washing to a stringency of 1xSSC at 50^{o}C should reduce but not completely eradicate the number of false positive clones in this type of experiment.

Analysis of the two real positive clones (bZ46J2 and bZ74M9) by BLAST showed that they also contained sequences orthologous to four other genes not previously detected by the reduced-stringency hybridisation method. These are dJ1139I1.CX.1, dJ876A24.CX.3, UPF3B and NDUFA1, and the previous negative hybridisation results therefore appear to be false. Analysis of the level of identity between the genomic sequences of the human and the zebrafish in the region of each STS showed that this was higher (above 75%) for those STSs that detected the presence of the orthologous gene by hybridisation, than those that failed to do so (less than 60%) (see Figure 6.4).

(a



(b)



(c

Gene	% Identity	Probe Length
UBE2A	79.2	125
ANT2	80.4	256
RPL39	81.5	81
dJ327A19.CX.3	80.8	99
dJ1139I1.CX.1	57.0	105
dJ876A24.CX.3	48.5	101
dJ327A19.CX.1	not matched in sequence	94
NDUFA1	51.4	107

Figure 6.4: *Evaluation of false negatives. (a) Position of the STS (blue bar) designed to human UBE2A (coding exons shown as red bars linked by black lines). The alignment of the human and zebrafish genomic sequence shows the two regions are 79.2% identical over 125 bp. (b) Position of the STS (blue bar) designed to human dJ1139I1.CX.1 (coding exons shown as red bars linked by black lines). The alignment of the human and zebrafish genomic sequence shows the two regions are 57% identical over 105 bp. (c) A table showing the percentage identity between the sequence in human and zebrafish for each STS and the length for the four true positives (names shown in red) and the four false negatives (names shown in blue). The two sequences for each STS are greater than 75% identical for the true positives, and less than 60% identical for the false negatives.*

In summary the technique described here for the identification of zebrafish orthologues of human genes in BAC clones is able to detect the presence of the orthologue in some instances, but the technique is dependent upon the sequence similarity between the probe and the genomic sequence being greater than approximately 75% identical. A reduction in the stringency might reduce the false negative level but would also result in a significant rise in the false positive rate.

6.4. Identification of BAC clones using orthologous zebrafish EST sequence.

At this time, a more detailed radiation hybrid (RH) map of the zebrafish genome, containing the locations of a large number of zebrafish EST sequences was made available (courtesy of Steve Johnson, Washington University, St Louis). Analysis of the region of interest between HPR6.6 and ZNF-Kaiso, showed that two genes appeared to be orthologous to zebrafish EST sequences. dJ876A24.CX.3 matched to EST wz3779 and dJ327A19.CX.3 matched to EST wz8217 (information for each zebrafish EST can be obtained from http://www.genetics.wustl.edu/fish_lab/cgi-bin/display.cgi). These ESTs were positioned at the same point on zebrafish linkage group (LG) fourteen at 14:56 centiRays (cR). A further twenty-one zebrafish ESTs have been mapped to this position in the RH map, but comparison of these sequences with the other human genes in the region by BLAST revealed no other significant matches (data not shown). For the genes HPR6.6 and ZNF183, the potential orthologous zebrafish EST sequences, have been mapped to LG20 and LG7 respectively.

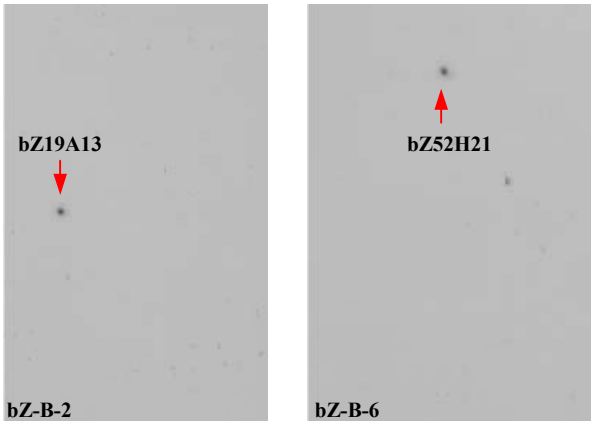
The identification of two zebrafish EST sequences orthologous to the human genes dJ876A24.CX.1 and dJ327A19.CX.3 enabled the use of the method described in chapter 5 for the identification of zebrafish BAC clones (see Figure 6.5). A probe from each zebrafish gene was produced by amplification for the PCR using primer pairs predicted to be within a single exon (based on an alignment of human and zebrafish sequences). The probes were hybridised as a pool to the zebrafish BAC library (RPCI-71). In this case, a total of six BACs were obtained by hybridisation, of which all six were confirmed by PCR. As expected the STSs mapped to the two contigs previously shown in Section 6.3 to contain the zebrafish orthologues of dJ876A24.CX.3 and dJ327A19.CX.3. Three additional clones, bZ52H21, bZ67M17, and bZ62I22 were identified by this method as opposed to the reduced stringency hybridisation method and incorporated into the contigs by fingerprinting. An example of one contig can be seen in Figure 6.5c.

It has been suggested that large regions of the ancestral zebrafish genome may have undergone either total or partial genome duplications (Barbazuk, W. B.*, et al.*, 2000). These two zebrafish-specific STSs were used as probes to identify bacterial clones that assembled into two contigs, one contig containing all the clones positive for one STS, the other contig containing all those positive for the other STS. The zebrafish EST sequences to which the two STSs were designed, have been positioned at LG14:56. There is no evidence from these data that the region containing the two zebrafish EST sequences, syntenic to the region in human containing dJ876A24.CX.3 and dJ327A19.CX.3, is present more than once in the zebrafish genome. However, it is still possible that a duplication of the region has taken place but that the sequence of one copy has diverged sufficiently so as not to be detected by the method described.

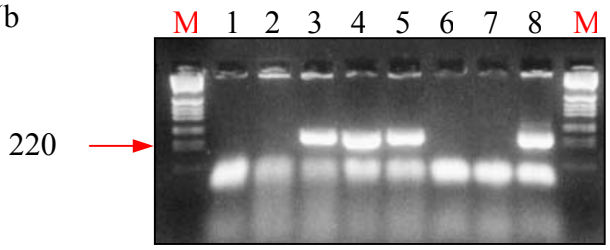
Analysis of the complete sequence of the zebrafish genome will enable a more detailed study of the region and an exhaustive search for other homologous regions at a lower stringency for evidence of a possible duplication in the zebrafish that was followed by substantial sequence divergence.

