Chapter 3

Cloning and homology of the zebrafish *jam* family

Summary

In this chapter I describe the evolutionary relationships between all members of the zebrafish *jam* family. Using BLAST searching of the zebrafish genome, I identified an additional two members of the family and cloned them by RT-PCR and 3′ RACE. I used the amino acid sequences of the conserved immunoglobulin-like domains from all of the zebrafish and mammalian JAM proteins to generate an alignment and a phylogenetic tree. This demonstrated that the zebrafish genome contains two orthologues of each of the three *JAM* genes in the mouse and human genomes. A cross-species analysis of local genome structure and evolutionarily conserved sequences indicate the genomic regions likely to have derived from genome duplication in zebrafish and which of those loci more closely resemble the ancestral loci.

3.1 Introduction

Mammalian genomes contain three *JAM* family members – *JAM-A*, *JAM-B* and *JAM-C*[†]. Prior to commencement of this project, four zebrafish *jam* family proteins had been identified. The zebrafish homologue of *JAM-A*, named *jama*; the *JAM-B* orthologues *jamb* and *jamb2*; and one *JAM-C* homologue, *jamc*, were identified from IMAGE consortium cDNA clones (Lennon *et al*, 1996). I sought to identify any other members of the family present in the zebrafish genome. The basic structural determinants of JAM family proteins are that they are type I transmembrane proteins with an N-terminal signal peptide, two immunoglobulin-like domains, a single transmembrane region and a short, apparently unstructured, cytoplasmic region ending in a type II PDZ-domain binding motif: ΦXΦ-COOH (figure 3.1).

The *JAM* genes have a conserved intron-exon structure over the regions encoding the extracellular protein domains, but the cytoplasmic domain-encoding exons of each *JAM* differ between family members (figure 3.2). Similarly, the amino acid sequence of the extracellular domains appears much more conserved than that of the cytoplasmic regions. Each immunoglobulin-like domain has a canonical disulfide bridge between B and F β-strands. Unusually, the membrane-proximal immunoglobulin-like domains of JAM-B and JAM-C each contain an additional, conserved, non-canonical disulfide bridge between A and G β-strands; the functional consequences of this feature are unknown. The JAM family proteins are predicted to be glycosylated and have conserved putative N-linked glycosylation sites: 'NX(S/T)'. The structures of recombinant ectodomains of murine and human JAM-A have been solved by X-ray crystallography (Kostrewa *et al*, 2001; Prota *et al*, 2003, respectively). In addition to the features already mentioned, both studies found the conformation of the immunoglobulin-like domains to be at an angle of 125° as a result of extensive hydrogen bonding between main chain atoms and hydrophobic interactions with the conserved linker peptide: 'VXV'. This conformation allows for the formation of homodimers *in cis*, which interact through the concave surface formed by the GFCC′ β-strands of the membrane-distal domain. An important motif within the

[†] There is considerable confusion of the nomenclature of the *JAM* family in the literature. For the sake of clarity, I have adopted the naming scheme suggested by Muller, 2003 that is now widely used by researchers. I have differentiated between zebrafish paralogues with a suffixed '*2*' for later discovered paralogues. This runs contrary to the guidelines given by the Zebrafish Nomenclature Committee, but is more useful for those researching the *JAM* family. Table 3.1 presents the gene names, aliases and Ensembl identifiers for each gene in the human, mouse and zebrafish genomes.

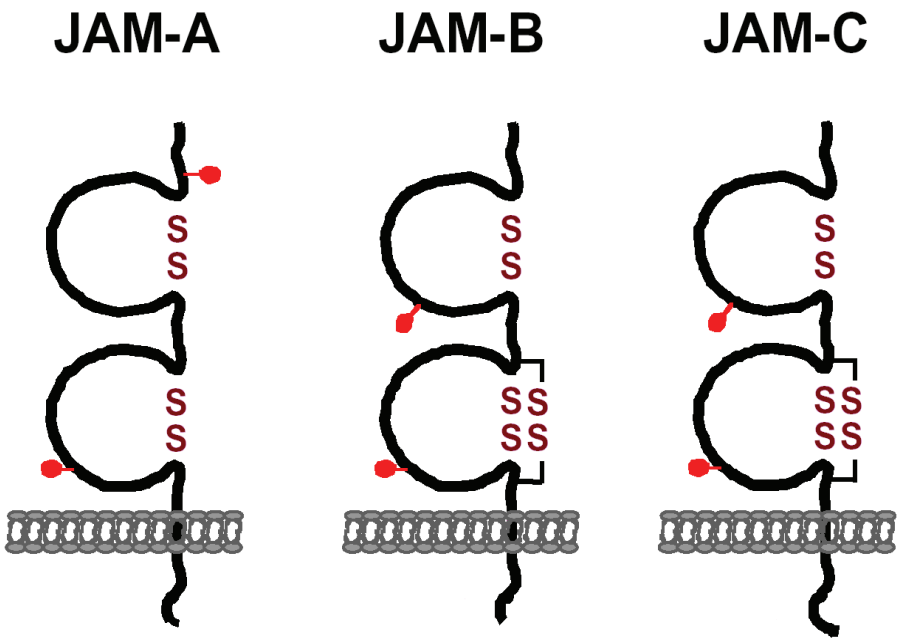


Figure 3.1 The mammalian JAM family.

Cartoon showing the basic structural features of all three mammalian JAM family proteins. Each JAM protein is a type I membrane protein with two extracellular, glycosylated, immunoglobulin-like domains and a short cytoplasmic domain ending in a type II PDZ-binding motif. Modified from Ebnet *et al* (2004), without permission.

Table 3.1 Nomenclature of the *JAM* family. Official gene symbols are marked in bold.

Name	Species	Synonyms	Gene identifier
JAM-A	Human	***F11R**, JAM-1, JAM, KAT, CD321, PAM1, JCAM*	ENSG00000158769
Jam-A	Mouse	***F11r**, Jam, Jcam, Jam-1, Ly106*	ENSMUSG00000038235
jama	Zebrafish	***f11r**, jam*	ENSDARG00000017320
jama2	Zebrafish		ENSDARG00000068114
JAM-B	Human	***JAM2**, CD322, VE-JAM, PRO245*	ENSG00000154721
Jam-B	Mouse	***Jam2**, Jam3, Vejam, Jcam2*	ENSMUSG00000053062
jamb	Zebrafish	***jam2**, vejam, cd322*	ENSDARG00000058996
jamb2	Zebrafish		ENSDARG00000079071
JAM-C	Human	***JAM3***	ENSG00000166086
Jam-C	Mouse	***Jam3**, Jam2*	ENSMUSG00000031990
jamc	Zebrafish	***jam3**, jam3b*	ENSDARG00000061794
jamc2	Zebrafish		ENSDART00000092689

