Chapter 3

IGH locus structure and evolution in the Cyprinodontiformes

First Draft, Wednesday 16th January, 2019

3.1 Introduction

The native structure of the immunoglobulin heavy chain (*IGH*) locus determines the state space of antibody heavy-chain diversity in a species, including the range of VH, DH and JH segment choices available in VDJ recombination [1], the relationship between VDJ recombination and isotype choice [2], and the ability of processes such as gene conversion [3] and class-switch recombination [4, 5] to affect the diversity and functionality of the antibody repertoire. The diversity produced by VDJ recombination, junctional diversity and secondary diversification processes in this locus are responsible for the majority of variation in antigen-specificity within a B-cell population, while the choice of isotype among the available *IGH* constant regions determines the antibody's effector function and relationship with the rest of the immune system [6]. Understanding the native structure of the *IGH* locus is therefore essential for understanding how the adaptive immune system functions in a given vertebrate species, while comparing loci between species enables the evolutionary history of adaptive immunity across lineages to be analysed, providing crucial insight into the complex history of this essential biological system. Last but not least, by providing thorough documentation of the *IGH* gene segments present in a species, characterising the *IGH* locus in a species is an essential forerunner to quantitative analysis of adaptive immunity using immunoglobulin sequencing.

Previous work has characterised *IGH* locus structure in a number of teleost species, including [2] zebrafish [7], medaka [8], stickleback [9, 10], rainbow trout [11], fugu [12], and Atlantic salmon [13]. This work has revealed remarkable diversity in the size, structure and functionality of teleost *IGH* loci. However, the number of loci characterised is very small compared to the evolutionary diversity of known teleost species, and is mainly confined to major aquaculture species (trout, catfish, salmon) or research models (zebrafish, stickleback, medaka), with characterised species often quite distantly related to one another across the teleost evolutionary tree. This relatively sparse sampling of teleost *IGH* loci has left wide swathes of teleost diversity without any characterised *IGH* loci, and has prevented higher-resolution analysis of locus structural evolution across closely related species.

In this chapter, therefore, I present complete characterisations of the *IGH* loci of two important model organisms from the Cyprinodontiformes, a diverse clade of primarily freshwater teleost fishes. *Nothobranchius furzeri*, the turquoise killifish, has recently emerged as an important model system for ageing research [14, 15], and is also of evolutionary and ecological interest due to its short lifespan, extreme natural environment, and unusual life history [16]. The southern platyfish *Xiphophorus maculatus*, meanwhile, is an important model organism in evolutionary ecology and population genetics [17]. Comparing the *IGH* loci of these two closely-related species reveals dramatic and unexpected differences in immune structure, which when combined with information from previously-published loci from related species suggest unexpected patterns of locus evolution within this group of teleost fishes. Most strikingly, the specialised mucosal antibody isoform *IGHZ* appears to have been independently lost in multiple closely-related lineages.

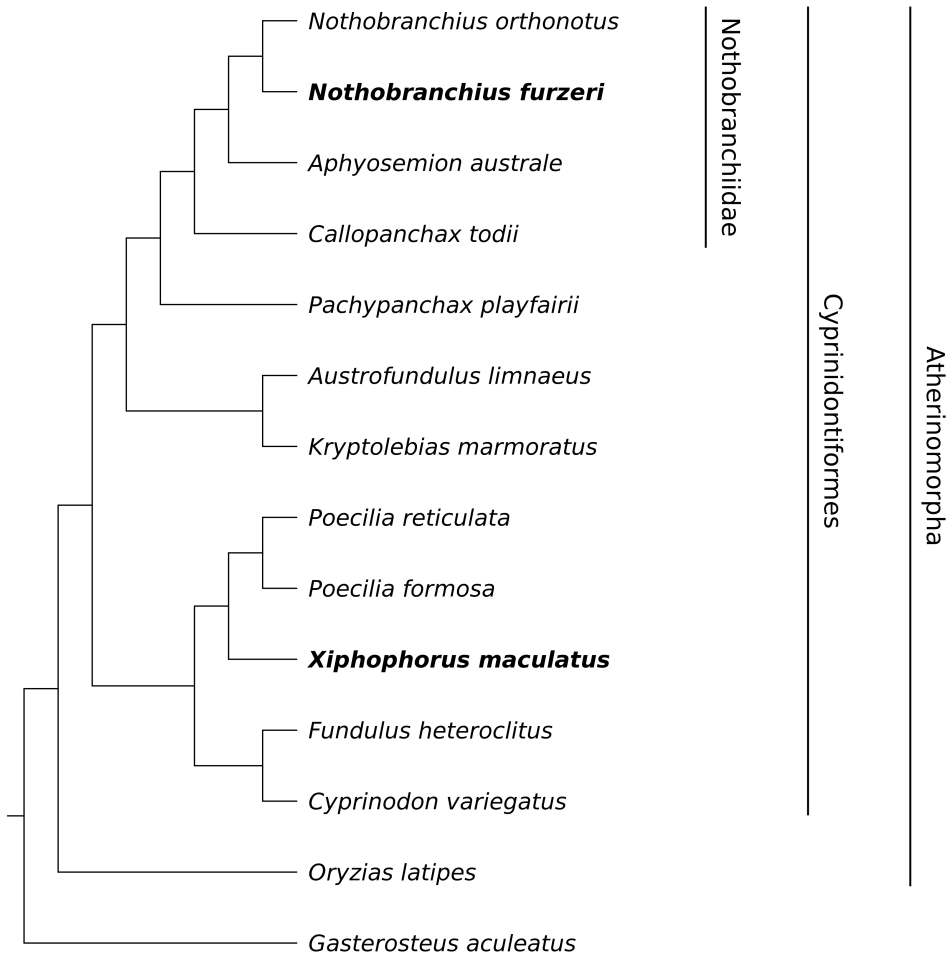


Figure 3.1: Cladogram of species included in this analysis. Boldface type indicates species for which new, complete *IGH* locus assemblies were generated for this study; other species were either previously-characterised reference species (*G. aculeatus*, *O. latipes*) or underwent constant-region characterisation only (all other species). Labelled vertical bars designate higher taxa of interest.

To further investigate this and other surprising features of *IGH* locus evolution in this lineage, I performed a partial reconstruction and analysis of the *IGH* loci from ten further cyprinodontiform species (Figure 3.1), as well as from a new and improved genome assembly of medaka (*Oryzias latipes*), with a focus on the constant-region exons present in each species. Phylogenetic analysis of these data confirms the repeated independent loss of *IGHZ* in this lineage and provides evidence for multiple, independent *IGHZ* subclasses present ancestrally in the clade. Taken together, this analysis significantly extends our knowledge of *IGH* locus and especially constant-region diversity in teleost fish, and establishes the cyprinodontiforms, and especially the African killifishes, as a highly promising collection of model systems for comparative evolutionary immunology.

3.2 The *IGH* locus of *Nothobranchius furzeri*

3.2.1 Assembling the *N. furzeri IGH* locus

In order to locate and characterise the *Nothobranchius furzeri IGH* locus, databases of VH, JH, and CH exon sequences were collated from the published locus sequences of three reference species (zebrafish [7], three-spined stickleback [9, 10], and medaka [8]) and aligned to the *Nothobranchius furzeri* genome (NFZ2.0) with `BLAST` [18, 19]. Genome scaffolds with high-confidence alignments to at least two distinct segment types or covering at least 1% of the scaffold's total length were retained for downstream analysis as potential locus candidates. In total, one chromosome (chr6) and 6 unincorporated scaffolds were identified as potentially covering part of the locus sequence (Table 3.1), with chromosome 6 bearing the majority of identified gene segments.

Table 3.1: *N. furzeri* genome scaffolds containing putative *IGH* locus fragments.

Scaffold	Total length (kb)	V	J	C_μ	C_δ	C_ζ	Included in locus?
chr6	6195.6	15	7	5	11	0	Yes
scf10901	1.4	0	0	0	3	0	Yes
scf21863	13.5	1	0	0	0	0	No
scf35954	16.3	3	0	0	0	0	No
scf36277	18.9	2	1	0	0	0	No
scf37083	17.7	1	0	0	0	0	No
scf9157	7.2	0	7	4	0	0	Yes

Table 3.2: *N. furzeri* BAC-library inserts containing putative *IGH* locus fragments.

BAC ID	Insert length (kb)	V	J	C_μ	C_δ	C_ζ	Included in locus?
154G24	106.6	17	1	0	0	0	No
162F04	119.4	5	1	0	0	0	No
165M01	110.7	15	1	0	0	0	Yes
206K13	106.7	17	1	0	0	0	No
208A08	103.2	17	1	0	0	0	Yes
209K12	133.0	1	8	4	20	0	Yes
220O06	104.8	4	1	0	0	0	No
223M21	99.3	17	1	0	0	0	No
248A22	47.3	7	0	0	0	0	No
276N03	127.9	7	0	0	0	0	Yes
277J10	120.8	17	1	0	0	0	Yes

In order to determine which of the putative locus scaffolds were in fact part of the *IGH* locus, integrate these into a contiguous locus sequence, and provide additional information on any missing gene segments, bacterial artificial chromosome insert sequences from the killifish genome project

BAC library [20] were included in the locus assembly. BAC candidates, whose ends had already been sequenced as part of the genome project, were identified as potentially containing part of the locus sequence on the basis of their ends aligning to promising candidate scaffolds from a previous genome assembly (first round) or to the insert sequences of previously sequenced BAC inserts (second round). Once identified, BAC candidates were isolated from culture by alkaline lysis, sequenced on an Illumina MiSeq sequencing machine, and assembled and scaffolded with `SPAdes` and `SSPACE`, respectively. Complete BAC insert assemblies were generated from these scaffolds by manual alignment to overlapping genome scaffolds and other BAC inserts, combined with PCR and Sanger sequencing of intervening sequences.

Finally, the assembled BAC inserts were screened for *IGH* locus segments in the same manner described for genome scaffolds, and passing insert sequences (Table 3.2) were aligned to and integrated with the identified candidate scaffolds to produce a contiguous locus assembly. To minimise the probability of losing relevant gene segments to assembly errors, priority in the event of a sequence conflict between BACs and scaffolds was given first to any sequence containing a segment missing from the other; if neither the BAC assembly nor the genome scaffolds met this condition, priority was given to the genome scaffold over the BAC assembly. In total, 3 candidate scaffolds (including chromosome 6) and 5 BAC inserts were included in the final locus assembly, while 4 scaffolds and 6 BACs were excluded as likely representing isolated *IGH* orphon segments elsewhere in the genome. The correspondence between the final locus sequence and the sequences used to construct it is shown in Figure 3.2.

3.2.2 Overall locus structure

The turquoise killifish genome contains a single *IGH* locus approximately 306 kilobases in length, located on chromosome 6 of the *N. furzeri* genome (Figure 3.3A). This locus comprises two complete subloci, *IGH1* (155 kb) and *IGH2* (118 kb), present in tandem and each occupying a classic VH-DH-JH-CH translocon configuration. This modified translocon structure, with multiple translocon subloci present in tandem, has been observed in a number of teleost *IGH* loci including catfish, medaka and stickleback [2]. Unusually, however, the smaller *IGH2* sublocus in *Nothobranchius furzeri IGH* is present in antisense relative to the larger *IGH1*, with the two subloci beginning at opposite ends of the locus and facing each other in the middle (Figure 3.3B). Such a multi-orientation locus structure has only previously been observed in medaka, the closest relative of the turquoise killifish to have its locus characterised prior to the present study [8]; it is interesting to see this unusual feature reproduced here, raising the question of whether that this ideosyncracy is homologous between the two loci.

Compared to other closely-related loci, the killifish locus is relatively sparse and simple, with comparatively low functional complexity relative to its overall size. For example, whereas the stickleback locus fits four subloci, 49 V segments and 10 constant regions into c. 200 kb [9, 10], the killifish locus, despite being 50% longer, contains only 2 subloci, four constant regions and 24 V

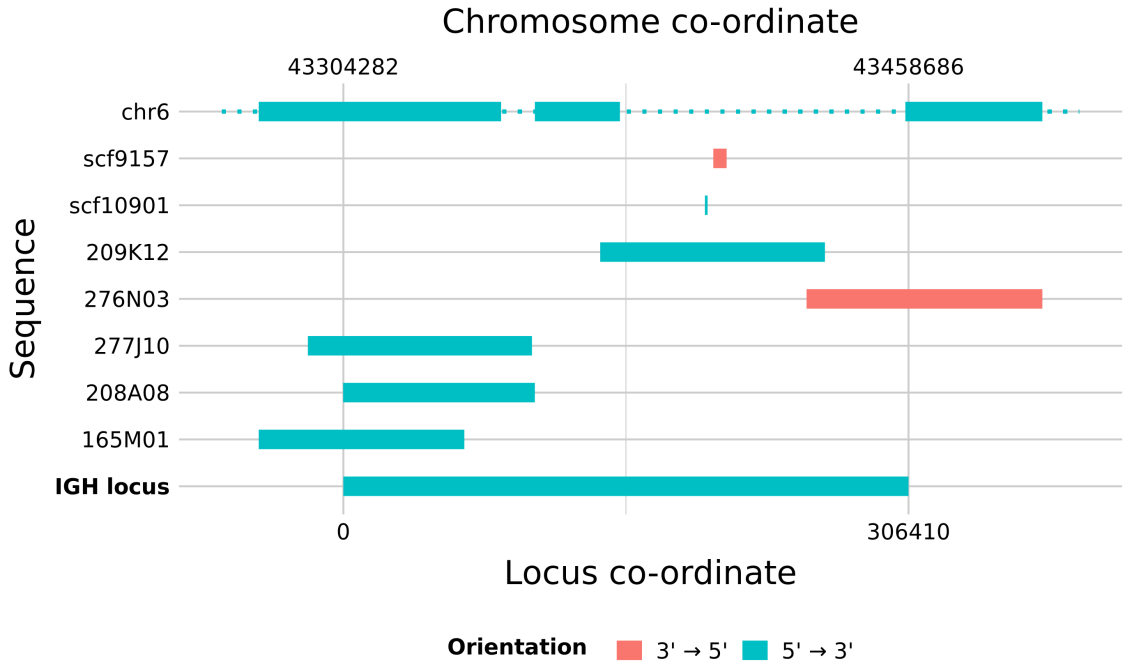


Figure 3.2: Assembling the *N. furzeri IGH* locus: Schematic of genome scaffolds and BAC inserts contributing to the *N. furzeri IGH* locus sequence, with their corresponding place within the locus sequence (bottom axis). Internal gaps with dotted lines indicate locus regions with no corresponding locus sequence, as a result of intercalation of BAC or scaffold sequences.

segments (including pseudogenised Vs). This difference results from the unusually large amount of nonfunctional sequence padding the killifish locus, resulting in large gaps between variable segments and in some cases between constant-region exons (Figure 3.3B); this high prevalence of repetitive DNA is consistent with the rest of the TK genome, which comprises more than 60% repetitive sequence, compared to just over 15% in stickleback [21].

The two subloci in the turquoise killifish locus are generally highly similar in their functional sequence, with a high degree of synteny between their functional regions (Figure 3.4). The greatest degree of divergence occurs in the VH and DH regions, with what appear to be repeated deletion events in IGH2 resulting in a substantially lower number of VH and DH segments compared to IGH1; conversely, the JH and constant regions are almost identical between two subloci. These patterns are discussed in more detail in Sections 3.2.3 and 3.2.4.

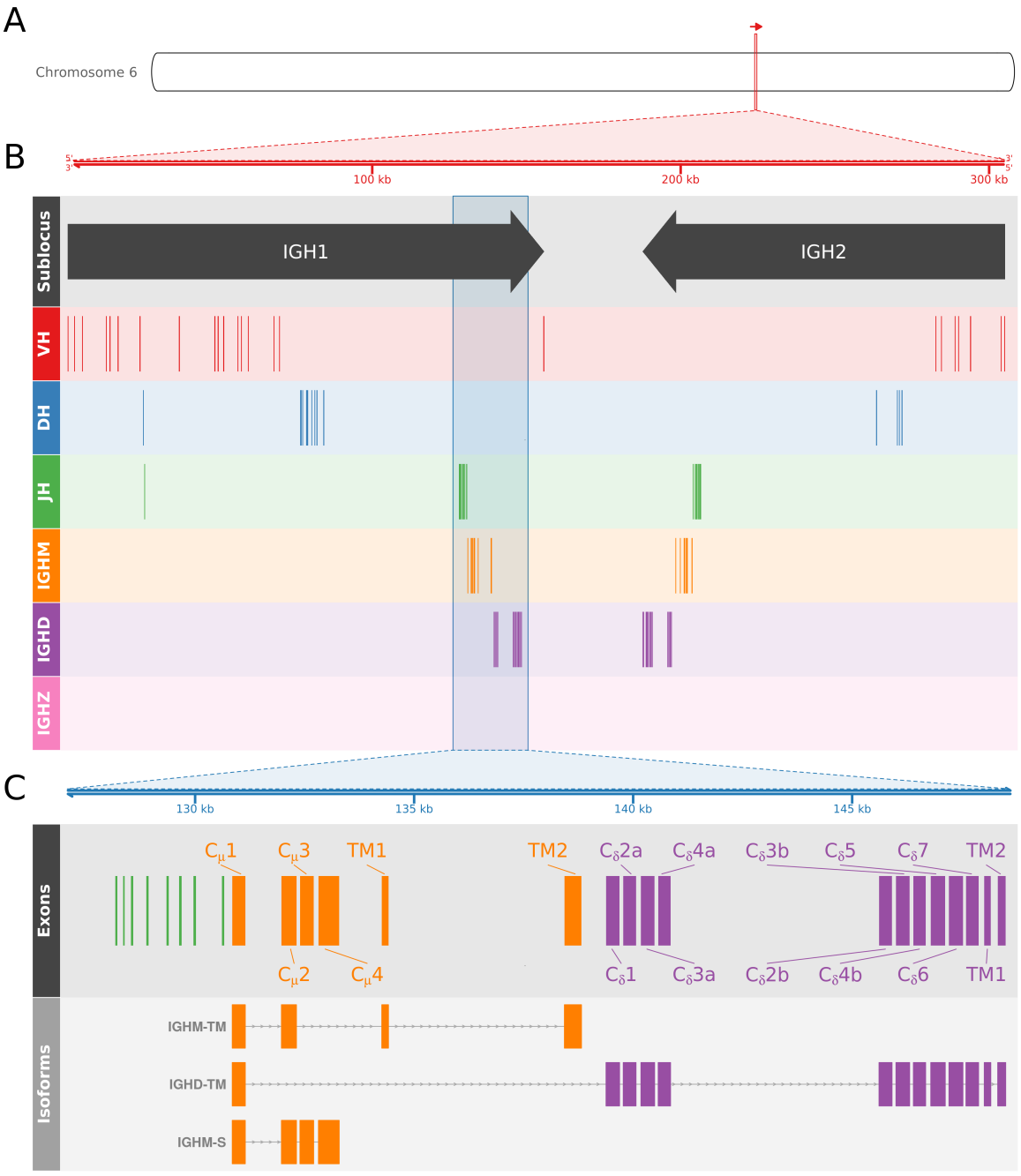


Figure 3.3: The immunoglobulin heavy chain (*IGH*) locus in *Nothobranchius furzeri*: (A) Position of the *IGH* locus on chromosome 6 of the *N. furzeri* genome. (B) Arrangement of VH, DH, JH and constant-region gene segments on the *N. furzeri IGH* locus. All segments follow the orientation of their parent sublocus, indicated in the uppermost track. (C) Detailed map of the constant regions of the *IGH1* sublocus, indicating the position and identity of the constant-region exons and the exon composition of expressed *IGH* isoforms in the turquoise killifish.

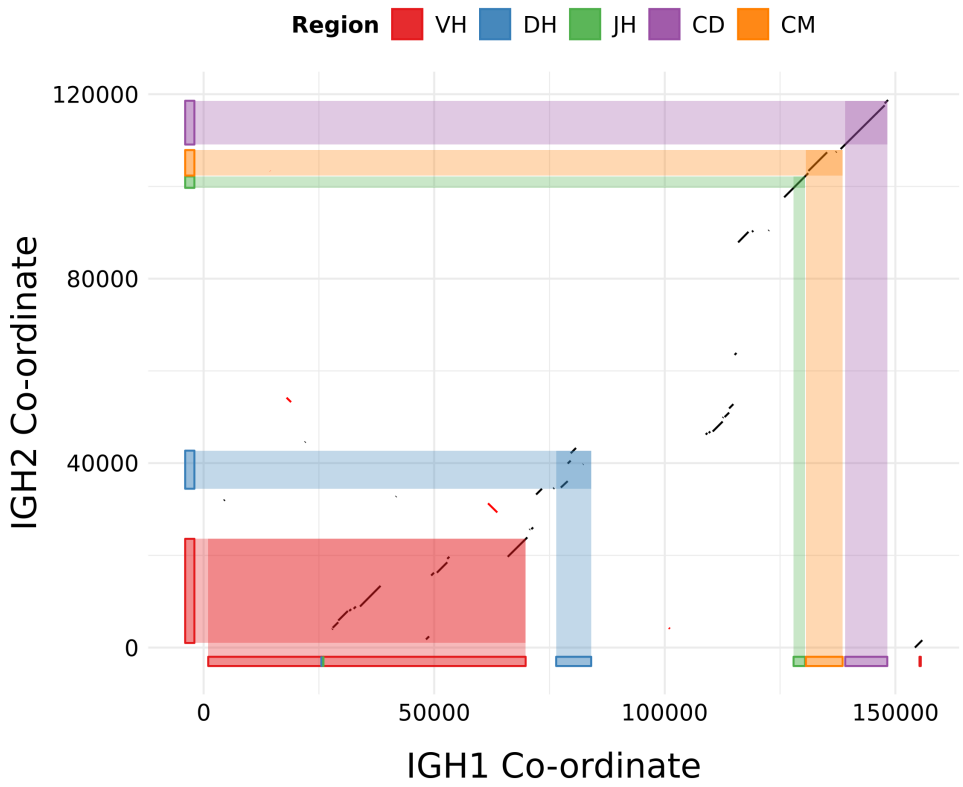


Figure 3.4: Sequence homology between subloci in *N. furzeri IGH*: Synteny plot of sequential best matches between *IGH1* and *IGH2* subloci, with gene segment regions indicated by coloured rectangles along each axis.

3.2.3 Constant regions

The *isotype* (also known as the *class*) of an antibody determines its functional role within the immune system, including its possible effector functions and whether it can be secreted [6]. Three antibody isotypes have been identified to date in teleost fishes: *IGHM*, *IGHD* and *IGHZ* (a.k.a. *IGHT*, *IGHT/Z* or *IGHZ/T*) [2, 22, 23]. Of these, *IGHM* and *IGHD* are highly primitive within the jawed vertebrates and found in most or all other vertebrate groups; within the teleosts, both appear to be universal [22]. Conversely, *IGHZ* is a teleost-specific isotype which is absent in other vertebrate taxa; within the teleosts, most characterised *IGH* loci possess *IGHZ*, but at least two (medaka and channel catfish) have been found to lack it [2, 22]. In rainbow trout, *IGHZ* has been found to play a specialised mucosal role in the immune system analagous to that of *IGHA* in mammals [24, 25], and it is widely assumed to play this specialised role throughout the teleosts; it is as yet unclear how mucosal immunity is effected in species lacking *IGHZ*.

In order to investigate constant regions in the *Nothobranchius furzeri IGH* locus, putative exon sequences were identified using `BLAST` alignments to the reference sequence databases described in Section 3.2.1, and intron/exon bounderies were refined through alignment of published RNA-sequencing data from killifish gut ([26], BioProject accession PRJNA379208, young and old untreated

groups) using STAR Figure 3.6. Strikingly, the *Nothobranchius furzeri IGH* locus appears to completely lack any *IGHZ* constant region, with no C_ζ exons or *IGHZ* transmembrane exons being found on either *IGH1* or *IGH2*. Given the widespread prevalence and specialised mucosal role of *IGHZ* in teleosts, its surprising absence in turquoise killifish (Figure 3.3B) immediately raises questions about the nature, kinetics and efficacy of mucosal adaptive immunity in this species. The similar absence of *IGHZ* in medaka, which again is the closest relative of *N. furzeri* with a characterised locus, raises further questions about the evolutionary history of *IGHZ* in the Atherinomorpha: does the shared absence of *IGHZ* in these species indicate a single ancestral deletion event, or parallel loss of this important isoform within both the Cyprinodontiformes (including the turquoise killifish) and Beloniformes (including medaka)? This latter question requires higher phylogenetic resolution to address effectively, and is investigated further in Section 3.3 and Section 3.4.