Figure 6.5: *(see over) Identification of BAC clones using an STS designed to the zebrafish EST wz3779 (a) Autoradiograph of two of the six filters (bZ-B-2 and bZ-B-6) after hybridisation of a pool of two STSs, stwz3779.1 and stwz8217.1, and washing to 0.5xSSC at 65°C, showing two positives bZ19A13 and bZ52H21. (b) Colony PCR of the positives from the hybridisation with the stwz3779.1 showing the positive clones bZ46J2, bZ19A13 and bZ52H21. The clones in lanes 1, 2, and 6 were shown to be positive with stwz8217.1. M = Marker. (c) FPC diagram showing the clones identified with stwz3779.1 assembled together by fingerprinting (highlighted in green). One other clone is also present in the contig. bZ11A23 was identified with an STS designed to the human UBE2A (indicated by a red arrow). The fingerprints of the clones are also shown. Bands in the fingerprint for bZ52H21 matching other bands in other lanes are shown in blue, and those matching bands in the other lanes are shown in red.*

(a

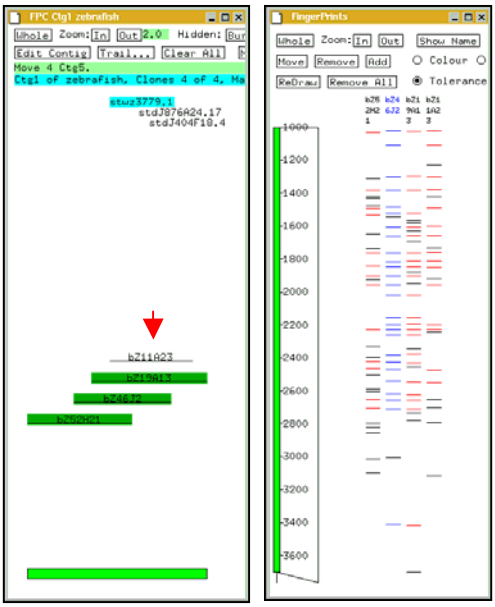


(b



1. bZ67m17
2. bZ74M9
3. bZ46J2
4. bZ52H21

(c



6.5 Sequence Analysis

As discussed in Sections 6.3 and 6.4, bZ46J2 and bZ74M9 were shown by BLAST to contain eight sequences, identified as being orthologous to genes in the region between HPR6.6 and ZNF-Kaiso in human. A more complete analysis of the sequence of bZ46J2 and bZ74M9 has been carried out using a combination of sequence similarity searches and *de novo* gene prediction. The analysis identified a total of twelve predicted genes (see Figure 6.6), but no cDNA-based experimental confirmation of gene structures was carried out due to the lack of available zebrafish cDNA resources at the time.

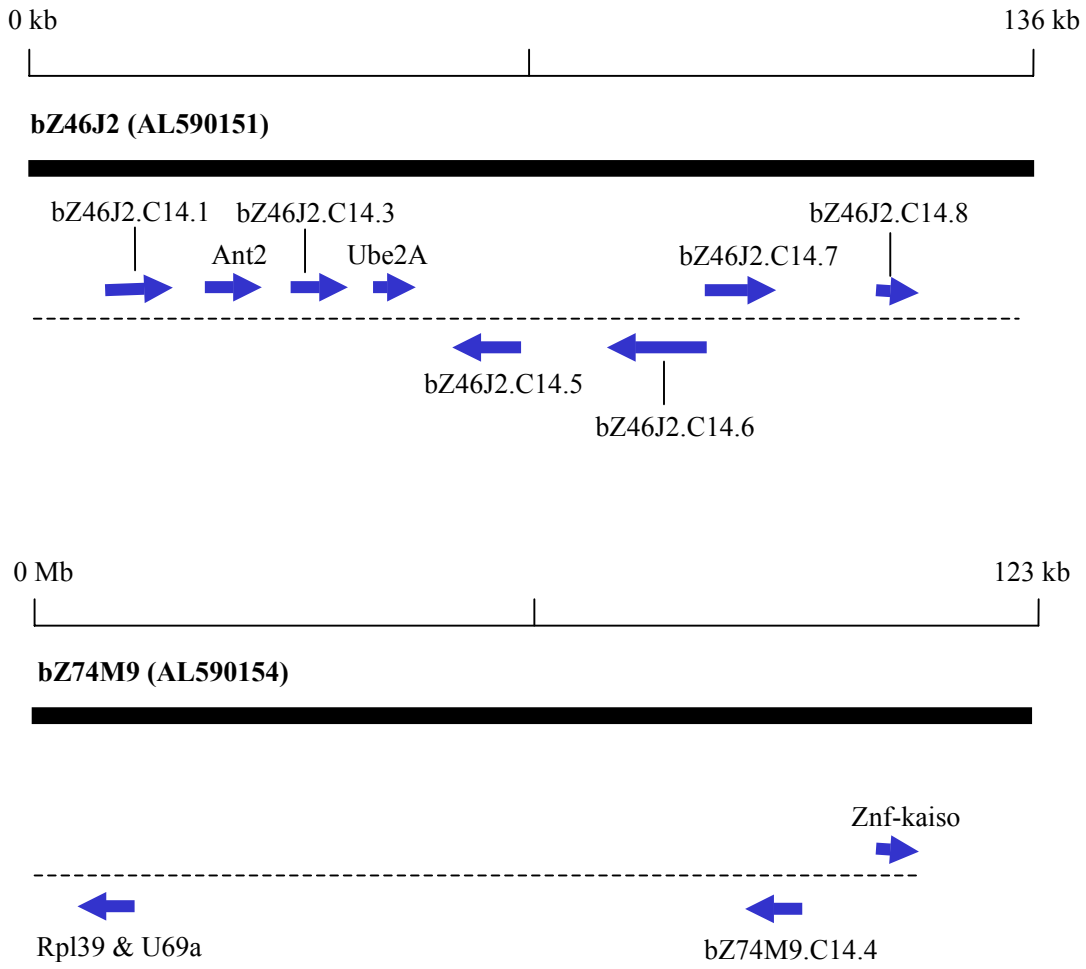


Figure 6.6: *Summary of the gene map constructed in zebrafish. The black bars indicate the finished sequence of the two clones analysed with the accession numbers in brackets. A scale is given in kilobase pairs (kb). Predicted genes are indicated by blue arrows, the direction of each arrow reflects the direction of transcription. Genes on the plus strand are positioned above the dotted line, genes on the minus strand are positioned below the dotted line.*

Comparison of the twelve predicted zebrafish genes with the genes in human between HPR6.6 and ZNF-Kaiso, showed that eight were newly identified orthologous pairs, based on their position, similarity at the nucleotide and protein level and similar gene structures (see Figure 6.7 – see also appendix at the end of this chapter for comparison of all eight orthologous genes in human, mouse and zebrafish). For instance, the ANT2 gene in human and mouse has been compared to the newly identified zebrafish orthologue (see Figure 6.8). There is good conservation between the sizes of exons, but less conservation in the sizes of introns. In general, the intron sizes are smaller in mouse and zebrafish along with the distances between genes, which may reflect the differences in the size of the respective genomes (3 Gb in human and mouse, and 1.7 Gb in zebrafish).