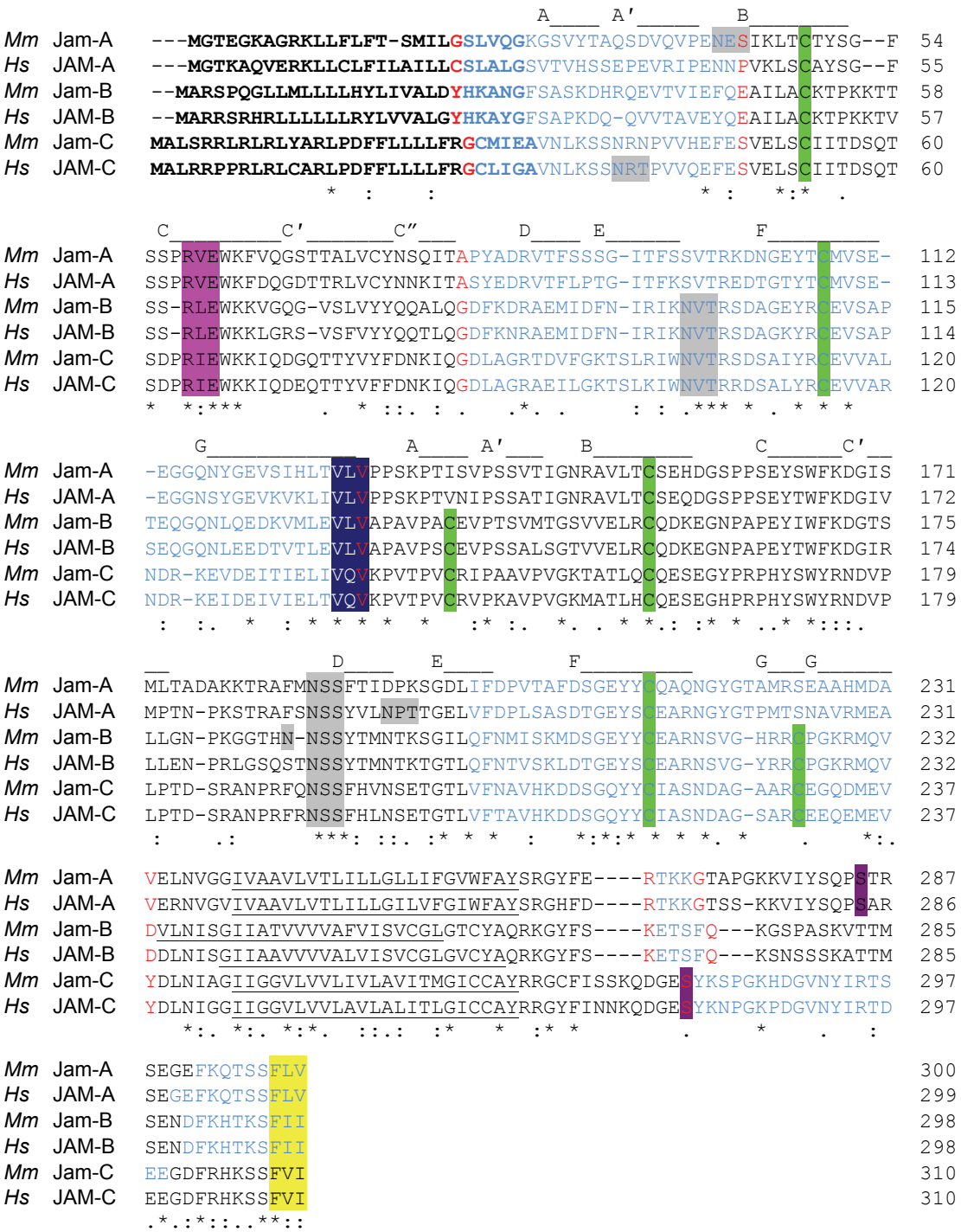


Figure 3.2 Conserved protein features of the mammalian JAM family.

ClustalW alignments of all human (*Hs*) and mouse (*Mm*) JAM family proteins, with key features highlighted. Exons – alternating blue/black colours, red – a cross-exon codon; bold – predicted signal peptide; green – disulfide bridge forming cysteines; lilac – binding interface residues; grey – putative N-linked glycoslyation sites; dark blue – linker sequence; underlined – predicted transmembrane helices; purple – phosphorylated serine; yellow – type II PDZ domain binding motif. The β-strands are indicated above the alignments.

C β-strand of this surface is 'R(V/I/L)E … Y' as these residues are important for forming salt bridges between monomers.

The cytoplasmic domain of JAM-A contains putative phosphorylation sites and some evidence for *in vivo* modification exists in activated platelets (Sobocka *et al*, 2000). Localisation of JAM-C to tight junctions seems to be regulated by phosphorylation of serine-281 in a cancer cell line (Mandicourt *et al*, 2007). The role of post-translational modification in the function of JAM-A, or the relevance to other members of the family, remains unexplored.

With the possibility of additional zebrafish *jam* family members that might be redundant with *jamb* and *jamc*, I searched the zebrafish genome for sequences with similarity to paralogues previously identified. I found two additional *jam* family genes and established their homology to mammalian *JAM-A* and *JAM-C* through sequence alignments and synteny.

3.2 Identification and cloning of *jama2* and *jamc2*

Putative paralogues of *jama* and *jamc* were identified in the zebrafish genome using TBLASTN searching of the zebrafish genome at the Ensembl website. The amino acid sequences of the extracellular immunoglobulin-like domains of Jama and Jamc were used as queries, as these regions were expected to be the most conserved between paralogues.

The best candidate paralogue of *jama* was a predicted gene found to be approximately 5.5 kbp upstream of *jama* on the same strand of chromosome 5. The protein sequence identity between the predicted gene product (hereafter referred to as Jama2) and Jama was very high across the immunoglobulin-like domains: approximately 79% by ClustalW alignment. However, careful manual searching of the genomic region downstream of *jama2* failed to reveal potential transmembrane and cytoplasmic domains. In order to establish that this putative paralogue is transcribed during development and to confirm the structure of the gene, 3′ RACE was performed, using nested primers specific to the predicted 5′ UTR of *jama2* (figure 3.3) and cDNA constructed from RNA extracted from 24 h. p. f. wildtype embryos, primed using a 3′ RACE primer. The major PCR product, approximately 1.5 kbp long, was purified, subcloned and sequenced. This *jama2* sequence included a small portion of the predicted 5′ UTR, the full open reading frame, including a stop codon, and 3′ UTR (figure 3.4). This sequence was compared to the genome and translated *in silico* and aligned against Jama (figure 3.5). The immunoglobulin-like domains of Jama2 are very closely matched to those of Jama and retain important protein

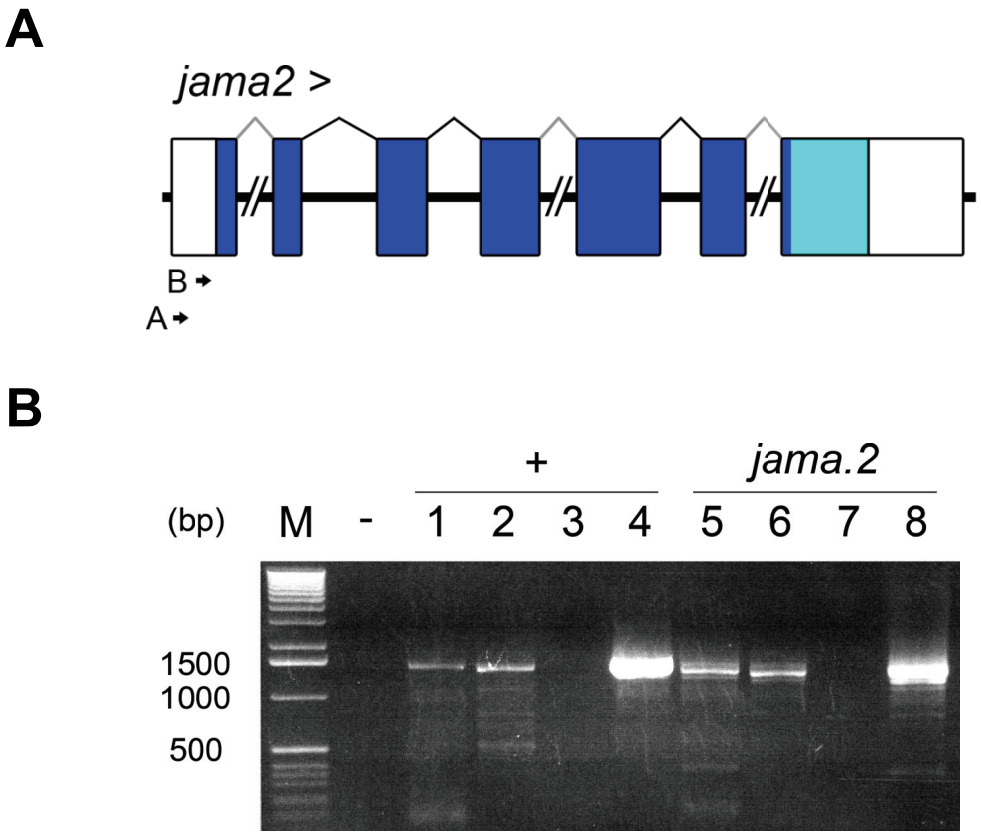


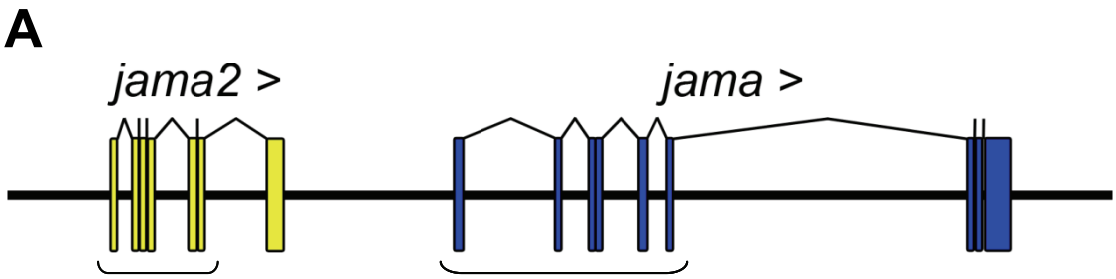
Figure 3.3 *jama2* is expressed at 24 h. p. f. as determined by 3′ RACE.