While *IGHZ* is completely missing from the *N. furzeri IGH* locus, *IGHM*, the most primitive and widely-found isotype in jawed vertebrates, is present in its expected location, immediately downstream of the main JH-region in both subloci. This constant region occupies the standard six-exon configuration, with four C_μ exons and two transmembrane exons present in series on the chromosome (Figures 3.3B, 3.3C and 3.5A, Table 3.4). As with other species, both secreted and transmembrane isoforms of *IGHM* are present in the transcriptome, with secreted *IGHM* (*IGHM-S*) consisting of C_μ1-4 (Figures 3.3C, 3.5B and 3.6A); however, the exon configuration of transmembrane *IGHM* (*IGHM-TM*) deviates from both that seen in mammals (in which exon TM1 is spliced to a cryptic splice site within C_μ4) and most teleosts (in which the canonical splice site following C_μ3 is used and C_μ4 is excised) [2]. Rather, turquoise-killifsh *IGHM-TM* resembles that of medaka, in which both C_μ3 and C_μ4 are excluded and the canonical splice site at the end of C_μ2 is spliced directly to TM1 (Figures 3.5C to 3.5E). This similarity to medaka again raises the possibility that this unusual feature may be a conserved feature of both lineages; however, the underlying mechanism giving rise to this difference in splicing behaviour is unknown.

Unlike *IGHM*, the exon structure of *IGHD* is highly variable across the teleosts, ranging from roughly 7-17 C_δ exons in addition to the transmembrane domains [2]. The core structure of *IGHD* comprises seven C_δ exons (C_δ1-7), but some subset of these exons may be missing or duplicated in any given species – in medaka, for example, C_δ5 is missing in all subloci [8], while in many species (e.g. zebrafish, salmon, and channel catfish) C_δ2-4 are duplicated in two or more tandem blocks [2]. This latter configuration is also observed in turquoise killifish, in which the *IGHD* constant region immediately follows *IGHM* in both subloci and has a

C_δ1-(C_δ2-C_δ3-C_δ4)$_2$-C_δ5-C_δ6-C_δ7-TM1-TM2

configuration, for a total of 12 exons per *IGHD* constant region (Figures 3.3B and 3.3C, Table 3.4). All of these exons appear to be expressed in tandem, resulting in a much longer transcript than is observed for any isoform of *IGHM* (Figures 3.3C and 3.6B). As in other teleost species, *IGHD* in the

Table 3.3: Cross-sublocus sequence similarity in constant-region exons in *N. furzeri*: Percentage sequence identities of pairwise Needleman-Wunsch global alignments between nucleotide (NT) or amino-acid (AA) sequences of corresponding exons from the two subloci of *N. furzeri IGH*.

Isotype	Exon	NT	AA
M	1	99.66	100.00
M	2	100.00	100.00
M	3	100.00	100.00
M	4	100.00	100.00
M	TM1	99.34	98.00
M	TM2	91.67	100.00
D	1	99.03	97.06
D	2A	98.97	98.96
D	3A	98.72	97.09
D	4A	99.65	98.92
D	2B	100.00	100.00
D	3B	98.72	96.12
D	4B	99.64	98.91
D	5	99.09	99.08
D	6	100.00	100.00
D	7	100.00	100.00
D	TM1	97.99	97.96
D	TM2	99.44	100.00

turquoise killifish includes a chimeric $C_\mu 1$ exon at the 5' end of the constant-region transcript, for a total of 13 exons per *IGHD-TM* mRNA (Figure 3.6B).

While the best-known form of *IGHD* in teleosts is transmembrane, secreted *IGHD* has been observed in at least two teleost species, with different mechanisms used in each case: in channel catfish, one dedicated sublocus has a dedicated IgD secretory exon in place of the transmembrane exons [27], while in rainbow trout (and possibly some other species like Atlantic salmon and cod) a run-on event at the end of $C_\delta 7$ results in the production of a secretory tail in a manner similar to secretory IgZ [28]. However, neither a specialised secretory exon nor a $C_\delta 7$ secretory tail could be detected in turquoise killifish, suggesting that IgD may only be expressed in transmembrane form in this species.

In the case of both *IGHM* and *IGHD*, the constant regions are present in their expected configuration in each sublocus and are highly similar in sequence between the subloci, with an average of 98.4% nucleotide sequence identity for corresponding IgM exons and 99.3% for corresponding IgD exons (Figure 3.7 and Table 3.3) in pairwise Needleman-Wunsch alignments [29]. This high level of similarity indicates either a very recent duplication event to produce the second sublocus or a high level of sequence conservation in both subloci, with the latter explanation suggesting that both subloci continue to be functional and active in the immune system.

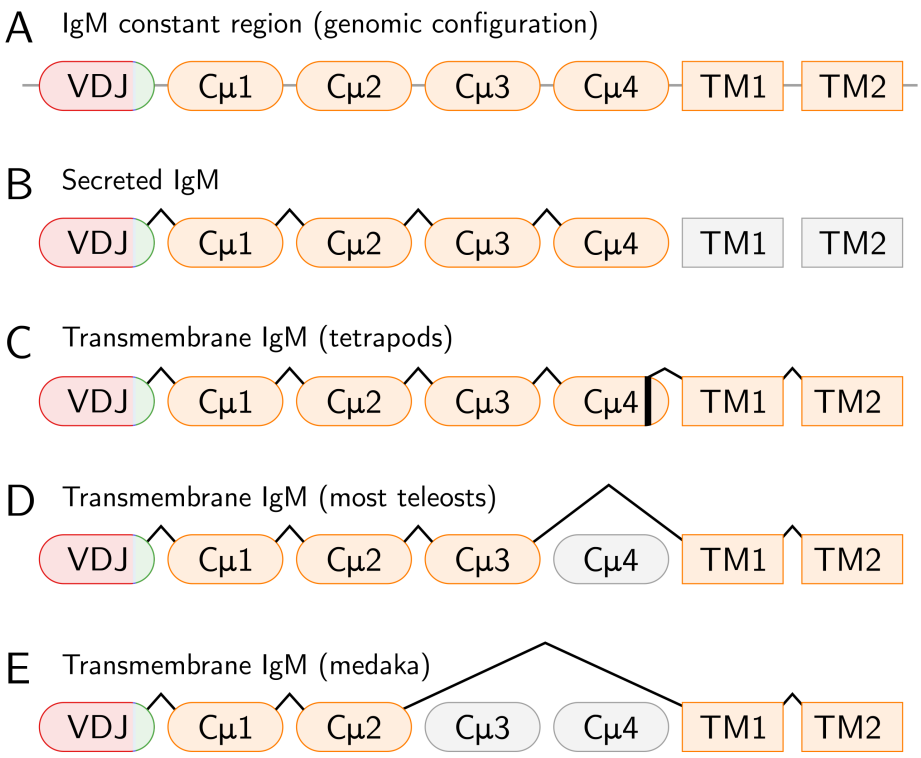


Figure 3.5: *IGHM* exon usage in other vertebrates: Schematic of *IGHM* splice patterns in different isoforms and taxonomic groups; (A) standard genomic (pre-splicing) configuration of *IGHM*, following VDJ recombination; (B) exon configuration of secreted *IGHM* (*IGHM-S*) in tetrapods and teleosts; (C) exon configuration of transmembrane *IGHM* (*IGHM-TM*) in tetrapods, demonstrating the use of a cryptic splice site in $C_{\mu}4$; (D) standard *IGHM-TM* exon configuration in teleosts, demonstrating the direct splicing of $C_{\mu}3$ to TM1 and exclusion of $C_{\mu}4$; (E) unusual *IGHM-TM* exon configuration observed in medaka, in which both $C_{\mu}3$ and $C_{\mu}4$ are excluded. Figure adapted from Fillatreau *et al.* (2013).

3.2.4 Variable regions

Variable-region gene segments in the killifish *IGH* locus were identified with a variety of methods, depending on the type of gene segment being analysed. VH candidates were identified probabilistically using Hidden Markov Models constructed by `nhmmer` [30] from `PRANK` [31] multiple-sequence alignments of reference sequences, with the 3'-ends of each V-exon identified by the presence of a recombination-signal sequence (RSS) [6] and the 5'-ends refined using `IMGT-DomainGapAlign` [32]. JH candidates were also identified using `nhmmer`, with segment ends identified by the presence of an RSS (5') and a `GTA` splice-site motif (3') [8]. Finally, DH-segments, being too short and variable in sequence for HMM-based approaches to be effective, were identified by searching for pairs of flanking RSS sequences in opposite orientation, using fuzzy pattern-matching (with EMBOSS `FUZZNUC` [33]) to conserved RSS sequence motifs.

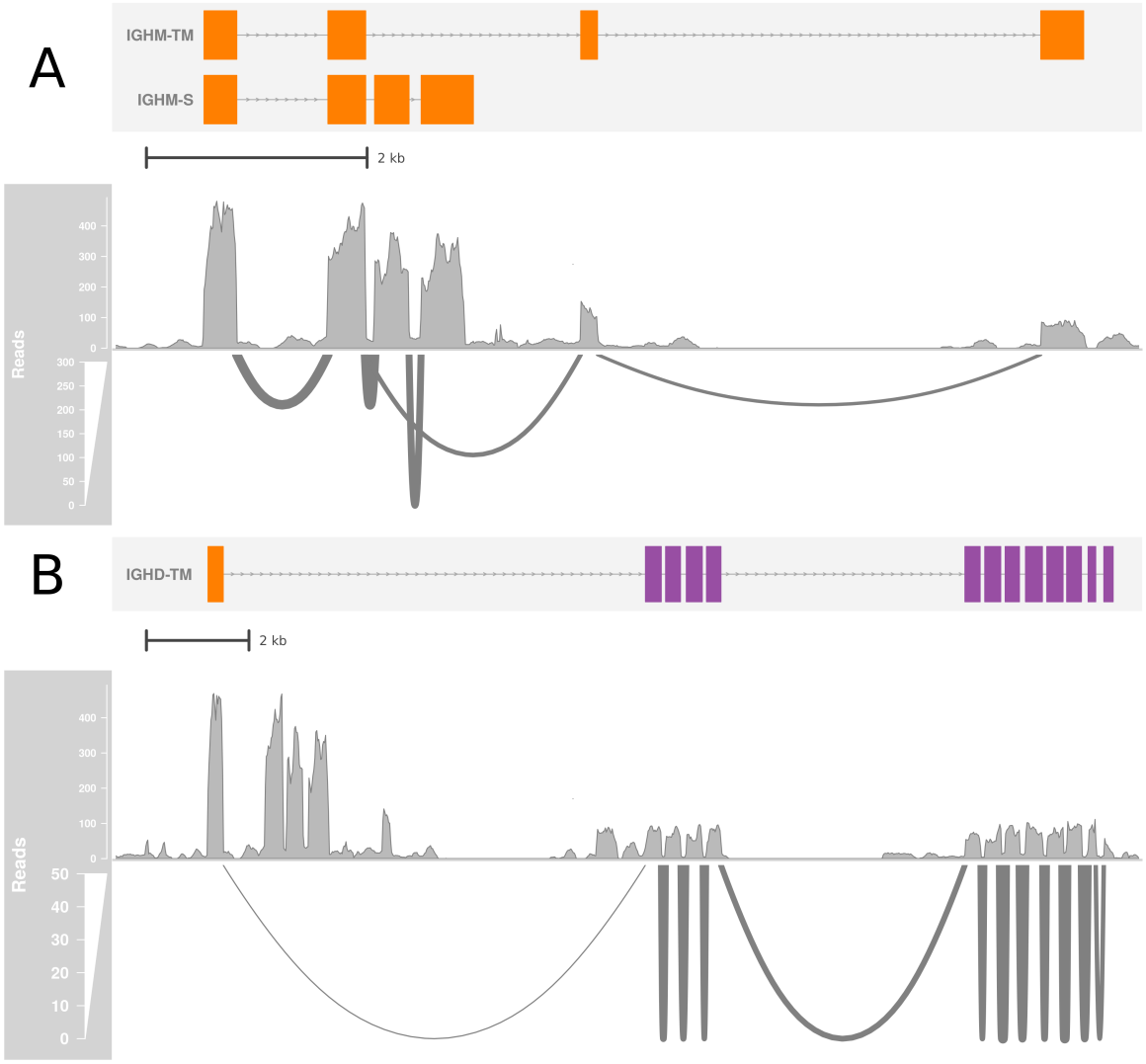


Figure 3.6: Constant-region isoforms in *N. furzeri*: Coverage and sashimi plots of `STAR`-aligned RNA-seq reads from *N. furzeri* gut samples [26], demonstrating the splicing behaviour of *IGH1* constant-region isoforms. (A) *IGHM* exon splicing, showing alternative splicing patterns of *IGHM-TM* and *IGHM-S*; (B) *IGHD* exon splicing, including chimeric splicing of $C_\mu 1$ to $C_\delta 1$.

In total, 24 VH-segments, 14 DH-segments and 17 JH-segments were identified in the *N. furzeri* locus (Tables 3.6, 3.7 and 3.10), of which the majority (17 VH, 10 DH and 8 JH) were present in *IGH1*. Of the VH segments identified, three contain premature STOP codons, though none is out-of-frame; conversely, all the DH and JH segments identified appear to be in-frame and functional, with no premature STOP codons. However, in all cases a minority of segments contain RSS sequences that deviate significantly from the expected consensus sequence (Tables 3.6, 3.8 and 3.9 and **??**); it is unclear whether these sequences can recombine to successfully produce mature VDJ sequences *in vivo*. In the case of the VH segments, of the six sequences without clearly functional RSS sequences,

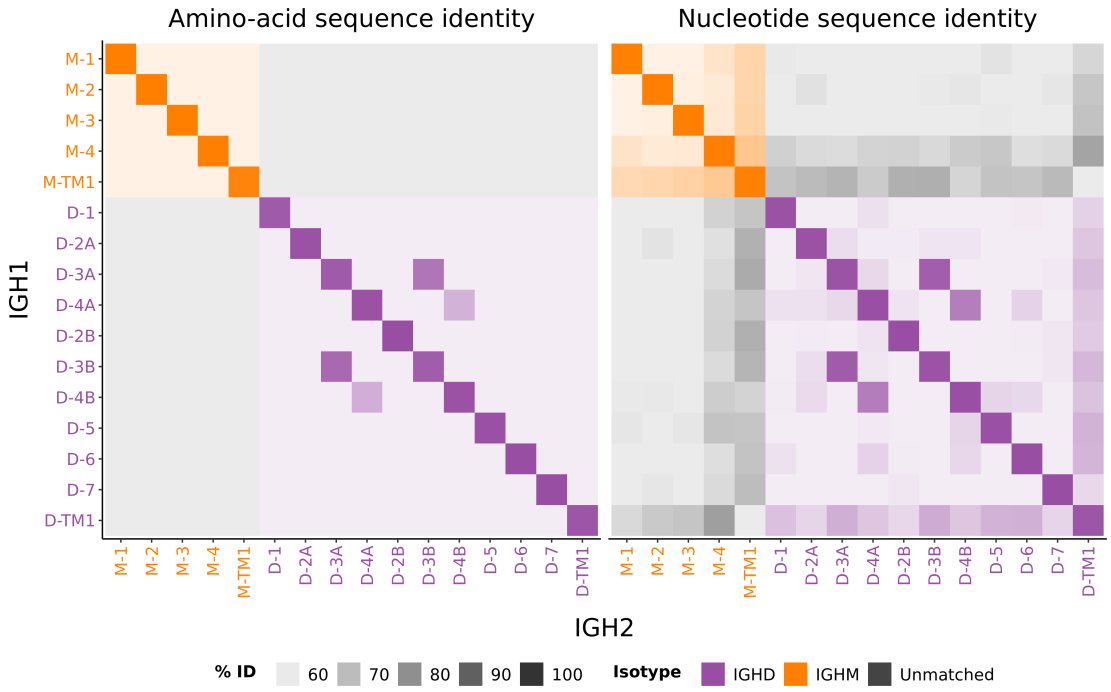


Figure 3.7: Cross-sublocus sequence similarity in constant-region exons in *N. furzeri*: Heatmap of percentage sequence identity between amino-acid (right) and nucleotide (left) sequences of constant-region exons (excluding *IGHM-TM2* and *IGHD-TM2*) from the two subloci of *N. furzeri IGH*, calculated using pairwise Needleman-Wunsch global alignments..

three also contain premature STOP codons, suggesting the changes to the RSS in these cases may arise from relaxed purifying selection on already-pseudogenised sequences.

Apart from these few exceptions, however, the recombination signal sequences (RSS) marking the ends of the VH, DH and JH gene segments in the *N. furzeri* locus otherwise strongly resemble those of other characterised teleosts, which in turn resemble those of non-teleost loci (Figures 3.10 and 3.11). The overall heptamer and nonamer consensus sequences (`CACAGTG` for heptamers and `ACAAAAACC` for nonamers) closely matched those expected from the literature [6], while in 88% of cases the spacer region was within 1bp of the expected length (12bp for D-RSSs, 23bp for V- and J-RSSs); unexpectedly, the greatest number of VH-RSSs had a 22bp (rather than 23bp) spacer, but this is unlikely to interfere with RSS functionality. Overall, the RSSs in the turquoise killifish appear to be supporting the normal operation of VDJ-recombination in this species.

Of the VH, DH and JH segments identified, all but one of each type of segment is located within contiguous V-, D-, and J-regions within each sublocus, supporting a modified translocon configuration for killifish *IGH*. The exceptions to this are *IGH1D01* and *IGH1J01*, which are embedded within the *IGH1* V-region, and a single VH segment located in between the *IGHD* contant regions of the

two subloci (Figure 3.3B). The unusual location of *IGH1D01* and *IGH1J01* may represent the result of a transposition event within the *IGH* locus; however, their close colocalisation and 5' position within the *IGH1* sublocus, as well as the fact that neither has a close paralogue in *IGH2* (Figure 3.9B), suggest that they may instead represent the remnant of a formerly present *IGHZ* constant region, as these typically have dedicated D/J segments independent of those serving *IGHM*. Given its forward orientation, meanwhile, the orphaned VH-segment was assigned to the *IGH1* sublocus as *IGH1V1-07*; however, if annotated correctly, it is unlikely to successfully recombine with segments in either sublocus due to its unusual location.

VH sequences within an *IGH* locus are conventionally grouped into families on the basis of nucleotide sequence identity, with a typical identity cutoff of 80% [23]. In order to group the *N. furzeri* VH genes into families, pairwise Needleman-Wunsch global alignments were performed on each pair of VH sequences to obtain pairwise identity scores, followed by single-linkage clustering on the resulting identity matrix. Cutting the dendrogram at 80% sequence identity revealed a total of six VH families, of which four contained more than one VH segment (Figure 3.8); this number of VH families in the *N. furzeri* locus is roughly in line with those found in related species (Table 3.5). Of these, V1 and V2 make up the bulk (42% and 29% respectively) of the VH segments in the locus. V2 and V4 are highly similar, and all the members of V4 are pseudogenised by premature STOP codons; it may therefore be more appropriate to regard V4 as a pseudogenised subfamily of V2 than as a VH family in its own right.

The total number of functional VH segments in the killifish locus is unusually small in comparison to the total numbers observed in many other teleost species (Table 3.5); however, the number of segments per sublocus is in line with the numbers seen in closely-related species (2 to 12 in medaka, 6 to 18 in stickleback), with the overall difference mainly arising from a difference in the number of subloci per locus. A similar pattern is observed with DH and JH segments, with similar numbers of segments per sublocus in killifish and closely-related species, especially medaka. It therefore appears that the per-sublocus segment diversity available to the turquoise killifish is similar to that of previously characterised species, with any difference in total available diversity at this level arising from differences in the number of functional subloci rather than the size of the V/D/J-regions *per se*.

As can be seen from Figure 3.4, much of the V-, D- and especially J-region sequence in the *N. furzeri* locus is syntenic between the two *IGH* subloci, with downstream portions of the *IGH2* V-region corresponding to downstream parts of the *IGH1* region. Of the seven VH segments in *IGH2*, six have a corresponding segment on *IGH1* with which they share at least 97% sequence identity (Figure 3.8), and these partner segments are largely (though not entirely) colinear in their ordering between the two subloci. A similar pattern can be observed for the D- and (especially) the J-regions: of the four DH segments detectable in *IGH2*, three (*IGH2D02* to *IGH2D04*) are identical with another block of adjacent DH segments in *IGH1* (*IGH1D05* to *IGH1D07*), while the JH-regions exhibit almost

complete sequence identity between the eight JH segments of the main JH region in *IGH1* and the eight JH segments in *IGH2* (Figure 3.9).

Nevertheless, as is clear from Figure 3.4, there are large portions the *IGH1* variable region, including the first 25 kilobases of the V-region, for which no corresponding sequence exists in *IGH2*, and there are many VH and DH segments in *IGH1* (and a much smaller number in *IGH2*) for which no close homologue exists in the other sublocus. Taken together, these data are consistent with a model in which *IGH2* was produced via duplication and inversion of all or part of *IGH1*, followed by subsequent deletion events in the redundant, and structurally volatile, *IGH2* VH and DH regions. However, it is not clear at present how to distinguish between this model and an alternative one of expansion in *IGH1*, or to identify why the JH region is so much more conserved between subloci than either the VH or JH regions.

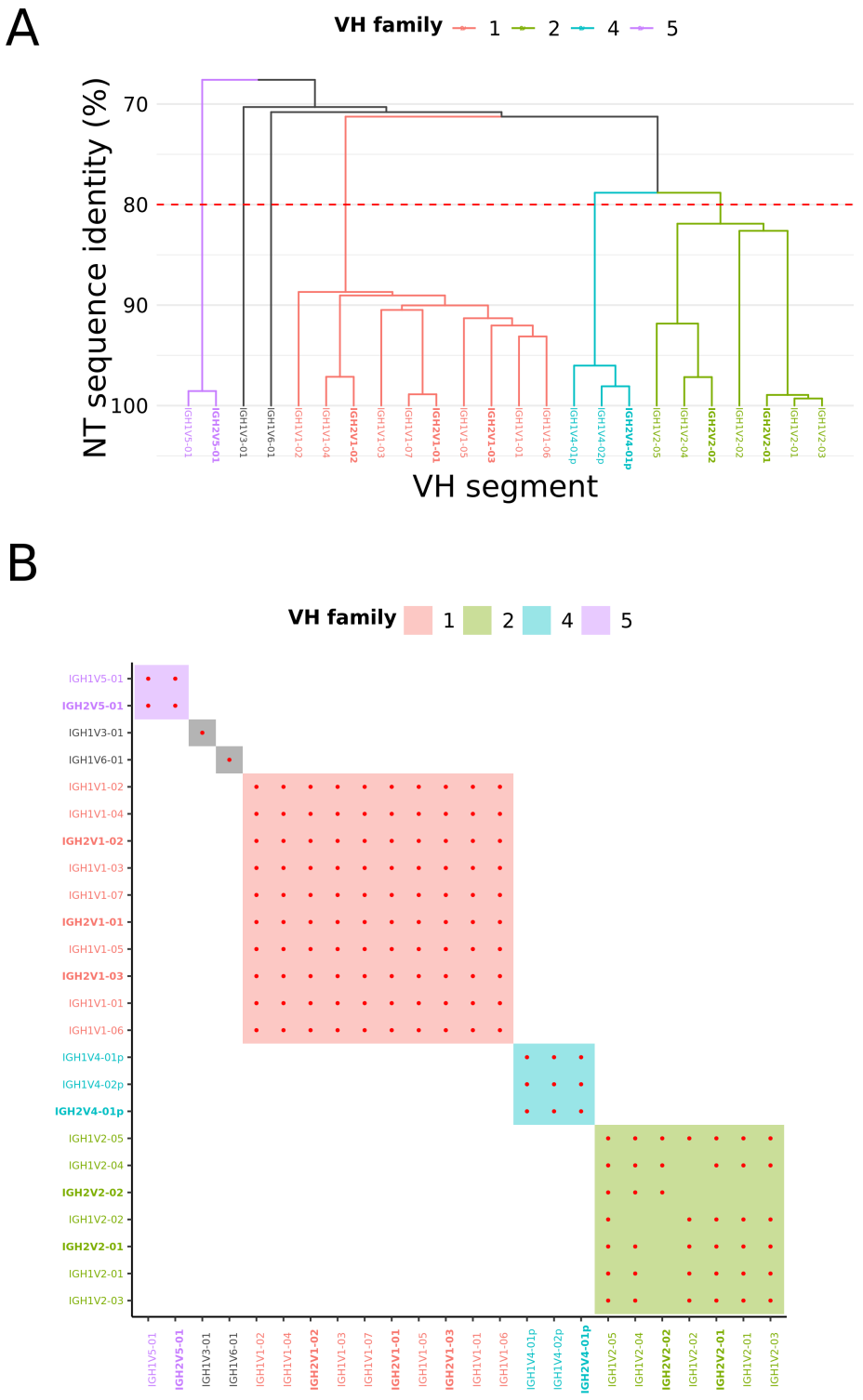


Figure 3.8: VH families in the in *N. furzeri IGH* locus: (A) Dendrogram of sequence similarity of VH segments in the *N. furzeri IGH* locus, arranged by single-linkage clustering on nucleotide sequence identity. The red line indicates the 80% cutoff point for family assignment, while branch colour indicates family membership. (B) Heatmap of family relationships among *N. furzeri* VH segments, with coloured shading indicating families and red dots indicating pairwise nucleotide sequence identity of at least 80%. In both subfigures, VH families containing multiple segments are coloured, single-segment families are in grey, and segments from the *IGH2* sublocus are displayed in **boldface**.

