博士論文

Studies on the immunoglobulin genes in torafugu

(トラフグ免疫グロブリン遺伝子に関する研究)

付希

フ　シ

2015

Studies on the immunoglobulin genes in torafugu

(トラフグの免疫グロブリン遺伝子に関する研究)

A Thesis in

Aquatic Bioscience

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

At

The University of Tokyo

Graduate School of Agricultural and Life Sciences

By

付希

フ シ

May 2015

Declaration

I, Fu Xi, hereby declare that the thesis titled “Studies on the immunoglobulin genes in torafugu” is an authentic record of the work done by me and it contains no material that has been presented for the awards of any other degree or diploma in any university.

May 15, 2015

Fu Xi

Laboratory of Aquatic Molecular Biology and Biotechnology

Graduate School of Agricultural and Life Sciences

The University of Tokyo

Acknowledgments

I first would like to thank my supervisor who guides me on the road to research, Professor Shuichi Asakawa at the University of Tokyo, for giving me the opportunity, and the encouragement, to apply what I knew about bioinformatics to what was an entirely new field for me. I am really grateful to him for the confidence he has shown in me and all the inspiration he has given to me.

I would like to thank all the other Professors: Prof. Shugo Watabe, Prof. Hideki Ushio, and Prof. Shigeharu Kinoshita for their unfailing support and guidance that helped me make through this study.

I would like to thank the China Scholarship Council (CSC) for supporting me financially over the last three years.

I want to express my appreciation for all the kind help offered by our laboratory members contributed to my education: Dr. Masako Nakaya, for being very cooperative and supportive in all matters; Dr. Engkong Tan, Dr. Yoji Igarashi, and FengJun Zhang for being helpful in data analysis and patient guidance provided to me; Dr. Lu Wang for her kind support in numerous ways.

A special thanks to my lovely parents for all the sacrifices that you've made on my behalf.

At the end I would like to express appreciation to my beloved husband who incented me to strive towards my goal and was always my support.

Contents

Abbreviations

AIS	**Adaptive immune system**
bp	**Base pair**
BCR	**B-cell receptor**
C region	**Constant region**
CDR	**Complimentarity-determining region**
cDNA	**Complementary deoxyribonucleic acid**
D segment	**Diversity segment**
DNA	**Deoxyribonucleic acid**
EST	**Expressed Sequence Tag**
FR	**Framework region**
H chain	**Heavy chain**
Ig	**Immunoglobulin**
IMGT	**International ImMunoGeneTics information system**
J segment	**Joining segment**
kb	**Kilo-base pair**
L chain	**Light chain**
mRNA	**Messenger ribonucleic acid**
NGS	**Next-generation sequencing**
NJ	**Neighbor joining**

ORF	**Open reading frame**
P	**Pseudogene**
PE	**Paired-end**
PCR	**Polymerase chain reaction**
PEAR	**Paired-End reAd mergeR**
RAG	**Recombination activating gene**
RSS	**Recombination signal sequences**
RT-PCR	**Reverse transcription polymerase chain reaction**
V region	**Variable region**

List of Tables

List of Figures

Abstract

Potent adaptive immune system (AIS) is fundamentally reliant on the generation of a diverse repertoire of B-cell antigen receptors (BCRs, or immunoglobulins (Igs)). The sequencing of genomes from almost every major class of vertebrate has greatly furthered the understanding of the diversity and evolutionary origins of Igs. Torafugu is a good model organism for comparative genome studies. Despite efforts made in understanding the nature of the torafugu immune system, the full picture of toragugu Ig genetic features such as the organization of Ig gene segments as well as the vast majority of Ig diversity are unknown. In this study, we aim to investigate the torafugu Ig heavy (H) and light (L) chain genes and to provide annotation maps of the two loci. The present study also took advantage of next-generation sequencing (NGS) capturing relevant Ig coding region sequences to characterize the Ig repertoire of torafugu.

The analysis of *IGH* genes (Chapter 2) revealed three new IGHV gene families (IGHV3–IGHV5). The interspersion of IGHV1 and IGHV2 family members and that they often intermingled with each other, while other IGHV family members were further interspersed was observed. Conservation of the promoter and recombination signal sequences (RSS) was observed in a family-specific manner. In addition to known variable region genes present on chromosome 5, we found 34 additional *IGHV* genes on scaffold 287 and 3 novel potentially functional *IGHD* genes on scaffold 483. In total, the variable region of the torafugu IGH locus consists of at least 48 *IGHV* genes, 7 *IGHD* genes, and 6 *IGHJ* genes. We confirmed the expression of newly identified IGHV3 family sequences in the spleen and kidney of adult torafugu and found a favorable IGHV segment usage by IgM and IgT. Possible structural variation in the IGHδ locus was observed based on the current torafugu assembly.

Genome builds of torafugu (assembly v4, October 2004 and assembly v5, January 2010) were searched to locate *IGL* genes (Chapter 3). This search resulted in the identification of 76 IGL gene segments to be localized as multiple clusters to three different chromosomes (chromosome 2, 3, and 5) and 38 different genomic scaffolds. Comparisons of the torafugu V segments revealed three distinct groups (designated IGLV1, IGV2, and IGLV3). The phylogenetic analysis of the IGLC sequences showed an unreported IGL isotype (type 3 (L3)) in torafugu. Collectively, the torafugu L2 locus contains 22 sequences matching IGLV segments, 8 IGLJ, and 11 IGLC segments scattered through 21 scaffolds. The type 1 (L1) *IGL* genes are located on at least 7 genomic scaffolds and they might operate as seven loci.

Based on the characterized IGH locus, we designed specific primers to isolate the relevant IgH sequences to profile the Ig repertoire of torafugu (Chapter 4). In the present study, we focused on analyzing the complimentarity-determining region (CDR3) of the H chain, because it is the most diverse component. We performed deep sequencing of amplicon library pooled from 3 torafugu using Illumina NGS platform (Miseq). The Paired-End reAd mergeR (PEAR) tool was used in the generation of consensus reads. Final quality filtering discarded low-quality consensus reads. This process generated ~1,000,000 high-quality consensus sequences (947,833-1,188,573 sequences) per sample. All the resultant consensus sequences were aligned first to germline V segment to find the optimal alignment and then aligned to all J segments to determine corresponding genomic V-J. As a result, we observed a preference for the IGHV1 family genes used by IgM, while IGHV2 family genes are associated with IgT, in most cases. The Jμ segments were also found be utilized in a preferential way by IgM. The length of CDR3 varies from 19 to 69 nt with a peak at 43 nt.

List of publication

Xi Fu, Hong Zhang, Engkong Tan, Shugo Watabe, Shuichi Asakawa. Characterization of the toraf ugu (*Takifugu rubripes*) immunoglobulin heavy chain gene locus. Immunogenetics, 67,179-193, 2015.

chapter 1: General Introduction

1.1 Background

1.1.1 General features of immunoglobulins

Adaptive immune system (AIS) is fascinating with its incredibly diverse ability to fight pathogens and has a memory. As the term suggests, the response is "adapted" to the particular pathogen and is thus specific. Potent AIS fundamentally rely on the generation of a diverse repertoire of B-cell antigen receptors (BCRs, or immunoglobulins (Igs)) (Alberts B, et al., 2002; Mutoloki, et al., 2014). A classic Ig molecule of the vertebrate AIS comprises two identical heavy (H) chains and two identical light (L) chains (Schroeder Jr and Cavacini, 2010). The H chain is composed of Ig superfamily (Igsf) domains, one N-terminal variable (V) domain, and two to six constant (C) domains, while the L chain always consists of one V and one C Igsf domain (Criscitiello and Flajnik, 2007). The L and H chains are covalently joined by a disulfide bond between the C_L and the CH1 domains, and the two V domains provided by each of the chains are associated non-covalently (Fig. 1-1). Generally, the paired V domains form the antigen-binding site to confer antigenic specificity, while the H chain C domains define different Ig isotypes (classes) (Schroeder Jr and Cavacini, 2010).

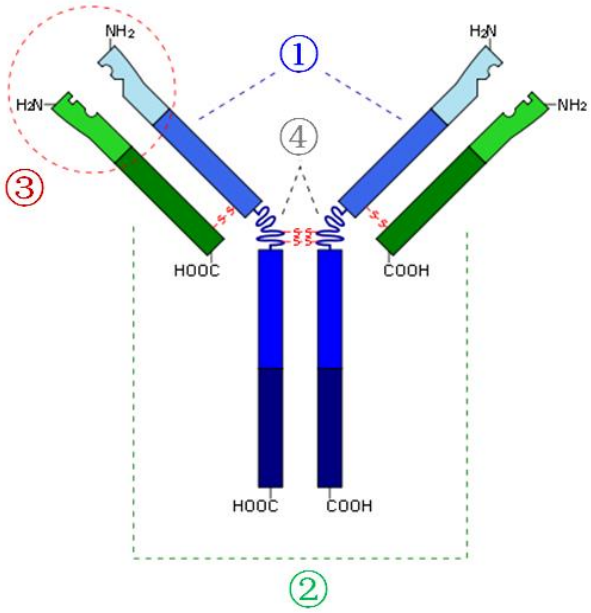


Fig. 1-1 Structure of a typical Ig molecule. ① Heavy chain (blue), ② Light chain (green), ③ Antigen binding site, ④ Hinge regions. Adapt from Wikipedia.

1.1.2 VDJ recombination

Functional *IGHV* genes are generated by random somatic rearrangement from a finite set of variable (V), diversity (D) and joining (J) germline gene segments [the L chain domain by V and J segments]. The rearrangement process, termed VDJ recombination, is initiated by the recombination activating gene 1 (*RAG1*) and *RAG 2* under the accessibility of proper RSS. Highly conserved RSS is composed of conserved a heptamer and a nonamer motif and is separated by a relatively non-conserved spacer of either 12 or 23 base pairs (bp) (Fig. 1-2), which defines the 12RSS or 23RSS.

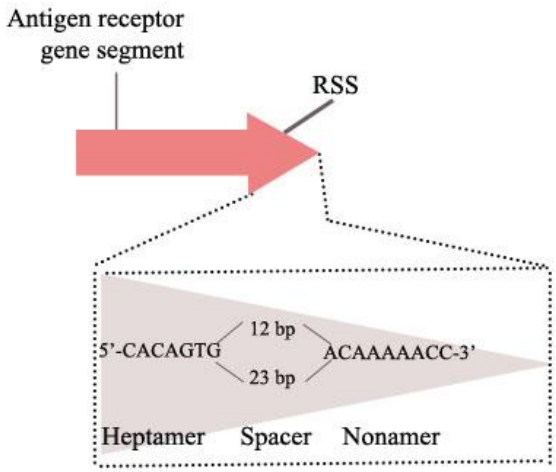


Fig. 1-2 The RSS lie immediately adjacent to each Ig gene segment and contain two well-conserved DNA elements (the heptamer and the nonamer) separated by a pacer region. Adapt from (Schatz and Ji, 2011).