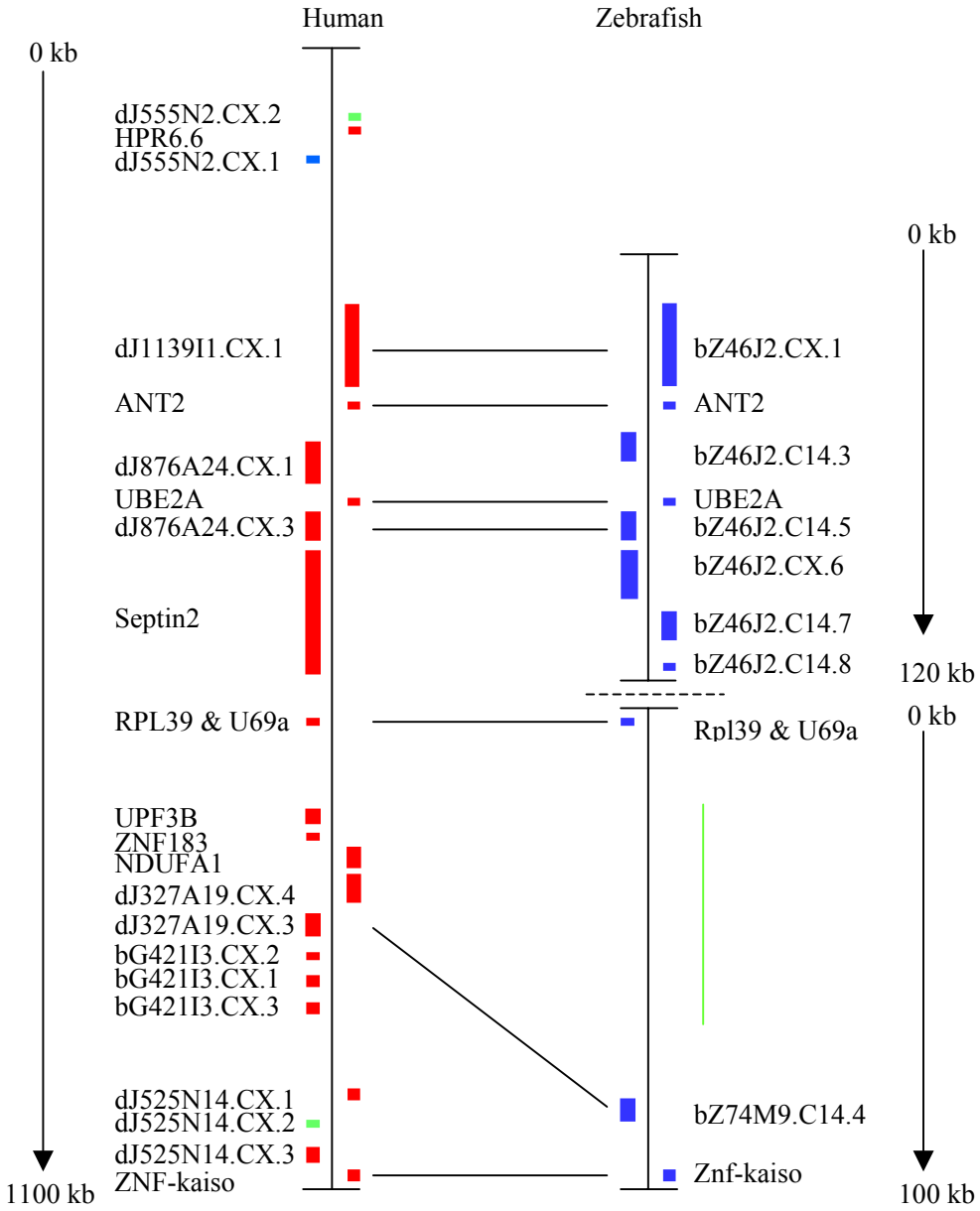
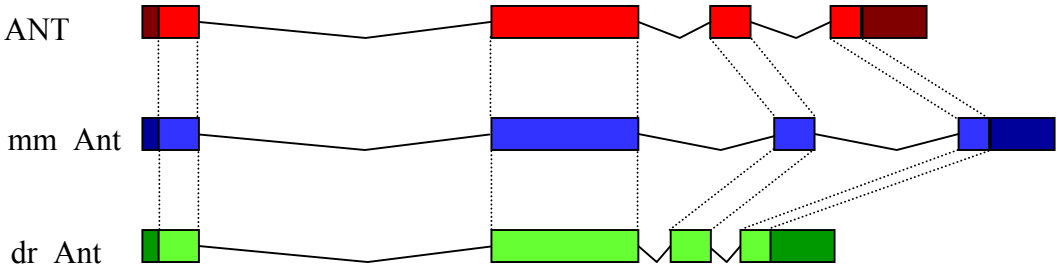


Figure 6.7: *Comparison of the genes identified in zebrafish (on the right) with the genes in the region of interest between HPR6.6 and ZNF-Kaiso in human (on the left). A vertical bar represents the extent of the sequence and genes are shown as bars (red = genes confirmed by cDNA, blue = predicted genes, green = pseudogene). Horizontal black lines link predicted orthologous pairs. A vertical green line indicates the region containing 6 direct repeats. The size of each region is indicated and suggests a tighter clustering of genes in zebrafish than was observed in human.*

(a



(b

Gene Name	ANT2	mm_ANT2	dr_ANT2
5'UTR	70	85	87
exon 1	111	111	111
2	487	487	487
3	141	141	141
4	158	158	158
3'UTR	258	263	305
Total Coding	897	897	897
Total cDNA size	1225	1245	984
intron 1	1034	1001	1015
2	225	406	80
3	387	510	84
Total Intron size	1646	1917	1179
Genomic coverage	2871	3162	2163

(c

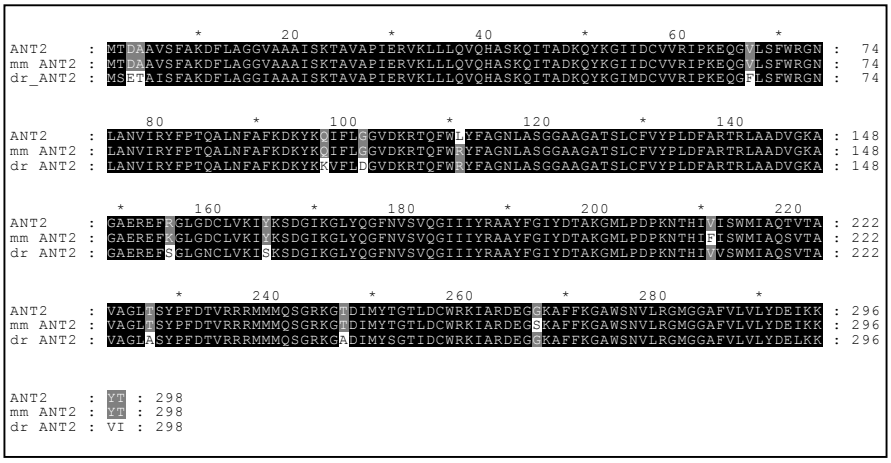


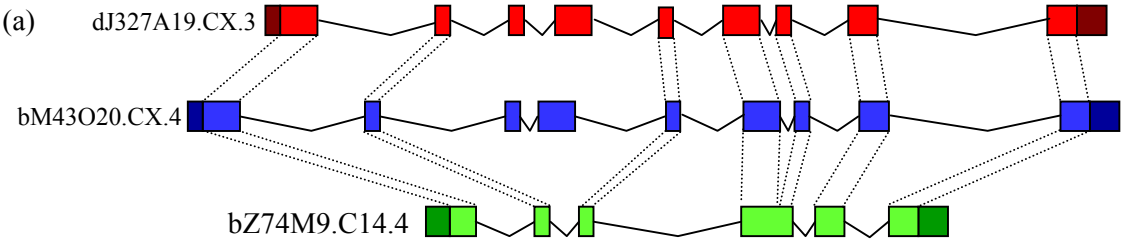
Figure 6.8: *Analysis of orthologues in human, mouse and zebrafish (1) (a) A schematic representation of the ANT2 gene (exons are shown as bars, introns by v-shaped lines) in human (red) mouse (blue) and zebrafish (green). Untranslated regions are shown darker. Dotted lines indicate equivalent exons, based on sequence similarity and exon size. (b) Comparison of exon and intron sizes of the three genes, showing good continuity between sizes of the coding exons. (c) An alignment of the predicted protein sequence of the ANT2 genes in human (top),*

mouse (middle) and zebrafish (bottom). Amino acids identical in genes from all three species are shaded in black.