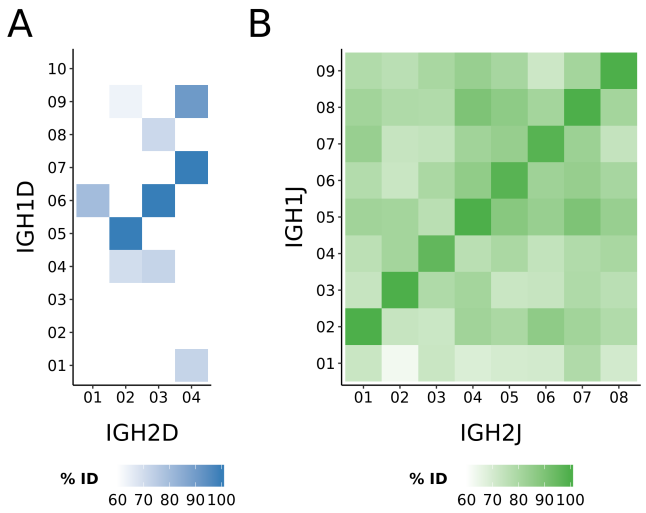


Figure 3.9: Cross-sublocus sequence similarity in DH and JH gene segments in *N. furzeri*: Heatmap of percentage nucleotide sequence identities of Needleman-Wunsch global alignments between (A) DH and (B) JH gene segments in IGH1 vs IGH2, revealing syntenic runs of highly similar sequences across both subloci.

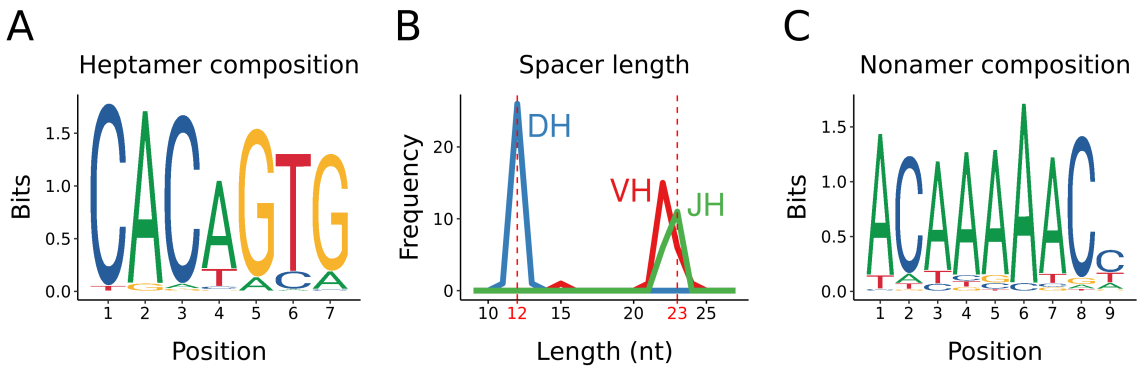


Figure 3.10: Recombination signal sequences in the *N. furzeri IGH* locus: (A) Sequence composition of conserved heptamer sequences across all *N. furzeri* heavy-chain RSSs; (B) length distribution of unconserved spacer sequences in *N. furzeri* heavy-chain RSSs; (C) sequence composition of conserved heptamer sequences across all *N. furzeri* heavy-chain RSSs.

Table 3.4: Co-ordinate table of constant-region exons in the *Nothobranchius furzeri IGH* locus.

Name	Isotype	Start	End	Length	Strand
IGH1M-1	M	130848	131144	297	+
IGH1M-2	M	131971	132312	342	+
IGH1M-3	M	132394	132705	312	+
IGH1M-4	M	132816	133288	473	+
IGH1M-TM1	M	134262	134413	152	+
IGH1M-TM2	M	138431	138819	389	+
IGH1D-1	D	139381	139689	309	+
IGH1D-2A	D	139774	140064	291	+
IGH1D-3A	D	140178	140489	312	+
IGH1D-4A	D	140572	140853	282	+
IGH1D-2B	D	145613	145909	297	+
IGH1D-3B	D	146000	146311	312	+
IGH1D-4B	D	146398	146676	279	+
IGH1D-5	D	146795	147124	330	+
IGH1D-6	D	147210	147527	318	+
IGH1D-7	D	147598	147885	288	+
IGH1D-TM1	D	148016	148164	149	+
IGH1D-TM2	D	148323	148504	182	+
IGH2D-TM2	D	187624	187803	180	-
IGH2D-TM1	D	187963	188111	149	-
IGH2D-7	D	188658	188945	288	-
IGH2D-6	D	189016	189333	318	-
IGH2D-5	D	189419	189748	330	-
IGH2D-4B	D	189867	190145	279	-
IGH2D-3B	D	190232	190543	312	-
IGH2D-2B	D	190636	190932	297	-
IGH2D-4A	D	195644	195925	282	-
IGH2D-3A	D	196008	196319	312	-
IGH2D-2A	D	196433	196723	291	-
IGH2D-1	D	196808	197116	309	-
IGH2M-TM2	M	198315	198506	192	-
IGH2M-TM1	M	199834	199985	152	-
IGH2M-4	M	200953	201425	473	-
IGH2M-3	M	201536	201847	312	-
IGH2M-2	M	201929	202270	342	-
IGH2M-1	M	203549	203845	297	-

Table 3.5: Number of functional VH-segments and VH-families in other teleost species.

Common Name	Species	# Functional VH Segments	# VH Families	Source
Zebrafish	*Danio rerio*	39	13 [1]	[23]
Grasscarp	*Ctenopharyngodon idella*	8	5 [2]	[34]
Fugu	*Takifugu rubripes*	34	3	[23]
Medaka	*Oryzias latipes*	35	6	[2, 8]
Stickleback	*Gasterosteus aculeatus*	49	4	[23]
Turquoise killifish	*Nothobranchius furzeri*	21 [3]	6	–

[1] VH families in zebrafish identified based on 70% (rather than 80%) sequence identity.
[2] It is not clear what clustering method or threshold was used to identify VH families in grasscarp.
[3] Excluding VH segments with nonsense or frameshift mutations, but not those with uncertain or missing RSS sequences.

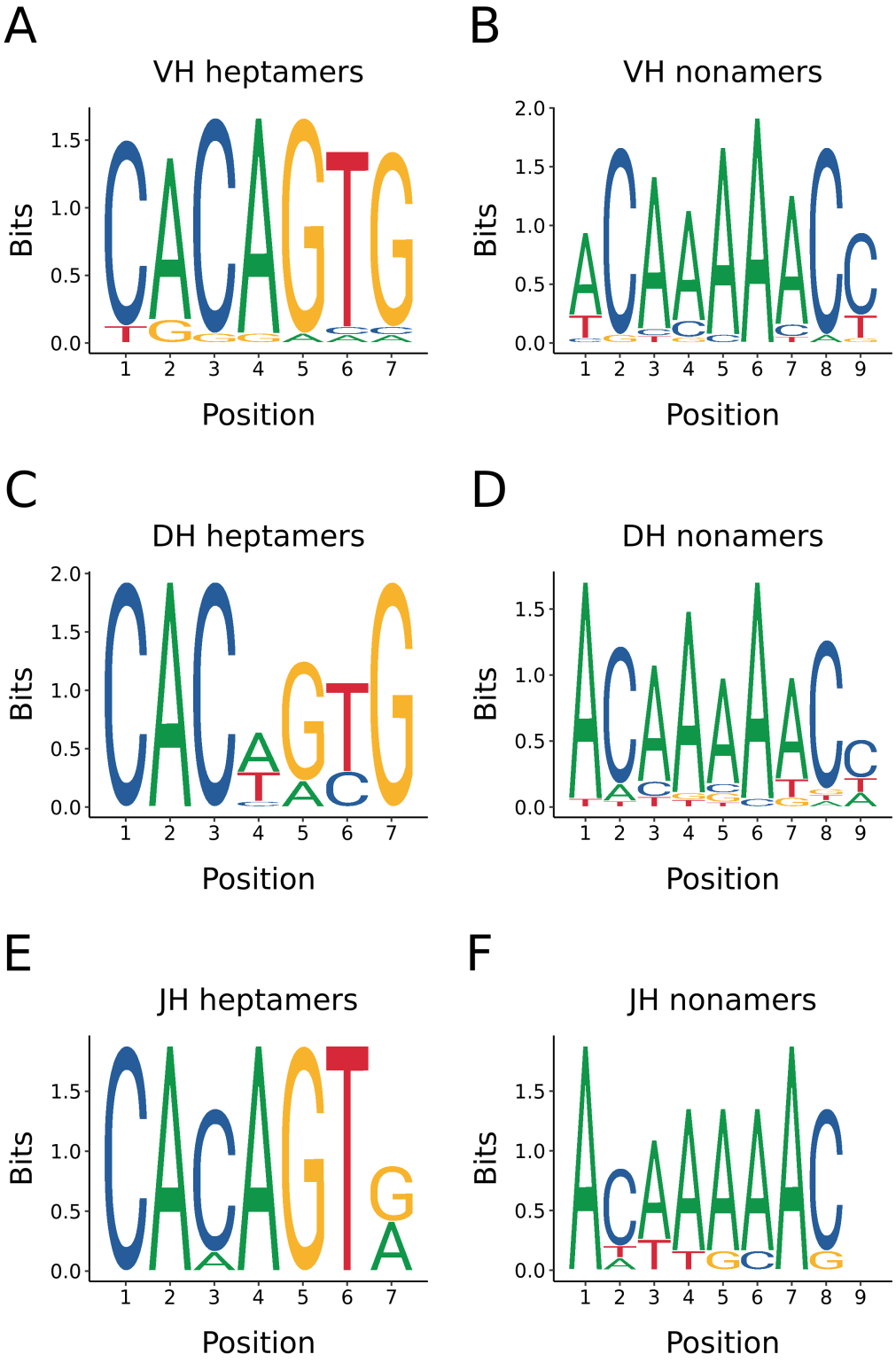


Figure 3.11: *N. furzeri* recombination signal sequences by segment type: Sequence composition of conserved heptamer (A,C,E) and nonamer (B,D,F) sequences from *N. furzeri* heavy-chain RSSs associated with VH (A,B), DH (C,D) or JH (E,F) gene segments.

Name	Start	End	Length	Strand	RSS Start	Heptamer	Spacer Length	Nonamer	RSS End	RSS Length	Comment
IGH1V1-01	1252	1540	289	+	1541	CACAGTG	22	ACAAAAACC	1578	38	
IGH1V1-02	3365	3656	292	+	3657	CACAGTG	22	ACAAAAACC	3694	38	
IGH1V2-01	5907	6201	295	+	6202	CACAGAA	15	ACAAAAACT	6232	31	
IGH1V1-03	13690	13964	275	+	13965	CACAGTG	22	ACAAAAACC	14002	38	
IGH1V3-01	14862	15162	301	+	15163	CACAGTG	23	ACAAAAACC	15201	39	
IGH1V2-02	17433	17730	298	+	17731	CACAATG	23	ACAAAAACC	17769	39	
IGH1V4-01p	24566	24837	272	+	24838	CGCAGTG	22	CCACAAACC	24875	38	Nonsense mutation
IGH1V1-04	37305	37596	292	+	37597	CACAGTG	22	ACAAAAACC	37634	38	
IGH1V2-03	48845	49139	295	+	49140	CACAGTG	23	TCAAAAACT	49178	39	
IGH1V1-05	49909	50197	289	+	50198	CACAGTG	22	ACAAAAACC	50235	38	
IGH1V5-01	51710	51998	289	+	51999	CACAGTG	22	ACAAAAACT	52036	38	
IGH1V2-04	56322	56616	295	+	56617	CACAGTG	23	ACAAAAACC	56655	39	
IGH1V6-01	57465	57762	298	+	57763	CACAGTG	21	ACTAAATCT	57799	37	
IGH1V1-06	59678	59966	289	+	59967	CACAGTG	22	ACAAAAACC	60004	38	
IGH1V4-02p	68017	68288	272	+	68289	TGCAGTG	22	TCACAAACC	68326	38	Nonsense mutation
IGH1V2-05	69787	70084	298	+	70085	CACAGTG	23	ACAAAAACC	70123	39	
IGH1V1-07	155485	155763	279	+	155764	CACAGTG	22	TCAAAACCC	155801	38	
IGH2V2-02	282620	282914	295	-	282915	CACAGTG	23	ACAAAAACC	282953	39	
IGH2V4-01p	284404	284675	272	-	284676	TGCAGTG	22	TCACAAACC	284713	38	Nonsense mutation
IGH2V5-01	288808	289096	289	-	289097	CACAGTG	22	ACAGAAACT	289134	38	
IGH2V1-03	289977	290271	295	-	290272	CACAGTG	22	ACAAAAACC	290309	38	
IGH2V1-02	293835	294126	292	-	294127	CACAGTG	22	ACAAAAACC	294164	38	
IGH2V2-01	303780	304074	295	-	304075	CAGGGCC	24	AGCACAAAG	304114	40	
IGH2V1-01	304926	305204	279	-	305205	CACAGTG	22	TCAAAACCC	305242	38	

Table 3.6: Co-ordinate table of VH segments in the *Nothobranchius furzeri IGH* locus.

Table 3.7: Co-ordinate table of DH segments in the *Nothobranchius furzeri IGH* locus.

Name	Start	NT Sequence	End	Length	Strand
IGH1D01	25782	ATACGTACTTTCGTGGTATATAGAGA	25807	26	+
IGH1D02	76700	GATATCTGGGTGGGGG	76715	16	+
IGH1D03	77027	TGAAATGATTAC	77038	12	+
IGH1D04	77476	TCGCGTAGCGGC	77487	12	+
IGH1D05	78717	GAAACCACGGCAGC	78730	14	+
IGH1D06	79049	TTTATAGCGGCTAC	79062	14	+
IGH1D07	80417	CAGACTGGAGA	80427	11	+
IGH1D08	81362	TTCATGGCAGCCAC	81375	14	+
IGH1D09	82067	CAGACTGGAGC	82077	11	+
IGH1D10	84282	TGGGGTGGCAGC	84293	12	+
IGH2D04	263497	CAGACTGGAGA	263507	11	-
IGH2D03	270243	TTTATAGCGGCTAC	270256	14	-
IGH2D02	270878	GAAACCACGGCAGC	270891	14	-
IGH2D01	271749	GACTTTTACTAC	271760	12	-

Table 3.8: Co-ordinate table of DH 5'-RSSs in the *Nothobranchius furzeri IGH* locus.

Name	5'-RSS Start	Nonamer	Spacer Length	Heptamer	5'-RSS End	Length
IGH1D01	25754	GGTTGTTGT	12	CACTGTG	25781	28
IGH1D02	76672	AGTTTTTGA	12	CACAGTG	76699	28
IGH1D03	76999	TGTTGTTGT	12	CACAGTG	77026	28
IGH1D04	77448	AGTTTTTGT	12	CACGGTG	77475	28
IGH1D05	78688	GATGTTTTT	13	CACAGTG	78716	29
IGH1D06	79021	TGTTTTTGT	12	CGCTGTG	79048	28
IGH1D07	80389	AGTTTTGGT	12	CACAGTG	80416	28
IGH1D08	81334	TGTTTTTGT	12	CGCTGTG	81361	28
IGH1D09	82039	AGTTTTGGT	12	CACAGTG	82066	28
IGH1D10	84254	TCATTCATT	12	CACTGTG	84281	28
IGH2D04	263469	AGTTTTGGT	12	CACAGTG	263496	28
IGH2D03	270215	TGTTTTTGT	12	CGCTGTG	270242	28
IGH2D02	270850	TGTTTTTGT	12	CACAGTG	270877	28
IGH2D01	271721	AGTTTTTAT	12	CATGGTG	271748	28

Table 3.9: Co-ordinate table of DH 3'-RSSs in the *Nothobranchius furzeri IGH* locus.

Name	3'-RSS Start	Heptamer	Spacer Length	Nonamer	3'-RSS End	Length
IGH1D01	25808	CACAGTG	12	ACAAAAACC	25835	28
IGH1D02	76716	CACAGTG	12	ACAAAAACC	76743	28
IGH1D03	77039	CACTGTG	11	AATATAACC	77065	27
IGH1D04	77488	CACAGCG	12	ACATAAAAC	77515	28
IGH1D05	78731	CACAGCG	12	ACAAAAGCC	78758	28
IGH1D06	79063	CACTGTG	12	ACAAGATCC	79090	28
IGH1D07	80428	CACAACG	12	ACAAAAACC	80455	28
IGH1D08	81376	CACTGTG	12	ACAAAATCC	81403	28
IGH1D09	82078	CACAATG	12	ACAAAAACC	82105	28
IGH1D10	84294	CACAGTG	12	ACAAAAACC	84321	28
IGH2D04	263508	CACAACG	12	ACAAAAACC	263535	28
IGH2D03	270257	CACTGTG	12	ACAAGATCC	270284	28
IGH2D02	270892	CACAGCG	12	ACAAAAGCC	270919	28
IGH2D01	271761	CACAATG	12	ACAAAAACC	271788	28

Name	Start	NT Sequence	AA Sequence	End	Length	Strand
IGH1J01	26187	GTGCTTTAGACAACTGGGGAAAAGGAACGGAGGTTACTGTTCAACCTG	ALDNWGKGTEVTVQP	26234	48	+
IGH1J02	128176	ATGACTACTTTGACTACTGGGGAAAAGGAACAATGGTGACGGTCACATCAG	DYFDYWGKGTMVTVTS	128226	51	+
IGH1J03	128354	ACCGTGGGGTAAAGGGACAACAGTCACGGTCAAAACAG	PWGKGTTVTVKT	128391	38	+
IGH1J04	128533	ACGGTGCTCTTGACTACTGGGGTAAAGGGACCGCAGTCACTGTAACATCAG	GALDYWGKGTAVTVTS	128583	51	+
IGH1J05	128887	ACAACGCTTTTGACTACTGGGGAAAAGGAACAACGGTCACCGTCACTTCAG	NAFDYWGKGTTVTVTS	128937	51	+
IGH1J06	129346	CTACGATGCTTTTGACTACTGGGGGAAAAGGACGATGGTCACGTCACTTCAG	YDAFDYWGKRTMVTSLQ	129397	52	+
IGH1J07	129635	TTAACTGGGCTTTCGACTACTGGGGAAAAGGGACGATGGTAACGGTGACTTCAG	NWAFDYWGKGTMVTVTS	129688	54	+
IGH1J08	129965	TTACCACGCAGCTTTGGACTACTGGGGAAAAGGGACGACGGTCACCGTCACCTCAG	YHXALDYWGKGTTVTVTS	130020	56	+
IGH1J09	130612	TCTACGCTGCTTTTGACTACTGGGGTAAAGGTACAACGGTAACCGTTTCATCAG	YAAFDYWGKGTTVTVSS	130665	54	+
IGH2J08	204031	TCTACGCTGCTTTTGACTACTGGGGTAAAGGTACAACGGTAACCGTTTCATCAG	YAAFDYWGKGTTVTVSS	204084	54	-
IGH2J07	204673	TTACCACGCAGCTTTGGACTACTGGGGAAAAGGGACGACGGTCACCGTCACCTCAG	YHXALDYWGKGTTVTVTS	204728	56	-
IGH2J06	205005	ATAACTGGGCTTTCGACTACTGGGGAAAAGGGACGATGGTAACGGTGACTTCAG	NWAFDYWGKGTMVTVTS	205058	54	-
IGH2J05	205296	CTACGATGCTTTTGACTACTGGGGGAAAAGGACGATGGTCACGTCACTTCAG	YDAFDYWGKRTMVTSLQ	205347	52	-
IGH2J04	205756	ACAACGCTTTTGACTACTGGGGAAAAGGAACAACGGTCACCGTCACTTCAG	NAFDYWGKGTTVTVTS	205806	51	-
IGH2J03	206111	ATGGTGCTTTTGACTACTGGGGTAAAGGGACCGCAGTCACTGTAACATCAG	GAFDYWGKGTAVTVTS	206161	51	-
IGH2J02	206303	ACCGTGGGGTAAAGGGACAACAGTCACGGTCAAAACAG	PWGKGTTVTVKT	206340	38	-
IGH2J01	206466	ATGACTACTTTGACTACTGGGGAAAAGGAACAATGGTGACGGTCACATCAG	DYFDYWGKGTMVTVTS	206516	51	-

Table 3.10: Co-ordinate table of JH segments in the *Nothobranchius furzeri IGH* locus.

Name	RSS Start	Nonamer	Spacer Length	Heptamer	RSS End	RSS Length
IGH1J01	26196	TGTTTTTGT	23	CACTGTG	26186	39
IGH1J02	128188	AGTGTTTGT	23	CACTGTG	128175	39
IGH1J03	128353	TGTTTATTT	23	CACTGTG	128353	39
IGH1J04	128545	GGTTTTTGT	23	CACTGTG	128532	39
IGH1J05	128899	GGTTTTAGT	23	TACTGTG	128886	39
IGH1J06	129360	TCTTCTTGT	22	TACTTTG	129345	38
IGH1J07	129650	AGTTTTTGT	23	TACTGTG	129634	39
IGH1J08	129983	AGTTTTAGT	22	TACTGTG	129964	38
IGH1J09	130628	CGTTTTTAT	22	CACTGTG	130611	38
IGH2J08	204047	CGTTTTTAT	22	CACTGTG	204030	38
IGH2J07	204691	AGTTTTAGT	22	TACTGTG	204672	38
IGH2J06	205020	AGTTTTTGT	23	TACTGTG	205004	39
IGH2J05	205310	TCTTCTTGT	22	TACTTTG	205295	38
IGH2J04	205768	GGTTTTAGT	23	TACTGTG	205755	39
IGH2J03	206123	GGTTTTTGT	23	CACTGTG	206110	39
IGH2J02	206302	TGTTTATTT	23	CACTGTG	206302	39
IGH2J01	206478	AGTGTTTGT	23	CACTGTG	206465	39

Table 3.11: Co-ordinate table of JH RSSs in the *Nothobranchius furzeri IGH* locus.

3.3 The *IGH* locus in *Xiphophorus maculatus*

The turquoise killifish *IGH* locus shares many features with other characterised teleost loci, including a modified tandem-translocon configuration with intact VH, DH, JH and constant regions (Figure 3.3B), a four-exon secreted configuration of *IGHM* (Figure 3.6A), an expanded *IGHD* constant region with tandem C_δ-exon block repeats (Figures 3.3B and 3.6B,), a conserved RSS structure (Figure 3.10), and a chimeric $C_\mu 1$ in *IGHD* (Figure 3.6B). However, it also exhibits many ideosyncratic features that differ from those observed in most characterised teleost loci, including an unusually small number of VH segments (Figure 3.8 and Table 3.5), a four-exon $C_\mu 1$-$C_\mu 2$-TM1-TM2 configuration of transmembrane *IGHM* (Figure 3.6A), an inverted sublocus present in antisense (Figure 3.3B), and a complete absence of *IGHZ*.