There is a strong preference for recombination between a 12RSS and a 23RSS, a restriction known as the 12-23 rule (Mansilla-Soto and Cortes, 2003). The 12-23 rule enables VDJ recombination properly. For example, wasteful V-to-V or J-to-J rearrangements are prevented because all the V and J segments within a given locus are flanked by RSS of the same spacer length. Within the IGH locus, RSS are located at the 3′ end of each VH segment, 3′ and 5′ ends of each DH segment, and at the 3′ end of each JH segment (Fig. 1-3). Since direct VH-to-JH rearrangement is prohibited, the RSS structure ensures utilization of DH segment and, consequently, enhances IgH repertoire diversification. Thus, the 12/23 rule dictates the vast majority of IGH locus rearrangements involving VH gene segments that have a DH gene segment sandwiched between a VH and a JH (Mansilla-Soto and Cortes, 2003; Schatz and Ji,

2011).

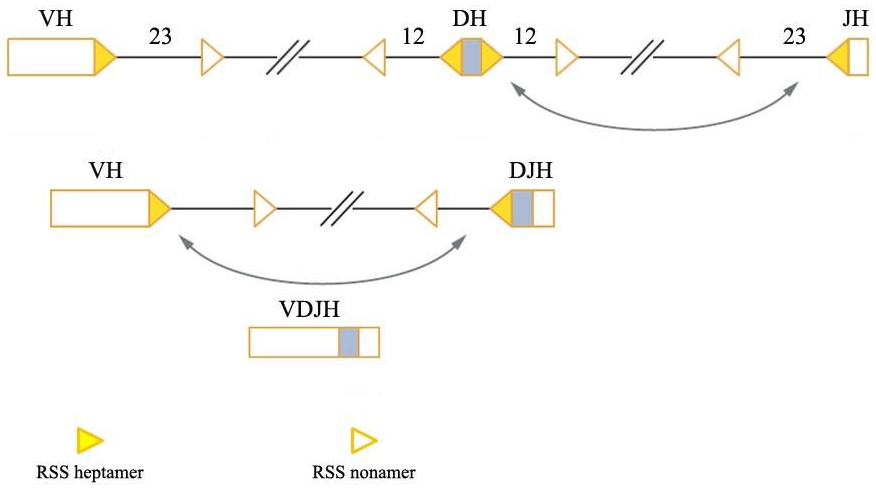


Fig. 1-3 The VDJ recombination adheres to 12/23 rule. RSS heptamer is depicted as yellow triangle, and RSS nonamer is depicted as white triangle. Spacer lengths are indicated above the different RSS. Adapt from (Jung, et al., 2006).

1.1.3 Generation of the Ig repertoire

During B cell development, recombination of H chain gene typically occurs before that of L chain. If both IgH and IgL genes are productively rearranged, the assembled antibody heterodimer is expressed on the surface of the B cell. In B cells containing productively rearranged Igs, the process of allelic exclusion (and locus exclusion for IgL) ensures expressing of a single Ig on each B cell. After several developmental checkpoints, newly generated mature $IgM^{+}IgD^{+}$ B cells form the naïve Ig repertoire (Georgiou, et al., 2014). In most cases, the diversity of the naïve Ig repertoire is concentrated in the three loops of the domain that bind to antigen, called the complementarity-determining regions (CDR). Among the three CDRs, CDR1 and CDR2 are encoded by the V gene itself, whereas CDR3 is encoded by

the VDJ or V-J ligation part and thus is the most diverse component (Ippolito, et al., 2006; Xu and Davis, 2000).

Proliferation of B cells will be induced if they encounter antigen wherein requisite co-stimulatory signals and T-cell help exist. This process, called B-cell clonal expansion, occurs primarily in germinal ceters, *i.e.* spleen. Subsequent somatic hypermutation of the V domains further increases affinity of B cells bearing somatic mutations for antigen compared to other B cells. As a result, B cells containing the highest affinity antibodies go through preferential expansion and survival. Taken together, diversity in the Ig repertoire comes from (1) the allelic diversity in Ig gene segments, (2) combinatorial diversity from somatic recombination, (3) junctional diversity caused by the recombination process, wherein the existing V gene segment is replaced with another.

1.1.4 Immunoglobulins in teleost fish

Teleost fish are the most primitive vertebrates that contain Igs. Throughout evolutionary time, Igs diversified into different isotypes with specialized roles in the mucosal and systemic compartments (Flajnik and Kasahara, 2010). Immunoglobulins can be divided into different isotypes and sybtypes based on the nature of the C domains of their H chains (Schroeder Jr and Cavacini, 2010). In mammals, five Ig isotypes - IgM, IgD, IgG, IgA, and IgE have been identified. Among these Igs, IgM is the most ancient and the only isotype functionally conserved in all gnathostomes (jawed vertebrates). IgD has been found in all gnathostomes, except birds. Until recently, teleost fish were thought to contain only two isotypes of Ig, IgM and IgD. It was also generally deemed that IgM was the only Ig isotype responding to antigenic challenges both in systemic and mucosal compartments, and teleost fish were

thus thought to be devoid of an Ig specialized in mucosal surfaces. Recent studies overturned this view as IgT/IgZ has been discovered in rainbow trout (*Oncorhynchus mykiss*) (IgT) (Hansen, et al., 2005) and zebrafish (*Danio rerio*) (IgZ) (Danilova, et al., 2005). This fish-specific Ig expressed in several teleost fish species with the exception of channel catfish (*Ictalurus punctatus*). To avoid a mixed terminology, throughout the thesis we will use the term 'IgT' to refer to IgT/IgZ, and we will use 'τ' for the gene encoding its H chain.

During the last ten years, IgT has been characterized at the gene level in many teleosts species, such as torafugu (*Takifugu rubripes*) (Savan, et al., 2005), three-spined sticklebacks (*Gasterosteus aculeatus*) (Bao, et al., 2010; Gambon-Deza, et al., 2010), and grass carp (*C. idella*) (Savan, et al., 2005). Further research demonstrated unique aspects of IgT that specializes in mucosal immunity and that it has anti-pathogenic function only in the gut, similar to IgA in warm-blooded animals and IgX in amphibians (Zhang, et al., 2010). Thus, these findings have challenged the paradigm that specialization of Ig isotypes into mucosal and systemic areas arose during tetrapod evolution.

1.2 Objectives of the study

Torafugu are a good model system for studying the AIS because in evolutionary terms they possess the earliest recognizable AIS. Despite efforts made in understanding the nature of the torafugu immune system, the full picture of the genetic features of toragugu Ig such as the organization of Ig gene segments as well as the vast majority of Ig diversity are unknown. Here, studies were carried out to understand the following concerns:

- To elucidate the full structure of the gene locus on both Ig H chain and L chain of the torafugu, to provide insights into the genetic basis of diversity in the Ig response of this species
- To understand what fraction of the potential IgH repertoire is expressed in an individual through analyses of VDJ segment use and the size distribution of CDR3 region.

1.3 Outline of the thesis

The thesis is composed of genetal introduction (Chapter 1), three research sections (Chapter 2, 3, and 4), and general discussion (Chapter 5).

Chapter 1 reviewed the general features of Igs and mechanisms that ensure the generation of productive Igs as well as the vast diverse repertoire of Igs. Studies concerning the fish Igs were also introduced in this part.

Chapter 2 describes the investigation of the torafugu Ig H chain gene locus wherein an annotated map of the IGH locus and the rearrangements of *IGH* genes were provided.

Chapter 3 provides the analysis of the torafugu Ig L chain gene locus.

Chapter 4 gives the profiling of the expressed IgH repertoire in torafugu by using next generation sequencing (NGS) technology.

Chapter 5 provides conclusions based on the findings in the present study and aims achieved with the prospects for future directions.

chapter 2: Organization of the torafugu immunoglobulin heavy chain gene locus

The content of this chapter was published in:

Xi Fu, Hong Zhang, Engkong Tan, Shugo Watabe, Shuichi Asakawa. Characterization of the torafugu (*Takifugu rubripes*) immunoglobulin heavy chain gene locus. Immunogenetics, 67,179-193, 2015.

2.1 Introduction

The IGH locus of teleost fish was initially thought to be a simplified version of the translocon organization in mammals (Bengtén, et al., 2000; Warr, 1997). However, as knowledge of IgH genes in teleost fish has accumulated, this picture has become more complicated. The blueprint of the overall organization of the IGH locus in teleost fish has been better understood since the discovery of the IgT isotype in zebrafish and rainbow trout.

The IGH translocon configuration in mammals consists of VH, DH, JH, and CH regions although the organizations varied among mammalian species as follows:

Human, VH-DH-JH-Cμ-Cδ-Cγ3-Cγ1-ψε-Cα1Cγ2-Cγ4-Cε-Cα2;

Mouse, VH-DH-JH-Cμ-Cδ-Cγ3-Cγ1-Cγ2b-Cγ2a-Cε-Cα;

Cattle, VH-DH-JH-Cμ-Cδ-Cγ3-Cγ1-Cγ2-Cε-Cα (Stavnezer and Amemiya, 2004).

The archetypal structure of the IGH loci in teleost fish follows a pattern of translocon organization with a region containing VH genes in 5′, followed by units comprising several DH, JH, and then CH region genes in 3' : [VH-DH-JH-Cτ-(VH)-DH-JH-Cμ-Cδ] (Flajnik and Kasahara, 2010). There are differences in Cτ gene locations between two teleost groups: (1) In rainbow trout, the Cτ gene is located within the VH region (Hansen, et al., 2005), (2) the Dτ-Jτ-Cτ cluster(s) encoding IgT specific genes are located between the region of *IGHV* genes and the locus encoding IgM/D. This structure is found for example in the zebrafish, three-spined stickleback, and torafugu (Danilova, et al., 2005; Gambon-Deza, et al., 2010; Savan, et al., 2005). Such genomic structure has allowed the prediction of the existence of two mutually exclusive B cell lineages expressing either IgT or IgM since the recombination of *IGHV* to *Dμ* deletes the Dτ-Jτ-Cτ region (s).

Although the zebrafish possesses only a single copy IGH locus, there is more than one copy of this locus in most teleost fish, and in some cases isoloci can be found on different chromosomes, *i.e.* Atlantic Salmon and rainbow trout possess two IgH loci (IgHA and IgHB) due to the tetraploidization of Salmonidae (Yasuike, et al., 2010), and three loci of IGHM-IGHD in catfish (including pseudo-Cμ genes) (Bengtén, et al., 2006; Bengtén, et al., 2002). These H chain gene clusters probably arose from a massive local duplication event. Thus, it seems that the Ig gene configuration in teleosts has acquired through its evolution a modified type of translocon organization.

Torafugu is a good model organism for comparative genome studies (S.Brenner, et al., 1993). With the development of powerful molecular biology tools, our knowledge of the torafugu IGH locus has accumulated (Aparicio, et al., 2002; Peixoto and Brenner, 2000; Saha, et al., 2004; Saha, et al., 2005; Savan, et al., 2005). In summary, these studies showed that the torafugu IGH locus exhibits its germline organization on chromosome 5 in a 130-kilobase (kb) region, similar to that of the rainbow trout (125 kb), but more compact than other teleosts described. To date, the exact number and location of *IGHV* genes in torafugu are not known since they only have been partially characterized by shotgun sequencing of one cosmid clone. In addition, torafugu *IGHC* genes were characterized based on cDNA-derived sequences (Saha et al. 2004; Saha et al. 2005; Savan et al. 2005); thus, sequence data for the CH locus is incomplete.