The genes dJ327A19.CX.3 (human), bM43O20.CX.4 (mouse) and bZ74M9.C14.4 (zebrafish) have also been classed as orthologous counterparts of each other (see Figure 6.9). Alignment of the predicted protein sequences of the three genes shows that the encoded human and mouse proteins are 94% identical to each other, and only 60% identical to the zebrafish protein. The 3' ends of the three genes are the part of this homologous set that are most similar to each other. Two exons in the human and mouse genes (exons 6 and 7) are present as a single exon in zebrafish (exon 4). Exons 3 and 4 in mouse do not appear to be present in the zebrafish gene. At this point there is no information regarding the possible function of these proteins, and so no conclusions can be drawn about the effect the amino acid sequence encoded by the extra exons in human and mouse will have on the function of the respective proteins.

Figure 6.9: *(see over) Analysis of orthologues in human, mouse and zebrafish (2) (a) A schematic representation of three genes, dJ327A19.CX.2, bM43O20.CX.4 and bZ74M9.C14.4 (exons are shown as bars, introns by v-shaped lines) in human (red) mouse (blue) and zebrafish (green). Untranslated regions are shown darker. Dotted lines indicate equivalent exons based on sequence similarity and exon size. (b) Comparison of exon and intron sizes of the three genes. (c) An alignment of the predicted protein sequence of the three genes in human (top), mouse (middle) and zebrafish (bottom). The three genes are similar at the 3' end, less similar at the 5' end, and the human and mouse genes encode extra amino acids in the middle of the protein.*

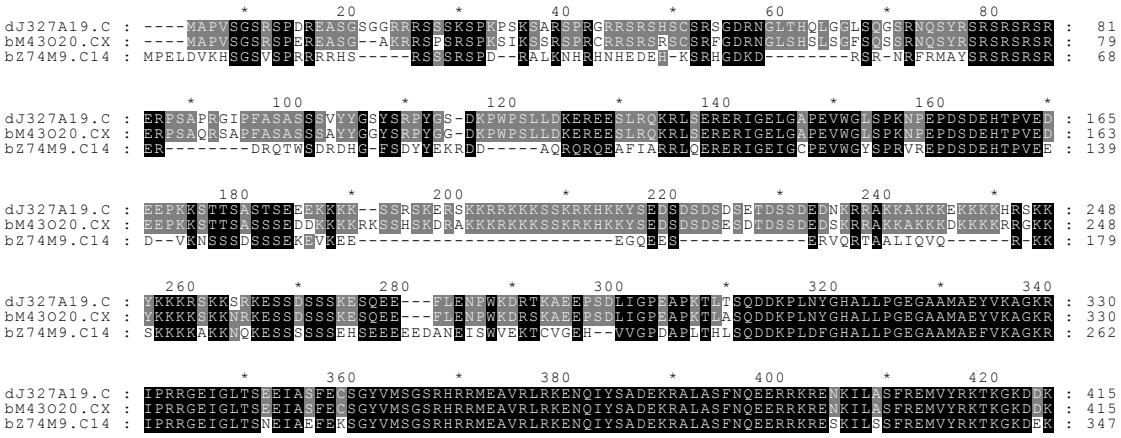
(a)



(b)

Gene Name	dJ327A19.CX.3	bM43O20.CX.4	dr_bZ74M9.C14.4
5'UTR	38	184	262
exon 1	386	380	308
2	81	81	81
3	71	71	64
4	135	141	189
5	64	64	150
6	110	110	175
7	76	76	
8	150	150	
9	175	175	
3'UTR	161	317	276
Total Coding	1248	1248	967
Total cDNA size	1447	1749	1505
intron 1	4409	3602	887
2	2047	5453	541
3	180	169	3319
4	1758	2784	93
5	2277	1897	625
6	74	89	
7	1770	992	
8	4621	4736	
Total Intron size	17136	19722	5465
Genomic coverage	18583	21471	6970
Distance to next gene	67048		

(c)



Four zebrafish genes do not appear to be orthologous to any of the other human genes in the region between HPR6.6 and ZNF-Kaiso, using the criteria described in Section 5.3. It is not unexpected that, due to the evolutionary distance between human and zebrafish, genes located in different regions of the human genome are present in the same region in zebrafish. However three of the four genes, bZ46J2.C14.6, bZ46J2.C14.7 and bZ46J2.C14.8 do match known genes located elsewhere in the human genome. bZ46J2.C14.6 shows similarity to members of the arrestin family of proteins, which are involved in the inactivation of rhodopsin and other heptahelical receptors. A comparison of the predicted protein sequence of bZ46J2.C14.6 with available protein sequences in EMBL using BLAST, shows that the most similar protein in human is β-Arrestin-1 (Sw:P49407) which is 46.15% identical. The predicted protein product of the zebrafish gene bZ46J2.C14.7 is approximately 60% identical to the human Inositol polyphosphate phosphatase-like 2 (INPPL2) protein, and the predicted protein encoded by the gene bZ46J2.C14.8 is approximately 61% identical to a human purinergic receptor (P2RY2). The three human genes are located on human chromosome 11q13 (data taken from ENSEMBL), within a 4 Mb region (82.9 Mb to 86.3 Mb). This analysis would indicate the presence of a syntenic block between human chromosome 11q13 and a region on zebrafish linkage group 14.

The one remaining zebrafish gene, bZ46J2.C14.3 did not match any sequence from any other organism currently available. bZ46J2.C14.3 has four exons and has a predicted mRNA size of 565 bp. The predicted protein is 184 amino acids in length and analysis of the protein in INTERPRO (http://www.ebi.ac.uk/INTERPRO) failed to identify any match to known protein domains.

The zebrafish genes, Rpl39 and bZ74M9.CX.4 are further apart than their predicted human orthologues RPL39 and dJ327A1.CX.3 (see Figure 6.7). Analysis of the genomic sequence in between the two genes in zebrafish reveals a region that may have undergone expansion due to the presence of five zebrafish-specific direct repeats (see Figure 6.10).