Many of these peculiarities, including the unusual *IGHM-TM* splicing pattern, inverted sublocus, and lack of *IGHZ*, are shared with the *IGH* locus of medaka (*Oryzias latipes*), which is the closest relative of *Nothobranchius furzeri* to have its immunoglobulin heavy chain locus characterised prior to this study [8]. Given the close relationship between the two species, the shared unusual features of their *IGH* loci suggested a common origin of these traits in the common ancestor of both species. If this hypothesis were correct, one would expect *IGHZ* to also be absent in any other descendents of this common ancestor, including other cyprinodontiform species.

To investigate this hypothesis further, I performed a complete characterisation of the *IGH* locus in the platyfish *Xiphophorus maculatus*, another cyprinodontiform species that has seen widespread use as a model organism [17]. Surprisingly, the *X. maculatus* locus shared none of the unusual features shared between the turquoise-killifish and medaka loci, strongly suggesting independent loss of *IGHZ* in both groups and implying a high level of volatility in *IGH* locus structure within the Atherinomorpha.

3.3.1 Overall structure

As was the case with the *N. furzeri IGH* locus, candidate genome scaffolds from the most recent *Xiphophorus maculatus* genome assembly (Genbank accession GCA_002775205.2) were identified by alignment to *IGH* gene segments from zebrafish, stickleback and medaka, supplemented in this case with segments from the newly-characterised *N. furzeri* locus itself. In contrast to the more fragmented results in *N. furzeri*, this process identified a single sequence region on one chromosome of the *X. maculatus* locus, which was extracted and characterised as described for the assembled *N. furzeri* locus (Section 3.2) without the need for further sequencing or assembly.

The *X. maculatus IGH* locus so identified occupies roughly 293 kb on chromosome 16 (scaffold NC_036458.1; Figure 3.13A). Unlike in turquoise killifish and medaka, all identified gene segments share a common orientation; no evidence of a second sublocus in antisense could be identified. In

Table 3.12: Sequence similarity between *IGHZ* constant-regions in *X. maculatus*: Percentage sequence identities of pairwise Needleman-Wunsch global alignments between nucleotide (NT) or amino-acid (AA) sequences of corresponding C_ζ exons from the two *IGHZ* constant regions of *X. maculatus IGH*.

Isotype	Exon	NT	AA
Z	1	59.14	44.57
Z	2	63.93	53.41
Z	3	66.19	43.48
Z	4	65.15	50.49

stark contrast with both killifish and medaka, the single “sublocus" comprising *X. maculatus IGH* contains not one but two *IGHZ* constant regions, along with a hugely extended V-region extending over almost 250 kb and containing more than 120 VH-segments. This enormous VH-diversity greatly exceeds that of any characterised teleost *IGH* locus except perhaps that of rainbow trout [22], while the presence of multiple *IGHZ* constant regions without intervening *IGHM* or *IGHD* is also highly unusual [2].

Even cursory examination of the *X. maculatus IGH* locus is therefore sufficient to reveal a unique and highly interesting structure with many unexpected differences from both turquoise killifish and medaka (Figure 3.14, columns 1-5). In particular, since *X. maculatus* is more closely related to *N. furzeri* than either is to medaka, the presence of *IGHZ* in the former strongly suggests at least two independent loss events in the Atherinomorpha, indicating an unexpected level of volatility in the evolution of this important isotype.

3.3.2 Constant regions

As discussed briefly in Section 3.3.1, the *X. maculatus IGH* locus contains two distinct *IGHZ* constant regions: one in the usual position immediately preceding the *IGHM*-associated D- and J-regions, the other, unexpectedly, at the far 5’-extremity of the locus (Figure 3.13B). Both *IGHZ* constant regions occupy the expected configuration, with four C_ζ exons, two transmembrane exons, and a secretory tail (Figure 3.13C, Table 3.13). However, in contrast to the duplicate constant regions in *N. furzeri*, the two *IGHZ* constant regions in *X. maculatus* are quite distinct from each other in sequence, with an average of only 64 % nucleotide and 48 % amino-acid sequence identity between corresponding C_ζ exons (Figure 3.15, Table 3.12). This unexpectedly high level of sequence divergence suggests a relatively ancient duplication event, and raises the possibility that the lineage giving rise to *N. furzeri* may have lost not one, but two distinct *IGHZ* constant regions.

While the state of *IGHZ* constant regions differs markedly between *X. maculatus* and *N. furzeri*, the configurations of the *IGHM* and *IGHD* constant regions of the two species are quite similar, with a C_μ1-C_μ2-C_μ3-C_μ4-TM1-TM2 configuration for *IGHM* and a C_δ1-(C_δ2-C_δ3-C_δ4)$_2$-C_δ5-C_δ6-C_δ7-TM1-TM2 configuration for *IGHD* (Figures 3.13B and 3.13C, Table 3.13). In the *X. maculatus* locus,

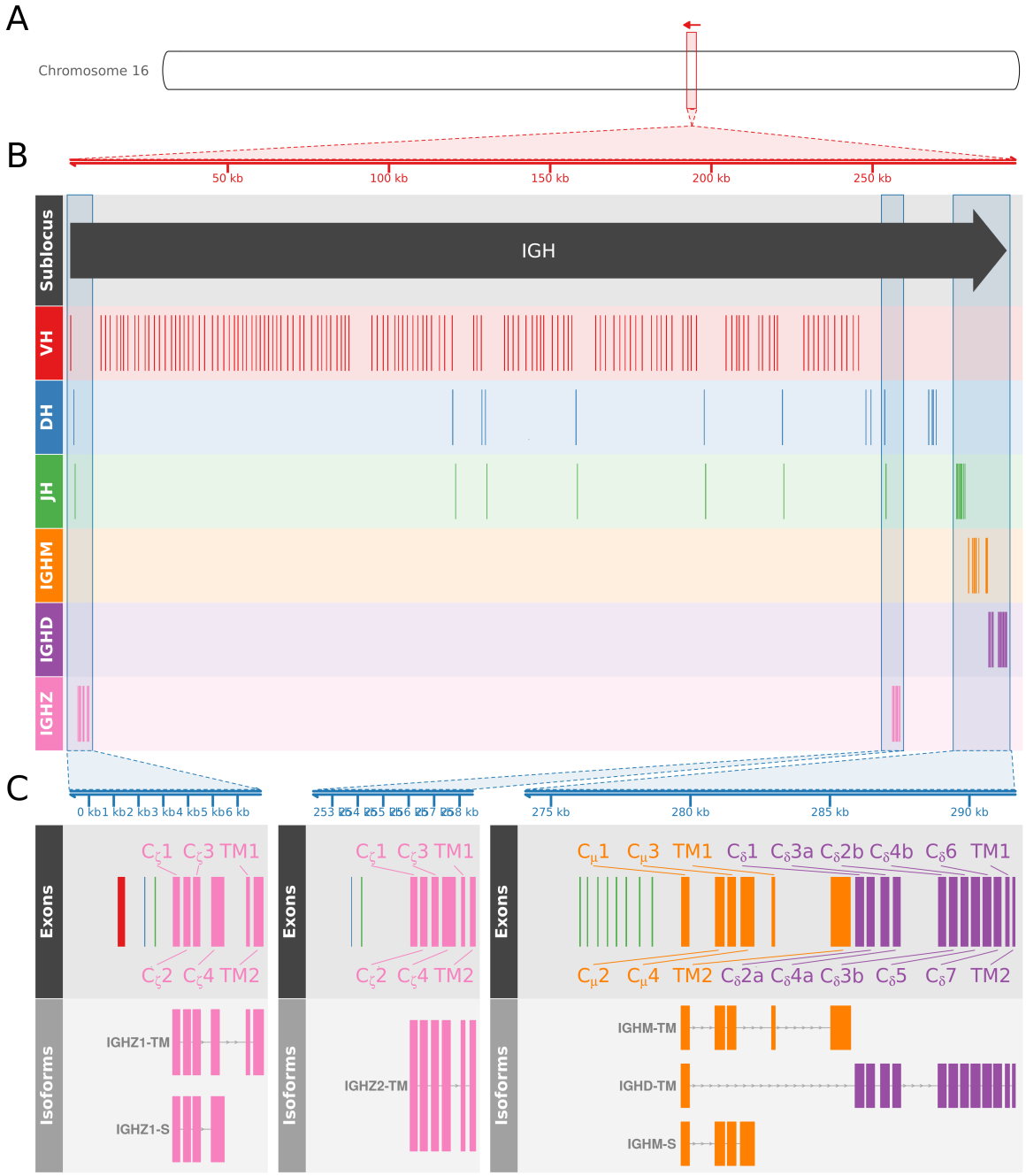
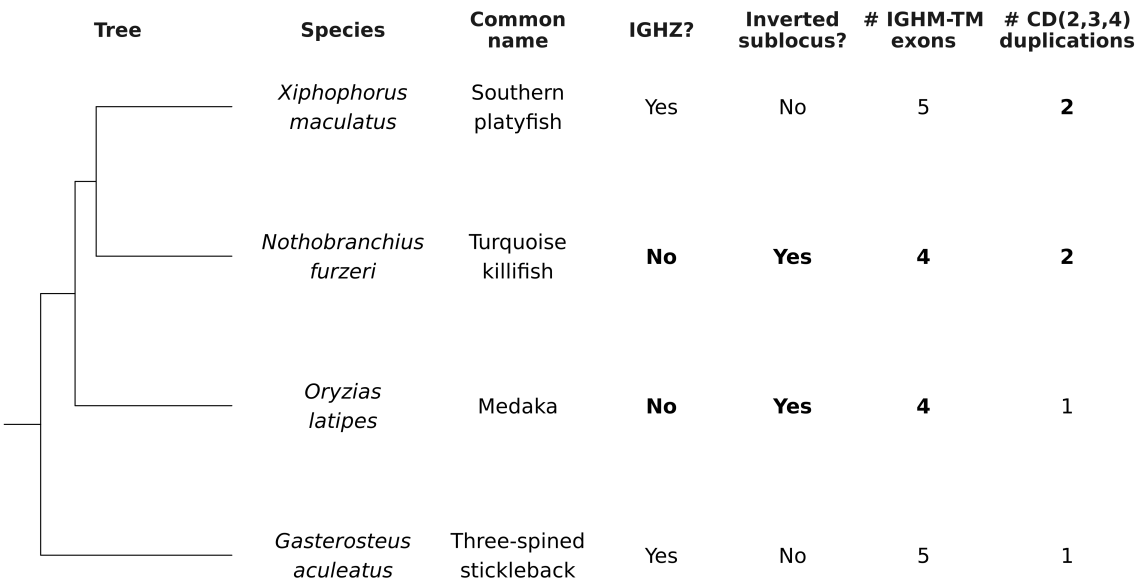


Figure 3.12: The immunoglobulin heavy chain (*IGH*) locus in *Xiphophorus maculatus*: (A) Position of the *IGH* locus on chromosome (group) 16 of the *X. maculatus* genome. (B) Arrangement of VH, DH, JH and constant-region gene segments on the *X. maculatus IGH* locus. (C) Detailed map of the *IGHZ1*, *IGHZ2* and *IGHM/D* constant regions, indicating the position and identity of the constant-region exons and the exon composition of expressed *IGH* isoforms in *X. maculatus*. Note change of orientation between subfigures (A) and (B-C).



Tree	Species	Common name	IGHZ?	Inverted sublocus?	# IGHM-TM exons	# CD(2,3,4) duplications
	Xiphophorus maculatus	Southern platyfish	Yes	No	5	**2**
	Nothobranchius furzeri	Turquoise killifish	**No**	**Yes**	**4**	**2**
	Oryzias latipes	Medaka	**No**	**Yes**	**4**	1
	Gasterosteus aculeatus	Three-spined stickleback	Yes	No	5	1

Figure 3.14: Summary of important *IGH* phenotypes in killifish, platyfish, and medaka: Cladogram of the evolutionary relationship between southern platyfish (*Xiphophorus maculatus*), turquoise killifish (*Nothobranchius furzeri*) and medaka (*Oryzias latipes*), with three-spined stickleback (*Gasterosteus aculeatus*) as an outgroup. The state of various *IGH* phenotypes of interest are annotated to the right of the tree; states deviating from the expected teleost configuration are in bold.

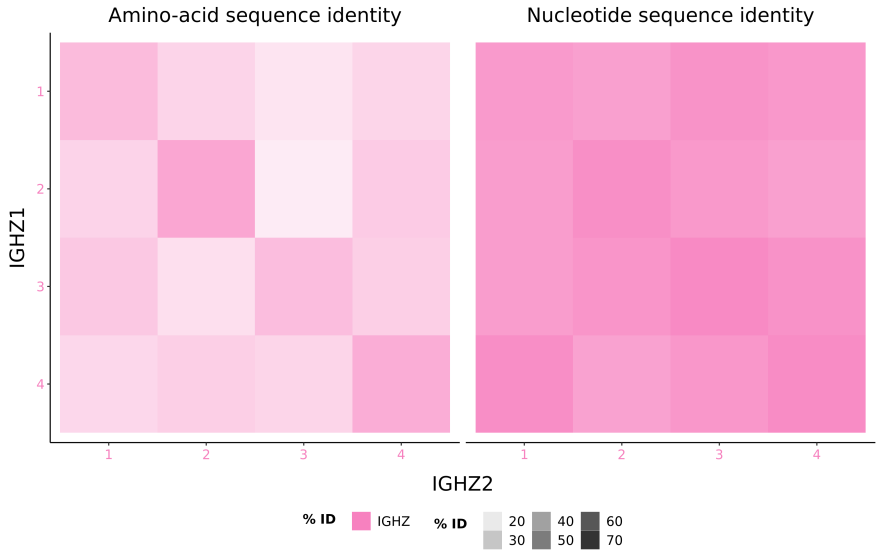


Figure 3.15: Sequence similarity between *IGHZ* constant-regions in *X. maculatus*: Heatmap of percentage sequence identity between amino-acid (right) and nucleotide (left) sequences of C_ζ exons from the two *X. maculatus IGHZ* constant regions, calculated using pairwise Needleman-Wunsch global alignments.

Table 3.13: Co-ordinate table of constant-region exons in the *Xiphophorus maculatus IGH* locus.

Name	Isotype	Start	End	Length	Strand
IGHZ1-1	Z	3380	3667	288	+
IGHZ1-2	Z	3814	4098	285	+
IGHZ1-3	Z	4195	4497	303	+
IGHZ1-4	Z	4934	5263	330	+
IGHZ1-S	Z	5264	5459	196	+
IGHZ1-TM1	Z	6345	6490	146	+
IGHZ1-TM2	Z	6645	7043	399	+
IGHZ2-1	Z	256059	256337	279	+
IGHZ2-2	Z	256453	256734	282	+
IGHZ2-3	Z	256893	257171	279	+
IGHZ2-4	Z	257319	257636	318	+
IGHZ2-S	Z	257637	257850	214	+
IGHZ2-TM1	Z	258059	258213	155	+
IGHZ2-TM2	Z	258410	258629	220	+
IGHM-1	M	279664	279960	297	+
IGHM-2	M	280880	281224	345	+
IGHM-3	M	281321	281629	309	+
IGHM-4	M	281789	282291	503	+
IGHM-TM1	M	282910	283034	125	+
IGHM-TM2	M	285028	285740	713	+
IGHD-1	D	285902	286219	318	+
IGHD-2A	D	286310	286597	288	+
IGHD-3A	D	286814	287128	315	+
IGHD-4A	D	287250	287534	285	+
IGHD-2B	D	288876	289166	291	+
IGHD-3B	D	289262	289576	315	+
IGHD-4B	D	289680	289964	285	+
IGHD-5	D	290052	290381	330	+
IGHD-6	D	290472	290789	318	+
IGHD-7	D	290865	291152	288	+
IGHD-TM1	D	291286	291434	149	+
IGHD-TM2	D	291541	291642	102	+

these constant regions and *IGHZ1* adopt the standard configuration seen in comparatively simple teleost *IGH* loci like those of zebrafish and fugu, with a VH-DH-JH-**CZ**-DH-JH-CM-**CD** arrangement that allows the choice between *IGHZ* and *IGHM/D* usage to be made via the choice of DH segment during VDJ-recombination. However, whether such a mechanism is also responsible for the choice between these constant regions and *IGHZ1*, which lies more than 200 kb away and upstream of the great majority of VH segments in the locus (Section 3.3.3) is questionable.

In order to investigate the expressed isoforms present in *X. maculatus*, published RNA-sequencing reads from various platyfish tissues (BioProject accession PRJNA420092, all libraries). were aligned together to the *IGHZ* and *IGHM/D* constant regions with `STAR`. The results indicate the expected six-exon transmembrane configuration in both *IGHZ1* and *IGHZ2*, as well as a secretory form of *IGHZ1* comprising $C_\zeta 1$ to $C_\zeta 4$ plus a 23 bp secretory tail formed by a transcriptional run-on event from $C_\zeta 4$ (Figures 3.16A and 3.16B). However, while an in-frame secretory tail of similar length (20 bp) can be found in *IGHZ2*, it does not appear to be expressed in the read sets analysed here, indicating that *IGHZ2* may only be expressed in transmembrane form in the individuals sampled (Figure 3.16B).

Meanwhile, the results for *IGHD* (Figure 3.16D) indicate a similar configuration to that observed in turquoise killifish, with a chimeric $C_\mu 1$ followed by 10 C_δ exons and two transmembrane exons; as in *N. furzeri*, neither a dedicated *IGHD* secretory exon nor a post-$C_\delta 7$ secretory tail was identified, suggesting that *IGHD* may be produced solely in transmembrane form in this species. Secretory *IGHM* (*IGHM-S*) was also found to occupy the same four-exon configuration seen in turquoise killifish and elsewhere. However, the configuration observed for transmembrane *IGHM* (*IGHM-TM*) did not correspond to the four-exon structure shared between turquoise killifish and medaka (Figure 3.6A); rather, *IGHM-TM* in *X. maculatus* occupies the five-exon configuration seen in most characterised teleosts (Figure 3.5D). This surprising difference indicates that two different splice configurations of *IGHM-TM* persist in the cyprinodontiform lineage, and raises the question of what, if any, functional difference arises from the presence or absence of $C_\mu 3$ in transmembrane *IGHM* in different species. However, it remains unclear whether this pattern of exon usage (Figure 3.14) is the result of independent changes in medaka and turquoise killifish or of a reversion in *X. maculatus* to the primitive teleost configuration.

3.3.3 Variable regions

In total, 125 VH segments, 14 DH segments and 15 JH segments were identified in the *X. maculatus* IGH locus (Figure 3.13B). Of these, exactly one VH (*IGHV01-01*), DH (*IGHDZ01*) and JH (*IGHJZ01*) lie upstream of the *IGHZ1* constant region, indicating that the variable-region sequence diversity available to this isotype is limited to a single VDJ combination. In contrast, the variable region between the end of *IGHZ1* and the start of *IGHZ2* is highly expanded, with 124 tightly-clustered VH segments – more than five times the total number seen in *N. furzeri*, and more than seven times the number in the largest *N. furzeri* sublocus. Of these 124 VH segments, 106 (86 %) are apparently functional, with the remainder pseudogenised by a variety of frameshift mutations, nonsense mutations, or truncation events (Tables 3.14 to 3.18); it remains to be seen whether *IGHV01-01* is also capable of recombining with DH segments downstream of the *IGHZ1* constant region, and so constitutes part of the range of VDJ combinations available to the other constant regions. The VH sequences in the *X. maculatus*

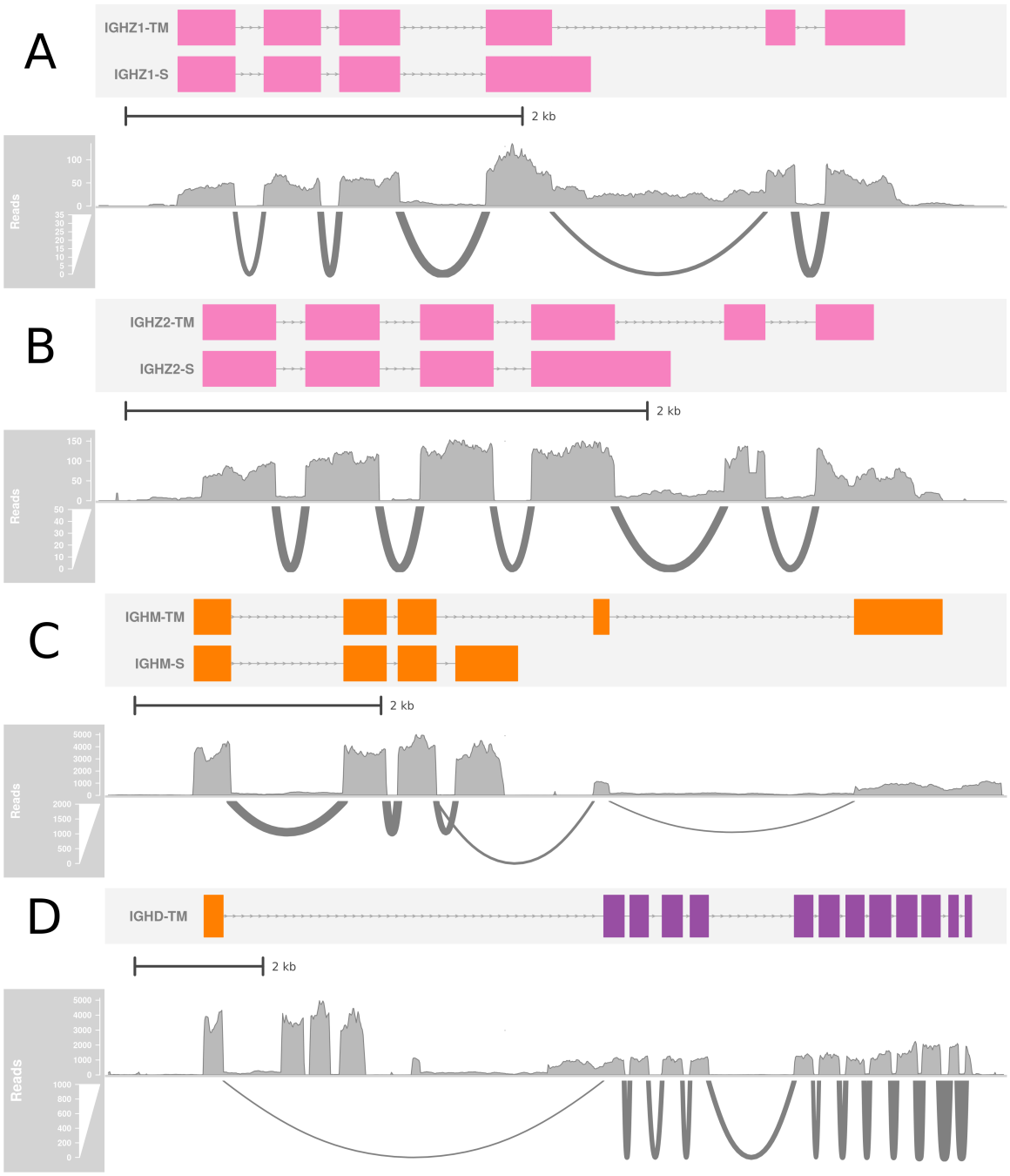


Figure 3.16: Constant-region isoforms in *X. maculatus*: Coverage and sashimi plots of STAR-aligned RNA-seq reads from *X. maculatus* samples, demonstrating the splicing behaviour of *IGH* constant-region isoforms. (A) *IGHZ1* exon splicing, showing alternative use of the $C_{\zeta}4$/TM1 splice junction and the post-$C_{\zeta}4$ secretory tail; (B) *IGHZ2* exon splicing; (C) IGHM exon splicing, showing alternative splicing patterns of IGHM-TM and IGHM-S; (D) IGHD exon splicing, including splicing of $C_{\mu}1$ to $C_{\delta}1$.

locus are much more tightly packed than in the *N. furzeri* locus, consistent with a lower overall prevalence of repetitive regions (21%) in the *X. maculatus* genome [21].