In the present study, we set out to define the structure of the IGH locus in torafugu. A search in the torafugu Expressed Sequence Tag (EST) database (http://www.fugu-sg.org/) generated only one EST for IGHV1 family sequences and six ESTs for IGHV2 family sequences, highlighting the lack of information on the torafugu IGHV locus. Recently, our laboratory reported additional genome assemblies generated by NGS technology (Illumina GA IIx platform) (Zhang, et al., 2013; Zhang, et al.,

2014), which were beneficial to our research on the characterization of torafugu IGH gene locus. Here, we investigated the torafugu IGH locus from available sources, including the Fifth Fugu Genome Assembly (v5 assembly) (Kai, et al., 2011) (http://www.fugu-sg.org/) and assembly of sequences generated by our laboratory. Our studies resulted in an annotated map of the torafugu IGH locus and showed the expansion of *IGHV* genes and *IGHD* genes than demonstrated in previous studies. Additionally, we confirmed the expression of newly identified IGHV3 family sequences in adult torafugu and found a favorable usage of IGHV segment by IgM and IgT.

2.2 Materials and methods

2.2.1 Identification and sequencing of IGH genes

A complete gene search was conducted to identify all the *IGH* genes in the fifth genome assembly of torafugu (assembly v5, January 2010) and the sequences generated by our laboratory. We performed TBlASTN (cutoff *E*-value of 10^{-15}) searches with published torafugu cDNA sequences of μ, δ, and τ against available genome databases. The query data set aligned to the same genomic region due to their similarity to one another. Thus, we only extracted non-overlapping genomic sequences that generated alignments with the lowest *E*-values. Sequences of chromosome 5 in the current assembly and three scaffolds (scaffold 287, 483, and 1358) from our database were identified that span the IGH locus. All identified sequences were retrieved and analyzed in detail. All scaffolds were found to harbor the VH domain, while sequences in scaffold 483 were identified to mainly span the CH domain. We did not find any Cμ sequence on chromosome 5, probably due to gaps. In order to determine chromosome 5 genomic regions that correspond to each scaffold, alignments were performed using the LAST program

(Kielbasa, et al., 2011) (http://last.cbrc.jp/) and confirmed by manual inspection of overlaps. Some of the gaps on chromosome 5 were closed by the identified scaffolds. The remaining gaps in the IGH locus were solved by chain-termination sequencing using primers (Table 2-1) specific for flanking region of each gap for polymerase chain reaction (PCR) amplification.

Table 2-1 Primers for PCR study for gap closing

Primer name	Primer sequence (5′---------3′)	Length (bp)
IGHV1S7_f	CGCAGTTCTGCTCTTCACAG	20
IGHV1S7_r	GAGCAGAGAGGAGGACAGAG	20
IGHV2S5p_f	TTAATCCTGGTGTGACAGCGC	21
IGHV2S5p_r	TGGAGTGGATTGGATGGCAT	20
IGHV1S8_f	CAGTTTTCAGTGTGCAGTTCTG	22
IGHV1S8_r	TCCCTGGGCATGTTCTAAGG	20
IGHV2S7_f	CACCAATCCACTCCAGTCCT	20
IGHV2S7_r	TGTGGACAGAATAGTTGGTGC	21
IGHV2S8_f	CTGCCTCAATGTTGTTGTCA	20
IGHV2S8_r	TCTGACCATCACCTGCCAG	19
IGHV1S10p_f	TGACCCAGCTCATCCAGTAG	20
IGHV1S10p_r	GCACACACCAGAACCCTTTT	20
IGHV2S9_f	TGAGAGAATAGGAGACCTGAC	21
IGHV2S9_r	TTCCCACTGCGTTTTGTCTG	20

IGHV1S11p_f	GTCTGTGTTCACTTCTGCTCT	21
IGHV1S11p_r	AGCAGCTACTACATGGCCTG	20
IGHV2S10_f	TGGTGATGATGTGTTGGTGC	20
IGHV2S10_r	ACCATCAGACTGTCAGTGGAG	21
IGHV1S12_f	ACCACTGATCCACTCCAGTC	20
IGHV1S12_r	ACGCACATCATCCAAAGACG	20
IGHV1S13orf_f	CACAGACCAGAATGAGACATCC	22
IGHV1S13orf_r	TCATTGGCTCACATCCTCC	19
IGHV2S13p_f	GGCTGCACATTCACTCCATT	20
IGHV2S13p_r	TGGTGGTCAGAACAGTTGGT	20
IGHV2S14p_f	GTCCAATCCACTCCAGTCCT	20
IGHV2S14p_r	ACAGTGGCAGGAAAACCTTC	20
IGHV1S15_f	GTGTTCAGGCTGCAGTTCTGC	21
IGHV1S15_r	AGAGGACAGAGAAGAGGGGAC	21
IGHV2S15_f	TGTTGATGCTCTGAGGGGTT	20
IGHV2S15_r	GGGAAAGGACTGGAGTGGAT	20
IGHV1S16_f	CCCAGTACATCCAGTCGTCA	20
IGHV1S16_r	CTGCTGAGTGTGAAGTGTGG	20
IGHV2S16_f	CCTGTCCAATCCACTCCAGT	20
IGHV2S16_r	GGCTCTATTTGATCAACAGTG	21
IGHV2S17_f	CGTAGAGGGAACAGTCAGGG	20

IGHV2S17_r	AGCAGAATAAGAGAAGCTCCCAT	23
IGHV1S17_f	TGAACGTGAATCCAGAGGCT	20
IGHV1S17_r	ACTGAGGAGGAGTTGATGCA	20
IGHV1S18_f	GTCAGTATTCACTTCTGCTC	20
IGHV1S18_r	ACACACAGTTGATCATGGTGG	21
IGHV2S19_f	GGTGATGTTGTGTTGGTGCT	20

2.2.2 Mapping and annotation of the IGH locus

CH exons were determined by conventional BLASTX approach (National Center for Biotechnology Information (NCBI)). Sequences of Igs that have not yet been published were deduced by FGENESH (http://linux1.softberry.com/). In order to confirm the exon ends, predicted messenger RNA (mRNA) from identified gene sequences was compared with the torafugu EST database.

IGHV genes were searched in both chromosome 5 and the three scaffolds that contain IGHV gene segments. Each matching *IGHV* gene sequence was manually analyzed as follows: (a) gene structure prediction was performed with FGENESH, refined by homology-based prediction with FGENESH+ and splice site prediction (SPL), (b) exon ends were determined by the presence of functional RSS and the canonical "tattattgt" nonamer sequences (allowing 2 nucleotide substitutions) that are located 6–9 nucleotides upstream of RSS (Jung, et al., 2006), and (c) CDR and FR delimitations were determined according to IMGT/V-QUEST alignment software (Teleostei unit) (Brochet, et al., 2008).

Torafugu IGHD and IGHJ gene sequences were obtained from Savan et al. (Savan, et al., 2005). Additional previously unreported *IGHD* genes were identified by searching for pairs of RSS

(CACAGTG-N11-13-ACAAAAACC, allowing up to 4 substitutions) spaced by no more than 40 nucleotides (Jung, et al., 2006). During the identification of *IGHJ* genes, regions between the 3′ of *IGHD* and the 5′ of *IGHC* genes were manually analyzed, taking into account the RSS at the 5′ end of the *IGHJ* gene.

Annotation of the complete IGH locus, including the transcription factor binding site at each *IGHV* gene flanking region, start and end positions of the leader sequence, the VH exon, and RSS location, was performed manually with the annotation software Artemis (Carver, et al., 2008). To confirm the expression of identified genomic sequences, we performed TBLASTX searches against the NCBI EST database.

2.2.3 Nomenclature

Identified genes were named according to IMGT®, the international ImMunoGeneTics information system® at http://www.imgt.org (Lefranc, 2007). *IGHV* genes were grouped into families using a 80% identity threshold as originally defined in the mouse and human IGH loci (Brodeur and Riblet, 1984). For the *IGHV* genes, any retrieved sequence that had an intact exon-intron structure, a proper RSS, had no frameshift mutation and/or premature stop codons in the leader exon and the V-exon, and possessed key amino acids (1st-Cys 23, conserved-Trp 41, and 2nd-Cys 104) in FR1, FR2, and FR3 regions, respectively, was deemed a functional V gene. All other sequences were regarded as pseudogenes.

2.2.4 Phylogenetic studies

The phylogenetic tree was constructed using the program MEGA5 (Tamura, et al., 2011) from multiple sequence alignments generated by the ClustalW software tool at the European Bioinformatics Institute (http://www.ebi.ac.uk/Tools/msa/).

2.2.5 Cloning and sequencing of VH cDNAs

Three adult torafugu weighing between 800 and 900g were obtained from Fish Interior (Tokyo, Japan). Fish were maintained in tanks with aerated seawater and maintained at 20°C. Adult fish were euthanized, followed by rapid dissection of tissues. Spleen and trunk kidney were collected and directly fixed in RNAlater (Applied Biosystems, Warrington, UK). Total RNA was extracted from spleen and trunk kidney using TRIzol reagent (Invitrogen, Carlsbad, CA). Purified total RNA (1.0μg) was reverse transcribed with SuperScript® III (Invitrogen, Carlsbad, CA) using oligo (dT)20 primer in 20μl reactions. RNase H (Invitrogen, Carlsbad, CA) was added to each reaction to remove RNA at the end of the cDNA synthesis step. All enzyme concentrations, reaction volumes and incubation temperature were according to the manufacturer's protocol. Equal amounts of each cDNA were combined and the mixture used as PCR template.

The torafugu IGH locus was described in this study. The consensus leader sequences for 32 functional IGHV gene segments of the torafugu were used to design the 6 forward primers (as part of a family). The reverse primers were derived from the first exon of Cμ and the second exon of Cμ and Cτ (Table 2-2). PCRs were performed using Platinum® Taq DNA Polymerase (Invitrogen, Carlsbad, CA) with an initial denaturation of 2 minutes at 94°C, followed by 30 cycles of denaturation at 94°C for 30 s, annealing of

primer to DNA at 55°C for 30 s, and extension at 72°C for 1 minute. PCR products were cloned using TOPO® TA Cloning Kit for sequencing (Invitrogen, Carlsbad, CA). Twelve (12) clones from each positive PCR product were picked and plasmids were purified and sequenced. The cDNAs sequences were BLAST against the IGHV and IGHJ gene sequences to identify their presence in the clones.

Table 2-2 Primers used in expression study

Primer name	Primer sequence (5′--------3′)	Length (bp)
FVH1-1	GGACAGGACTGCTGCTTCTAAC	22
FVH1-2	CTTCTAACTGTCTGCTGGGCA	21
FVH2-1	AGCTCTGCTGCTGCTGTTG	19
FVH2-2	TTCTCTGCAGCTGTGGTGC	19
FVH3-1	CAGAGGTTTACTGATCATTGTC	22
FVH3-2	TCTTCAGTGTCTGGTGGACG	20
Cτ2	GTGATCAGACACACAAGAGTGACG	24
Cμ2	TCTCAGATATTTTGGAGGTCACC	23
Cμ1	AGGGCTACCGTCCCAGTCCTGT	22

2.3 Results

2.3.1 Identification of three novel IGHV gene families

The *IGHV* genes were assigned to families as originally suggested by Brodeur and Riblet (Brodeur and

Riblet, 1984). In this study, we classified members of a torafugu IGHV gene family as having at least 80% identity at the nucleotide level over the coding exon sequence (excluding the leader exon). In addition to the two previously identified IGHV families (IGHV1 and IGHV2) (Peixoto and Brenner, 2000), we identified 3 novel families (IGHV3, IGHV4, and IGHV5), of which IGHV4 and IGHV5 consist of only *IGHV* pseudogenes. Each *IGHV* gene was numbered based on the location of the most 5′ family member along the locus.