A. Scale diagram of the *jama2* loci, as determined sequence comparison between the 3′ RACE product and the zebrafish genome. Arrows indicate the position of the primers used in 3′ RACE experiments. Portion of gene cloned and used for protein production is indicated in dark blue. Introns larger than 500 bp were truncated for clarity, as indicated. **B.** Gel of 3′ RACE experiment using 24 h. p. f. cDNA. Lane descriptions as follows: '-' – negative control (no primers); *igsf11* positive control experiments: 1 & 2 - 5′ PCR primers, 3 – nested PCR control (no template), 4 – nested PCR; *jama2* experiments: 5 - 5′ primer A, 6 – nested 5′ primer B, 7 – nested PCR control (no template), 8 – nested PCR. M – DNA ladder, size of selected bands (in bp) shown to the left of the gel.

```
        10        20        30        40        50        60        70        80
cgggtaacatttgaaacgcataccgtggaaaaccttctatcatttcagactggaaATGTTGACTTTAGTCTTTGTGTGTC
                                                      M  L  T  L  V  F  V  C

        90       100       110       120       130       140       150       160
TCTCTTTTTCACTCACAGGCCTACATGCTTCCTTTTCAGTGGCTGTTAATGGTCCCGTAGTAAAAGTGAAGGAGAATGAG
L  S  F  S  L  T  G  L  H  A  S  F  S  V  A  V  N  G  P  V  V  K  V  K  E  N  E

       170       180       190       200       210       220       230       240
GGAGTTGACTTGCAATGTTCCTACACCGCTGACTTTGGAGCAACACCCAGAGTAGAATGGAAGTTCAGAAATCTGAAGGG
 G  V  D  L  Q  C  S  Y  T  A  D  F  G  A  T  P  R  V  E  W  K  F  R  N  L  K  G

       250       260       270       280       290       300       310       320
CTTTCAGTATTTCATCTACTTTAATAACAAACCAACTGTTGAATATGAACAGCGCATCACTGTGTACGCTGGAGGACTGA
  F  Q  Y  F  I  Y  F  N  N  K  P  T  V  E  Y  E  Q  R  I  T  V  Y  A  G  G  L

       330       340       350       360       370       380       390       400
GATTTCAAAAAGTAACGCGAGCAGACGCTGGAGATTATAACTGTGAGGTTTCTGGAAACGGTGGATATGGAGAGAATACC
R  F  Q  K  V  T  R  A  D  A  G  D  Y  N  C  E  V  S  G  N  G  G  Y  G  E  N  T

       410       420       430       440       450       460       470       480
ATCAAACTTGTAGTCTCTGTTCCTCCTTCCAAGCCTGTATCCAGCATTCCTTCATCAGTCACAACAGGCAGTAACGTCCG
 I  K  L  V  V  S  V  P  P  S  K  P  V  S  S  I  P  S  S  V  T  T  G  S  N  V  R

       490       500       510       520       530       540       550       560
CCTGACTTGCTTTGACCCAGTTGGCTCTCCTCCATCCACCTATGAGTGGTACAAAGACAACAACCTCCTCCCTGAGGACC
  L  T  C  F  D  P  V  G  S  P  P  S  T  Y  E  W  Y  K  D  N  N  L  L  P  E  D

       570       580       590       600       610       620       630       640
CAACCAAGTTTCCCATTTTTAAGAACCTCACATATAAGATGAATGCTTTCAATGGAAACCTGGAGTTCTTGAGTGTGTCT
P  T  K  F  P  I  F  K  N  L  T  Y  K  M  N  A  F  N  G  N  L  E  F  L  S  V  S

       650       660       670       680       690       700       710       720
AAGTGGGATGCTGGCTCATATTTTTGTGTGGCCAGTAATGAAAACGGTGTCTCTCAGCATGGTGATGCAGTGAAGATGGA
 K  W  D  A  G  S  Y  F  C  V  A  S  N  E  N  G  V  S  Q  H  G  D  A  V  K  M  E

       730       740       750       760       770       780       790       800
AGTTTATGATGTAGACAGCAGTcaagtgctggatgtgaagagcaacttgagcatggagacacacaacattccaggcaaga
  V  Y  D  V  D  S  S  Q  V  L  D  V  K  S  N  L  S  M  E  T  H  N  I  P  G  K

       810       820       830       840       850       860       870       880
tcaccaacagccacataatggaaaaacagtatggtgtgttcatgttgcaggaggtgaaactaaaactagagatccagaaa
I  T  N  S  H  I  M  E  K  Q  Y  G  V  F  M  L  Q  E  V  K  L  K  L  E  I  Q  K

       890       910       920       930       940       950       960       970
ctggaattagaagtgaccaagctaaagctggagctgcaaaaacttggacatgaagtgtagatgatatcattcatcattac
 L  E  L  E  V  T  K  L  K  L  E  L  Q  K  L  G  H  E  V  *

       980       990      1000      1010      1020      1030      1040      1050
tgctataagtcaaagaacttatttcatgtgctgatttcagatgttattgtaattacatttgtttttatacagctggggtc

      1060      1070      1080      1090      1100      1110      1120      1130
tgtttttcatacctcacttaaaagatctgtgatgatttgacagatatggatttttttcaggatatccgtgtggtcttaaa

      1140      1150      1160      1170      1180      1190
gtcttaaatctcaaaaactcaaatttaagccttaaagtgtctttaatcttctaaaaaaaa
```

Figure 3.4 Sequence of *jama2* mRNA as determined by 3′ RACE.

The mRNA sequence of the *jama2* 3′ RACE product and translation of the open reading frame. The 5′ and 3′ UTR elements, as determined by genomic alignment, are indicated in orange. Alternate exons within the open reading frame are indicated by alternate black and blue text. The region of cDNA cloned for protein expression is indicated by underlined capital letters.



B

```
Dr Jama     MLTFAFVCFSLFVTGIHG-FQVTVTSP-VKVKENEGVDLQCSYTSDFGATPRVEWKFKDL   58
Dr Jama2    MLTLVFVCLSFSLTGLHASFSVAVNGPVVKVKENEGVDLQCSYTADFGATPRVEWKFRNL   60
            ***:.***:*: :**:*. *.*:*..* ****************:************::*

Dr Jama     KGSQTLVYFDGKPTGQYTGRVTMYDKGLRFNKVTRADTGDYDCEVSGSGGYGENTIKLTV   118
Dr Jama2    KGFQYFIYFNNKPTVEYEQRITVYAGGLRFQKVTRADAGDYNCEVSGNGGYGENTIKLVV   120
            ** * ::**:.*** :*  *:*:*  ****:******:***:*****.**********.*

Dr Jama     LVPPAKPVSRIPSSVTTSSNVRLTCFDPVGSPPSTYKWYKDNTPLPEDPTKFPAFKNLTY   178
Dr Jama2    SVPPSKPVSSIPSSVTTGSNVRLTCFDPVGSPPSTYEWYKDNNLLPEDPTKFPIFKNLTY   180
             ***:**** *******.******************:*****. ********* ******

Dr Jama     KMNVFNGNLEFPSVSKMDTGSYFCEASNGEGVPQRGDEVKMEVRDLN---VGGIVAGVIV   235
Dr Jama2    KMNAFNGNLEFLSVSKWDAGSYFCVASNENGVSQHGDAVKMEVYDVDSSQVLDVKSNLSM   240
            ***.******* **** *:***** *** :**.*:** ***** *::   * .: :.: :

Dr Jama     ALLAVGLLLFGLWYASKKGYLPKLSETKQKPQAVYTQPQTDDVDEANGFRQKSSFVV      292
Dr Jama2    ETHNIPGKITNSHIMEKQYGVFMLQEVKLK---LEIQKLELEVTKLKLELQKLGHEV      294
                :   : .    .*:  :  *.*.* *   :  *    :* : :   ** .. *
```

Figure 3.5 Genomic alignment of *jama2* cDNA sequence and comparison of the translated open reading frame with Jama.