In total, the VH regions in *X. maculatus IGH* fall into 23 families, of which eight contain multiple segments (Figures 3.17 and 3.18); strikingly, the single VH segment serving IGHZ (*IGHV01-01*) represents a separate family which is distinct from any other segment in the locus. To further investigate the evolutionary history of these families, the VH segments from both the *X. maculatus* and *N. furzeri IGH* loci were aligned together with `PRANK`, and the resulting alignment was used to construct a phylogenetic tree with `RAxML` [35, 36, 37]; the resulting tree (Figure 3.19) revealed a clear interrelationship between the largest families in both loci (*X. maculatus* V02 and *N. furzeri* V1), with a similar relationship observed for the second-largest families (*X. maculatus* V03 and *N. furzeri* V2). In accordance with the close sequence relationship noted in Section 3.2.4, *N. furzeri* V4 falls comfortably within the V03/V2 subtree, supporting its status as a pseudogenised subfamily of *N. furzeri* V2.

In addition to its highly expanded VH region, the variable region of the *X. maculatus* locus is unusual in the arrangement of its DH and JH segments (Tables 3.20 and 3.22): in addition to the relatively densely-packed blocks of four DH and eight JH regions between *IGHZ1* and *IGHM*, and the smaller groups of three DH segments and one JH segment between the last VH segment and *IGHZ1*, small numbers of DH and JH segments are interspersed between blocks of VH segments in the extended V-region between *IGHZ1* and *IGHZ2* (Figure 3.13B). Many of these segments are arranged such that groups of one or two DH segments are closely associated with a single JH segment, raising the possibility of a more cluster-like behaviour in which each VDJ group acts as a distinct recombination unit. However, the presence of larger D- and J-regions more closely upstream of the constant regions suggests a more conventional translocon behaviour; it remains to be seen which of these traditional models of antigen-receptor structure more closely matches the *in vivo* recombination behaviour of this locus.

Finally, as is the case with *N. furzeri*, the recombination signal sequences (RSSs) in *X. maculatus* IGH correspond closely to the standard expectations across the vertebrates, with the expected heptamer and nonamer consensus sequences and spacer length distributions (Figures 3.20 and 3.21, 97.6% of RSS spacers within 1 bp of the expected conserved length).

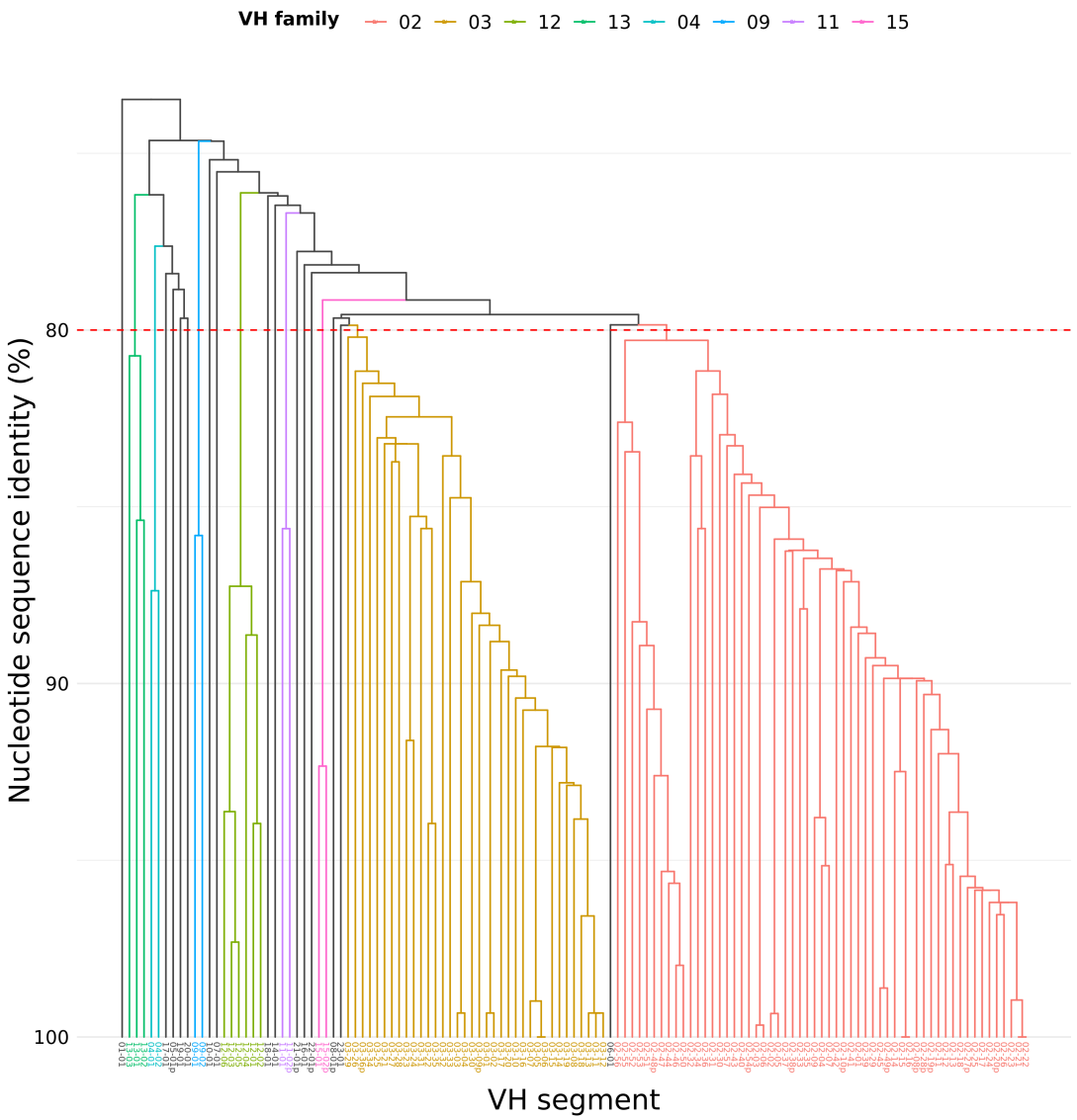


Figure 3.17: Dendrogram of VH families in the in *X. maculatus* (*IgH*) locus: Dendrogram of sequence similarity of VH segments in the *X. maculatus* locus, arranged by single-linkage clustering on nucleotide sequence identity. The red line indicates the 80% cutoff point for family assignment, while branch colour indicates family membership: VH families containing multiple segments are uniquely coloured, while single-segment families are in grey.

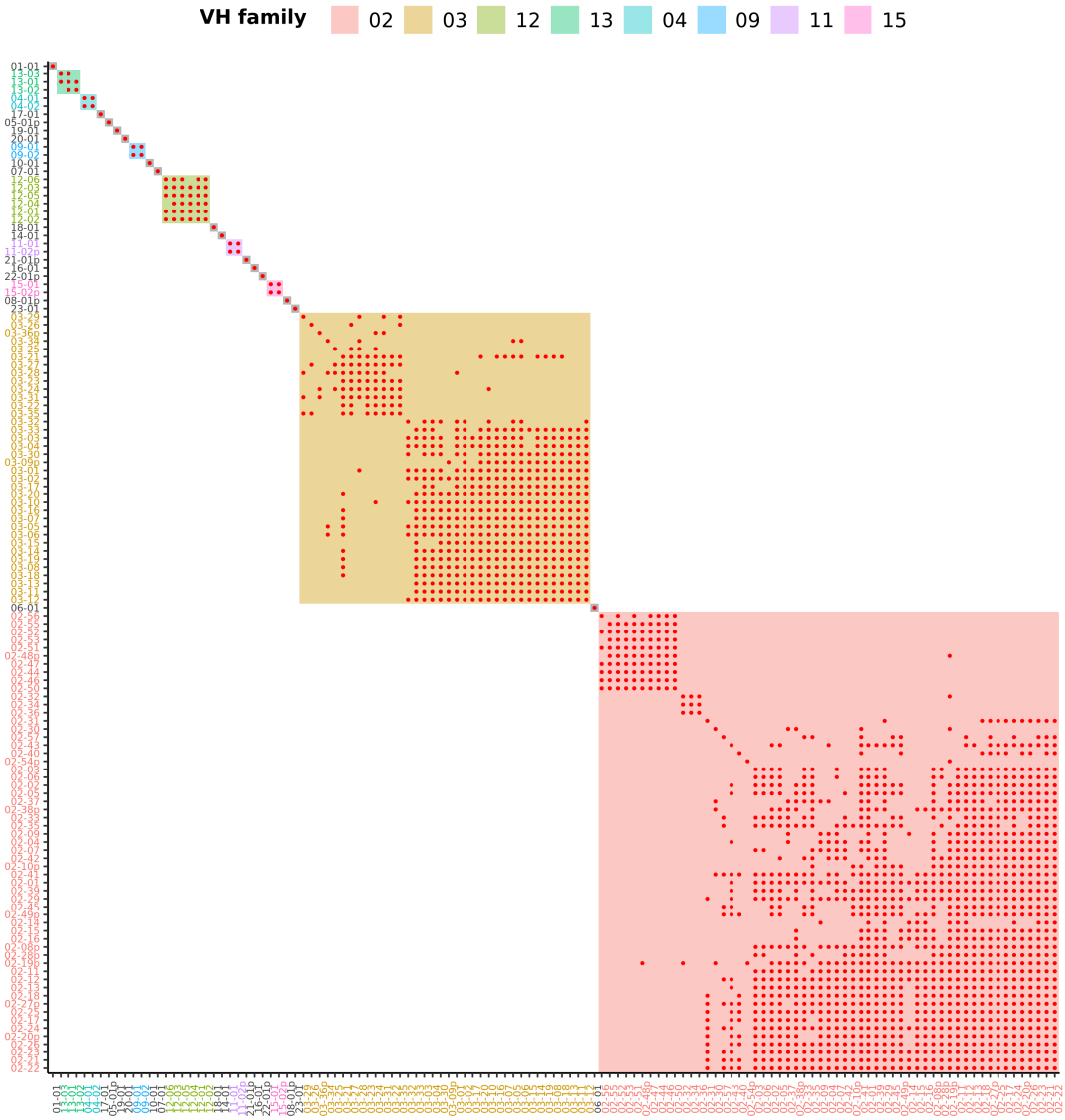


Figure 3.18: Heatmap of VH families in the in *X. maculatus* (*IgH*) locus: Heatmap of family relationships among *X. maculatus* VH segments, with coloured shading indicating families and red dots indicating pairwise nucleotide sequence identity of at least 80%. VH families containing multiple segments are uniquely coloured, while single-segment families are in grey.

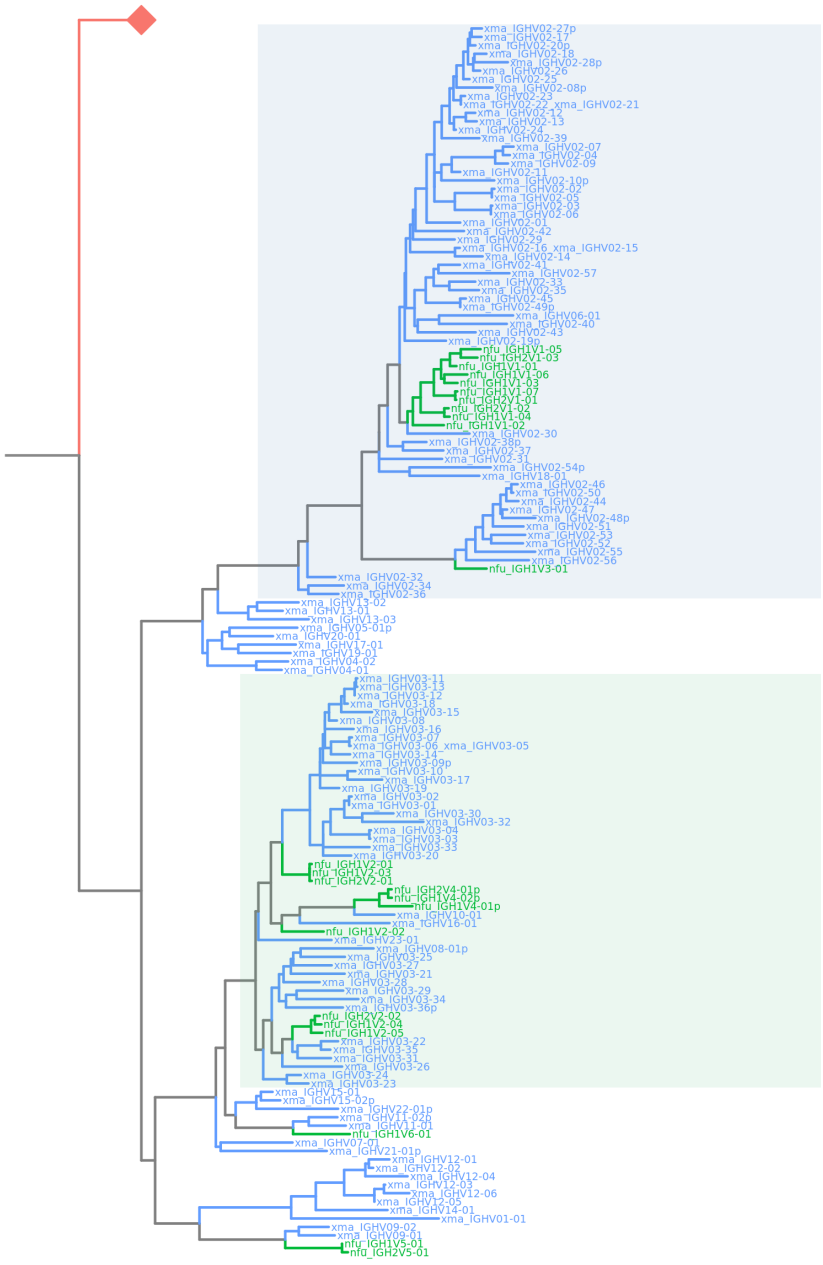


Figure 3.19: Evolutionary relationships between VH families in *X. maculatus* and *N. furzeri*: Phylogenetic tree of evolutionary relationships between *IGH* VH segments in *N. furzeri* and *X. maculatus*, as inferred from the nucleotide sequences of VH segments from both loci. Note the close interrelationship between the largest (blue zone) and second-largest (green zone) families in each species. The red diamond indicates the location of the outgroup, which is composed of zebrafish *TRB* V-segments.

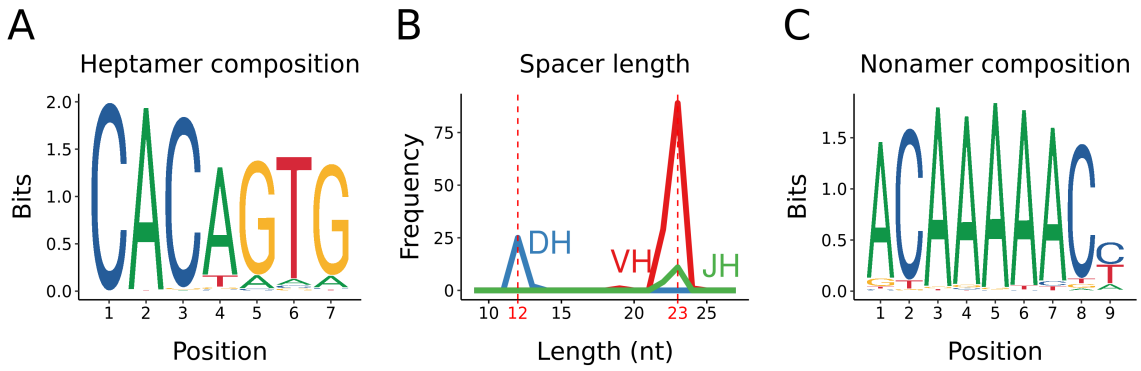


Figure 3.20: Recombination signal sequences in the *X. maculatus IGH* locus: (A) Sequence composition of conserved heptamer sequences across all *X. maculatus* heavy-chain RSSs; (B) length distribution of unconserved spacer sequences in *X. maculatus* heavy-chain RSSs; (C) sequence composition of conserved heptamer sequences across all *X. maculatus* heavy-chain RSSs.

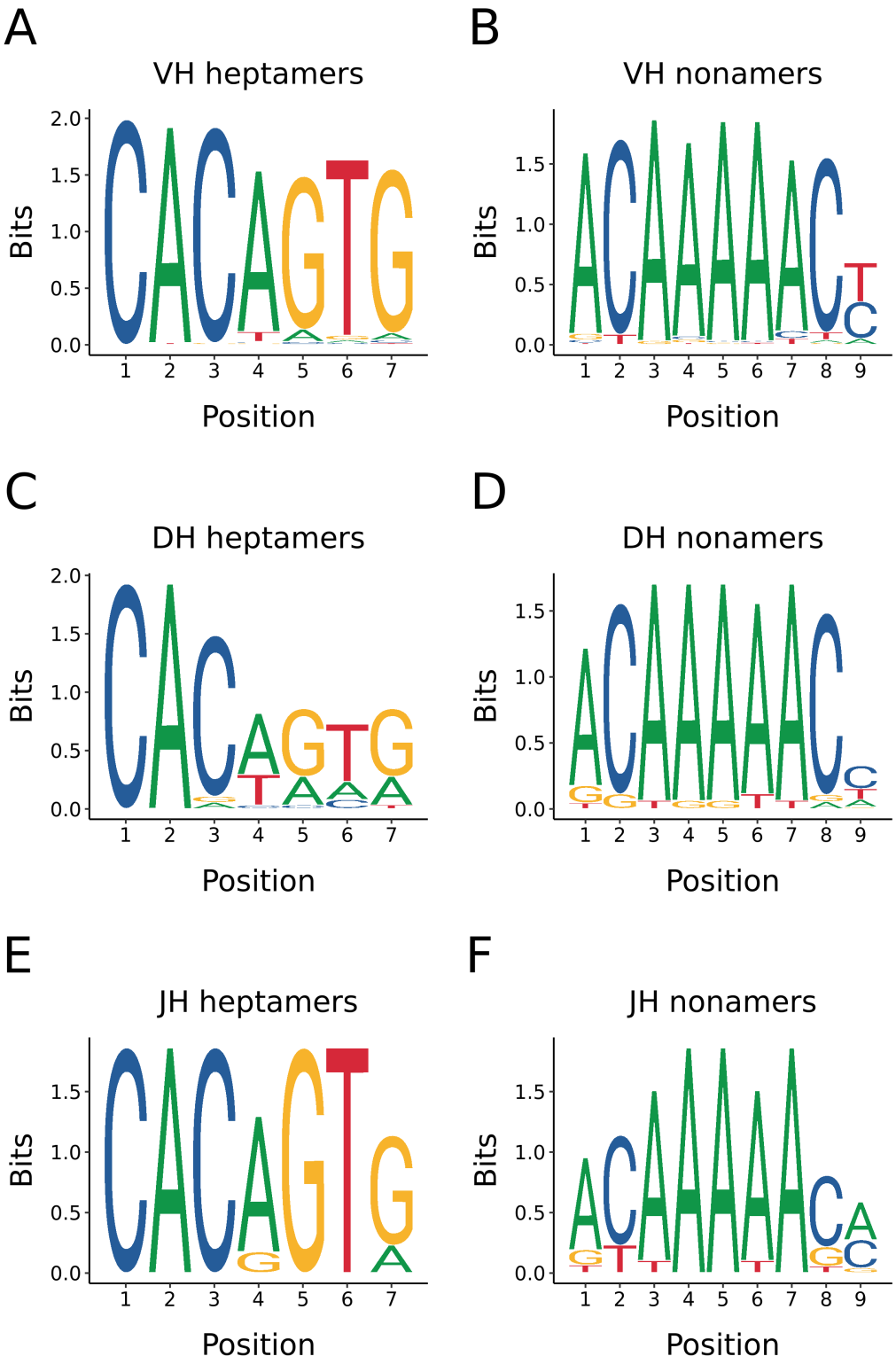


Figure 3.21: *X. maculatus* recombination signal sequences by segment type: Sequence composition of conserved heptamer (A,C,E) and nonamer (B,D,F) sequences from *X. maculatus* heavy-chain RSSs associated with VH (A,B), DH (C,D) or JH (E,F) gene segments.

Name	Start	End	Length	Strand	RSS Start	Heptamer	Spacer Length	Nonamer	RSS End	RSS Length	Comment
IGHV01-01	1159	1450	292	+	1451	CACAGTG	23	GTAAAAACC	1489	39	
IGHV02-01	10534	10825	292	+	10826	CACAGTG	23	ACAAAACCC	10864	39	
IGHV02-02	11961	12261	301	+	12262	CACTGTG	23	ACAAAAACT	12300	39	
IGHV02-03	13319	13616	298	+	13617	CACAGTG	23	ACACAAACT	13655	39	
IGHV03-01	15440	15734	295	+	15735	CACAGTG	22	ACAAAAACT	15772	38	
IGHV02-04	16618	16908	291	+	16909	CACAGTG	23	ACAAAAACC	16947	39	
IGHV02-05	17522	17822	301	+	17823	CACTGTG	22	ACAAAAACT	17860	38	
IGHV02-06	18881	19178	298	+	19179	CACAGTG	23	ACACAAACT	19217	39	
IGHV03-02	21000	21294	295	+	21295	CACAGTG	22	ACAAAAACT	21332	38	
IGHV02-07	22179	22467	289	+	22468	CACAGTG	23	ACAAAAACC	22506	39	
IGHV02-08p	24234	24514	281	+	24515	CACAGTG	23	ACAAAAACT	24553	39	Frameshift
IGHV04-01	25359	25659	301	+	25660	CACAGTG	23	ACAAAAACT	25698	39	
IGHV04-02	27066	27366	301	+	27367	CACAGTG	23	ACAAAAACA	27405	39	
IGHV02-09	28669	28958	290	+	28959	CACAGTG	23	ACAAAAACC	28997	39	
IGHV02-10p	30460	30741	282	+	30742	CACAATG	23	ACAAAACTC	30780	39	Frameshift
IGHV02-11	32395	32681	287	+	32682	CACAGTG	23	ACAAAAACC	32720	39	
IGHV03-03	33663	33957	295	+	33958	CACTGTG	22	ACAAAAACT	33995	38	
IGHV02-12	35012	35299	288	+	35300	CACAGTG	23	ACAAAAACC	35338	39	
IGHV03-04	36281	36575	295	+	36576	CACTGTG	22	ACAAAAACT	36613	38	
IGHV02-13	37639	37931	293	+	37932	CACAGTG	23	ACAAAAACT	37970	39	
IGHV02-14	39019	39311	293	+	39312	CACAGTG	23	ACAAAAACT	39350	39	
IGHV03-05	41008	41302	295	+	41303	CACAGTG	22	ACAAAAACT	41340	38	
IGHV02-15	42660	42952	293	+	42953	CACAGTG	23	ACAAAAACT	42991	39	
IGHV03-06	45081	45375	295	+	45376	CACAGTG	22	ACAAAAACT	45413	38	
IGHV02-16	46732	47024	293	+	47025	CACAGTG	23	ACAAAAACT	47063	39	

Table 3.14: Co-ordinate table of VH segments in the *Xiphophorus maculatus IGH* locus, part 1.