There is a novel gene, which does not meet the criteria for inclusion in any of the IGHV families. This gene contains an identifiable leader sequence, a VH region exon, a functional RSS, and conserved splice junctions, but does not have >80% identity to any other known *IGHV* gene. Its closest homology is to *Anamin-IGHV3S1* (the *IGHV3S1* gene of spotted wolfish based on the IMGT database) (77.43% identity at the nucleotide level when only using the exon sequence). We have named this novel gene *IGHV3S2* in accordance with the existing nomenclature. Two other IGHV homologs that showed 97% and 99% of nucleotide identity with the novel gene, respectively, were classified into the IGHV3 family (*IGHV3S1p* and *IGHV3S3orf*).

Pseudogenes were grouped to specific IGHV family if they had over 80% nucleotide homology to functional genes from that family. Although some pseudogenes were clearly recognized as *IGHV* gene remnants, they could not be assigned to any IGHV family due to their low homology (<80%) to grouped IGHV gene sequences. We assigned these segments to 2 different IGHV families based on nucleotide identity, with one family containing 2 members (*IGHV4S1p* and *IGHV4S2p*) and another family with only one member (*IGHV5S1p*). Furthermore, the leader sequences of the novel IGHV gene families are conserved within each family and are not identical to those in the reported families (IGHV1 and IGHV2).

It is worthy to note that two potentially functional *IGHV* genes, *IGHV3S2* and *IGHV3S3orf*, which have

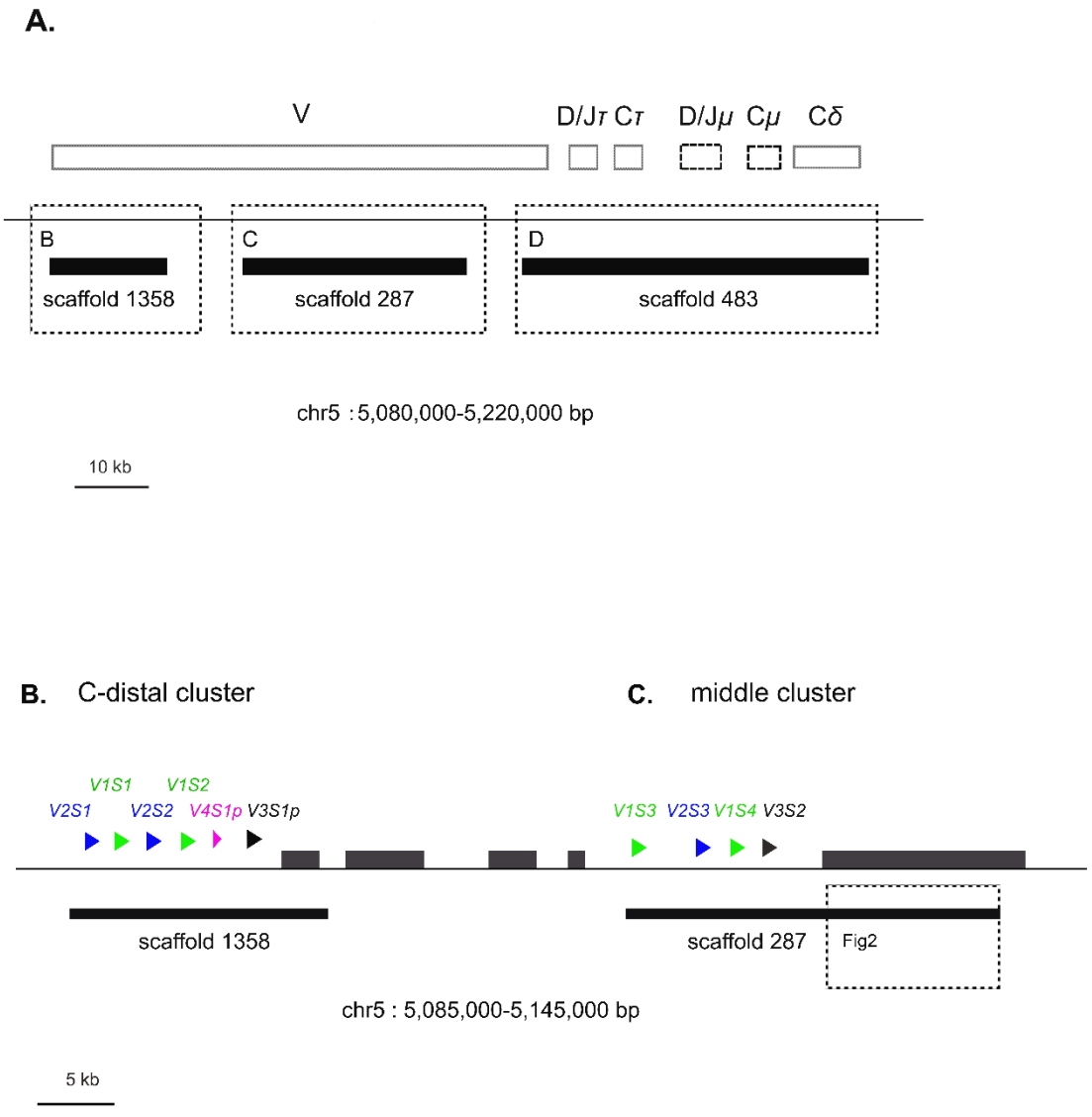
not been reported in the torafugu EST database, have been identified.

2.3.2 Genomic organization of the torafugu IGHV locus

The torafugu IGHV locus (referring to the *IGHV*-gene-containing region of the IGH locus in this study) spans approximately 67 kb on chromosome 5 and is organized in three clusters containing 14 IGHV gene segments. These segments, orientated in the same direction, are separated by 480 bp to 22 kb. The 5′ portion of the IGHV locus is more compact compared to the 3′ portion, with an average intergenic distance (between *IGHV* genes and/or pseudogenes) less than 2 kb, whereas an approximately 5 kb distance appears at the 3′ end. Members of IGHV1 and IGHV2 families are completely interspersed and they often mix with each other. In contrast to the 2 families, members of newly identified families (IGHV3–IGHV5) are more spatially separated over the locus. The number of IGHV sequences identified per family varies greatly, from only one member in the IGHV5 family to 21 members in both the IGHV1 and IGHV2 families.

In order to establish a detailed chromosomal map of the IGHV locus, including the exact location and structure of individual genes, we analyzed chromosome 5 sequences and the three scaffolds containing *IGHV* genes based on the characteristic features of *IGHV* genes (Fig. 2-1a). The C-distal cluster (farther to the constant regions compared to other clusters), approximately 50 kb upstream, contains 6 *IGHV* genes, of which two are pseudogenes (Fig. 2-1b). Two of the pseudogenes have premature stop codons (*IGHV2S21p* and *IGHV5S1p*), one is frameshifted due to a nucleotide deletion within FR2 (*IGHV4S1p*). The IGHV3S1p segment, which loses the RSS by replacement of an unknown sequence, was not identified in the torafugu EST database and, therefore, is classified as a pseudogene. Four *IGHV*

genes (*IGHV1S3*, *IGHV2S3*, *IGHV1S4*, and *IGHV3S2*) were commonly found in both chromosome 5 and scaffold 287. They formed the middle cluster (Fig. 2-1c) and were all defined as functional *IGHV* genes. The C-proximal (closer to the constant regions compared to other clusters) cluster contains four *IGHV* genes, of which two are pseudogenes (Fig. 2-1d).



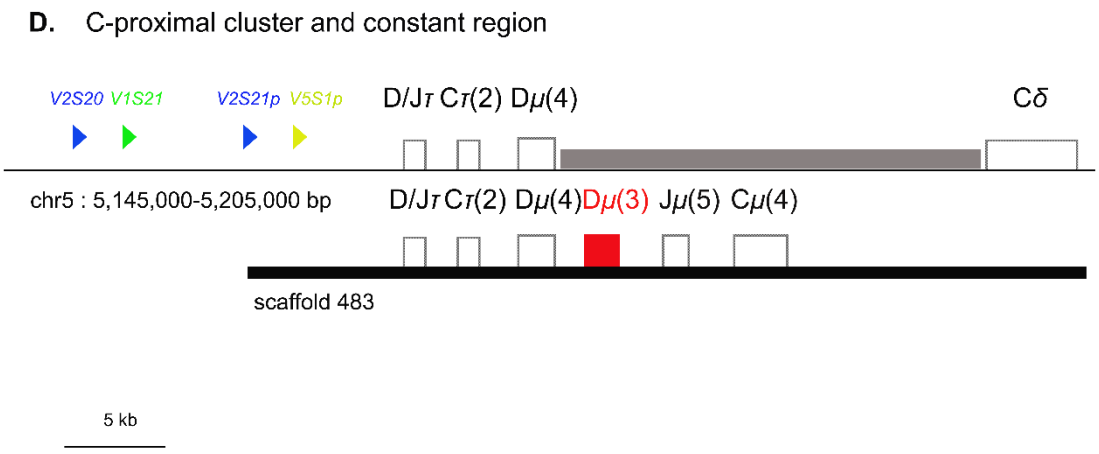


Fig. 2-1 Torafugu IGH locus

a. Alignment of scaffold1358, scaffold287, scaffold483, and chromosome 5 containing IgH gene segments. Dotted open symbols D/Jμ and Cμ represent their absence on chromosome 5. **b.** Organization of the C-distal cluster on chromosome 5, to scale. The *IGHV* genes are depicted as triangles pointing in the direction of their orientation and numbered based on the location of their most 5′ end family member. Members of the same IGHV family share same colors. Pseudogenes are indicated with a "*p*" after the family number. Gaps in the present assembly are marked by solid gray bars. **c.** Scale map of *IGHV* genes in the middle cluster on chromosome 5. Scaffold 287 extends into the gap region downstream of the *IGHV*-gene-containing region to potentially reveal more *IGHV* genes. Additional *IGHV* genes revealed by scaffold 287 are depicted in Figure 2-2. **d.** The C-proximal cluster of *IGHV* genes and adjacent IgT-encoding locus as well as *IGHC* genes located immediately downstream. Scaffold 483 reveals 3 novel *Dμ* gene segments as shown in red filled symbols. Sequences of scaffold 483 and the complete sequence information for the *μ* gene, including J*μ* and C*μ* exons, are missing in the present assembly.

Potential additional *IGHV* genes were found in scaffold 287 (Fig. 2-2), of which twelve are pseudogenes and two genes were classified as open reading frames (ORFs) (*IGHV1S13orf* and *IGHV3S3orf*). One of the ORFs (*IGHV1S13orf*), which possesses all the features of a functional *IGHV* gene except for the absence of RSS, is found in cDNA and EST libraries of torafugu, and is, therefore, categorized as a coding gene, even though it may not be capable of rearrangement or transcription. Another coding sequence, *IGHV3S3orf*, loses the octamer motif due to a 5′ truncation and may not be efficiently expressed, but it seems otherwise intact and is designated as an ORF gene. Other sequences were found only as fragments and could not be amplified by PCR when using specific primer(s).