A. The cDNA sequence of the *jama2* 3′ RACE product was compared to the zebrafish genome using BLAT at Ensembl. The cDNA matches the genomic sequence very close to *jama,* as shown in this schematic of the loci. The exons encoding the immunoglobulin-like domains in both genes are bracketed. The figure is drawn to scale. **B.** ClustalW alignment of the amino acid sequence of Jama and Jama2, as translated from the cDNA sequence. 58% of residues are identical over the whole alignment, rising to 79% over the immunoglobulin domains alone. Feature annotations as in figure 2.

features (such as signal peptide, cysteine residues for disulfide bridges, salt bridge residues of the dimerisation interface and glycoslyation site). The predicted intron-exon structure also matches very closely with that of *jama*, except just after the final immunoglobulin-like domain encoding exon. Unsurprisingly, a hidden Markov model analysis of the protein sequence (TMHMM v2.0; Krogh *et al*, 2001) failed to identify any possible transmembrane region within the *in silico* predicted protein, suggesting that Jama2 is a secreted protein (data not shown). In the absence of identification of any transmembrane or cytoplasmic domain, and for the sake of consistency, I chose to clone only the immunoglobulin-like-domain-encoding region of the *jama2* gene for later expression and analysis (figure 3.4, underlined; see Materials and Methods for experimental details).

The most likely candidate paralogue of *jamc* was found on chromosome 15; a predicted gene with a similar intron-exon structure and a closely matching predicted protein sequence (approximately 63% over the two immunoglobulin-like domains). The different predictions and EST evidence were contradictory in parts and even included a small portion of a likely downstream gene. To resolve the structure of the open-reading frame, I designed pairs of primers within different regions of the gene predictions and EST evidence and attempted to amplify the gene by RT-PCR from cDNA prepared from RNA extracted from 24 h. p. f. wild-type embryos (figure 3.6). Having confirmed the structure of the *jamc2* gene, I subsequently amplified the cDNA using a primer within the 5′ UTR region and the 3′ primer at the end of the coding sequence. A single product of approximately 1 kbp was purified, subcloned and fully sequenced. Comparing this sequence to the zebrafish genome identified some 5′ UTR and nearly 900 bases of open reading frame (figure 3.7). This was translated *in silico* and used in alignments with Jamc (figure 3.8). The high level of amino acid identity and the predicted intron-exon structure confirms this gene as a member of the *jam* family and the paralogue of *jamc* and is therefore referred to as *jamc2*. Analysis of the translated sequence predicted a transmembrane domain (TMHMM v2.0 Krogh *et al*, 2001), followed by a short cytoplasmic domain (figure 3.8). The cytoplasmic domain of Jamc2 contains conserved residues that may represent a type II PDZ motif, if they are indeed the C-terminal residues, as predicted from sequence alignment. The cDNA product did not include a stop codon, although there is a stop codon present in the genome immediately after the end of the aligned sequence, suggesting it is the true 3′ end of the gene. The complete extracellular domain-encoding region of *jamc2* was cloned for later expression and analysis (figure 3.7; see Materials and Methods for experimental details).

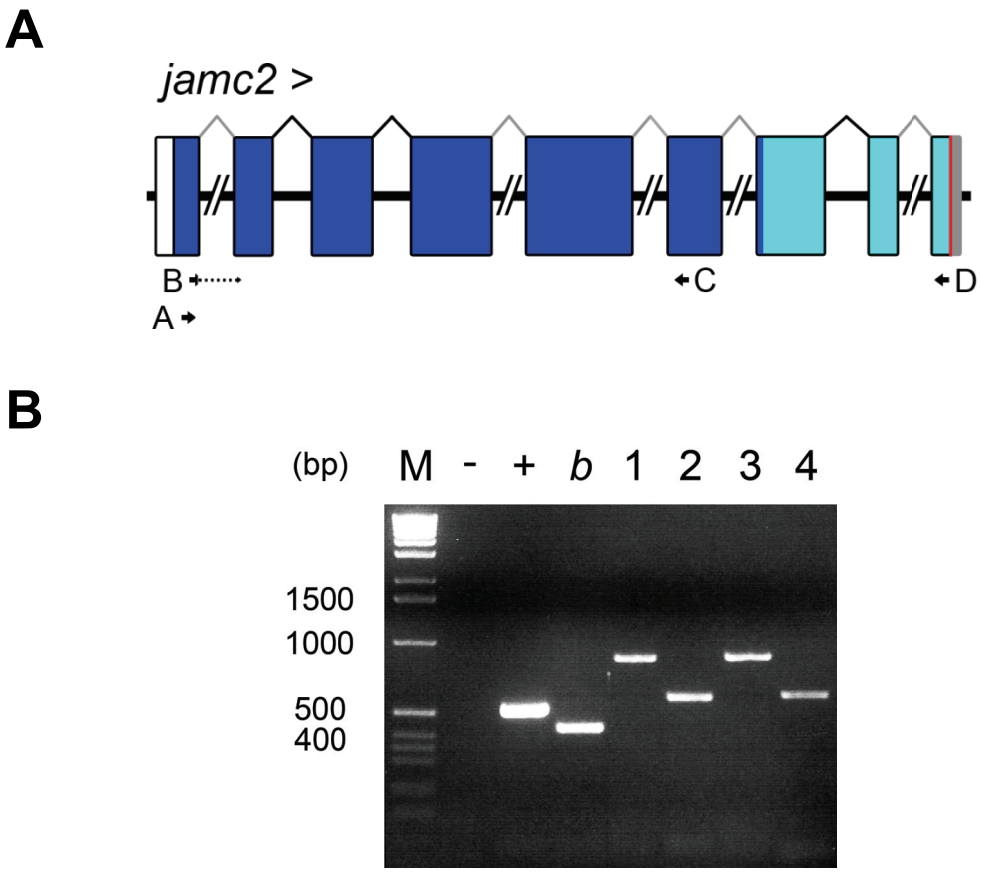


Figure 3.6 Structure of *jamc2* mRNA as determined by RT-PCR.