Name	Start	End	Length	Strand	RSS Start	Heptamer	Spacer Length	Nonamer	RSS End	RSS Length	Comment
IGHV03-07	48618	48912	295	+	48913	CACAGTG	22	ACAAAAACT	48950	38	
IGHV02-17	50323	50611	289	+	50612	CACAGTG	23	ACAAAAACC	50650	39	
IGHV03-08	51890	52184	295	+	52185	CACAGTG	22	ACAAAAACT	52222	38	
IGHV03-09p	53026	53274	249	+	53275						3'-truncated, no RSS
IGHV02-18	54462	54747	286	+	54748	CACAGTG	23	ACAAAAACC	54786	39	
IGHV02-19p	55729	55866	138	+	55867	CACAGTG	23	ACAAAAACC	55905	39	3'-truncated
IGHV03-10	57371	57662	292	+	57663	CACAGTG	22	ACAAAAACT	57700	38	
IGHV02-20p	58698	58986	289	+	58987	CACAGTG	23	ATAAAAACC	59025	39	Nonsense mutation
IGHV03-11	59940	60234	295	+	60235	CACAGTG	22	ACAAAAACT	60272	38	
IGHV02-21	61249	61537	289	+	61538	CACAGTG	23	ATAAAAACC	61576	39	
IGHV03-12	62491	62785	295	+	62786	CACAGTG	22	ACAAAAACT	62823	38	
IGHV02-22	63801	64089	289	+	64090	CACAGTG	23	ATAAAAACC	64128	39	
IGHV03-13	65043	65337	295	+	65338	CACAGTG	22	ACAAAAACT	65375	38	
IGHV02-23	66354	66640	287	+	66641	CACAGTG	23	ACAAAAACT	66679	39	
IGHV03-14	68452	68743	292	+	68744	CACTATG	22	ACAAAACTC	68781	38	
IGHV02-24	70101	70389	289	+	70390	CACAGTG	23	ACAAAAACC	70428	39	
IGHV03-15	72206	72501	296	+	72502	CACAGTG	22	ACAAAAACT	72539	38	
IGHV02-25	73484	73772	289	+	73773	CACAGTG	23	ACAAAAACC	73811	39	
IGHV03-16	75799	76090	292	+	76091	CACAGTG	22	ACAAAAACT	76128	38	
IGHV03-17	77773	78067	295	+	78068	CACAGTG	22	ACAAAAACT	78105	38	
IGHV02-26	79001	79289	289	+	79290	CACAGTG	23	ACAAAAACC	79328	39	
IGHV03-18	80492	80784	293	+	80785	CACAGTG	22	ACAAAAACT	80822	38	
IGHV02-27p	81799	82082	284	+	82083	CACAGTG	23	ACAAAAACC	82121	39	Frameshift
IGHV03-19	83736	84030	295	+	84031	CACAGTG	22	ACAAAAACT	84068	38	
IGHV02-28p	85093	85381	289	+	85382	CACAGGG	23	GCAAAAACC	85420	39	Nonsense mutation

Table 3.15: Co-ordinate table of VH segments in the *Xiphophorus maculatus IGH* locus, part 2.

Name	Start	End	Length	Strand	RSS Start	Heptamer	Spacer Length	Nonamer	RSS End	RSS Length	Comment
IGHV02-29	86225	86505	281	+	86506	CACAGTG	23	ATAAAAACC	86544	39	
IGHV03-20	87419	87713	295	+	87714	CACAGTG	22	ACAAAAACT	87751	38	
IGHV03-21	94532	94826	295	+	94827	CACAGTG	23	ACAAAAACC	94865	39	
IGHV03-22	96192	96489	298	+	96490	CACAGTG	23	ACAAAAACC	96528	39	
IGHV03-23	98068	98368	301	+	98369	CACAGTG	23	ACAAAAACC	98407	39	
IGHV03-24	99482	99779	298	+	99780	CACAGTG	23	ACAAAAACC	99818	39	
IGHV03-25	101639	101936	298	+	101937	CACAGTG	23	ACAAAAACC	101975	39	
IGHV05-01p	102818	103096	279	+	103097	CAGAAGC	0	ACAAAAACT	103112	16	Frameshift
IGHV03-26	104098	104389	292	+	104390	CACAGTG	23	ACAAAATCC	104428	39	
IGHV06-01	105551	105831	281	+	105832	CACAGTG	23	ACAAAAACC	105870	39	
IGHV03-27	107274	107571	298	+	107572	CACAGTG	23	ACAAAAACC	107610	39	
IGHV03-28	108775	109072	298	+	109073	CACAGAG	23	ACAAAAACC	109111	39	
IGHV03-29	110372	110672	301	+	110673	CACAGTG	23	ACAAAAACC	110711	39	
IGHV07-01	111565	111856	292	+	111857	CACAATG	23	ACAAAAACT	111895	39	
IGHV08-01p	113033	113330	298	+	113331	CACAGAG	23	CCAAGAACC	113369	39	Nonsense mutation
IGHV09-01	115512	115800	289	+	115801	CACAGTG	22	ACAAAAACT	115838	38	
IGHV10-01	117078	117379	302	+	117380	CACAGTG	22	ACATAAACT	117417	38	
IGHV11-01	119462	119760	299	+	119761	CACAGTG	23	ACAAAAACT	119799	39	
IGHV03-30	126125	126416	292	+	126417	CACAGTG	22	ACAAAAACC	126454	38	
IGHV03-31	127109	127400	292	+	127401	CACAGTG	23	GCAAAAACC	127439	39	
IGHV12-01	128489	128786	298	+	128787	CACAGTG	23	ACAAAAACC	128825	39	
IGHV02-30	135711	136000	290	+	136001	CACAGTG	22	ACAAAAACA	136038	38	
IGHV13-01	136757	137057	301	+	137058	CACAGTG	23	ACAAAAACT	137096	39	
IGHV02-31	138344	138637	294	+	138638	CACAGTG	23	ACAAAAATC	138676	39	
IGHV02-32	140024	140315	292	+	140316	CACTGTG	23	ACAAAAACT	140354	39	

Table 3.16: Co-ordinate table of VH segments in the *Xiphophorus maculatus IGH* locus, part 3.

Name	Start	End	Length	Strand	RSS Start	Heptamer	Spacer Length	Nonamer	RSS End	RSS Length	Comment
IGHV02-33	142332	142620	289	+	142621	CACAGTG	23	ACAAAAACA	142659	39	
IGHV02-34	144334	144625	292	+	144626	CACAGTG	23	ACAAAAACT	144664	39	
IGHV02-35	145740	146031	292	+	146032	CACAGTG	23	ACAAAAAAT	146070	39	
IGHV02-36	146903	147194	292	+	147195	CACAGTG	23	ACAAAAACT	147233	39	
IGHV02-37	147839	148138	300	+	148139	CACAGTG	23	ACAAAAATC	148177	39	
IGHV02-38p	150504	150797	294	+	150798	CACAATA	23	ACAAAAACC	150836	39	Nonsense mutation
IGHV02-39	152249	152537	289	+	152538	CACAGTA	23	ACAAAAACC	152576	39	
IGHV14-01	154075	154374	300	+	154375	CACAGTG	23	ACAAAAAGT	154413	39	
IGHV02-40	155433	155709	277	+	155710	CACAGTG	23	ACAAAAACC	155748	39	
IGHV02-41	156583	156870	288	+	156871	CACAGTG	23	ACAAAAACC	156909	39	
IGHV02-42	163977	164269	293	+	164270	CACAGTG	23	ACAAAACCC	164308	39	
IGHV03-32	165416	165708	293	+	165709	CACAGTG	22	ACAAAAACA	165746	38	
IGHV02-43	166994	167293	300	+	167294	CACAATG	23	ACAGAAACT	167332	39	
IGHV12-02	169602	169900	299	+	169901	CACAGTG	23	ACAAAAACC	169939	39	
IGHV02-44	171452	171752	301	+	171753	CACTGTG	23	GCAAAAACT	171791	39	
IGHV02-45	173096	173384	289	+	173385	CTCAGTG	23	ACAAAAACC	173423	39	
IGHV02-46	174714	175009	296	+	175010	CACAGTG	23	ACAAAAACT	175048	39	
IGHV02-47	176396	176697	302	+	176698	CACAGTG	23	ACAAAAACT	176736	39	
IGHV12-03	178422	178719	298	+	178720	CACAGTG	23	ACAAAAACA	178758	39	
IGHV12-04	181245	181543	299	+	181544	CACAGTG	23	ACAAAAACC	181582	39	
IGHV02-48p	182977	183236	260	+	183237	CACAGGT	8	ACAAAAACT	183260	24	5'-truncated
IGHV02-49p	184323	184611	289	+	184612	CACAGTG	23	ACAAAAACC	184650	39	Nonsense mutation
IGHV02-50	185946	186244	299	+	186245	CACAGTG	23	ACAAAAACT	186283	39	
IGHV02-51	187624	187925	302	+	187926	CACAGTG	23	ACAAAAACT	187964	39	
IGHV12-05	190987	191284	298	+	191285	CACAGTG	23	ACAAAAACA	191323	39	

Table 3.17: Co-ordinate table of VH segments in the *Xiphophorus maculatus IGH* locus, part 4.

Name	Start	End	Length	Strand	RSS Start	Heptamer	Spacer Length	Nonamer	RSS End	RSS Length	Comment
IGHV02-52	192570	192868	299	+	192869	CACAGTG	19	CTGAAAACC	192903	35	
IGHV12-06	193608	193906	299	+	193907	CACAGTG	23	ACAAAAACA	193945	39	
IGHV02-53	195271	195572	302	+	195573	CACAGTG	23	ACAAAAACC	195611	39	
IGHV15-01	204396	204693	298	+	204694	CACAATC	23	ACAAAAACT	204732	39	
IGHV13-02	206203	206503	301	+	206504	CACAGTG	23	ACAAAAACT	206542	39	
IGHV16-01	207726	208020	295	+	208021	CACAGTG	22	ACAAAAACT	208058	38	
IGHV13-03	208477	208777	301	+	208778	CACAGTA	23	ACAAAAACT	208816	39	
IGHV03-33	209921	210215	295	+	210216	CACGGTG	22	ACGAAAACT	210253	38	
IGHV17-01	211322	211625	304	+	211626	CACAGTA	23	ACAAAAACC	211664	39	
IGHV15-02p	214600	214860	261	+	214861						3’-truncated, no RSS
IGHV18-01	215671	215962	292	+	215963	CACACTG	23	ACAAAAACC	216001	39	
IGHV19-01	217874	218174	301	+	218175	CACAGTG	23	ACAAAAACT	218213	39	
IGHV03-34	219368	219668	301	+	219669	CACAGTG	23	ACAAAAACA	219707	39	
IGHV20-01	220329	220632	304	+	220633	CACAGTG	23	ACAAAAATT	220671	39	
IGHV02-54p	228547	228838	292	+	228839	CACACTG	23	ACAACCCCC	228877	39	Nonsense mutation
IGHV02-55	229963	230267	305	+	230268	CACAGCG	23	ACAAAAAAA	230306	39	
IGHV03-35	231630	231928	299	+	231929	CACAGTG	23	ACAAAAACC	231967	39	
IGHV21-01p	233069	233230	162	+	233231						Nonsense mutation, 3’-truncated, no RSS
IGHV22-01p	234954	235102	149	+	235103	CACAGTG	23	TCAAAAACT	235141	39	5’-truncated
IGHV02-56	236029	236330	302	+	236331	CACAGTG	23	ACAAATACT	236369	39	
IGHV03-36p	238122	238413	292	+	238414	CACAATG	23	ACAGAATCC	238452	39	Nonsense mutation
IGHV11-02p	240281	240579	299	+	240580	CACAGTG	24	ACAAAAACT	240619	40	Nonsense mutation
IGHV09-02	241878	242166	289	+	242167	CACAGTG	22	ACAAAAACT	242204	38	
IGHV23-01	243867	244164	298	+	244165	CACAGTG	23	ACAAAATCC	244203	39	
IGHV02-57	245524	245813	290	+	245814	CACCATA	22	ACAAAATCC	245851	38	

Table 3.18: Co-ordinate table of VH segments in the *Xiphophorus maculatus IGH* locus, part 5.

Table 3.19: Co-ordinate table of DH segments in the *Xiphophorus maculatus IGH* locus.

Name	Start	NT Sequence	End	Length	Strand
IGHDZ01	2243	GTGGGCAGGAGGCTATGC	2260	18	+
IGHDZ02	119768	AGG	119770	3	+
IGHDZ03	128794	ACTAAAGG	128801	8	+
IGHDZ04	129907	ATCGGG	129912	6	+
IGHDZ05	158017	ATATATGGGGG	158027	11	+
IGHDZ06	197791	ATATACTGGGGTGG	197804	14	+
IGHDZ07	222022	ATGGACTGGGGGG	222034	13	+
IGHDZ08	247941	GTGATTACGGCTACGGGGC	247959	19	+
IGHDZ09	249514	TTATGGGCTGGGGAG	249528	15	+
IGHDZ10	253752	TGGGTGGGGC	253761	10	+
IGHDM01	267392	TATACAGTGGCAAC	267405	14	+
IGHDM02	268498	CAGTATAGCAAC	268509	12	+
IGHDM03	268836	TACAATGGCAAC	268847	12	+
IGHDM04	269694	TAAACAGTGGCTAC	269707	14	+

Table 3.20: Co-ordinate table of DH 5'-RSSs in the *Xiphophorus maculatus IGH* locus.

Name	5'-RSS Start	Nonamer	Spacer Length	Heptamer	5'-RSS End	Length
IGHDZ01	2215	GGTTTTTGT	12	CACTGTG	2242	28
IGHDZ02	119739	TGTATTACT	13	CACAGTG	119767	29
IGHDZ03	128766	TTTACTTCT	12	CACAGTG	128793	28
IGHDZ04	129879	GGTTTTTGT	12	CACAGTG	129906	28
IGHDZ05	157989	AGTTTTTGT	12	CACAGTG	158016	28
IGHDZ06	197763	GGTTTTTGC	12	TACTGTG	197790	28
IGHDZ07	221994	GGTTTTTGT	12	CGCTGTG	222021	28
IGHDZ08	247913	TGTTTTTGT	12	ATCTGTG	247940	28
IGHDZ09	249486	AGTTTTTGT	12	TGTGGTG	249513	28
IGHDZ10	253724	AGTTTTTGT	12	TGTAGTG	253751	28
IGHDM01	267364	AGTTTTTGT	12	TACAGTG	267391	28
IGHDM02	268470	TGTTTTTGT	12	CACAGTG	268497	28
IGHDM03	268808	AGTTTTTGC	12	TACTGTG	268835	28
IGHDM04	269666	CGTTTTTGT	12	CATTGTG	269693	28

Table 3.21: Co-ordinate table of DH 3'-RSSs in the *Xiphophorus maculatus IGH* locus.

Name	3'-RSS Start	Heptamer	Spacer Length	Nonamer	3'-RSS End	Length
IGHDZ01	2261	CACTAAG	12	ACAAAAAGT	2288	28
IGHDZ02	119771	CAAAATG	13	ACAAAAACT	119799	29
IGHDZ03	128802	CAGAGAA	8	ACAAAAACC	128825	24
IGHDZ04	129913	CACAATG	12	TCAAAAACC	129940	28
IGHDZ05	158028	CACAGAG	12	ACAAAAACC	158055	28
IGHDZ06	197805	CACACAG	12	ACAAAAACC	197832	28
IGHDZ07	222035	CACAGAG	12	ACAAAAACC	222062	28
IGHDZ08	247960	CACAATA	12	ACAAAAACC	247987	28
IGHDZ09	249529	CACAATG	12	ACAAAAACC	249556	28
IGHDZ10	253762	CACAGTA	12	ACAAAAACC	253789	28
IGHDM01	267406	CACAGTG	12	GCAAAAACC	267433	28
IGHDM02	268510	CACAGTG	12	ACAGAAACC	268537	28
IGHDM03	268848	CACAGTG	12	ACAAAAACC	268875	28
IGHDM04	269708	CACTGTG	12	ACAAAATCA	269735	28

Name	Start	NT Sequence	AA Sequence	End	Length	Strand
IGHJZ01	2653	ATGCCTTAGATTACTGGGGTGAAGGGACCAGAGTCACAGTGACTTCAG	ALDYWGEGTRVTVTS	2700	48	+
IGHJZ02	120639	ATTACGCTCTTGACTACTGGGGAGCAGGAACCAAAGTTACTGTAAAGCCAG	YALDYWGAGTKVTVKP	120689	51	+
IGHJZ03	130376	ACTACGGCTTTGATTACTGGGGAGACGGAACTGAAGTTACTGTTGAACCAG	YGFDYWGDGTEVTVEP	130426	51	+
IGHJZ04	158408	AGATTTAGACTACTGGGGTAATGGAACAACAGTCACGGTTCTACCAG	DLDYWGNGTTVTVLP	158454	47	+
IGHJZ05	198186	ATTATGGTTTTGACTACTGGGGAGACGGAACCACAGTCACTGTTAGTCCAG	YGFDYWGDGTTVTVSP	198236	51	+
IGHJZ06	222417	ATGCTTTTGACGTCTGGGGTAAAGGAACCACAGTTACTGTTGTACCAG	AFDVWGKGTTVTVVP	222464	48	+
IGHJZ07	254130	ATGTTTTTGACTACTGGGGTAAAGGGACTGATGTCACAGTATCTCCAG	VFDYWGKGTDVTVSP	254177	48	+
IGHJM01	276014	ACGGCTACTTCGACTACTGGGGGAAAGGAACACAAGTCACAGTGACTTCTG	GYFDYWGKGTQVTVTS	276064	51	+
IGHJM02	276284	CCACTACTTTGACTACTGGGGAAAAGGAACCACGGTTACCGTCACTTCAG	HYFDYWGKGTTVTVTS	276333	50	+
IGHJM03	276654	ACAATGCTTTTGACTACTGGGGAAAAGGAACTACGGTAACAGTAACATCAG	NAFDYWGKGTTVTVTS	276704	51	+
IGHJM04	276999	ACTACGCTTTTGACTACTGGGGAAAAGGAACAATGGTCACTGTCACTTCAG	YAFDYWGKGTMVTVTS	277049	51	+
IGHJM05	277322	ACAACTGGGCTTTTGACTACTGGGGAGCAGGAACCATGGTAACAGTAACATCAG	NWAFDYWGAGTMVTVTS	277375	54	+
IGHJM06	277672	CTACGGTGCTTTTGACTACTGGGGTAAAGGGACTACAGTCACCGTCACTTCAG	YGAFDYWGKGTTVTVTS	277724	53	+
IGHJM07	278150	CTACGATGCTTTTGACTATTGGGGGAAAGGAACAACAGTCACCGTCATCACTTCAG	YDAFDYWGKGTTVTVITS	278205	56	+
IGHJM08	278606	TTACTACTACGCTTTTGACTATTGGGGAAAAGGGACAATGGTCACCGTCACTTCAG	YYYAFDYWGKGTMVTVTS	278661	56	+

Table 3.22: Co-ordinate table of JH segments in the *Xiphophorus maculatus IGH* locus.

Name	RSS Start	Nonamer	Spacer Length	Heptamer	RSS End	RSS Length
IGHJZ01	2662	TGTTTTTGT	23	CACTGTG	2652	39
IGHJZ02	120651	TGTTTTTGT	23	CACTGTG	120638	39
IGHJZ03	130388	TGTTTTTGT	23	CACCGTG	130375	39
IGHJZ04	158416	GGTTTTTGT	23	CACTGTG	158407	39
IGHJZ05	198198	GGTTTTTGT	23	CACTGTG	198185	39
IGHJZ06	222426	TGTTTTTGT	23	CACTGTG	222416	39
IGHJZ07	254139	GGTTTTTGT	23	CACTGTG	254129	39
IGHJM01	276026	TGTATTTGT	23	CACTGTG	276013	39
IGHJM02	276295	TATTTTTGC	23	CACCGTG	276283	39
IGHJM03	276666	TGTTTTTGT	23	TACTGTG	276653	39
IGHJM04	277011	TGTTTTAGT	23	TACTGTG	276998	39
IGHJM05	277338	GGTTTTTGT	22	TACTGTG	277321	38
IGHJM06	277687	GCTTTTTAT	22	CACTGTG	277671	38
IGHJM07	278168	CCTTTTTAC	22	CACTGTG	278149	38
IGHJM08	278624	GCTTTTTAA	22	CACTGTG	278605	38

Table 3.23: Co-ordinate table of JH RSSs in the *Xiphophorus maculatus IGH* locus.

3.4 *IGH* constant-region evolution in the Atherinomorpha

The characterised *IGH* loci of *Nothobranchius furzeri*, *Xiphophorus maculatus* and medaka together reveal a high degree of variability in structure and function across the Atherinomorpha, the parent clade of the Cyprinidontiformes (including *N. furzeri* and *X. maculatus*) and the Beloniformes (including medaka). Several unusual features (including the loss of *IGHZ*, an inverted sublocus, and an unusual splicing pattern of *IGHM-TM*) are shared between medaka and turquoise killifish, but absent in platyfish (Figure 3.14), indicating either an independent origin in the former two species or a reversion to the primitive state in the latter. In addition, the copy number, exon usage and orientation of constant regions of other isotypes differs among the three species, raising the further question of how, when and why these changes occurred.

In order to investigate a subset of these questions with a greater degree of phylogenetic resolution, *IGH* constant regions were identified and analysed in a further ten cyprinodontiform species (Figure 3.1, Table 3.24), as well as in a new and improved genome assembly of medaka (Genbank accession GCA_002234675.1), using the same methods described for *N. furzeri* and *X. maculatus* (Sections 3.2.3 and 3.2.4). The exons so identified (Figure 3.22) were then grouped by order, exon type and spatial proximity to identify contiguous constant-regions, enabling the presence/absence and number of constant regions of each isotype to be estimated for each species. The results of this analysis (Figure 3.23, Tables 3.25 to 3.27) demonstrate that every species investigated possesses at least one complete *IGHM* and *IGHD* constant region in tandem, with several species exhibiting multiple such regions. Apart from *N. furzeri* and medaka, only *Nothobranchius orthonotus* was identified as clearly possessing adjacent constant regions in opposite orientation, indicating the presence of at least one sublocus in antisense; however, the fragmented nature of the *IGH* locus assembly in many analysed species prevented a confident exclusion of such loci in other cases.

In addition to at least one *IGHM* and *IGHD* constant region, the majority of species analysed (8 out of 13) were also found to possess at least one complete *IGHZ* constant region; of the exceptions, *A. limnaeus* exhibits an orphaned, pseudogenised *IGHZ-TM1* exon but no $C_\zeta$exons in the current assembly (Figure 3.23, Table 3.26), while *O. latipes*, *A. australe*, *N. furzeri* and *N. orthonotus* display no *IGHZ* exons at all. Multiple duplications of *IGHZ* appeared even more common than for the other isotypes, with an average of 2.125 regions per *IGHZ*-bearing locus. Annotating the tree from Figure 3.1 with the *IGHZ* status of each species (Figure 3.24) confirms that the loss of *IGHZ* in turquoise killifish (and related species) and medaka represent two distinct deletion events, with *Austrofundulus limnaeus* potentially representing a third independent loss of *IGHZ* within the Atherinomorpha.

Apart from its repeated loss within the lineage, a second striking feature of *IGHZ* within the Atherinomorpha is its frequent presence in multiple copies per *IGH* locus; on average (geometric mean), the species analysed have approximately 1.62 *IGHZ* constant regions per *IGHM* constant

region, and the same ratio veforrsus *IGHD*, suggesting a more complex evolutionary history than can be captured by a simple presence/absence metric. Concordantly, phylogenetic analysis (Figure 3.25, tree built using `PRANK` and `RAxML` on of $C_\zeta 1$–$C_\zeta 4$ exon sequences) reveals three distinct lineages (or *subclasses*) of *IGHZ* constant regions in the Cyprinidontiformes, each of which is present in multiple different species and appears to have been present in the common ancestor of the eight *IGHZ*-bearing species analysed (Figure 3.26).

Only one *IGHZ* constant region from the analysed species could not be confidently assigned to one of these three subclasses, namely the single *IGHZ* of *Pachypanchax playfairii* (Figure 3.25). In order to more closely investigate the relationships of *IGHZ* in this species, the exon sequences of *P. playfairii* $C_\zeta 1$–$C_\zeta 4$ were separately aligned to the C_ζ exons of all other *IGHZ*-bearing species using Needleman-Wunsch global alignments, and the distribution of alignment scores was plotted in Figure 3.27 . The results show a striking difference in alignment behaviour between the exons, with $C_\zeta 1$ and $C_\zeta 2$ aligning much more strongly to exons from the B subclass and $C_\zeta 3$ and $C_\zeta 4$ showing more ambiguous affinity for either A- or C-subclass sequences. This unexpected behaviour indicates that the *P. playfairii IGHZ* sequence is the result of a deletion or fusion event combining the first two exons of a B-lineage *IGHZ* constant region with the latter exons of a constant region from another lineage, resulting in a chimeric gene with ambiguous ancestry.

In summary, in addition to the still-universal primitive antibody classes *IGHM* and *IGHD*, the cyprinodontiforms ancestrally possessed both at least three variants of *IGHZ*, giving rise to multiple

Genus	**Species**	**Common Name**	**GenBank Assembly Accession**
Nothobranchius	*furzeri*	Turquoise killifish	NA[1]
Xiphophorus	*maculatus*	Southern platyfish	GCA_002775205.2
Austrofundulus	*limnaeus*	–	GCA_001266775.1
Fundulus	*heteroclitus*	Mummichog	GCA_000826765.1
Poecilia	*formosa*	Amazon molly	GCA_000485575.1
Poecilia	*reticulata*	Guppy	GCA_000633615.1
Cyprinodon	*variegatus*	Sheepshead minnow	GCA_000732505.1
Kryptolebias	*marmoratus*	Mangrove rivulus	GCA_001649575.1
Aphyosemion	*australe*	Lyretail panchax	NA[2]
Callopanchax	*toddi*	–	NA[2]
Pachypanchax	*playfairii*	Golden panchax	NA[2]
Nothobranchius	*orthonotus*	Spotted killifish	NA[2]
Oryzias	*latipes*	Medaka	GCA_002234675.1

[1] Willemsen *et al.*, unpublished at time of writing
[2] Cui *et al.*, unpublished at time of writing

Table 3.24: Genome assemblies used to identify putative *IGH* locus sequences in cyprinodontiform fishes.