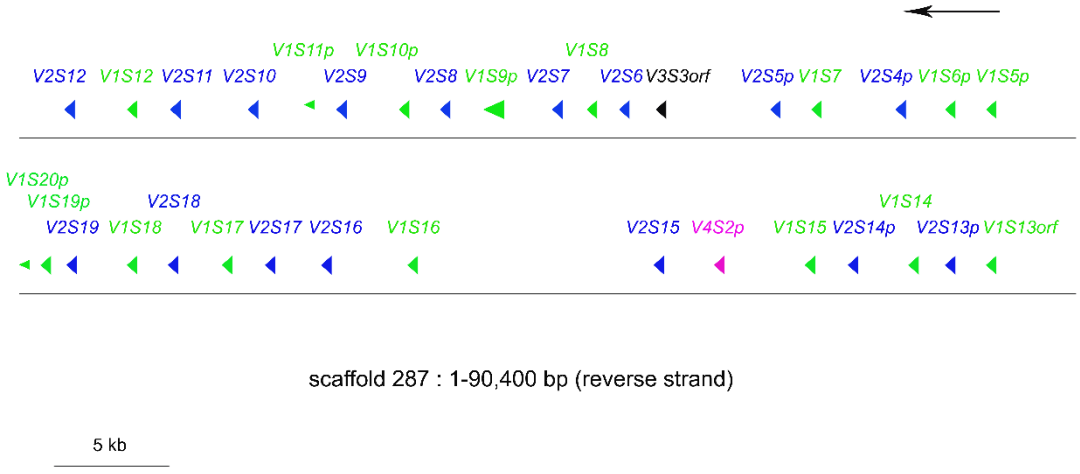


Fig. 2-2 Organization of the 34 additional *IGHV* genes present on scaffold 287

The *IGHV* genes are all in the direction of their reverse strand orientation. ORF genes are indicated with an "*orf*" after the family number.

2.3.3 Gap closing of the IGHV locus

The current torafugu genome assembly (the v5 assembly) has many gaps in the IGHV locus, partially due to the fact that this genomic region is comprised of many similar genes and pseudogenes, which are difficult to identify by shotgun sequencing. There are more than 10 large and small gaps encompassing approximately 22 kb of sequence in the IGHV locus on chromosome 5. After aligning with our scaffold data, an approximately 7 kb-gap-containing region has not yet been determined. This region is located between the C-distal cluster and the middle cluster as shown in Figure 2-1b and c. We did not identify any *IGHV* gene in this region and PCR amplification failed. However, the possibility of additional *IGHV* genes within this region cannot be ruled out considering the contiguous distribution of *IGHV* genes in each cluster.

The three scaffolds (scaffold 287 of 112,361 bp, scaffold 483 of 404,812 bp, and scaffold 1358 of 16,974 bp) were identified that span the IGH locus, including gap regions in the v5 assembly. We searched for *IGHV* genes among the three scaffolds to explore the presence of unidentified *IGHV* genes. This search resulted in the identification of 46 sequences as matching IGHV gene sequences (38 sequences in scaffold 287, 6 sequences in scaffold 1358, and 2 sequences in scaffold 483). As depicted in Figure 2-1a, the three scaffolds (on the reverse strand orientation) were all assigned to chromosome 5 in the IGH locus. Scaffold 1358 was aligned to chromosome 5 from the most upstream *IGHV* gene of the C-distal cluster (Fig. 2-1b). Scaffold 483 could map on the two 3′ *IGHV* genes of C-proximal cluster and further spans downstream of the constant region encoding locus. Apart from *IGHV* genes identified on chromosome 5, these two scaffolds did not reveal any additional *IGHV* gene, although they partially overlap the gap region of chromosome 5. On the other hand, alignment of scaffold 287 and chromosome

5 starts from the 5′ of the middle cluster (spans approximately a 10 kb region). There is a large gap (12 kb) on chromosome 5 immediately downstream of the middle cluster (Fig. 2-1c). By analyzing contiguous, overlapping sequences represented by scaffold 287 and chromosome 5, it might well be that this large gap can be partially filled by scaffold 287. In fact, scaffold 287 sequences cover from the 5' most *IGHV* gene of the middle cluster to part of the large gap region on chromosome 5 and could potentially reveal more *IGHV* genes.

2.3.4 Analysis of the RSS sequence

The RSS sequence is composed of conserved heptamer (CACAGTG) and nonamer (ACAAAAACC) sequences, separated by the 23 bp spacer nucleotides. It has been demonstrated that the first three nucleotides of the heptamer and the fifth and sixth positions of the nonamer are essential for efficient V-D-J recombination (Ramsden et al. 1996; Cuomo et al. 1996). Among the 48 IGHV gene segments characterized in this study, 44 have RSS located immediately downstream of the coding region sequence (Table 2-3). We observed that the conservation of RSS is family-specific and that they all maintain the complete 23 bp spacer nucleotides. The IGHV1 segments, except for *IGHV1S9p*, *IGHV1S20p*, and *IGHV1S13orf* that lose the RSS, maintain the highly conserved heptamer sequence and show some slight difference from the consensus in the seventh nucleotide (A) of the RSS nonamer, except for two segments. The *IGHV1S15* and *IGHV1S12* contain more mutated RSS nonamer (AACAAAAAC and TCAAAAACA). However, they both carry the five critical nucleotides and, therefore, may not affect the V-D-J recombination. The RSS features among IGHV3 segments are similar to that in the IGHV1 family,

where the consensus of the heptamer and one nucleotide difference of the nonamer at the first position (G) were observed. The fifth and seventh positions of the heptamer of most IGHV2 segments have diverged (ATA in the last three nucleotides), while their nonamer remains well conserved. In contrast, the RSS of the IGHV4 and IGHV5 segments have much more divergent nonamers, especially with a mutation in a critical site (the sixth position).

Table 2-3 Genomic features of the torafugu *IGHV* genes

IGHV	**Promoter**						**Gene**				**RSS**			
gene	Octamer	(nt)	TATA	(nt)	ATG	gt/ag	Leader (nt)	Intron (nt)	V-exon (nt)	Locus Position	7mer	spacer	9mer	
Functional *IGHV* genes on chromosome 5														
IGHV2S1	ATGCAAAT	25	TATAAA	55	+	+	40	89	293	5089399-5089820	CACAATA	23	ACAAAAACC	
IGHV1S1	ATGCAAAA	19	TACTTA	50	+	+	49	85	305	5090743-5091181	CACAGTG	23	ACAAAAACA	
IGHV2S2	ATGCAAAG	20	TATAAA	70	+	+	43	93	296	5092000-5092431	CACAATA	23	ACAAAAACC	
IGHV1S2	ATGCAAAA	19	TACTTA	50	+	+	49	85	305	5093382-5093820	CACAGTG	23	ACAAAAACA	
IGHV1S3	ATGCAAAA	19	TACTTA	50	+	+	49	85	305	5119668-5120106	CACAGTG	23	ACAAAAACA	
IGHV2S3	ATGCAAAT	25	TATAAA	55	+	+	40	89	293	5124533-5124954	CACAATA	23	ACAAAAACC	
IGHV1S4	ATGCAAAA	19	TACTTA	50	+	+	46	80	305	5125891-5126321	CACAGTG	23	ACAAAAACA	
IGHV3S2	ATGCAAAC	11	TATAAA	86	+	+	42	95	341	5127159-5127636	CACAGTG	23	GCAAAAACC	

IGHV2S20	ATGCAAAT	25	TATAAA	55	+	+	40	89	293	5148361-5148782	CACAATA	23	ACAAAAACC
IGHV1S21	ATGCAAAA	19	TACTTA	47	+	+	49	85	302	5149718-5150156	CACAGTG	23	ACAAAAACA
Functional *IGHV* genes on scaffold 287													
IGHV1S7	ATGCAAAA	19	TACTTA	50	+	+	49	85	305	81404-81860 R	CACAGTG	23	ACAAAAACA
IGHV2S6	ATGCAAAT	25	TATAAA	55	+	+	40	89	296	72785-73209 R	CACAATA	23	ACAAAAACC
IGHV1S8	ATGCAAAA	19	TACTTA	50	+	+	49	85	305	71319-71757 R	CACAGTG	23	ACAAAAACA
IGHV2S7	ATGCAAAG	20	TATAAA	70	+	+	43	93	296	69682-70113 R	CACAATA	23	ACAAAAACC
IGHV2S8	ATGCAAAG	20	TATAAA	70	+	+	43	93	296	65205-65636 R	CACAATA	23	ACAAAAACC
IGHV2S9	ATGCAAAT	25	TATAAA	55	+	+	40	89	293	61172-61593 R	CACAATA	23	ACAAAAACC
IGHV2S10	ATGCAAAT	25	TATAAA	50	+	+	40	89	296	57759-58183 R	CACAATA	23	AACAAAAAC
IGHV2S11	ATGCAAAT	25	TATAAA	55	+	+	40	81	296	53368-53784 R	CACAATA	23	ACAAAAACC
IGHV1S12	ATGCAAAA	19	TACTTA	50	+	+	49	85	305	51937-52375 R	CACAGTG	23	TCAAAAACA
IGHV2S12	ATGCAAAG	20	TATAAA	70	+	+	43	93	296	49421-49852 R	CACAATG	23	ACAAAAACC
IGHV1S14	ATGCAAAA	19	TACTTA	50	+	+	49	85	302	40201-40636 R	CACAGTG	23	ACAAAAACA
IGHV1S15	ATGCAAAA	19	TACTTA	47	+	+	49	85	308	35978-36381 R	CACAGTG	23	AACAAAAAC
IGHV2S15	ATGCAAAG	20	TATAAA	73	+	+	40	81	296	30265-30681 R	CACAATA	23	AACAAAAAC
IGHV1S16	ATGCAAAA	19	TACTTA	50	+	+	49	80	302	17156-17586 R	CACAGTG	23	ACAAAAACA
IGHV2S16	ATGCAAAT	-	TATAAA	55	+	+	40	87	293	12984-13403 R	CACAATA	23	ACAAAAACC
IGHV2S17	ATGCAAAG	20	TATAAA	73	+	+	43	93	293	10033-10461 R	CACAATA	23	ACAAAAACC
IGHV1S17	ATGCAAAA	19	TACTTA	50	+	+	49	85	302	8522-8957 R	CACAGTG	23	ACAAAAACA
IGHV2S18	ATGCAAAA	20	TATAAA	73	+	+	43	93	296	6542-7015 R	CACAATA	23	ACAAAAACC