A. Scale diagram of the jamc2 loci, predicted from EST and gene prediction data and subsequently determined by comparison of sequenced RT-PCR products and the zebrafish genome. The true 3′ end of the gene remains to be formally established, but an in-frame stop codon is present in the genome sequence, directly adjacent to the aligned RT-PCR product sequence (red line). The 3′ UTR is likely to extend beyond this (region in grey) as no splice donor site is present nearby. Arrows indicate the position of primers used in RT-PCR experiments. Portion of the gene cloned and used in protein production is highlighted in dark blue. Introns larger than 500 bp were truncated for clarity, as indicated. **B.** Gel showing products of RT-PCR experiments. Lane descriptions as follows: '-' – negative control (no template), '+' - *ef1α* positive control, *b* – *jamb* positive control, 1 – primers B and D, 2 – primers B and C, 3 – primers A and D, 4 – primers A and C. M – DNA ladder, size of selected bands (in bp) shown to the left of the gel.

```
        10        20        30        40        50        60        70        80
ctaaacctgcatgtggaaacagcggctcaaaATGGCGTTCGGCCGTCAAACGCTTTCCCTGGTGCTCTTCTGCTGGCTGT
                                  M  A  F  G  R  Q  T  L  S  L  V  L  F  C  W  L

        90       100       110       120       130       140       150       160
GTAACAGTGCTGCCTTTGCTGTAATACTCCGAACAACTGAGAAATCTGTGTGGGCAAATGAATTTGAGTCAATCGAACTG
C  N  S  A  A  F  A  V  I  L  R  T  T  E  K  S  V  W  A  N  E  F  E  S  I  E  L

       170       180       190       200       210       220       230       240
ACCTGCTTGATAGAGTCCATTTCTACAAACAATCCTCGAATTGAATGGAAGAAAATAAAAAACGGTGTACCCAGTTATGT
 T  C  L  I  E  S  I  S  T  N  N  P  R  I  E  W  K  K  I  K  N  G  V  P  S  Y  V

       250       260       270       280       290       300       310       320
GTACTTTCAAAACAAAATATCAGGTGACCTGGAGCACAGGGCTTTGCTGCGAGAACCTGCAAACCTTCTGATACTGAACG
  Y  F  Q  N  K  I  S  G  D  L  E  H  R  A  L  L  R  E  P  A  N  L  L  I  L  N

       330       340       350       360       370       380       390       400
CCAGCAGATCAGACACAGCACAGTATCGCTGCGAGGTGGCCGCCATTGATGACCAGAAGCCTTTTGACGAAATATTAATC
A  S  R  S  D  T  A  Q  Y  R  C  E  V  A  A  I  D  D  Q  K  P  F  D  E  I  L  I

       410       420       430       440       450       460       470       480
AGTCTAGCTGTAAGAGTGAAGCCGGTAATCCCCAGATGTAGTGTGCCAGATGCAGTTAATGTGGGTTCAAGCACTGAACT
 S  L  A  V  R  V  K  P  V  I  P  R  C  S  V  P  D  A  V  N  V  G  S  S  T  E  L

       490       500       510       520       530       540       550       560
GCGATGTATTGAGAACGAAGGCTTTCCTCAGTCACAGTACCAGTGGTTCAAAAACAGCGAGGAGCTGCCCGAGGACCCAA
  R  C  I  E  N  E  G  F  P  Q  S  Q  Y  Q  W  F  K  N  S  E  E  L  P  E  D  P

       570       580       590       600       610       620       630       640
AAACCAGCAGCAAGTTCTACAATTCCTCATACATCATGAACATTGAGACTGGCTCTCTGAAATTCCGGTCGGTAAAGAAA
K  T  S  S  K  F  Y  N  S  S  Y  I  M  N  I  E  T  G  S  L  K  F  R  S  V  K  K

       650       660       670       680       690       700       710       720
GAGGATGCGGGTGAATATTATTGCCAGGCCAGAAATGAAGCCGGATGGTCAAAATGTATTCGACAGAGCATGGAAGTGTA
 E  D  A  G  E  Y  Y  C  Q  A  R  N  E  A  G  W  S  K  C  I  R  Q  S  M  E  V  Y

       730       740       750       760       770       780       790       800
TGACTTGGACATTGTGGGaatatttctgaaggttttgggtggagttgcagcatttatttttgtcattgtgggaatttgtc
  D  L  D  I  V  G  I  F  L  K  V  L  G  G  V  A  A  F  I  F  V  I  V  G  I  C

       810       820       830       840       850       860       870       880
aaattcagaaaagtggttactgttcctgcaaagatcacagagaaaccaactacaaagtaccccaacatgaaaacaggatg
Q  I  Q  K  S  G  Y  C  S  C  K  D  H  R  E  T  N  Y  K  V  P  Q  H  E  N  R  M

       890       900       910       920       930
gagtacaccactccagatgagggacattttcgccacaaatcctccttcgtcatc
 E  Y  T  T  P  D  E  G  H  F  R  H  K  S  S  F  V  I
```

Figure 3.7 Sequence of *jamc2* mRNA as determined by RT-PCR.

The cDNA sequence of the *jamc2* RT-PCR product and translation of the open reading frame. The 5′ UTR, as determined by genomic alignment, is indicated in orange. Alternate exons within the open reading frame are indicated by alternate black and blue text. The region of cDNA cloned for protein expression is indicated by underlined capital letters.

Figure 3.8 Comparison between Jamc and Jamc2 highlights conserved features.

ClustalW alignment of the amino acid sequence of Jamc and Jamc2, as translated from the cDNA sequence. 59% of residues are identical over the whole alignment, rising to 70% over the immunoglobulin-like domains alone. Feature annotations as in figure 2.

3.3 Evolutionary relationships of zebrafish *jam* family orthologues

To confirm the identity and family membership of the newly identified genes, I performed a phylogenetic analysis. I generated a clustalW alignment of the amino acid sequences of the immunoglobulin-like domains of all zebrafish, mouse and human *JAM* family genes. The immunoglobulin-like domains were used exclusively because of uncertainty about the true 3′ sequences of some of the family members and the strongly divergent nature of the unstructured C-terminal cytoplasmic domains. The closely-related cell surface, two immunoglobulin-like-domain containing proteins ESAM (mouse and human), A33 (mouse and human), CAR (mouse, human and zebrafish) and Jam4 (mouse only) were also included to test the robustness of paralogue assignments. This alignment was then used to generate a phylogenetic tree (figure 3.9). As expected, each of the putative paralogues was confirmed as *JAM* family members, distinct from the other immunoglobulin superfamily members. Both Jama and Jamc were also confirmed as either an 'A' type or 'C' type JAM family member, respectively.

Comparing human, mouse and zebrafish *JAM-A* loci suggest that this very small region has undergone duplication in the teleost lineage (figure 3.10). Only the gene immediately upstream of *jama* and *jama2*, *usf1l*, appears conserved between fish and mammals[†]. The lack of a transmembrane domain and cytoplasmic region in *jama2* suggest it is the derived allele. Only the exon sequences of the immunoglobulin-like domains have been conserved.

In contrast, *jamc2* appears to have a considerable amount of conservation of local gene structure, with local genes, *igsf9b*, *vps26b*, *acad8* and *thyn1* present in the same order and orientation with respect to *Jam-C* in the mouse and chick genomes (figure 3.11). An intervening gene in the mammalian and avian genomes, *Ncapd3*, is apparently missing from the *jamc* loci in zebrafish, but is present elsewhere in the zebrafish genome. This suggests that it has been deleted or transposed in teleosts, or inserted into the *Jam-C* locus in the mammalian lineage. The first *JAM-C* paralogue to be identified, *jamc*, is close to an *igsf9b* orthologue but apparently no other genes that are present in the *jamc2*, mouse or chicken loci. The apparent lack of gene conservation suggests that *jamc* is the derived allele of an ancient duplication of the genomic region bounded at one end by the ancestral *jamc*. Despite

[†] The chicken genome was not included in this analysis as no homologue of *JAM-A* could be found by BLAST search. I attempted to find the appropriate locus by finding homologues of other genes at the human *JAM-A* locus, but without success.