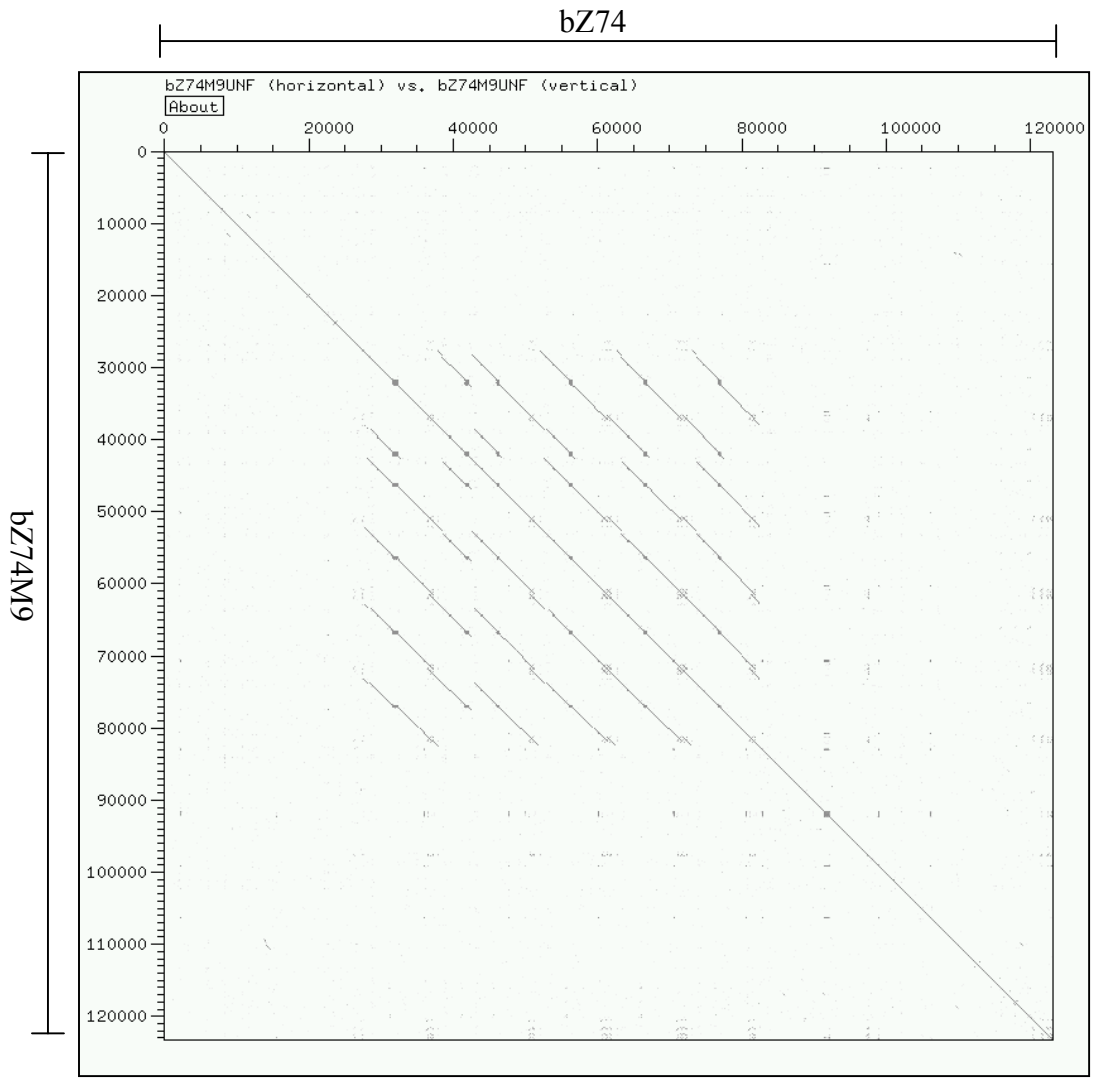


Figure 6.10: *A DOTTER of bZ74M9 against itself showing the presence of five copies of a direct repeat (indicated by red lines).*

6.6 Identification of 20 novel repeat elements in the zebrafish genome

The BAC clones identified in this chapter were among the first zebrafish clones to be sequenced by the Sanger Institute Sequencing teams. In order to further the understanding of the repeat content of the zebrafish genome the genomic sequence was analysed for the presence of repeats. At the time there were nine zebrafish repeat sequences listed in REPBASE (http://www.girinst.org) and these are summarised in Table 6.2.

Table 6.2: *Breakdown of known repeats*

Repeat Name	Repeat Type	Reference
ANGEL	DNA transposon	Izsvak, Z.*, et al.*, 1999
BHIKHARI	DNA retroposon	Vogel, A. M.*, et al.*, 1999
BHIKHARII	DNA retroposon	Vogel, A. M.*, et al.*, 1999
BRSATI	Satellite type I DNA	Ekker, M.*, et al.*, 1992
DANA	DNA retroposon	Izsvak, Z.*, et al.*, 1996
DRSATII	Satellite type II DNA	Ekker, M.*, et al.*, 1992
LINE_DR	LINE-like	direct submission
TDR1	Tc1-like element	Izsvak, Z.*, et al.*, 1995
TZF28	DNA transposon	direct submission

In an attempt to identify novel repeat sequences in the zebrafish genome, the draft sequence from the fourteen available BAC clones were compared to each other by BLAST and the clones analysed for regions of sequence that were present three or more times. These regions are candidates for novel repeat sequences and a consensus of each novel repeat region was generated (courtesy of Sarah Hunt). A total of twenty novel repeat sequences have been identified and these are summarised in Table 6.3.

Table 6.3: *Summary of novel repeat sequences in the zebrafish genome*

Repeat Name	Repeat Length	Matches in genome	Sequence contribution (kb)
DR_Rep1	1139	21	23.9
DR_Rep2	578	104	60.1
DR_Rep3	522	108	56.4
DR_Rep4	526	20	10.5
DR_Rep5	735	33	24.3
DR_Rep6	238	29	6.9
DR_Rep7	191	6	1.1
DR_Rep8	110	208	22.9
DR_Rep9	1407	100	140.7
DR_Rep10	670	75	50.3
DR_Rep11	198	150	29.7
DR_Rep12	485	16	7.7
DR_Rep13	391	6	2.3
DR_Rep14	593	3	1.7
DR_Rep15	908	75	68.1
DR_Rep16	375	210	78.7
DR_Rep17	110	67	7.3
DR_Rep18	1226	300	367.8
DR_Rep19	555	125	69.4
DR_Rep20	1117	222.5	2485.3
Total	-	-	3515.1

In order to get an estimate for the number of copies of each repeat in the zebrafish genome, each repeat was compared to the available zebrafish whole genome shotgun sequence using the Trace Archive available at http://www.ncbi.nlm.nih.gov/Traces/trace.cgi. At the time the analysis was carried out there were 4.2 million traces deposited which represents approximately 1.2 genome equivalents (assuming an average read length of 500 bp, and a genome size of 1.7 Gb). The number of copies for each repeat was calculated as being the number of different matches across the entire length of the repeat, divided by 1.2 (the genome equivalents available). The results are shown in Table 6.3, column 3.

Using this information, it is possible to estimate the amount of DNA sequence these novel repeats contribute to the zebrafish genome (see Table 6.3, column 4). Based on the length of each novel repeat and the number of copies in the genome, the data would suggest that these novel repeats contribute approximately 0.2% to the zebrafish genome size. A similar analysis was carried out using the previously known repeat sequences and this showed that they contributed 0.06% of the zebrafish genome.

The zebrafish genome is a half to two thirds smaller than the human genome which may be accounted for in part by a lower repeat content. However, the zebrafish genome is more than four times larger than *Fugu rubripes* genome (1.7 Gb compared to 0.4 Gb for fugu) and therefore may be expected to contain a greater number of repeat sequences. It is known that the genome of *Fugu rubripes* contains very few repeat sequences which is thought to account in part for the reduced genome size, along with reduced intron sizes (Elgar, G.*, et al.*, 1999). The number of repeats identified in this study and the estimates for the percentage of the zebrafish genome that is made up of repeat sequences is likely to be an underestimate because of the limitations of the analysis described here. Fourteen clones is a very small number to identify novel repeats in. These may be biased to regions of high or low repeat content. Also, the novel repeat sequences have been compared to short sequences of approximately 500 bp available as individual sequence reads from the whole genome shotgun of the zebrafish genome, which may generate errors in the analysis.