IGHV1S18	ATGCAAAA	19	TACTTA	47	+	+	49	85	306	5114-5607 R	CACAGTG	23	ACAAAAACA
IGHV2S19	ATGCAAAT	22	TATAAA	55	+	+	40	87	296	2214-2636 R	CACAATA	23	ACAAAAACC
ORF													
IGHV1S13orf	ATGCAAAA	19	TACTTA	50	+	+	49	81	302	s43272-43703 R	-	-	-
IGHV3S3orf	-	66	TATTAT	117	+	+	63	67	305	s73995-74429 R	CACAGTG	23	GCAAAAACC
Pseudogene													
IGHV4S1p[a]	ATGCAAAT	80	TATGTT	17	+	+	40	73	246	c5094614-5094972	CACTGGG	23	TGAAATTCC
IGHV3S1p[b]	ATGCAAAC	11	TATAAA	86	+	+	42	95	305	c5096701-5097142	-	-	-
IGHV1S5p[c]	ATGCAAAA	19	TACTTA	50	+	+	49	80	308	s89645-90081 R	CACAGTG	23	ACAAAAACA
IGHV1S6p[c]	ATGCAAAA	19	TACTTA	50	+	+	49	79	305	s87589-88021 R	CACAGTG	23	ACAAAAACA
IGHV2S4p[d]	-	-	-	-	-	-	-	-	293	s85034-85455 R	CACAATA	23	ACAAAAACC
IGHV2S5p[e]	ATGCAAAT	25	TATAAA	85	+	+	40	96	297	s79282-79798 R	CACAGTA	23	ACAAAAACC
IGHV1S9p[c]	ATGCAAAA	32	TTAAAT	15	+	+	43	46	210	s66737-67430 R	CACAGTG	23	ACAAAAACA
IGHV1S10p[f]	ATGCAAAA	19	TACTTA	50	+	+	49	80	305	s63806-64236 R	CACAGTG	23	ACAAAAACA
IGHV1S11p[d]	-	-	-	-	-	+	-	86	305	s59917-60221 R	CACAGTG	23	ACAAAAACA
IGHV2S13p[f]	ATGCAAAA	21	TATAAA	70	+	+	43	93	296	s41596-42024 R	CACAATA	23	ACAAAAACC
IGHV2S14p[c]	ATGCAAAT	21	TAATAA	22	+	+	59	122	296	s37521-37997 R	CACAATA	23	ACAAAAACC
IGHV4S2p[g]	ATGCAAAT	82	TATGAT	17	+	gt/tg	40	73	304	s32748-33158 R	CACACTG	23	TCTGAAATT
IGHV1S19p[h]	ATGCAAAA	22	TTATGT	47	+	+	49	85	330	s839-1302 R	CACAGTG	23	ACAAAAACA
IGHV1S20p[i]	ATGCAAAG	59	TATTAA	253	+	+	62	81	184	s3-327 R	-	-	-
IGHV2S21p[f]	ATGCAAAT	58	TTTCAA	8	+	+	40	67	289	c5155554-5155949	CACAATA	23	ACAAAAACC

IGHV5S1p[f]	ATGAAACA	24	TATAAA	99	+	+	36	78	302	c5156429-5156841	CACAGTG	23	GTAAACACT	

Because the sequence of the genes were partially or fully located in the gap regions, the position (underlined) indicates composite position based on the position in scaffold 287, the length of gene obtained by PCR amplification, and sequencing of each gene.

R, reverse strand; ORF, open reading frame; s, scaffold 287; c, chromosome 5

a: Fifty-three nucleotide (nt) deletion and frameshift from position 5094871, frameshift in FR2-IMGT; premature stop-codon in the V-exon

b: 3′ truncation of RSS and no expression data

c: Stop codon in L-PART1

d: L-PART1 is missing

e: One nt insertion and frameshift at position 79447 R

f: Premature stop codon in the V-exon

g: One nt deletion and frameshift from position 32987 R; premature stop-codon in the V-exon

h: Twenty-five nt insertion and frameshift from position 1078 R

i: Stop codon in L-PART1; 3′ truncation of V-exon

2.3.5 Features of *IGHV* gene 5′ flanking sequences

We examined 5′ flanking sequences for every functional *IGHV* gene to reveal possible regulatory features. The 5′ flanking region is of importance in the regulation of the VH gene expression. This region contains two conserved motifs, namely the octamer motif, which is critical to correct transcription of Ig genes, and the TATA box for the general transcription process (Falkner and Zachau 1984). As summarized in Table 2-3, all 5′ flanking sequences of functional *IGHV* genes exhibit considerable family-specific conservation. We found that all the functional or ORF segments of the IGHV1 family contain sequences completely identical to the octamer consensus (ATGCAAAA) and the TATA

consensus (TACTTA). Functional *IGHV* genes within the IGHV2 family show slightly less conserved octamer sequences and most functional IGHV2 members have single point variation (ATGCAAAT/G) in the octamer sequence, with the exception of the *IGHV2S18* gene, which presents the exact consensus octamer (ATGCAAAA). In addition, the TATA consensus (TATAAA) is well conserved across functional IGHV2 genes. The only functional member of the IGHV3 family, the *IGHV3S2* gene, shows similar features compared to that of the IGHV2 family (ATGCAAAC and TATAAA).

2.3.6 The interspersed repeats in the torafugu IGHV locus

Different repeats are known to be differentially associated with gene-rich active euchromatin, the short interspersed nuclear elements (SINEs) and gene-poor inactive heterochromatin, the long interspersed nuclear elements (LINEs) (Chen and Manuelidis, 1989). We detected the content and distribution of genome-wide repetitive elements in the IGHV-encoding region by RepeatMasker (Tarailo-Graovac and Chen, 2009). This analysis revealed that the entire torafugu IGHV locus contains a relatively low proportion of interspersed repeats (28.86%) compared to mammalian genome including the mouse genome (52.4%) and human genome (41.8%). The identified repeat elements are categorized into 7 LINE elements, 8 long terminal repeats (LTRs) elements and 2 DNA transposons. LTR elements are the largest contributor, constituting 21.48% of the locus. There are no SINEs identified in this region while the LINEs constitute 6.01% of the locus. The LINE-rich/SINE-poor structure has been proved to be characteristic of monoallelically expressed loci (Allen, et al., 2003).

2.3.7 Phylogenetic analysis between IGHV gene families

To assess sequence relatedness between IGHV gene families, we constructed a phylogenetic tree based on the amino acid sequence alignment of the 48 IGHV segments that have full-length exons. The amino acid sequence alignment and phylogenetic tree are shown in Figure 2-3 and Figure 2-4, respectively. The phylogenetic tree showed the presence of four distinct IGHV groups, corresponding to the four IGHV families (IGHV1-IGHV4). The *IGHV5S1p* segment appears to be independent from the four groups and it shows 76% homology to torafugu *IGHV1S4*01* gene from the IMGT database at the nucleotide level, which may indicate the possibility that this segment is a mutated IGHV1 member and the accumulation of mutations has decreased its overall homology to torafugu IGHV1 segments. The clustering of either functional *IGHV* genes or pseudogenes from the same IGHV family can be observed in the IGHV1–IGHV4 families. The segments of the four families are scattered along the locus, suggesting the existence of an ancestral family-specific IGHV segment and subsequent interspersion of duplicated copies along the locus.

```
                                FR1-IMGT            CDR1-IMGT       FR2-IMGT       CDR2-IMGT                  FR3-IMGT            CDR3-IMGT
                                (1-26)              (27-38)         (39-55)        (56-65)                    (66-104)            (105-117)
                              A           B           BC        C        C'         C'C"       C"        D          E          F    FG

                           (1-15)      (16-26)      (27-38)   (39-46)   (47-55)    (56-65)    (66-74)   (75-84)    (85-96)   (97-104)
                          ----------> ----------> ----------> ----------> ----------> ----------> ----------> ----------> ----------> ---------->
                          1        10  15 16    23 26 27        38 3941  46 47     55 56      65 66     74 75      84 85  89    96 97   104 105
                          |........|....| |......|..| |.........| |.|....| |.......| |........| |.......| |........| |...|.....| |......| |...
IGHV1S1                   AQTLTQSEP.VVKRP GESHTLTCSAS GFTF....SSYE MHWVRQAP GKGLEWIAY IYRS..GSSQ YYSESVK.G RFIISRDNNR AQLNLQMNSLKT EDSAVYYC ARE
IGHV1S2                   AQTLTQSEP.VVKRP GESHTLTCSAS GFTF....SSYE MNWVRQAP GKGLEWIAY ISSS..SSYK YYSESVK.G RFIISRDDNR AQLNLQMNSLKT EDSAVYYC ARE
IGHV1S3                   AQTLTQSEP.VVKRP GESHTLTCSAS GFTF....RSYW MVWVRQAP GKGLEWIAY ISSGS.GSSQ YYSESVK.G QFIISRDNNK AQLNLQMNSMKT EDSAVYYC AR
IGHV1S4                   AQTLTQSEP.VVKRP GESHTLTCSAS GFTF....SSYG MHWVRQAP GKGLEWISL IYYD..GSNK YYSESVK.G RFLISRDNNR AQLNLQMNSLKT EDSAVYYC ARW
IGHV1S5p                  AQTLTQSEP.VVKRP GESHTLTCSDS GFTF....SSYY MVWVRQAP GKGLEWISS ISSS.SGSST YYSESVK.G RFIISRDNNR AQLNLQMNSLKT EDSAVYYC ARW
IGHV1S6p                  AQTLTQSEP.VVKRP GESHTLTCSAS GFTF....SGYW MAWVRQAP GKGLEWISG IYTD..GSSQ HYSESVK.G RFIISRDNNR AQLNLQMNSLKT EDSAVYYC ARW
IGHV1S7                   AQTLTQSEP.VVKRP GESHTLTCSAS GFTF....SSYY MAWVRQAP GKGLEWIAY ISSG..GGSQ YYSESVK.G RFIISRDNNR AQLNLQMNSLKT EDSAAYYC ARE
IGHV1S8                   AQTLTQSEP.VVKRP GESHTLTCSAS GFTF....SSYN MAWVRQAP GKGLEWISA ISNG..GSSQ YYSESVK.G RFIISRDNNR AQLNLQMNSLKT EDSVVYYC ARW
IGHV1S9p                  .............. .......... ...F....SSYG MAWVRQAP GKGLEWIAY IYSS...GSS YYSESVK.G RFIISRDDNK AQLNLQMNSLKT EDSAVYYC ARW
IGHV1S10p                 AQTLTQSEP.VVKRP GESHKLTCSAS GFTF....SSYW MSWVRQAP GKGLEWISA IKYD..GSSS *YSESVK.G RFIISRDNNR AQLNLRMNSLKT EDSAVYYC ARW
IGHV1S11p                 DQTLTQSEP.VVKRP GESHTLTCSAS GFTF....SSYY MAWVRQAP GKGLEWISR ISSS..GGSQ YYSESVK.G RFIISRDNNR AQLNLQMNSLKT EDSAVYYC ARE
IGHV1S12                  AQTLTQSEP.VVKRP GESHTLTCSAS GFTF....SDHY MNWVRQAP GKGLEWISG ISSP.SGSNK HYSESVK.G RFIISRDNNR AQLNLQMNSLKT EDSAVYYC AR
IGHV1S13orf               AQTLTQSEP.VVKRP GESHTLTCSAS GFTF....SSYG MHWVRQAP GKGLEWISL IYYD..GSNK YYSESVK.G RFLISRDNNR AQLNLQMNSLKT EDSAVYYC AR
IGHV1S14                  AQTLTQSEP.VVKRP GESHTLTCSAS GFTF....SSYW MVWVRQAP GKGLEWIAY ITTS...SSP YYSESVK.G RFIISRDNNR AQLNLQINSLKT EDSAVYYC ARE
IGHV1S15                  AQTLTQSEP.VLKRP GESHTLTCSAS GFTF....RSYW MDWVRQAP GKGLEWIAY ISYTD.GSNK YYSESVK.G RFIISRDNNR AQLNLQMNSLKT EDSAVYYC ARE
IGHV1S16                  AQTLTQSEP.VVKRP GESHTLTCSAS GFTF....SDDW MYWVRQAP GKGLEWISA ISNG..GSST YYSESVK.G RFIISRNNNR AQLNLQMNSLKT EDSAVYYC AR
IGHV1S17                  AQTLTQSEP.VLKRP GESHTLTCSAS GFTF....SSYW MNWVRQAP GKGLEWISA IST...GSSQ YYTESVK.G RFLISRDNNR AQLNLQMNSLKT EDSAVYYC ARE
IGHV1S18                  AQTLTQSEP.VLKRP GESHTLTCSAS GFTF....SSYW MHWVRQAP GKGQEWIAY ISYPS.GSKK YYSESVK.G RFIISRDNNR AQLNLQMNSLKT EDSAVFYC ARE
IGHV1S19p                 AQTLTQSEP.VVKRP GESHTLTCSAS GFTF....SSYG MAWVRQAP GKGLEWISA ISYD..GSST YYSESVK.G RFLISRDDNR AQLNLQMNSLKT EDSAVYYC ARW
IGHV1S20p                 AQTLTQSEP.VVKRP GESHTLTCSAS GFTF....SSYR MNWVRQAP GKGLEWIAF ISNY..GGSI ......... ......... ............ ........ ...
IGHV1S21                  AQTLTQSEP.VVKRP GESHTLTCSAS GFTF....SDYG MHWVRQVP GKGLEWIAY IYSS..GSSQ YYSESVK.G RFIISRDNNR AQLNLQMNSLKT EDSAVYYC ARW
```