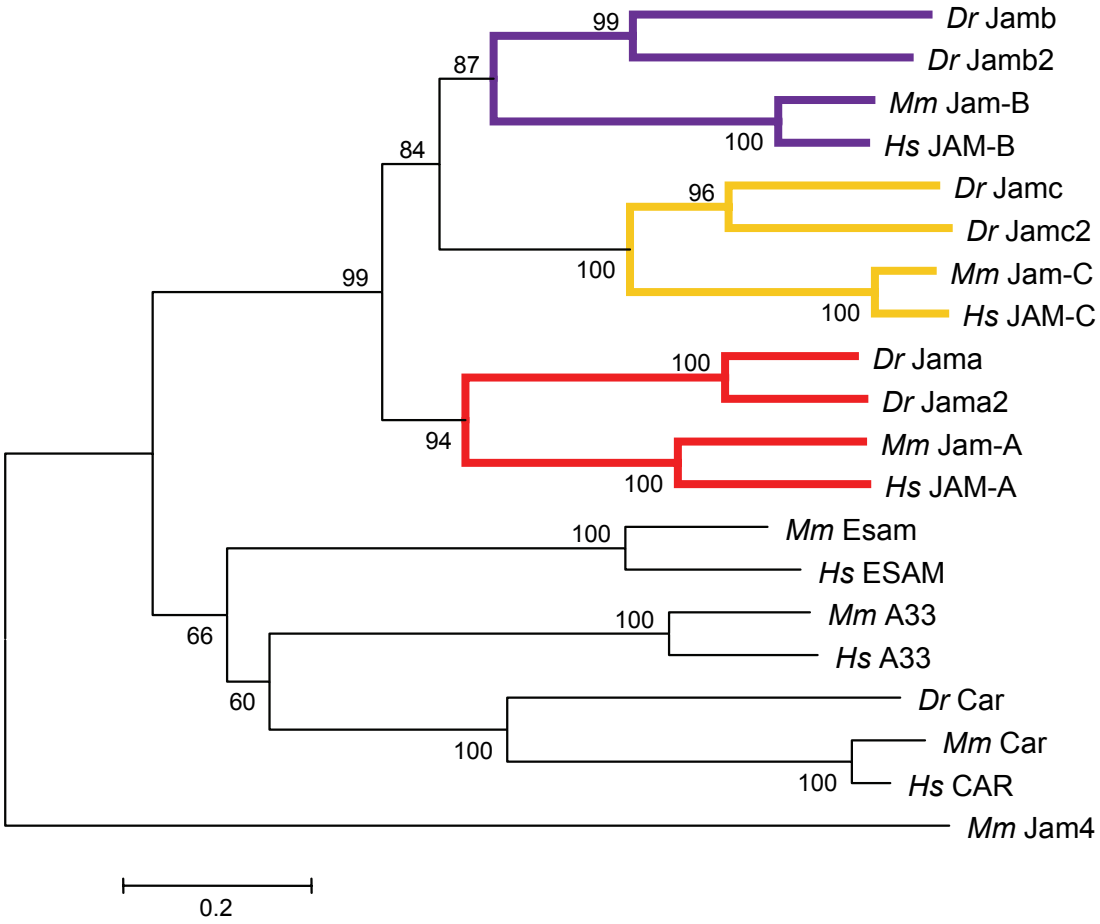


Figure 3.9 Zebrafish *JAM* family genes are distinct from related IgSF proteins and share a common ancestor with human and mouse *JAM* family genes.

Phylogeny generated by MEGA from ClustalW alignment of amino acid sequences of immunoglobulin-like domains from all human (*Hs*), mouse (*Mm*) and zebrafish (*Dr*) JAM family proteins and a selection of related transmembrane proteins. Values at nodes indicate percentage of bootstraps showing the same branching relationship (n = 500). Relative evolutionary distances are estimated by branch length; scale shown below tree.

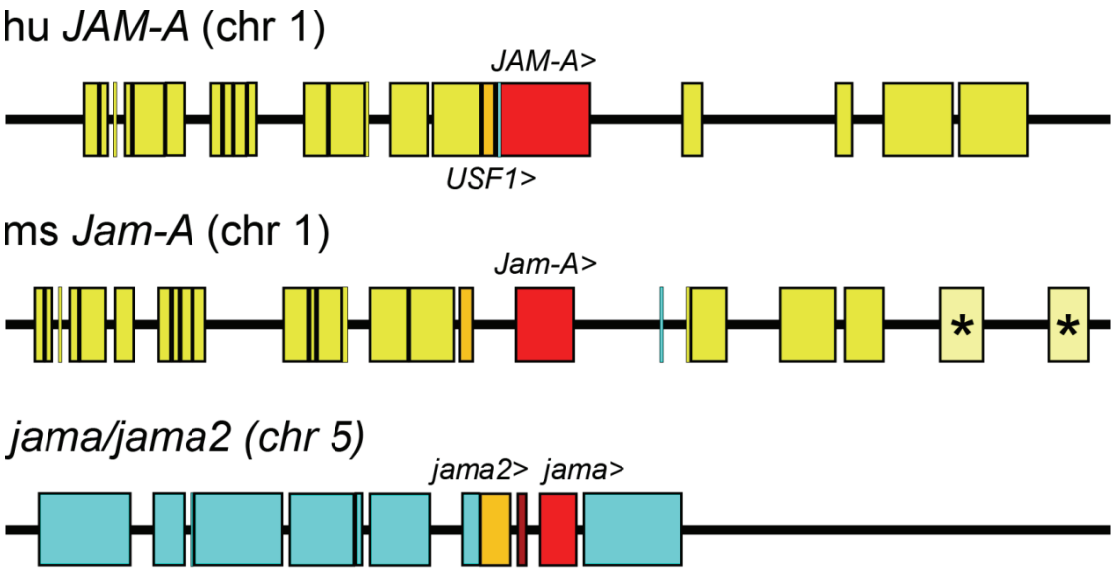


Figure 3.10 Multi-species comparison of *JAM-A* loci reveals limited conservation of local gene structure between zebrafish and mammals.

Schematic showing the arrangement of annotated genes (coloured boxes) within 0.5 Mb of sequence from the human, mouse and zebrafish genomes centered on *JAM-A* orthologues (red boxes). Genes conserved between these loci are highlighted by yellow boxes – only one example, *usf1l*, exists at the zebrafish *jama*/*jama2* loci. Genes that are present at only one locus are highlighted by blue boxes. The mouse genes highlighted with an asterisk (from left to right: *CD48* and *Slamf7*) are present at this loci in the human genome, but are outside the 0.5 Mb window presented here. The figure is drawn to scale.