Further investigation of these repeats on large stretches of finished zebrafish sequence from different regions will improve the analysis and the continuing efforts of sequence generation will allow this. Also, it is possible that the repeats contained within the zebrafish genome are highly diverged and re-iterative BLAST analyses with all the repeat sequences may identify more copies of each novel repeat sequence. This will allow for a more detailed study of the repeat content and enable comparisons to be made with other fish genomes, such as fugu whose repeat content is reported to be very low, and mammals such as human and mouse.

6.7 Multiple sequence analysis

Genomic sequence in mouse and zebrafish has been generated that appears to be syntenic to parts of the region between HPR6.6 and ZNF-Kaiso in human. A total of twelve genes have been identified in zebrafish, of which eight appear to be orthologous in three different species, human, mouse and zebrafish (see Figure 6.11). Pairwise analysis of the sequence in human and mouse (see chapter 5) revealed a total of twenty-nine novel conserved sequences predicted by at least one of PIPMAKER, VISTA or ungapped BLAST and fourteen of those were predicted by all three. The additional information provided by the zebrafish sequence generated in this chapter allows for a further evaluation of the conserved sequences in the region between HPR6.6 and ZNF-Kaiso. Comparisons were carried out between human and zebrafish using the same three methods described in the previous chapter. Conserved sequences can only be identified for part of the region in human between HPR6.6 and ZNF-Kaiso given the lack of sequence covering the entire syntenic region in zebrafish.

Given the increased evolutionary distance between human and zebrafish, the threshold for sequence similarity was reduced from 75% (used for human-mouse comparisons) to 50%. The results are shown in Figure 6.12 and summarised in Table 6.4.

Figure 6.11: *(see over) Comparison of genes identified in human (middle), mouse (right) and zebrafish (left). A vertical bar represents the extent of the sequence generated in each species and genes are shown as bars (red = genes confirmed by cDNA, blue = predicted genes, green = pseudogene). Horizontal black lines link predicted orthologous genes. The names of orthologous genes identified in all three species are given in red.*

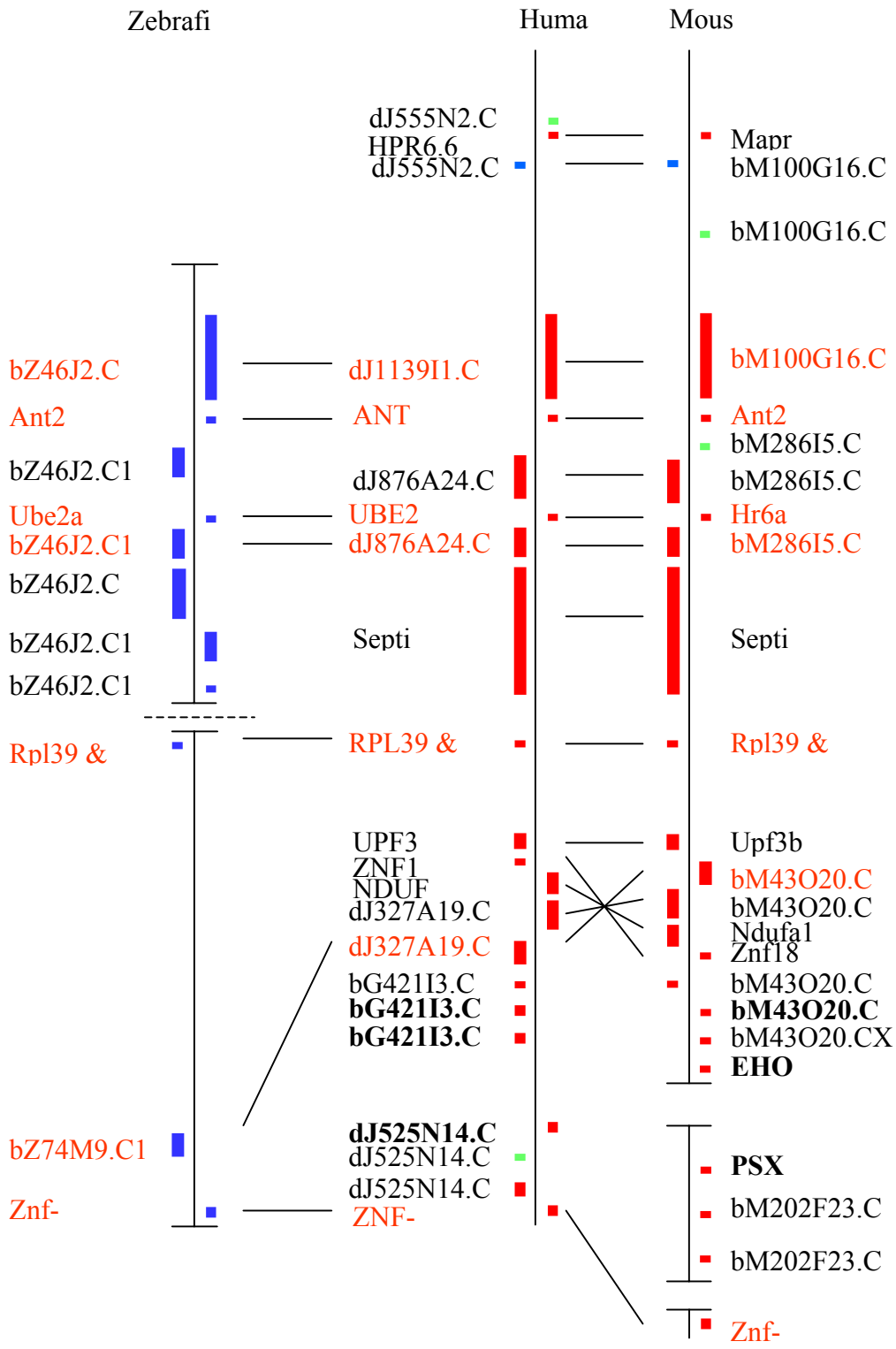

Zebrafi
Huma
Mous
dJ555N2.C
HPR6.6
dJ555N2.C
Mapr
bM100G16.C
bM100G16.C
bZ46J2.C
dJ1139I1.C
bM100G16.C
Ant2
ANT
Ant2
bM286I5.C
bZ46J2.C1
dJ876A24.C
bM286I5.C
Ube2a
UBE2
Hr6a
bZ46J2.C1
dJ876A24.C
bM286I5.C
bZ46J2.C
bZ46J2.C1
Septi
Septi
bZ46J2.C1
Rpl39 &
RPL39 &
Rpl39 &
UPF3
Upf3b
ZNF1
bM43O20.C
NDUF
dJ327A19.C
bM43O20.C
Ndufa1
dJ327A19.C
Znf18
bG421I3.C
bM43O20.C
bG421I3.C
bM43O20.C
bG421I3.C
bM43O20.CX
EHO
dJ525N14.C
bZ74M9.C1
dJ525N14.C
PSX
dJ525N14.C
bM202F23.C
Znf-
ZNF-
bM202F23.C
Znf-