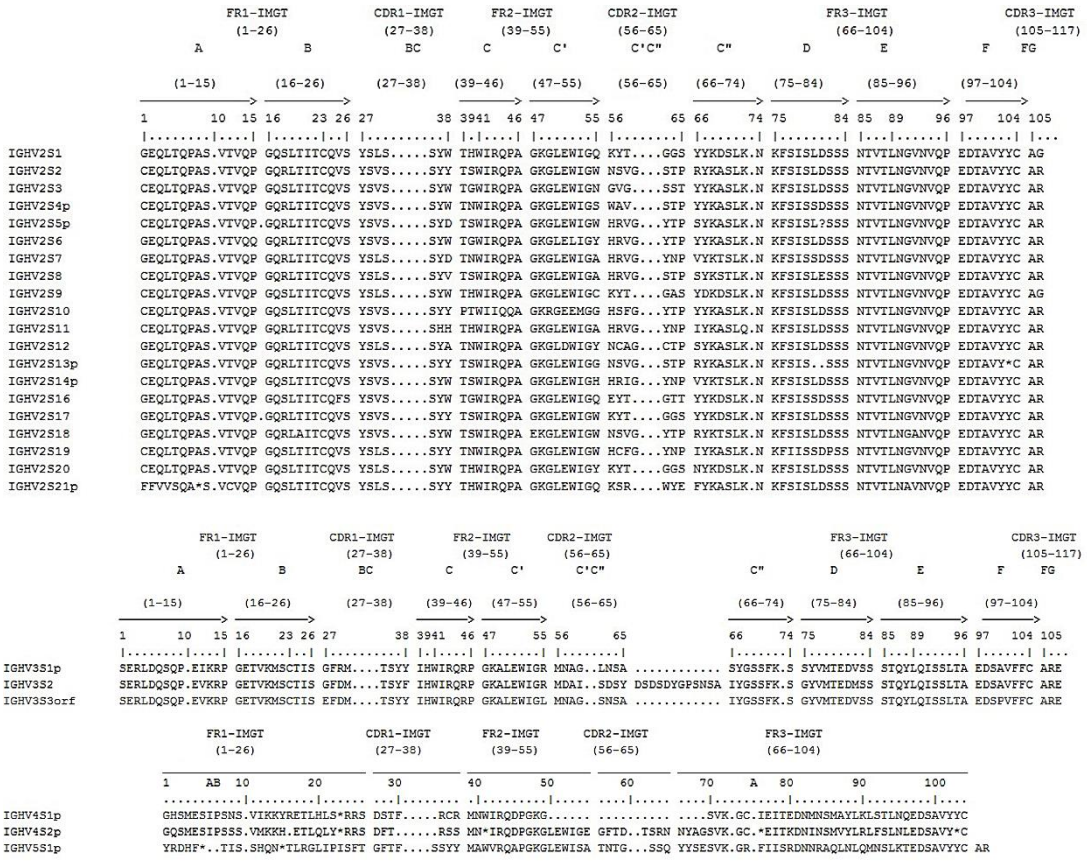


Fig. 2-3 Alignment of IGHV amino acid sequences

The protein display is shown using IMGT header (IMGT Repertoire, http://www.imgt.org). The IGHV segments are assembled according to the families. Naming was done using the nomenclature proposed by IMGT (Lefranc, et al., 2003). The newly identified IGHV families diverge more or less from the listed torafugu *IGHV* genes. Thus, the protein display of these genes is shown by using IMGT headers that most fit the IMGT rule.

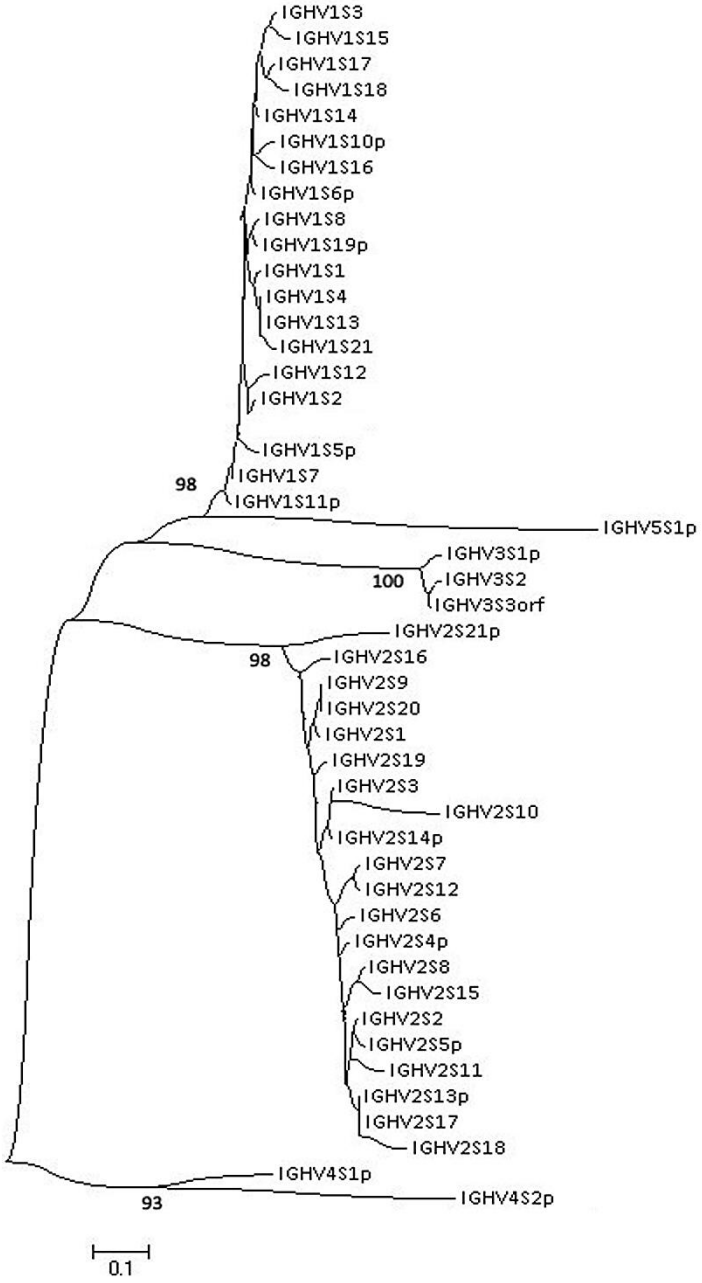


Fig. 2-4 Phylogenetic analysis of torafugu IGHV segment sequences

This tree was constructed with IGHV amino acid sequences (those with relative intact exons) using MEGA6 (Tamura, et al., 2013). Neighbor-Joining algorithm and Poisson correction were used to construct the tree. Bootstrapping was of 1000 replicates.

2.3.8 Phylogenetic relationships between IGHV sequences of torafugu and those of other vertebrates

To evaluate the relationship among IGHV sequences of torafugu and other vertebrates, we constructed a phylogenetic tree of representative sequences (Fig. 2-5). Sequences analyzed with ClustalW did not include the leader exon. As reported earlier, *IGHV* genes from vertebrates could be classified into five different phylogenetic classes (Ota and Nei, 1994). Groups A and B contained only tetrapod IGHV sequences; group C contained sequences from various vertebrates; group D contained only sequences from teleosts; and group E contained mostly cartilaginous fish sequences. IGHV sequences from two torafugu families (IGHV2 and IGHV3) fall into group D, and another sequence from the IGHV1 family falls into group C. The presence of gene from group C in torafugu indicates that the divergence of IGHV1 family sequences might occurred before teleosts and tetrapods.

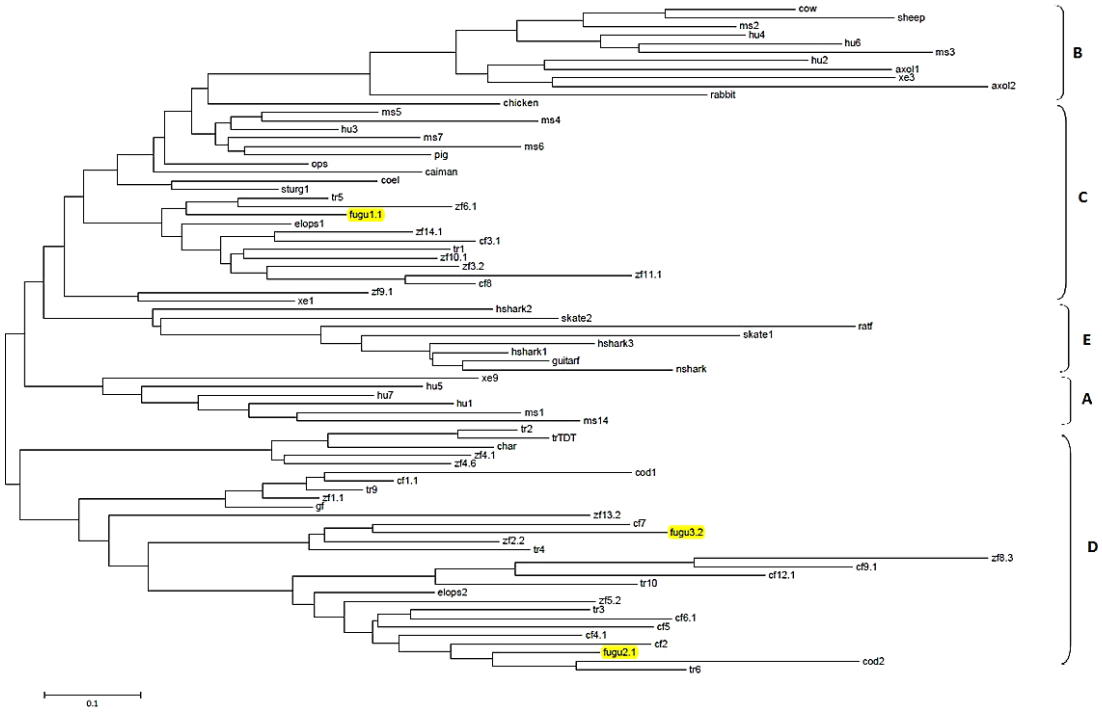


Fig. 2-5 NJ phylogenetic tree of vertebrate IGHV sequences

MEGA6 software (Tamura, et al., 2013) was used to construct the phylogenetic tree. The alignment of amino acid sequences was carried out by MUSCLE (Edgar, 2004). Neighbor-Joining method and Poisson correction were used to draw the tree. Bootstrapping was of 1000 replicates. Torafugu IGHV sequences are highlighted. The leader sequences have been omitted. Sequence designations and GenBank accession numbers are: axol, axolotll (X73553), 2 (X73554); char, cACVH3 (AJ000360); cod1 (X76510), 2 (X76507); caiman, C3 (M12768); cf, catfish VH1, VH1.1 (U09719), VH2, NG22 (M58670), VH3, VH3.1 (U09721), VH4, VH4.1 (U09722), VH5, NG66 (M58674), VH6, VH6.1 (U09724), VH7 (U67446), VH8.1 (AY238377), VH9.1, (AY238378), VH12.1, (AY238381); chicken VH1 (M30350); coel, coelacanth (X57354); cow (U55164); elops1 (M26182), 2 (M26579); gf, goldfish, gf99 (X61312); guitarf, guitarfish (*Rhinobatos productus*), (AY612426); hshark, horn shark (*Heterodontus francisci*), (M12195), 2 (Z11776), 3 (AY612427); hu, human VH1, DP-1 (Z12303), VH2, DP-26 (Z12328), VH3, A3-M13 (Z96969), VH4, DP-64 (Z12364), VH5, VH32 (U44509), VH6, DP-74