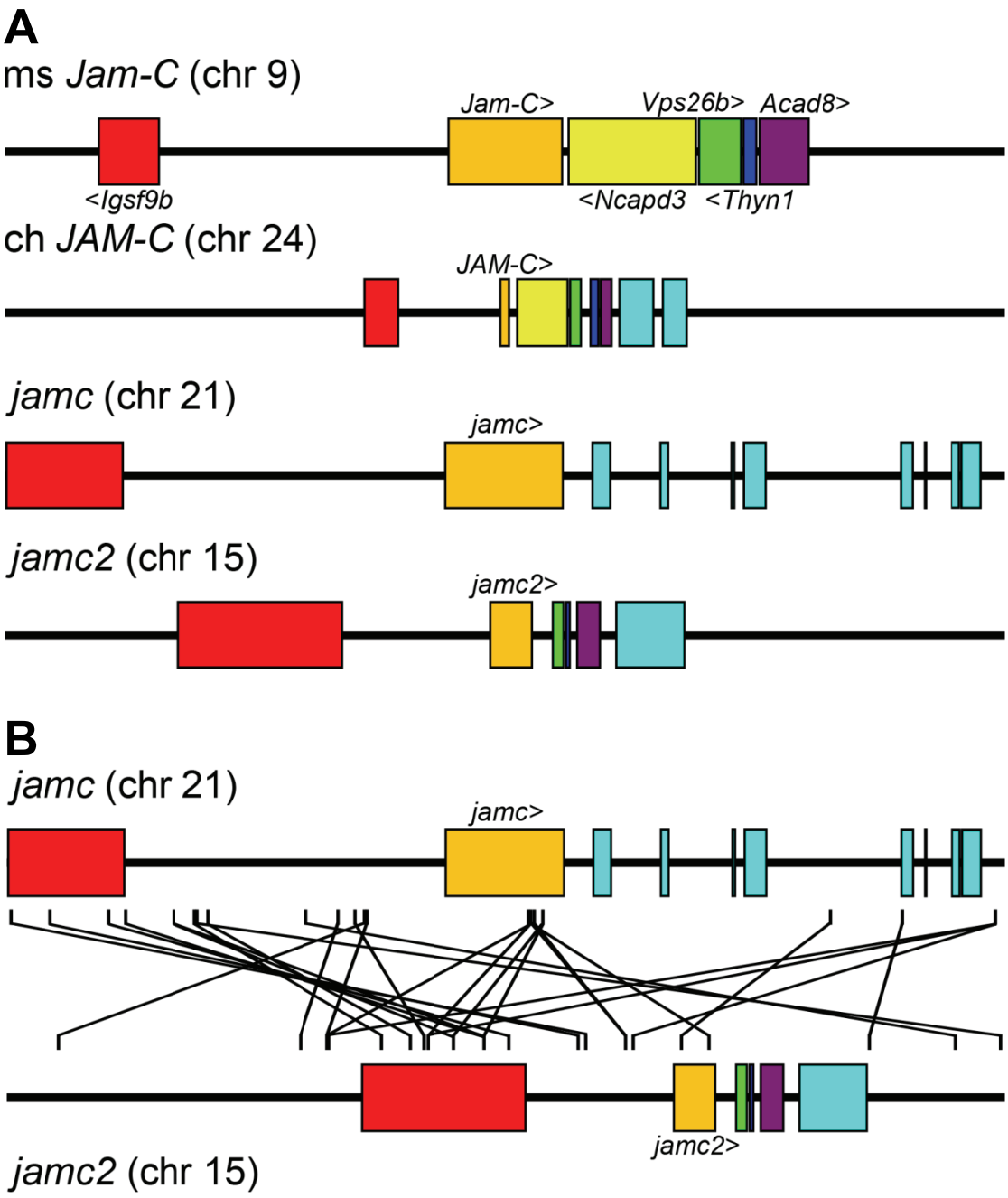


Figure 3.11 Multi-species comparison of *Jam-C* loci suggest *jamc* is the derived allele from an ancient genome duplication.

A. Schematic of the arrangement of annotated genes (coloured boxes) within 0.5 Mb of sequence from the mouse, chicken and zebrafish genomes centered on *Jam-C* orthologues (orange boxes). Genes conserved between these loci are colour-coded according to identity. Genes that are present at only one locus are highlighted by blue boxes. **B.** Schematic (as in A) of evolutionary conserved regions (ECRs) between zebrafish *jamc* and *jamc2* loci. ECRs are sequences at least 100 bp long with 70% identity or greater, as identified by BLASTz. The figure is drawn to scale.

this, the amino acid sequence of the extracellular domain of Jamc appears to be more conserved than that of Jamc2 in comparison to mouse Jam-C - 52% vs 45% identity respectively. I performed a more detailed comparison between the *jamc* and *jamc2* loci using zPicture (Ovcharenko *et al*, 2004), an interactive Blastz tool for genomic comparisons (figure 3.11). The resulting alignments suggest a more confusing relationship between the loci at the sequence level, with many sequences aligning in multiple positions within either region. As expected, the majority of the non-coding sequence alignments are between the region containing the *jamc* and *igsf9b* orthologues, although there are a few alignments outside this region. Many of these elements align with many different regions across the genome and might perhaps contain promoter or protein-binding sequences or as yet unidentified repetitive elements.

A multispecies comparison between zebrafish, chicken and mouse suggests a much less complicated evolution of the *Jam-B* loci (figure 3.12). The zebrafish *jamb* locus contains most of the genes present in the mouse and chicken loci (*adamts-1*, *cyyr1*, *appa*, *atp5j* and *mrpl39*), in the correct order and orientation. One clear difference is the apparent lack of an orthologue of *Gabpa*. In contrast, this gene is one of very few to be retained at the *jamb2* locus along with the *App* orthologue *appb*. It seems likely that the *jamb2* locus is derived from a duplication event and the *jamb* locus has retained ancestral characteristics. Both proteins are the least well conserved zebrafish Jams when compared by alignment (figure 3.13). Between paralogues, 46% of residues are identical across the immunoglobulin-like domains; this falls to 38% (Jamb) and 36% (Jamb2) in comparison to mouse Jam-B. Jamb2 also appears to lack all N-linked glycosylation sites, while Jamb has a novel site in the membrane-distal immunoglobulin-like domain, near the 'VLV' linker peptide. The cytoplasmic domains of the zebrafish JAM-B orthologues have diverged significantly from each other and the mammalian JAM-B. There appears to be an extra exon inserted into the cytoplasmic domain-encoding region of *jamb*, meaning the intracellular region is significantly longer than Jamb2 or mouse Jam-B (figure 3.13). Only the few residues close to the C-terminal PDZ-domain binding motif are conserved.

3.4 Discussion

Given that many orthologous genes in the zebrafish genome appear to have been duplicated (reviewed in Volff, 2005 and Ravi and Venkatesh, 2008), I undertook a search for any remaining unidentified *JAM* family genes. I identified two genes

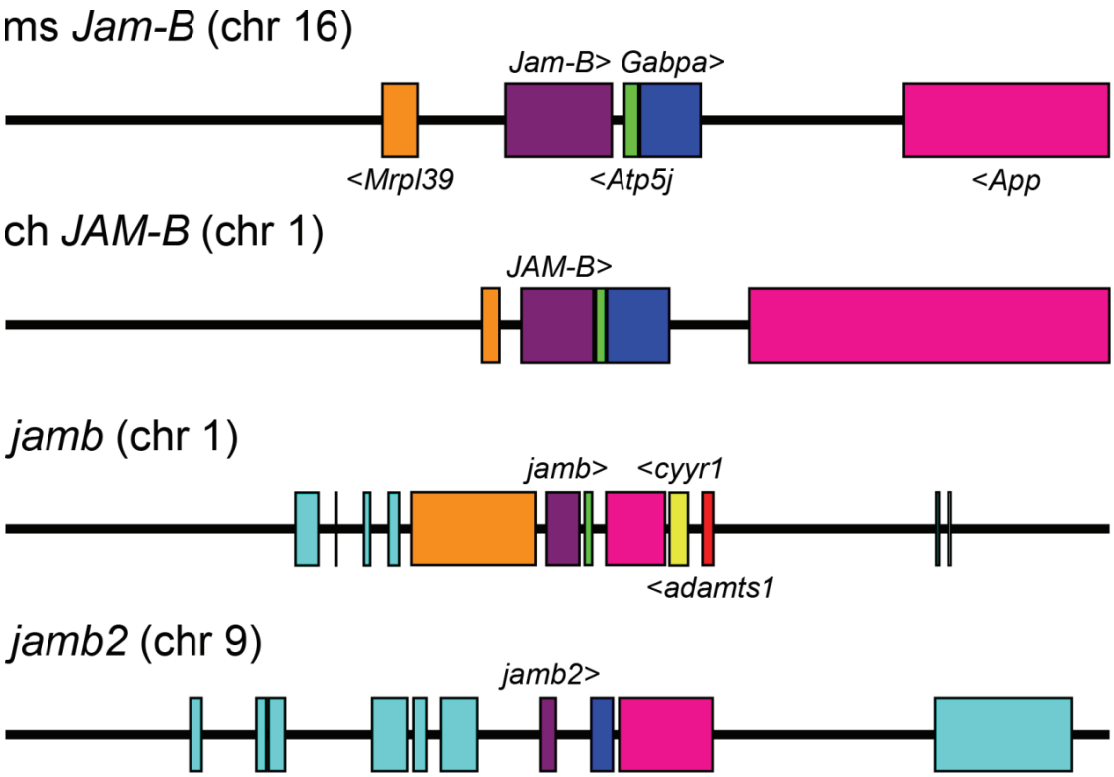


Figure 3.12 Multi-species comparison of *Jam-B* loci suggests *jamb2* is the derived allele of an ancient genome duplication event.

Schematic of the arrangement of annotated genes (coloured boxes) within 0.5 Mb of sequence from the mouse, chicken and zebrafish genomes centered on *Jam-B* orthologues (purple boxes). Genes conserved between these loci are colour-coded according to identity. Two genes present at the *jamb* locus, *cyyr1* (yellow) and *adamts1* (red), are conserved at the murine and chicken loci, but outside of the 0.5 Mbp genomic region presented here. Genes that are present at only one locus are highlighted by blue boxes. The figure is drawn to scale.