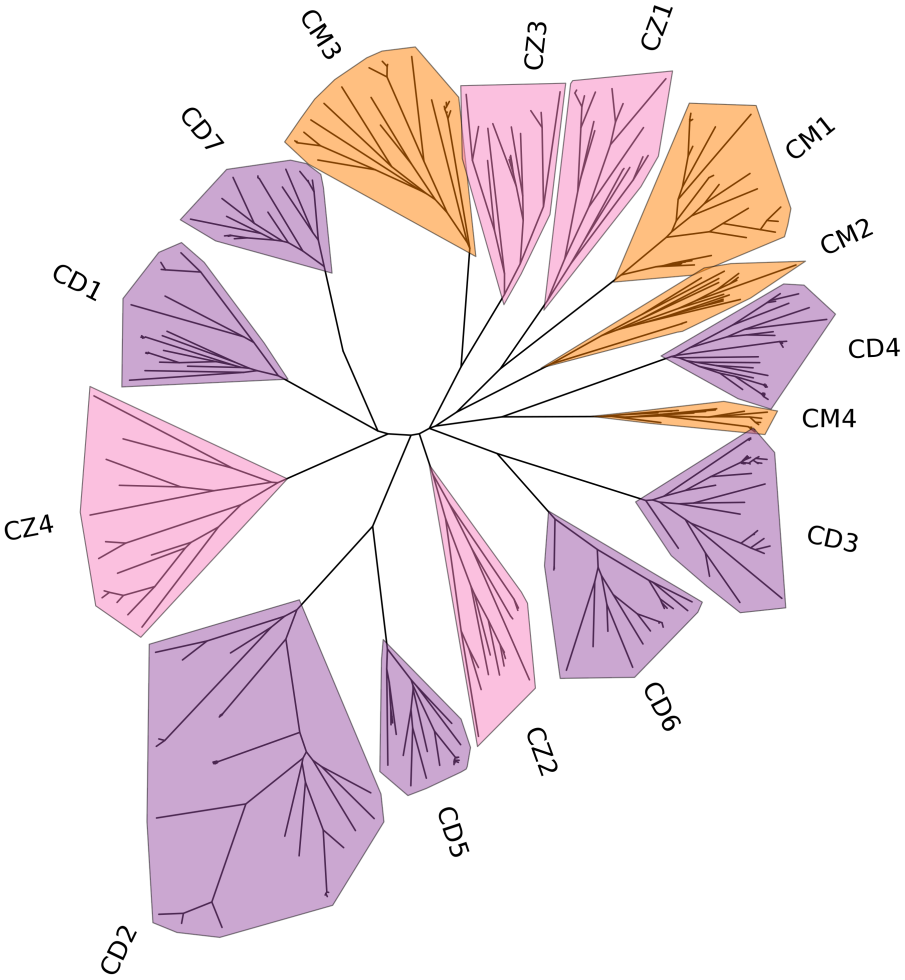


Figure 3.22: CH exons in the Atherinomorpha: Unrooted phylogram of CH exons in thirteen fish species from the Atherinomorpha, constructed using `PRANK` and `RAxML`. Each exon type is clustered separately in the tree topology, indicating that the types of the identified exons have all been correctly annotated.

subclasses of *IGHZ* constant regions evolving in parallel across the clade. Each of these subclasses appears to have been lost in multiple cyprinodontiform species, with different species showing distinct patterns of retention and loss, and in at least one lineage – that of *Pachypanchax playfairii* – two different *IGHZ* lineages have fused to produce a chimeric isotype. All three subclasses are missing from a subset of species in the Nothobranchiidae (including *Nothobranchius furzeri*), and also appear to have been independently lost in *Austrofundulus limnaeus*. Taken together, these data suggest a high degree of complexity and volatility in the evolution of mucosal adaptive immunity in the Cyprinodontiformes.

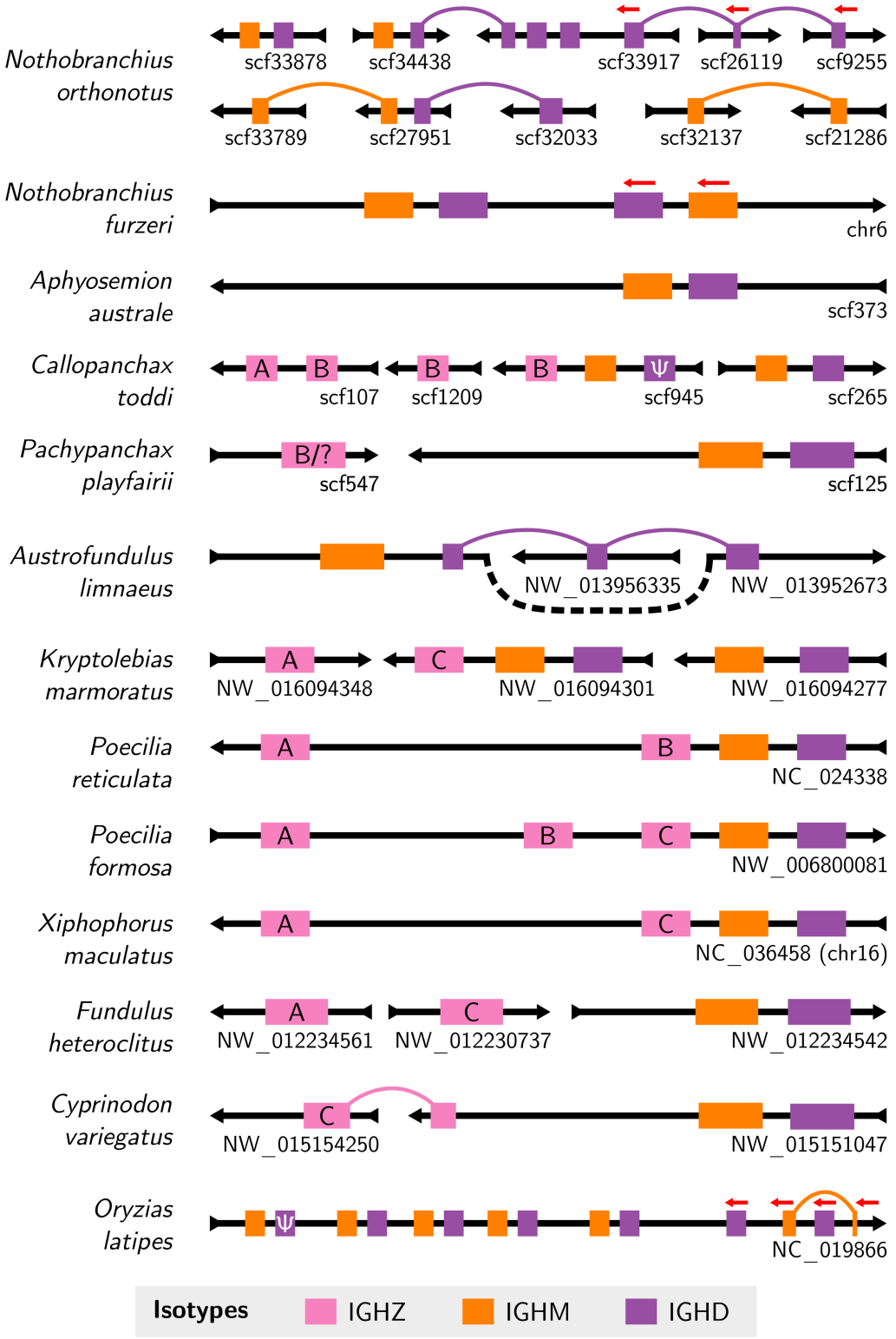


Figure 3.23: Constant-region organisation in the Atherinomorpha: Schematic of *IGH* constant regions in the genomes of thirteen species from the Atherinomorpha. Scaffold orientation is given by the black arrows; constant regions are oriented left-to-right unless otherwise specified (red arrows). Links between regions on different scaffolds indicate that exons from what appears to be the same constant region are distributed across multiple scaffolds in the order indicated; the order of unlinked scaffolds is arbitrary. The isotype of each region is given by its colour; *IGHZ* regions are further annotated with their subclass (Figure 3.24). Clearly pseudogenised constant regions are indicated by Ψ. Isotype length, scaffold length, and scaffold position are not to scale. Variable regions and lone, isolated constant-region exons are not shown.

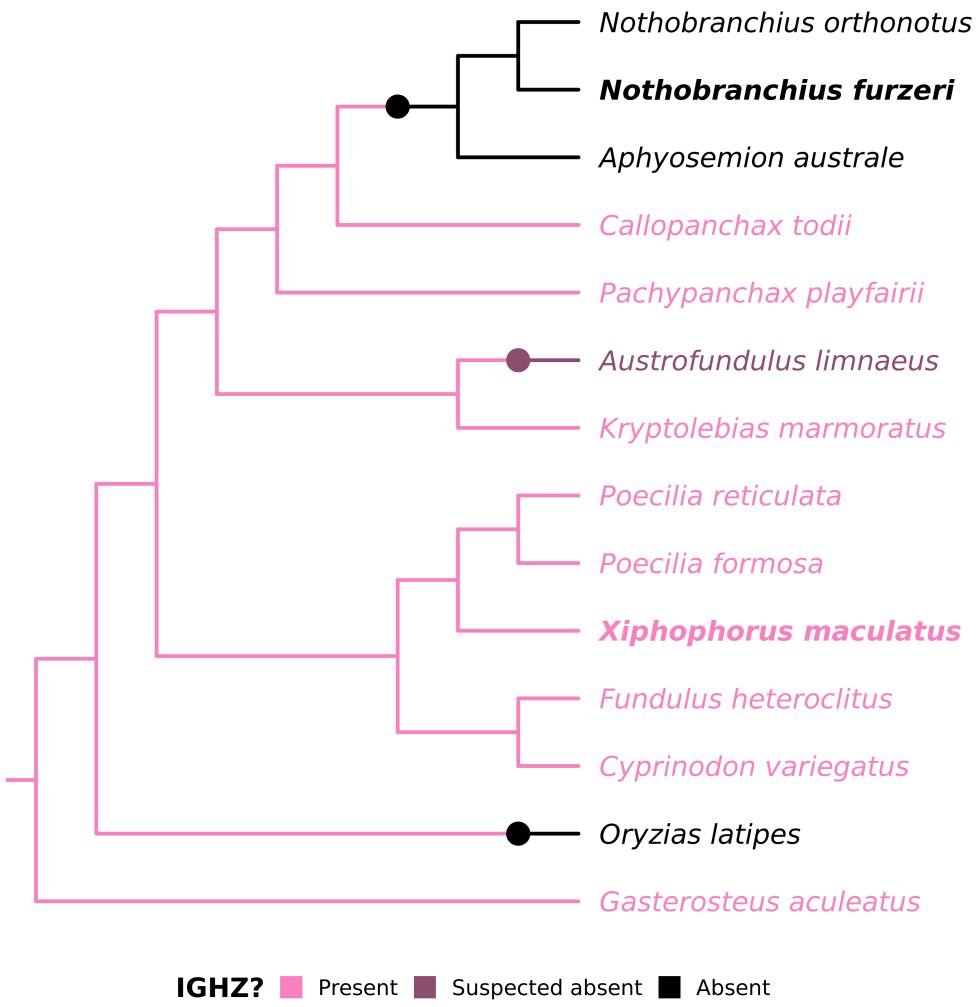


Figure 3.24: *IGHZ* in the Atherinomorpha has been lost multiple times independently: Cladogram of species reproduced from Figure 3.1, annotated according to the known (tip nodes) or inferred (internal nodes) presence or absence of intact *IGHZ* constant regions in each species. Large coloured points on the cladogram denote sites of hypothesised state changes; *IGHZ* is assumed to be primitively present in the clade and losses to be irreversible. The currently-available genome assembly of *A. limnaeus* (dark pink) contains one pseudogenised *IGHZ-TM1* exon and no C_ζ exons.

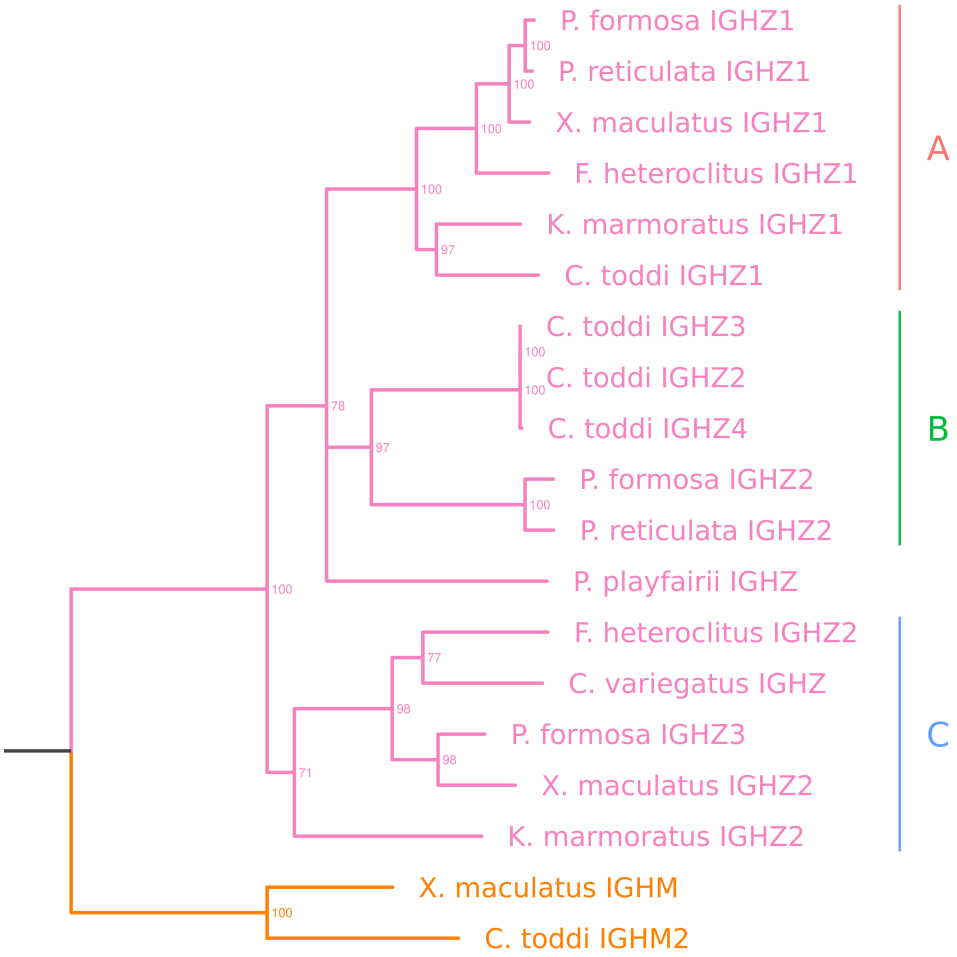
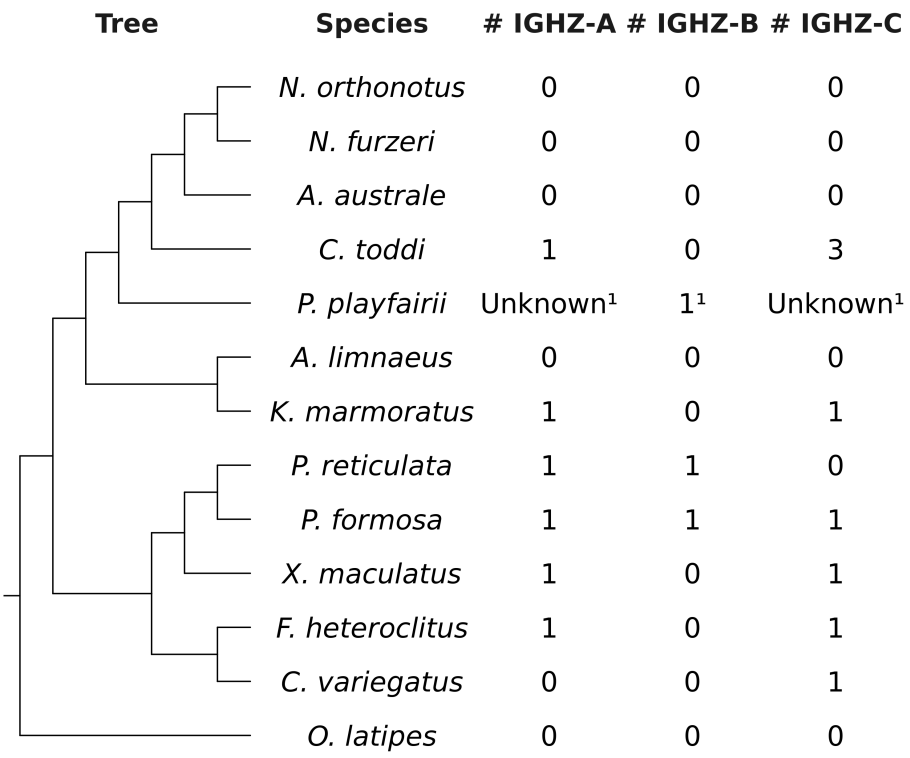


Figure 3.25: *IGHZ* constant regions constitute three distinct subclasses: Phylogram of concatenated $C_\zeta 1$–$C_\zeta 4$ nucleotide sequences from *IGHZ*-bearing species in Table 3.24, with $C_\mu 1$–$C_\mu 4$ sequences from two species as an outgroup (in orange). Major clades (A–C) are annotated on the right. Support values indicate the result of rapid bootstrapping by `RAxML` across 1000 replicates.



Tree	Species	# IGHZ-A	# IGHZ-B	# IGHZ-C
	N. orthonotus	0	0	0
	N. furzeri	0	0	0
	A. australe	0	0	0
	C. toddi	1	0	3
	P. playfairii	Unknown[1]	1[1]	Unknown[1]
	A. limnaeus	0	0	0
	K. marmoratus	1	0	1
	P. reticulata	1	1	0
	P. formosa	1	1	1
	X. maculatus	1	0	1
	F. heteroclitus	1	0	1
	C. variegatus	0	0	1
	O. latipes	0	0	0

[1] *P. playfairii IGHZ* $C_\zeta 1$–$C_\zeta 2$ appear to be derived from *IGHZ-B*, while the other exons are of uncertain subclass origin (Figure 3.27).

Figure 3.26: *IGHZ* subclasses in the Atherinomorpha: Cladogram of atherinomorph species with characterised *IGHZ* constant regions, annotated with the number of regions belonging to each *IGHZ* isotype in each species. All three subclasses are present in at least one species in both major branches of the cyprinodontiform clade, suggesting that they were all present in the common ancestor of this grouping.

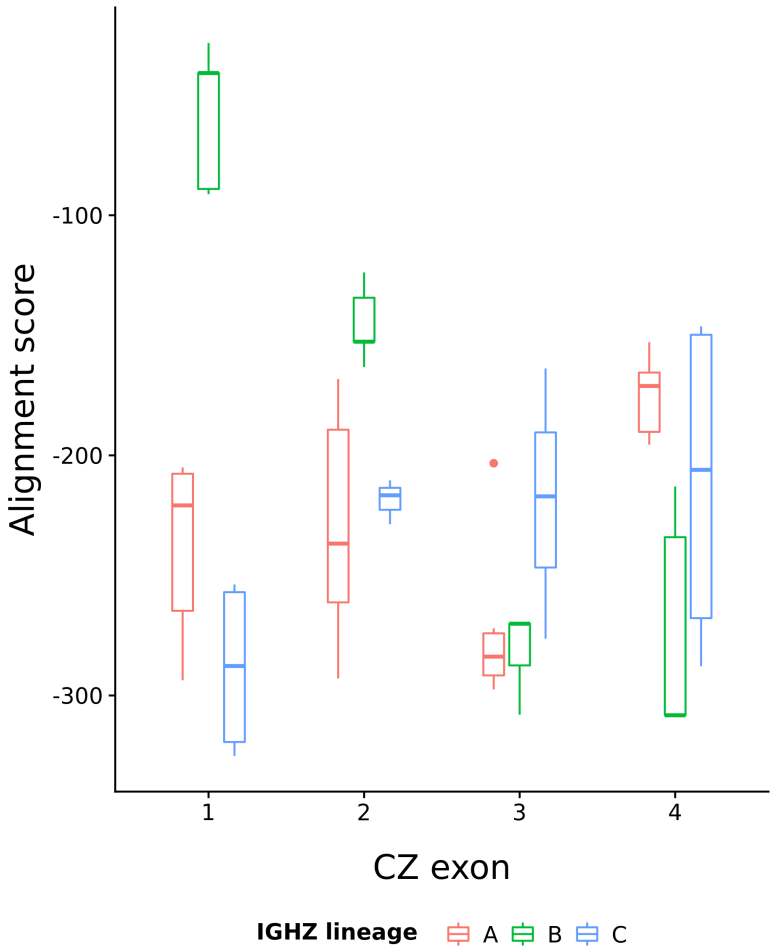


Figure 3.27: Subclass affinity of *Pachypanchax playfairii IGHZ*: Boxplots of Needleman-Wunsch alignment scores between the amino-acid sequences of *P. playfairii* C_ζ exons and those of seven other *IGHZ*-bearing cyprinodontiform species, demonstrating the differing affinity of diferent *P. playfairii* exons for each of the three *IGHZ* subclasses.

Species	Scaffold(s)	Region	Isotype	Known Exons [1]	Complete?	Pseudo-exons	Comments
Nothobranchius orthonotus	scf33878	IGHM1	M	1,2,3,TM1	**No**	–	CM4 missing (missing sequence)
Nothobranchius orthonotus	scf33878	IGHD1	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Nothobranchius orthonotus	scf34438	IGHM2	M	1,2,3,4,TM1	Yes	–	
Nothobranchius orthonotus	scf34438, scf33917	IGHD2	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Nothobranchius orthonotus	scf33917	IGHD3	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Nothobranchius orthonotus	scf33917	IGHD4	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Nothobranchius orthonotus	scf9255, scf26119, scf33917	IGHD5	D	3,4,2,3,4,5,6,7,TM1	**No**	–	CD1 & CD2A missing (missing sequence)
Nothobranchius orthonotus	scf27951, scf33789	IGHM3	M	1,2,3,4,TM1	Yes	–	
Nothobranchius orthonotus	scf27951, 32033	IGHD6	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Nothobranchius orthonotus	scf32137, scf21286	IGHM4	M	1,2,3,4,TM1	Yes	–	
Nothobranchius furzeri	chr6 [2]	IGH1M	M	1,2,3,4,TM1	Yes	–	
Nothobranchius furzeri	chr6 [2]	IGH1D	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Nothobranchius furzeri	chr6 [2]	IGH2M	M	1,2,3,4,TM1	Yes	–	
Nothobranchius furzeri	chr6 [2]	IGH2D	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Aphyosemion australe	scf373	IGHM	M	1,2,3,4,TM1	Yes	–	
Aphyosemion australe	scf373	IGHD	D	1,2,3,4,5,6,7,TM1	Yes	–	
Callopanchax toddi	scf107	IGHZ1	Z	1,2,3,4,TM1	Yes	–	
Callopanchax toddi	scf107	IGHZ2	Z	1,2,3,4,TM1	Yes	–	
Callopanchax toddi	scf1209	IGHZ3	Z	1,2,3,4,TM1	Yes	–	
Callopanchax toddi	scf1209	IGHM1	M	1	**No**	–	Isolated CM1 exon
Callopanchax toddi	scf945	IGHZ4	Z	1,2,3,4,TM1	Yes	–	
Callopanchax toddi	scf945	IGHM2	M	1,2,3,4,TM1	Yes	–	
Callopanchax toddi	scf945	IGHD1	D	1,2,3,4,5,6,7,TM1	Yes	1,4,5	Frameshift mutations in CD1, CD4 & CD5
Callopanchax toddi	scf265	IGHM3	M	1,2,3,4,TM1	Yes	–	
Callopanchax toddi	scf265	IGHD2	D	1,5,7,TM1	**No**	–	CD2-4 & CD5-6 missing (not in sequence)

[1] Excluding TM2 and secretory exons.