Figure 6.12: *(see over) Identification of conserved sequences. A schematic of the region in human between HPR6.6 and ZNF-Kaiso. A scale indicates the size of the region, and genes are shown as vertical lines or boxes (exons) linked by horizontal lines (introns). Genes transcribed on the plus strand are positioned above the horizontal line, and those transcribed on the minus strand are positioned below the line. The zebrafish orthologous counterpart has been identified for the genes shown in red. No orthologue has been identified for the genes shown in green. Each red exon indicates a region conserved in human, mouse and zebrafish. The results of three methods for identifying conserved sequences between human and mouse (discussed in chapter 5) and between human and zebrafish are shown. Red vertical lines/boxes indicate the identification of a known conserved sequence. Other coloured lines/boxes indicated the position of a novel conserved sequence predicted by either PIPMAKER (black), VISTA (blue) or BLAST (green). The position of the novel conserved sequence predicted by all three methods is shown by a dotted arrow.*

Figure 6.12 – fold out of conserved sequences

Table 6.4: *Summary of prediction of conserved sequences in human, mouse and zebrafish*

Method	Conserved sequence	Other sequences	Total sequences	Sensitivity	Specificity
PIPMAKER	29	3	32	0.69	0.90
VISTA	14	2	16	0.33	0.87
BLAST	26	1	27	0.62	0.96
Total	42	4	34	0.71	0.88

There are a total of 42 known sequences conserved between human, mouse and zebrafish, which are the exons of the orthologous genes. PIPMAKER identified 29 of the 42 (69%), VISTA identified 14 (33%) and BLAST identified 26 (62%). In

general, fewer conserved sequences were identified in the human-zebrafish comparison than with the human-mouse comparison because the percentage identity between human and zebrafish exons was much lower. Where the percentage identity remained high, the conserved sequences were identified. This was seen for the ANT2 gene where all three methods identified all four exons and the ANT2 gene is greater than 90% identical at the nucleotide level for the coding region between all three species. In contrast, the first coding exon of the human gene dJ327A19.CX.3 is 86% identical to the mouse orthologue bM43O20.CX.4, but only 48% identical to the zebrafish orthologue bZ74M9.C14.4, and the match was not detected in zebrafish by the methods used.

The specificity of the three methods increased significantly when comparing human and zebrafish sequence as opposed to human and mouse sequence. The specificity is calculated assuming that all of the novel conserved sequences between human and mouse, predicted by the three methods in the previous chapter that lay outside the known coding sequences, were false. PIPMAKER predicted only three novel conserved sequences, VISTA predicted two and BLAST predicted only one. One novel conserved sequence was predicted by all three methods (indicated in Figure 6.12), and further analysis of the region containing this feature in the human sequence showed that it lay within a predicted CpG island (predicted by CpGfinder). There was also a match to a human 5' EST sequence (Em:BG118506) that showed no evidence of splicing, and no apparent polyadenylation signal (see Figure 6.13). Previous analysis of predicted genes in this region (discussed in chapter 4) excluded this feature as a potential gene, as the supporting evidence was not sufficient to follow up given the guidelines outlined in Section 4.2. However, the apparent conservation of

the feature in human, mouse and zebrafish, provides more confidence that the feature may be functional. In an attempt to determine whether this does represent novel coding sequence in the region, an STS, stbK38K21.3 was designed and used to screen DNA pools representing 12 different human cDNA libraries (v1-12 see Section 2.8.3). No expression was detected in any of the available cDNA libraries. However, given the amount of supporting evidence for this region of sequence, it is likely that a novel functional unit has been identified.

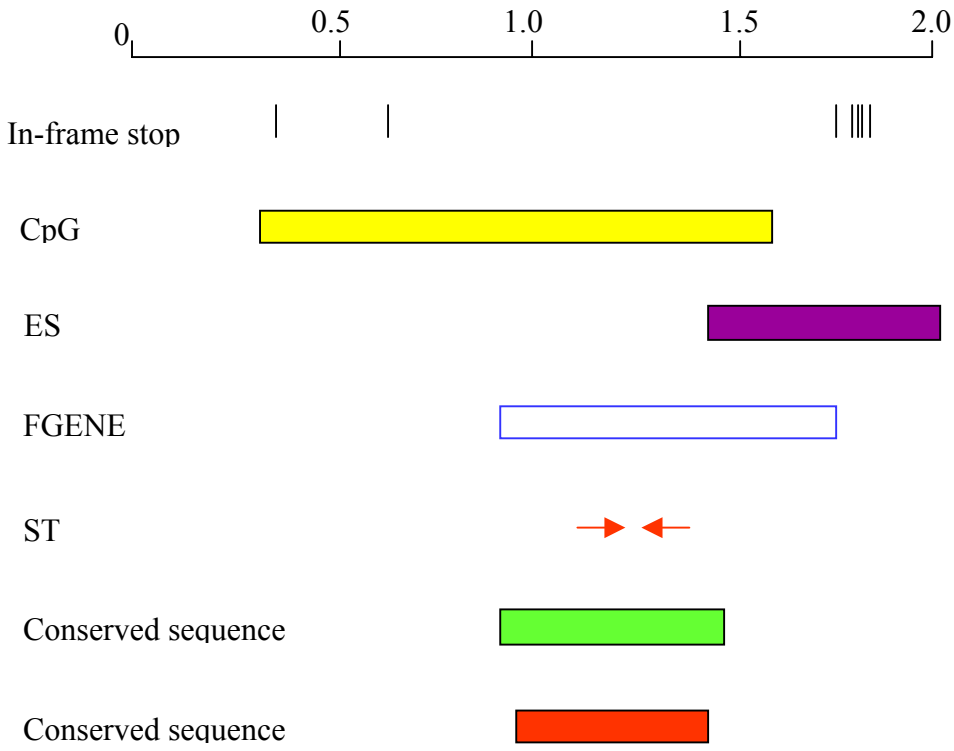


Figure 6.13: *Evidence of a novel conserved exon. Mouse DNA homology (green box) and zebrafish DNA homology (red box) positioned in same sequence as CpG island (yellow box), non-splicing EST (purple box) and FGENESH prediction (white box). The position of in-frame stop codons, based on the FGENESH prediction (shown as vertical black lines) delineates an ORF of 1.1 kb. An STS, stbK38K21.3 designed to part of the ORF (primers shown as red arrows) failed to identify any positive pools in the human cDNA libraries currently available.*

6.8 Discussion

Zebrafish bacterial clone isolation has been carried out using probes generated from STSs designed to human exons to screen a zebrafish BAC library by reduced stringency hybridisation. Analysis of the sequence of all the clones identified by reduced stringency hybridisation showed that only two of the fourteen clones contained orthologous sequences to the region of interest between HPR6.6 and ZNF-Kaiso in human. Evaluation of the method revealed two limitations. Firstly, a number of false positives were identified which did not appear to contain any sequence homologous to the human-specific sequence from which the probe was derived. Increasing the stringency of the washing after the hybridisation reduced the number of false positive clones identified, but also increased the risk of generating false negatives. Secondly, the method was not sufficiently sensitive to detect sequences less than 75% identical. Probes derived from STSs designed to four human genes did not identify clones that were later shown to contain the orthologous gene.

Recent progress has been made in both the sequencing of zebrafish ESTs, and the positioning of them within the zebrafish genome by RH mapping (Johnson, unpublished, see http://zfish.wustl.edu). Analysis of the available EST sequence data revealed two ESTs, localised to the same region on LG14 of the zebrafish genome, that are orthologous to two human genes in the region of interest between HPR6.6 and ZNF-Kaiso. This analysis reveals that the syntenic portion of part of the region of interest in human between dJ1139I1.CX.1 and ZNK-Kaiso is located on LG14 in zebrafish.