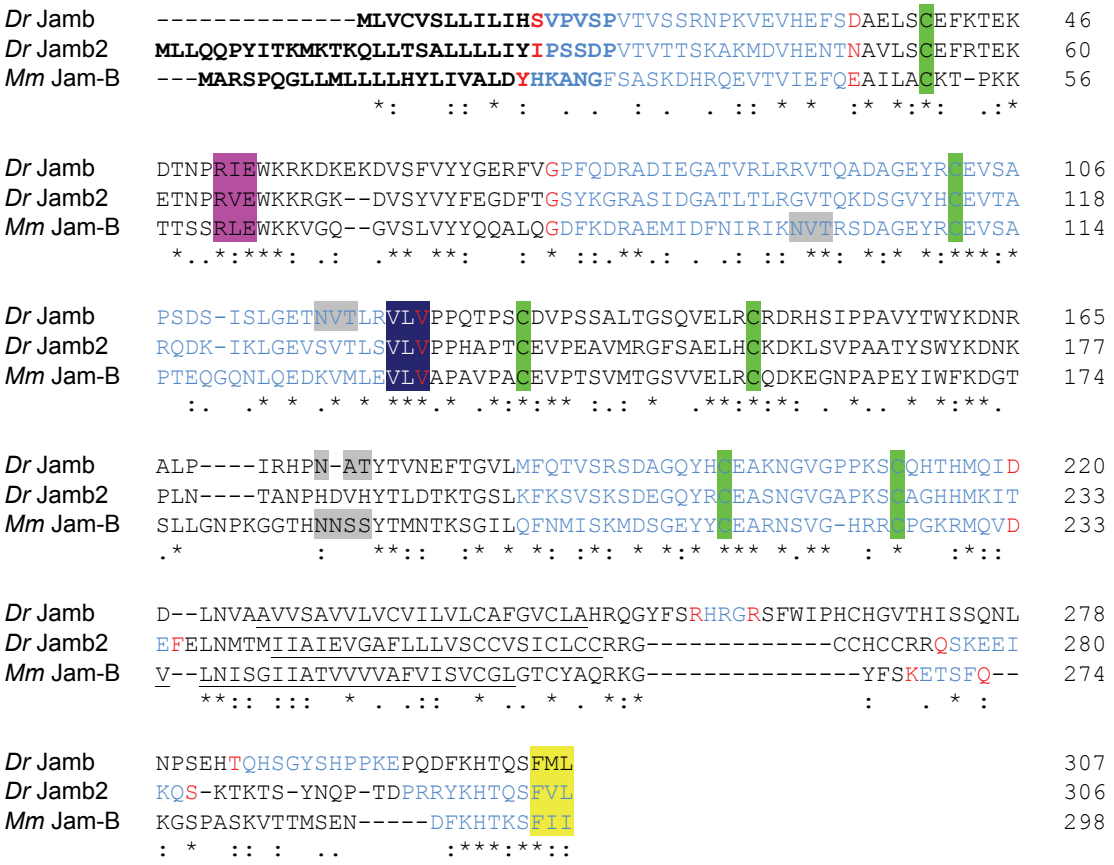


Figure 3.13 Comparison between Jamb, Jamb2 and Jam-B highlights conserved features and divergent cytoplasmic domains.

ClustalW alignment of the amino acid sequence of Jamb, Jamb2 and mouse Jam-B. 44% of residues are identical over the whole alignment of Jamb and Jamb2, rising to 46% over the immunoglobulin-like domains alone. 37% and 34% of residues are identical over the whole alignment between Jamb or Jamb2 and mouse Jam-B, respectively. Feature annotations as in figure 2.

encoding proteins with the appropriate structural features of JAM proteins and confirmed their relationship with the family by alignment, phylogeny and cross-species genome comparison. The presence of paralogous *JAM* genes has important implications for studying the function of any member of the family and relating those experimental results to human biology. Each paralogous gene may retain the same functions as the ancestral gene i.e. be functionally redundant; may have retained different ancestral functions, i.e. undergone sub-functionalisation; or may have diverged to take on different functions, i.e. neo-functionalisation. Gene expression may also be affected by duplication. Important promoter, enhancer and repressor sequences may also have been duplicated; may have been duplicated but subsequently mutated; or may have been lost from the duplicated gene entirely, resulting in a novel pattern of expression. Recombination may also bring the duplicated genes under the control of different regulatory elements. In this chapter, I have attempted to establish the evolutionary relationships between paralogues using a simple bioinformatics approach in order to address some of those questions.

There are two *JAM-C* orthologues in the zebrafish genome, the previously identified *jamc* on chromosome 21 and *jamc2* on chromosome 15. Both genes encode proteins with all the conserved features of JAM family proteins and the additional features of JAM-C - the additional disulfide bridge in the membrane-proximal immunoglobulin-like domain, both glycosylation sites and the transmembrane/cytoplasmic domain encoded by three exons instead of four. Comparing the extracellular domains of both zebrafish proteins with the mouse protein indicates that Jamc is better conserved. However, a multispecies analysis of synteny suggests that the *jamc* locus is derived from the ancestral duplication event. There is strong conservation of gene order at the *jamc2* locus in comparison to the chicken and mouse loci. Comparing both zebrafish loci using zBLAST demonstrates that there is good conservation of non-coding regions between the *jamc* and *igsf9b* orthologues, but few unique alignments outside of this region. This suggests that *jamc* marks the boundary of the duplicated region.

Similarly, there are two *JAM-B* orthologues in zebrafish: *jamb* on chromosome 1 and *jamb2* on chromosome 9. These are the least well conserved members of the zebrafish *jam* family. They retain the important characteristics of the JAMs, but have some interesting differences, in particular the divergent cytoplasmic domains. The C-terminal PDZ domain-binding motif remains intact in the paralogues, but the length and composition of the intracellular region is quite different, in terms of amino acid sequence and splicing of the appropriate regions of mRNA. The implication is a

divergence in intracellular function, but whether or not these changes represent sub-functionalisation is impossible to determine bioinformatically.

The existence of two *JAM-A* orthologues in the zebrafish genome is controversial. I identified a gene, close to *jama* on chromosome 5, encoding two immunoglobulin-like domains that are very highly conserved with those of Jama (79%) and reasonably well conserved with the human and mouse JAM-A (~40%). This appears to have arisen from a duplication of quite a small genomic region that may have included only part of the ancestral *jama* gene. The resulting duplicate acquired novel 3′ exons from local sequences, and as a result lacks any exons encoding a transmembrane domain or anything similar to the cytoplasmic domain of a JAM-A orthologue. As such, it lacks some important structural determinants of the JAM family; most notably, it does not encode a type I cell surface protein with a short cytoplasmic domain ending in a PDZ-domain binding motif. The protein is predicted to have a signal peptide and should therefore be secreted. There is no suggestion that such an arrangement exists in the *Gasterosteus aculeatus* (stickleback) or *Takifugu rubripes* (fugu) genomes (data not shown), implying that the duplication is restricted to the zebrafish sub-lineage of teleost fish. The amino acid sequence of the immunoglobulin-like domains of Jama2 are very well conserved, but there is little or no conservation of introns of *jama* and *jama2*. Given the level of conservation of exons and that the open reading frame was cloned from cDNA, it is unlikely that *jama2* is a pseudogene. It is possible that the *jama2* sequence is from a mis-spliced transcript, but extensive searching of the region 3′ to the immunoglobulin-like domain encoding exons revealed no candidate transmembrane or cytoplasmic domain exons. The significant change in the characteristics of *jama2* and the level of conservation suggest that the paralogue has taken on novel functions.

In summary, the zebrafish *jam* family has been duplicated in the teleost lineage, resulting in six *jam* genes with strong conservation of protein coding sequence between paralogues and orthologues in avians and mammals. The extracellular regions of the paralogues are well conserved, but the intracellular domains are much more divergent. Almost all loci have a local gene structure that is conserved through to mammals, except for *jama* and *jama2*, which are neighbouring genes in a small region with very little similarity to that of other species.