[2] Expanded *IGH* locus sequency from Section 3.2.

Table 3.25: *IGH* constant regions in cyprinidontiform fish, part 1.

Species	Scaffold(s)	Region	Isotype	Known Exons [1]	Complete?	Pseudo-exons	Comments
Pachypanchax playfairii	scf547	IGHZ	Z	1,2,3,4,TM1	Yes	–	
Pachypanchax playfairii	scf125	IGHM1	M	1,2,3,4,TM1	Yes	–	
Pachypanchax playfairii	scf125	IGHD	D	1,2,3,4,5,6,7,TM1	Yes	–	
Pachypanchax playfairii	scf547	IGHM2	M	1	**No**	–	Isolated CM1 exon
Austrofundulus limnaeus	NW_013954375.1	IGHZ	Z	TM1	**No**	TM1	Isolated TM1 exon with frameshift mutation
Austrofundulus limnaeus	NW_013952673.1	IGHM	M	1,2,3,4,TM1	Yes	–	
Austrofundulus limnaeus	NW_013952673.1, NW_013956335.1	IGHD	D	1,2,3,4,5,6,7,TM1	Yes	–	
Kryptolebias marmoratus	NW_016094348.1	IGHZ1	Z	1,2,3,4,TM1	Yes	–	
Kryptolebias marmoratus	NW_016094348.1	IGHZ2	Z	1,4,TM1	**No**	–	CZ2 & CZ3 missing (not in sequence)
Kryptolebias marmoratus	NW_016094301.1	IGHM1	M	1,2,3,4,TM1	Yes	–	
Kryptolebias marmoratus	NW_016094301.1	IGHD1	D	1,2,3,4,5,6,7,TM1	Yes	–	
Kryptolebias marmoratus	NW_016094277.1	IGHM2	M	1,2,3,4,TM1	Yes	–	
Kryptolebias marmoratus	NW_016094277.1	IGHD2	D	1,2,3,4,5,6,TM1	**No**	–	CD7 missing (not in sequence)
Poecilia reticulata	NC_024338.1	IGHZ1	Z	1,2,3,4	**No**	–	TM1 missing (missing sequence)
Poecilia reticulata	NC_024338.1	IGHZ2	Z	1,2,3,4,TM1	Yes	–	
Poecilia reticulata	NC_024338.1	IGHM	M	1,2,3,4,TM1	Yes	–	
Poecilia reticulata	NC_024338.1	IGHD	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Poecilia formosa	NW_006800081.1	IGHZ1	Z	1,2,3,4,TM1	Yes	–	
Poecilia formosa	NW_006800081.1	IGHZ2	Z	1,2,3,4,TM1	Yes	–	
Poecilia formosa	NW_006800081.1	IGHZ3	Z	1,2,3,4,TM1	Yes	–	
Poecilia formosa	NW_006800081.1	IGHM	M	1,2,3,4,TM1	Yes	–	
Poecilia formosa	NW_006800081.1	IGHD	D	1,2,3,4,5,6,7,TM1	Yes	–	
Xiphophorus maculatus	NC_036458	IGHZ1	Z	1,2,3,4,TM1	Yes	–	
Xiphophorus maculatus	NC_036458	IGHZ2	Z	1,2,3,4,TM1	Yes	–	
Xiphophorus maculatus	NC_036458	IGHM	M	1,2,3,4,TM1	Yes	–	

[1] Excluding TM2 and secretory exons.

Table 3.26: *IGH* constant regions in cyprinidontiform fish, part 2.

Species	Scaffold(s)	Region	Isotype	Known Exons [1]	Complete?	Pseudo-exons	Comments
Xiphophorus maculatus	NC_036458	IGHD	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Fundulus heteroclitus	NW_012234561.1	IGHZ1	Z	1,2,3,4,TM1	Yes	–	
Fundulus heteroclitus	NW_012230737.1	IGHZ2	Z	4,TM1	**No**	–	CZ1 to CZ3 missing (missing sequence)
Fundulus heteroclitus	NW_012234542.1	IGHM	M	1,2,3,4,TM1	Yes	–	
Fundulus heteroclitus	NW_012234542.1	IGHD	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Cyprinodon variegatus	NW_015154250.1, NW_015151047.1	IGHZ	Z	1,2,3,4,TM1	Yes	–	
Cyprinodon variegatus	NW_015151047.1	IGHM	M	1,2,3,4,TM1	Yes	–	
Cyprinodon variegatus	NW_015151047.1	IGHD	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHM1	M	1,2,3,4,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHD1	D	1,2,3,4,6,7,TM1	Yes	7	Nonsense mutation in CD7
Oryzias latipes	NC_019866.2	IGHM2	M	1,2,3,4,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHD2	D	1,2,3,4,6,7,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHM3	M	1,2,3,4,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHD3	D	1,2,3,4,6,7,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHM4	M	1,2,3,4,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHD4	D	2,7,TM1	**No**	–	CD1 & CD3-6 missing (not in sequence)
Oryzias latipes	NC_019866.2	IGHM5	M	1,2,3,4,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHD5	D	1,2,3,4,6,7,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHM6	M	1,2,3,4,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHD6	D	1,2,3,4,6,7,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHD7	D	1,2,3,6	**No**	–	CD4, CD5, CD7 and TM1 missing (not in sequence)

[1] Excluding TM2 and secretory exons.

Table 3.27: *IGH* constant regions in cyprinidontiform fish, part 3.

3.5 Discussion

The teleost fishes represent the largest and most diverse group of vertebrates, with over 26,000 species making up almost half of all vertebrate species diversity. Of this vast diversity, only a few species, mostly those used extensively in aquaculture or scientific research, have undergone extensive study with regard to their immunoglobulin gene loci. Studies in these model species have established a high level of structural diversity among teleost *IGH* loci, with huge variation in both size and organisation, as well as the existence of three distinct antibody isotypes in fish, of which two (*IGHM* and *IGHD*) are shared with tetrapods and one (*IGHZ*) appears to be teleost-specific. However, in many large and important teleost lineages, the genetic bases of humoral immunity remains unknown.

In this chapter, I presented two complete and ten partial assemblies of *IGH* loci from the Cyprinodontiformes, a diverse clade of primarily freshwater fishes for which no such loci have previously been characterised. The two complete assemblies were of the *IGH* loci of the turquoise killifish *Nothobranchius furzeri* and the southern platyfish *Xiphophorus maculatus*, two important species with an estimated divergence time of less than 80 Mya. Despite their close relationship, these species show radically different locus organisations, with huge differences in VDJ number (24 VH segments in *N. furzeri* versus 125 in *X. maculatus*), locus organisation (two small subloci in opposite sense in *N. furzeri*, one large “sublocus" in *X. maculatus*) and isotype availability (no *IGHZ* in *N. furzeri*, two distinct *IGHZ* regions in *X. maculatus*), as well as more subtle but important distinctions like differences in constant-region splicing behaviour (four exons in *N. furzeri IGHM-TM*, five in *X. maculatus*). These results are consistent with previous findings of highly-diverse teleost loci and support a process of highly rapid evolution in the *IGH* locus. Characterisation of the constant regions of a further ten cyprinodontiform species confirmed this impression, with several groups of closely-related species (e.g. *Nothobranchius furzeri*, *Nothobranchius orthonotus* and *Callopanchax Toddi)* showing highly divergent locus structures and constant-region availability.

It is interesting to speculate on the origins of this extremely rapid diversification in gene structure. Very little is known about the relationship between environmental context and immune locus structure; it is possible that part of the variety in *IGH* gene locus structure in the Cyprinodontiformes represents divergent adaptations to different immune environments. Alternatively, this diversification may be primarily the result of unusually high rates of stochastic, non-adaptive changes in gene structure in germline *IGH*. Finally, at least some of the difference between locus structures in different species is likely to be attributable to differences in assembly quality; for example, the characterisation of medaka constant regions presented here contains many fewer unusual or incomplete constant regions than that presented in the published medaka *IGH* locus, primarily due to the increased quality of the more recent medaka genome assemblies.

Before the publication of this work, only two teleost species (medaka and channel catfish) were known or thought to lack the *IGHZ* antibody isotype from their *IGH* loci, out of more than ten species

with published locus characterisations. This relative rarity of observed absence, combined with the apparent importance of *IGHZ* in teleost mucosal immunity, suggested that the loss of *IGHZ* was likely to be a rare and unusual event. However, in addition to confirming the absence of *IGHZ* in medaka, this study identified four new teleost species (*Nothobranchius furzeri*, *Nothobranchius orthonotus*, *Aphyosemion australe* and *Austrofundulus limnaeus*) that appear to lack *IGHZ* constant regions in their *IGH* loci, representing two distinct and previously unknown loss events independent from that affecting the closely-related medaka. This finding, which triples the number of known teleost species without *IGHZ* and doubles the number of known loss events, is even more striking when combined with the discovery that the cyprinidontiform common ancestory likely had no fewer than three distinct *IGHZ* constant regions, all of which had to be lost on the way to any given *IGHZ*-free lineage; many analysed species appear to have lost at least one of these *IGHZ* subclasses, in some cases (e.g. that of *Xiphophorus maculatus* or *Poecilia reticulata*) very recently. The high level of observed variability in *IGHZ* prevalence among the cyprinidontiforms suggests that the presence/absence of *IGHZ* in the wider teleost clade may be much more volatile than typically assumed, and raises the possibility that, given sufficiently-high-density analysis of other teleost lineages, a similar frequency of *IGHZ*-lacking species may also be found elsewhere. However, it may also be the case that the apparently high frequency of *IGHZ* loss events in the Atherinomorpha is a special case, arising from chance, an unusual selective environment, or limitations in the available genome assemblies.

The absence of *IGHZ* from so many species in the Atherinomorpha naturally raises the important question of how their mucosal adaptive immune system differs from that of their *IGHZ*-bearing relatives. Data from rainbow trout suggests that *IGHT* (an alternative name for *IGHZ* in some species) plays a specialised role in antibody immune responses at multiple mucosal surfaces, with increased prevalence of $IGHT^+$ B-cells and secreted IGHT antibodies relative to serum, a primarily *IGHT*-dependent response to mucosal infections, and a much higher rate of bacterial coating by IGHT in skin and gut flora relative to IGHM [24, 25]. If these findings hold for other teleost species, it is not clear how *IGHZ*-lacking teleost species carry out specialised immune functions at mucosal barriers: how, and to what extent, can *IGHM* compensate for the lack of a specialised mucosal isotype? This question is especially interesting in the case of *IGHZ*-lacking species with close *IGHZ*-bearing relatives (e.g. *Nothobranchius furzeri* and *Callopanchax toddi*, or perhaps *Austrofundulus limnaeus* and *Kryptolebias marmoratus*); if it is the case that mucosal immune responses differ systematically between these species, such that *IGHM* takes up some or all of the roles normally played by *IGHZ*, then uncovering the mechanisms by which this shift is regulated could reveal important new insights into decision-making and control of humoral adaptive immunity.

One important difference between the *X. maculatus* and *N. furzeri* loci whose evolution is more difficult to investigate from genomic data is the exon usage behaviour of the different splice isoforms present in the transcriptome of each species. In *X. maculatus*, transmembrane *IGHM* adopts the same configuration as that seen in most teleosts for which this has been investigated: a five-exon isoform in which the end of $C_\mu 3$ is spliced to the start of TM1 and $C_\mu 4$ is excised. Conversely,

in *N. furzeri IGHM-TM* adopts the same four-exon configuration observed in medaka, in which $C_\mu 3$ is also excluded. Given that *X. maculatus* adopts the primitive configuration, the recurrence of the same unusual configuration in both medaka and turquoise killifish is surprising, and indicates that both configurations are present in the Cyprinodontiformes; however, without more information about the mechanisms and genomic sequence correlates underlying this difference, it is impossible to distinguish an independent origin of the derived phenotype in medaka and killifish from a reversion to the primitive phenotype in medaka. It is also not clear at present what functional differences, if any, arise from this difference in exon usage, although it is unlikely that the shorter four-exon form of *IGHM-TM* would persist in multiple species if it prevented effective antibody development, selection, or antigen response.

As a result of the research findings presented in this chapter, a number of previously-uncharacterised teleost species now have databases of constant regions available. As a result, primer design for targeted RNA-sequencing of expressed antibody sequences is now possible for these taxa, enabling quantitative immune-repertoire sequencing approaches in a large number of closely-related cyprinidontiform species. In addition to the special interest of immune-repertoire data from any one of these new species (for example, the turquoise killifish for immune repertoire ageing research, or the platyfish or guppy for intersections of immune repertoire biology and ecology), the possibility of sequencing the repertoires of several related species in parallel adds an exciting comparative dimension to the investigation of adaptive-immune functionality and responses to different interventions and conditions. This comparative element is especially interesting in the context of comparing the repertoire responses of closely related species with *IGHZ*-bearing characteristics. In combination with the genomic and functional comparisons discussed above, the novel possibility of large-scale comparative repertoire studies arising as a result of this research establishes the Cyprinidontiformes, and especially the African killifishes, as a highly-promising group of model species for comparative evolutionary immunology.

References

[1] David Jung, Cosmas Giallourakis, Raul Mostoslavsky, and Frederick W. Alt. "Mechanism and Control of V(d)j Recombination at the Immunoglobulin Heavy Chain Locus". *Annual Review of Immunology* 24.1 (2006), pp. 541–570. DOI: 10.1146/annurev.immunol.23.021704.115830.

[2] S. Fillatreau, A. Six, S. Magadan, R. Castro, et al. "The astonishing diversity of Ig classes and B cell repertoires in teleost fish". *B Cell Biology* 4 (2013), p. 28. DOI: 10.3389/fimmu.2013.00028.

[3] L. J. Wysocki and M. L. Gefter. "Gene Conversion and the Generation of Antibody Diversity". en. *Annu. Rev. Biochem.* 58.1 (1989), pp. 509–527.

[4] B. G. Magor. "Antibody Affinity Maturation in Fishes—Our Current Understanding". *Biology* 4.3 (July 2015), pp. 512–524. ISSN: 2079-7737. DOI: 10.3390/biology4030512.

[5] B. Patel, R. Banerjee, M. Samanta, and S. Das. "Diversity of Immunoglobulin (Ig) Isotypes and the Role of Activation-Induced Cytidine Deaminase (AID) in Fish". en. *Molecular Biotechnology* (Apr. 2018), pp. 1–19. ISSN: 1073-6085, 1559-0305. DOI: 10.1007/s12033-018-0081-8.

[6] H. W. Schroeder and L. Cavacini. "Structure and function of immunoglobulins". *Journal of Allergy and Clinical Immunology*. 2010 Primer on Allergic and Immunologic Diseases 125.2, Supplement 2 (Feb. 2010), S41–S52. ISSN: 0091-6749. DOI: 10.1016/j.jaci.2009.09.046.

[7] N. Danilova, J. Bussmann, K. Jekosch, and L. A. Steiner. "The immunoglobulin heavy-chain locus in zebrafish: identification and expression of a previously unknown isotype, immunoglobulin Z". en. *Nature Immunology* 6.3 (Mar. 2005), pp. 295–302. ISSN: 1529-2908. DOI: 10.1038/ni1166.

[8] S. Magadán-Mompó, C. Sánchez-Espinel, and F. Gambón-Deza. "Immunoglobulin heavy chains in medaka (Oryzias latipes)". en. *BMC Evolutionary Biology* 11.1 (June 2011), p. 165. ISSN: 1471-2148. DOI: 10.1186/1471-2148-11-165.

[9] Y. Bao, T. Wang, Y. Guo, Z. Zhao, et al. "The immunoglobulin gene loci in the teleost Gasterosteus aculeatus". *Fish & Shellfish Immunology* 28.1 (Jan. 2010), pp. 40–48. ISSN: 1050-4648. DOI: 10.1016/j.fsi.2009.09.014.

[10] F. Gambón-Deza, C. Sánchez-Espinel, and S. Magadán-Mompó. "Presence of an unique IgT on the IGH locus in three-spined stickleback fish (Gasterosteus aculeatus) and the very recent generation of a repertoire of VH genes". *Developmental & Comparative Immunology* 34.2 (Feb. 2010), pp. 114–122. ISSN: 0145-305X. DOI: 10.1016/j.dci.2009.08.011.

[11] J. D. Hansen, E. D. Landis, and R. B. Phillips. "Discovery of a unique Ig heavy-chain isotype (IgT) in rainbow trout: Implications for a distinctive B cell developmental pathway in teleost fish". en. *Proceedings of the National Academy of Sciences* 102.19 (May 2005), pp. 6919–6924. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0500027102.

[12] R. Savan, A. Aman, K. Sato, R. Yamaguchi, et al. "Discovery of a new class of immunoglobulin heavy chain from fugu". en. *European Journal of Immunology* 35.11 (2005), pp. 3320–3331. ISSN: 1521-4141. DOI: 10.1002/eji.200535248.

[13] M. Yasuike, J. de Boer, K. R. von Schalburg, G. A. Cooper, et al. "Evolution of duplicated IgH loci in Atlantic salmon, Salmo salar". *BMC Genomics* 11 (Sept. 2010), p. 486. ISSN: 1471-2164. DOI: 10.1186/1471-2164-11-486.

[14] I. Harel, B. A. Benayoun, B. Machado, P. P. Singh, et al. "A Platform for Rapid Exploration of Aging and Diseases in a Naturally Short-Lived Vertebrate". *Cell* (2015). ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.01.038.

[15] D. R. Valenzano, B. A. Benayoun, P. P. Singh, E. Zhang, et al. "The African Turquoise Killifish Genome Provides Insights into Evolution and Genetic Architecture of Lifespan". English. *Cell* 163.6 (Dec. 2015), pp. 1539–1554. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2015.11.008.

[16] A. Cellerino, D. R. Valenzano, and M. Reichard. "From the bush to the bench: the annual Nothobranchius fishes as a new model system in biology". en. *Biological Reviews* 91.2 (May 2016), pp. 511–533. ISSN: 1469-185X. DOI: 10.1111/brv.12183.

[17] M. Schartl, R. B. Walter, Y. Shen, T. Garcia, et al. "The genome of the platyfish, *Xiphophorus maculatus*, provides insights into evolutionary adaptation and several complex traits". en. *Nature Genetics* 45.5 (May 2013), pp. 567–572. ISSN: 1546-1718. DOI: 10.1038/ng.2604.

[18] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, et al. "Basic local alignment search tool". *Journal of Molecular Biology* 215.3 (Oct. 1990), pp. 403–410. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(05)80360-2.

[19] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". en. *Nucleic Acids Research* 25.17 (Sept. 1997), pp. 3389–3402. ISSN: 0305-1048. DOI: 10.1093/nar/25.17.3389.

[20] K. Reichwald, A. Petzold, P. Koch, B. R. Downie, et al. "Insights into Sex Chromosome Evolution and Aging from the Genome of a Short-Lived Fish". English. *Cell* 163.6 (Dec. 2015), pp. 1527–1538. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2015.10.071.

[21] Z. Yuan, S. Liu, T. Zhou, C. Tian, et al. "Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments". *BMC Genomics* 19 (Feb. 2018). ISSN: 1471-2164. DOI: 10.1186/s12864-018-4516-1.

[22] E. Bengtén and M. Wilson. "Antibody Repertoires in Fish". en. In: *Pathogen-Host Interactions: Antigenic Variation v. Somatic Adaptations*. Results and Problems in Cell Differentiation. Springer, Cham, 2015, pp. 193–234. ISBN: 978-3-319-20818-3 978-3-319-20819-0. DOI: 10.1007/978-3-319-20819-0_9.

[23] S. Magadan, O. J. Sunyer, and P. Boudinot. "Unique Features of Fish Immune Repertoires: Particularities of Adaptive Immunity Within the Largest Group of Vertebrates". en. In: *Pathogen-Host Interactions: Antigenic Variation v. Somatic Adaptations*. Results and Problems in Cell Differentiation. Springer, Cham, 2015, pp. 235–264. ISBN: 978-3-319-20818-3 978-3-319-20819-0. DOI: 10.1007/978-3-319-20819-0_10.

[24] Y.-A. Zhang, I. Salinas, J. Li, D. Parra, et al. "IgT, a primitive immunoglobulin class specialized in mucosal immunity". en. *Nature Immunology* 11.9 (Sept. 2010), pp. 827–835. ISSN: 1529-2908. DOI: 10.1038/ni.1913.

[25] Z. Xu, D. Parra, D. Gómez, I. Salinas, et al. "Teleost skin, an ancient mucosal surface that elicits gut-like immune responses". en. *Proceedings of the National Academy of Sciences* 110.32 (Aug. 2013), pp. 13097–13102. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1304319110.

[26] P. Smith, D. Willemsen, M. Popkes, F. Metge, et al. "Regulation of life span by the gut microbiota in the short-lived African turquoise killifish". en. *eLife* 6 (Aug. 2017), e27014. ISSN: 2050-084X. DOI: 10.7554/eLife.27014.

[27] E. Bengtén, S. Quiniou, J. Hikima, G. Waldbieser, et al. “Structure of the catfish IGH locus: analysis of the region including the single functional <Emphasis Type="Italic">IGHM</Emphasis> gene”. en. *Immunogenetics* 58.10 (Oct. 2006), pp. 831–844. ISSN: 0093-7711, 1432-1211. DOI: 10.1007/s00251-006-0139-9.

[28] F. Ramirez-Gomez, W. Greene, K. Rego, J. D. Hansen, et al. “Discovery and Characterization of Secretory IgD in Rainbow Trout: Secretory IgD Is Produced through a Novel Splicing Mechanism”. en. *The Journal of Immunology* 188.3 (Feb. 2012), pp. 1341–1349. ISSN: 0022-1767, 1550-6606. DOI: 10.4049/jimmunol.1101938.

[29] S. B. Needleman and C. D. Wunsch. “A general method applicable to the search for similarities in the amino acid sequence of two proteins”. *Journal of Molecular Biology* 48.3 (Mar. 1970), pp. 443–453. ISSN: 0022-2836. DOI: 10.1016/0022-2836(70)90057-4.

[30] T. J. Wheeler and S. R. Eddy. “nhmmer: DNA homology search with profile HMMs”. en. *Bioinformatics* 29.19 (Oct. 2013), pp. 2487–2489. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt403.

[31] A. Löytynoja. “Phylogeny-aware alignment with PRANK”. en. In: *Multiple Sequence Alignment Methods*. Ed. by D. J. Russell. Methods in Molecular Biology. Totowa, NJ: Humana Press, 2014, pp. 155–170. ISBN: 978-1-62703-646-7. DOI: 10.1007/978-1-62703-646-7_10.

[32] F. Ehrenmann and M.-P. Lefranc. “IMGT/DomainGapAlign: IMGT Standardized Analysis of Amino Acid Sequences of Variable, Constant, and Groove Domains (IG, TR, MH, IgSF, MhSF)”. en. *Cold Spring Harbor Protocols* 2011.6 (June 2011), pdb.prot5636. ISSN: 1940-3402, 1559-6095. DOI: 10.1101/pdb.prot5636.

[33] P. Rice, I. Longden, and A. Bleasby. “EMBOSS: The European Molecular Biology Open Software Suite”. *Trends in Genetics* 16.6 (June 2000), pp. 276–277. ISSN: 0168-9525. DOI: 10.1016/S0168-9525(00)02024-2.

[34] F. S. Xiao, Y. P. Wang, W. Yan, M. X. Chang, et al. “Ig heavy chain genes and their locus in grass carp Ctenopharyngodon idella”. *Fish & Shellfish Immunology* 29.4 (Oct. 2010), pp. 594–599. ISSN: 1050-4648. DOI: 10.1016/j.fsi.2010.06.004.

[35] A. Stamatakis. “RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies”. en. *Bioinformatics* 30.9 (May 2014), pp. 1312–1313. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu033.

[36] A. Stamatakis, T. Ludwig, and H. Meier. “RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees”. en. *Bioinformatics* 21.4 (Feb. 2005), pp. 456–463. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti191.

[37] A. Stamatakis. “RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models”. en. *Bioinformatics* 22.21 (Nov. 2006), pp. 2688–2690. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl446.