The zebrafish EST sequence data provides a more reliable method for the generation of bacterial clone contigs covering regions syntenic to human in the zebrafish and the strategy relies on the identification of orthologous gene sequences between the two organisms. An STS assay can be designed using this sequence and used to produce a probe that can then be labelled and hybridised to gridded arrays of zebrafish bacterial clones. Despite its higher accuracy, this approach is obviously limited by the availability of zebrafish specific cDNA or EST sequences. So in this case, only two human genes could be used to identify clones in zebrafish, as orthologous zebrafish EST sequences were only identified for dJ876A24.CX.1 and dJ327A19.CX.3. This method is applicable to comparative analysis between any two organisms for which orthologous sequences have been identified and is currently being applied to identify zebrafish BAC clones containing sequences orthologous to human chromosome seven (E. Green unpublished).

The two clones isolated by reduced stringency hybridisation, bZ46J2 and bZ74M9, appear to represent a region in zebrafish that is syntenic to a portion of human Xq24 between HPR6.6 and ZNF183. The evidence is based on the identification of zebrafish genes that are predicted to be the orthologues of eight of the human genes. A further four genes were identified in bZ46J2 and bZ74M9, three of which show similarity at the protein level to genes in human 11q13, suggesting a possible novel syntenic block.

A combination of *de novo* gene prediction and similarity searches predicted twelve genes in bZ46J2 and bZ74M9. Carrying out this type of analysis to identify genes in zebrafish is more difficult than in human and mouse, the other organisms studied here. The twelve genes identified in this study in zebrafish could not be confirmed by publicly availably cDNA sequence (as was the case for some of the genes described in chapter 4), and only had partial confirmation by EST sequence from zebrafish. In the cases where genes are predicted based on their similarity to sequences (both cDNA, protein and genomic) from distantly related organisms it can be very difficult to identify the exact exon boundaries as the level of similarity can be low (50-60% identity).

The region in human between HPR6.6 and ZNF-Kaiso has been compared to that in mouse and zebrafish. The availability of sequences in human, mouse and zebrafish thought to have descended from the same region in a common ancestor allows for analysis of the sequences conserved between them. Sequences that have maintained the same function in all three species are likely to be conserved. Comparisons of the conserved sequences identified in this chapter and chapter 5 shows that human and zebrafish sequences are mainly conserved in regions predicted to be coding. This compares to human and mouse sequences which show conservation outside the coding regions. The presence of a sequence conserved in all three organisms in the region between HPR6.6 and ZNF-Kaiso suggests the presence of a novel functional element. One possibility is that it is a novel exon and further screening in a wider variety of cDNA libraries (derived from human, mouse and zebrafish resources) may identify a cDNA clone to confirm this.

Multiple sequence analysis carried out in this chapter has utilised three methods, PIPMAKER, VISTA and BLAST. However, in reality the analysis that has actually been performed is a series of pairwise comparisons with the data being presented in a single view. This does not take into account conservation in the three species directly, although the results from the pairwise comparisons were considered together manually (as shown in Figure 6.12). It is not currently possible to carry out this type of analysis automatically. The sequence of the human genome is nearing completion and the sequencing of both the mouse and the zebrafish genome is underway. The sequencing of other genomes such as *S. cerevisiae* and *D. melanogaster* are available and sequencing of other vertebrate genomes is being discussed. Improvements in the tools to analyse these long sequences are required in order to extract the maximum amount of information from comparative sequence analysis.

6.9 Appendix

Table 6.5: *Comparison of orthologous genes in human, mouse and zebrafish (a '+' indicates UTR that spans multiple exons)*

Gene Name	dJ1139I1.CX.1	bM100G16.CX.4	bZ46J2.C14.1
5'UTR	0	0	75
exon 1	152	275	272
2	242	242	242
3	173	173	173
4	135	135	135
5	201	57	212
3'UTR	351	464	0
Total Coding	903	882	1034
Total cDNA size	1254	1346	1109
intron 1	6781	8114	5105
2	3488	5393	97
3	41646	17541	984
4	721	1264	990
Total Intron size	52636	32312	7176
Genomic coverage	53890	33658	8285
Gene Name	**ANT2**	**Ant2**	**Ant2**
5'UTR	70	85	87
exon 1	111	111	111
2	487	487	487
3	141	141	141
4	158	158	158
3'UTR	258	263	305

Total Coding	897	897	897
Total cDNA size	1225	1245	984
intron 1	1034	1001	1015
2	225	406	80
3	387	510	84
Total Intron size	1646	1917	1179
Genomic coverage	2871	3162	2163
Gene Name	**UBE2A**	**Hr6a**	**bZ46J2.C14.4**
5'UTR	174	128	124
exon 1	44	44	44
2	81	81	81
3	26	26	26
4	90	90	90
5	89	89	89
6	129	129	129
3'UTR	1162	1075	14
Total Coding	459	459	459
Total cDNA size	1795	1662	597
intron 1	145	181	219
2	393	375	401
3	6106	66464	2323
4	991	842	2278
5	450	324	92
Total Intron size	8085	68186	5313
Genomic coverage	9880	69848	5910
Gene Name	**dJ876A24.CX.3**	**bM286I5.CX.6**	**bZ46J2.C14.5**
5'UTR	690+34	108+34	160+129+6
exon 1	109	97	2220
2	1958	1958	

3'UTR	1015	1015	418
Total Coding	2067	2055	2220
Total cDNA size	3082	3070	2638
intron 1	13217	10980	180
2	1061	1112	3321
Total Intron size	14278	12092	3501
Genomic coverage	17360	15162	6139
Gene Name	**RPL39**	**Rpl39**	**bZ74M9.C14.1**
5'UTR	67	275	-
exon 1	3	3	3
2	104	104	104
3	49	49	49
3'UTR	178	181	-
Total Coding	156	156	156
Total cDNA size	401	612	156
intron 1	1562	1036	201
2	3175	1235	1661
Total Intron size	4737	2271	1862
Genomic coverage	5138	2883	2018
Gene Name	**U69a**	**U69a**	**bZ74M9.C14.2**
5'UTR	-	-	-
exon 1	132	132	64
3'UTR	-	-	-
Total Coding	-	-	-
Total cDNA size	132	132	64
Genomic coverage	132	132	64
Gene Name	**dJ327A19.CX.3**	**bM43O20.CX.4**	**bZ74M9.C14.4**
5'UTR	38	184	262
exon 1	386	380	308

2	81	81	81
3	71	71	64
4	135	141	189
5	64	64	150
6	110	110	175
7	76	76	-
8	150	150	-
9	175	175	-
3'UTR	161	317	276
Total Coding	1248	1248	968
Total cDNA size	1447	1749	1506
intron 1	4409	3602	887
2	2047	5453	541
3	180	169	3319
4	1758	2784	93
5	2277	1897	625
6	74	89	
7	1770	992	
8	4621	4736	
Total Intron size	17136	19722	5465
Genomic coverage	18583	21471	6971
Gene Name	**ZNF-kaiso**	**Znf-kaiso**	**bZ74M9.C14.6**
5'UTR	134+2	185+2	0
exon 1	2019	2016	1875
3'UTR	321	453	0
Total Coding	2019	2016	1875
Total cDNA size	2340	2469	1875
intron 1	2427	2132	0
Total Intron size	2427	2132	0

Genomic coverage	4767	4148	1875