(Z12374), VH7 (X61012); ms, mouse VH1, J558 (D14634), VH2, Q52 (U53526), VH3, VH36-60 (K02786), VH4, X-24 (M22955), VH5, 7183.14 (AF290968), VH6, (U05819), VH7, S107 (U97569), VH11 (MMIGVHAB), VH14, SM7 (X55934); nshark, nurse shark (*Ginglymostoma cirratum*), (M92851); ops, opossum (*Monodelphis domestica*), MVH298 (AF007075); rabbit VH1 (M29947); sheep (X59994); skate, clearnose skate (*Raja eglanteria*), 1 (M29679), 2 (M29672); sturg, sturgeon (*Acipenser baeri*),VH1, ScH 16.1 (Y13260), VH2, ScH227 (AJ223052); pig (M81769); ratf, ratfish (*Hydrolagus colliei*), (AF003841); tr, rainbow trout (*Oncorhynchus mykiss*), VH1, RTVH253 (X92501), VH2a, cRTVH12 (X81509), VH2b, TDT271 (X81512), VH3, cRTVH19 (X81510), VH4 (L28744), VH5, cRTVH6 (X81513), VH6, cRTVH1 (X81481), VH9, cRTVH20 (X81504), VH10, cRTVH4 (X81508); xe, African clawed frog (*Xenopus laevis*),VH1 (M24673), VH3 (M24675), VH8 (X56862), VH9 (X56863).

2.3.9 Rearrangements

A set of *IGHV*-specific primers (Table 2-2) were designed to complement *IGHC* gene-specific primers and 86 clones from all positive PCR products were sequenced to identify expression and rearrangements. Of the 36 and 38 clones generated from Cμ-specific and Cτ-specific primers (both located in exon 2), respectively, 35 clones of Cμ and 36 and of the Cτ groups contained conserved sequences in exon 1 and were further analyzed. Confirmation of the IGHV segment was performed by searching in the IMGT/V-QUEST database and the use of the IGHJ segment was manually inspected. We observed a preference for the IGHV1 family genes by Cμ2 and Cτ2, ~70% of the time (Fig. 2-6). In addition, in 86 cDNA sequences, we found that Jμ was always expressed with Cμ and Jτ with Cτ. We were unable to

distinguish the distribution of IGHV sequences between Cμ and Cδ since PCR failed to amplify a product using a Cδ-specific reverse primer (either from the δ1, δ4, or δ7 domains) when paired with the forward primers (data not shown), in agreement with a previous report that a variable region could not be isolated from the torafugu IgD cDNA library (Saha, et al., 2004). However, the data still shows that the IGHV1 family genes are associated with Cμ1, rather than with Cτ2, in most cases (Fig. 2-6).

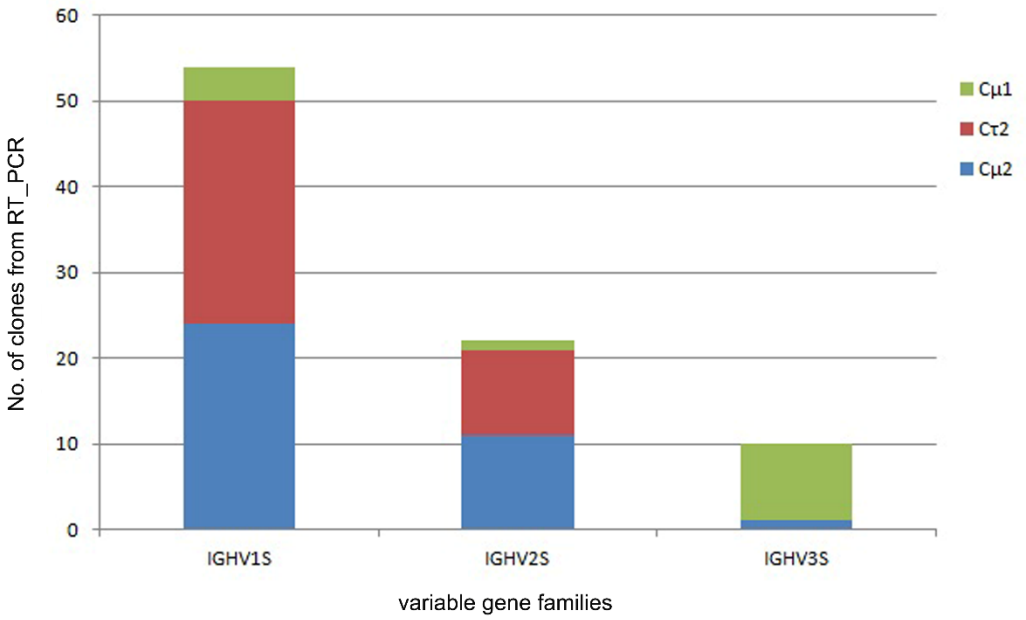


Fig. 2-6 Use of IGHV sequence families in rearrangements

Few cDNAs containing the IGHV3 family gene sequences were identified by using each C- specific reverse primer coupled with a combination of all the forward primers. Therefore, we performed another PCR experiment using only the IGHV3 primer set with a mixture of reverse primers, and the PCR products were cloned and sequenced. As a result, we confirmed the expression of IGHV3 family genes (Fig. 2-7), and found that the IGHV3 family members were combined with Cμ1, indicating their association with the expression of IgM and/or IgD. The IGHJ sequences were also found to be combined

with the IGHV3 gene segments in a biased manner, that is, the Jμ2 segment was preferably found in 80% of the 12 observed cDNA sequences (data not shown).

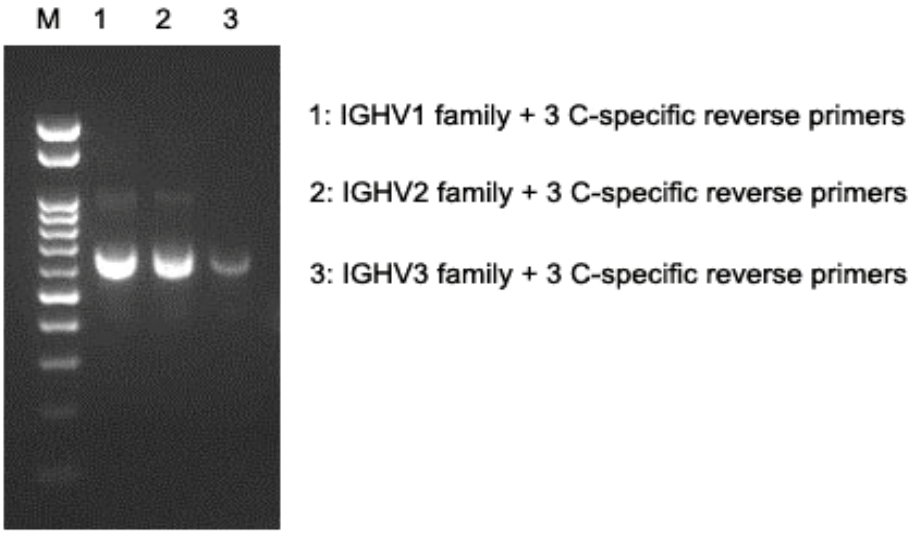


Fig. 2-7 Confirmation of the IGHV3 family segment expression

2.3.10 Characterization of *IGHD* genes

So far, a total of 5 *IGHD* genes have been described, including 4 *Dμ* segments and 1 *Dτ* segment (Savan, et al., 2005). According to conserved RSS motif specificity, we manually screened the genomic assembly and scaffolds to identify any additional *IGHD* gene. We identified 4 *Dμ* segments located 3.7 kb upstream of Cμ1 in scaffold 483 and in the same region of chromosome 5, which is consistent with a previous report (Savan, et al., 2005). We also found 3 previously unidentified *Dμ* in scaffold 483, which span a 700 bp region 5.9 kb upstream of Cμ1 (Table 2-4). These newly found *Dμ* segments are most likely functionally expressed since they possess the features of those in their mammalian counterparts

and teleosts, the ORF sequences flanked by functional RSS (12bp spacer) on both sides. In addition, these genes are located between the Cτ region and the *Jμ* segments, very close to the reported *Dμ* gene cluster, suggesting that they may effectively participate in the *VDJ* recombination process. There is a gap region in chromosome 5 (approximately 2 kb) upstream of the 4 reported *Dμ* genes. The nucleotide sequence of scaffold 483 overlaps chromosome 5 from the most 3′ *IGHV* gene in chromosome 5 and further downstream the 4 reported *Dμ* segments. This implies that scaffold 483 could represent a more complete IGHD region, in which the total number of *Dμ* segment is seven (Fig. 2-1d).

Table 2-4 Nucleotide sequences of DH gene segments

	Nonamer	**Spacer**	**Heptamer**	**DH region**	**Heptamer**	**Spacer**	**Nonamer**	**location**
DH1τ	AGTTTTTGT	12	CACAGTG	ATATTTTGGGCTGGC	CACAGAG	12	ACAAAAACC	chr5 5158747-5158817
DH1μ	GGTTTTGGT	12	CACTGTG	GTATTATTATAGCTCTG GTTACAG	CACAGTG	12	ACAAAAACC	sca483 383297-383376
DH2μ	GATTTTTGT	12	CAGTGTG	TAACAGTGGGAACA	CACAGTG	12	GCAAAAACT	sca483 383061-383130
DH3μ	GGTTTTTGT	12	CACAGTG	TACTATAACTAC	CACAGTA	12	TCAAAAACT	sca483 382692-382759
DH4μ	TGTTTTTGT	12	CACTGTG	TATATTTGGG GGATGGGG	CACAGTG	12	GCAAAAACC	sca483 381882-381955 (chr5:5165161-5165234)

DH5μ	GGTTTTTGT	12	CACAGTG	TACTATAACTAC	CACAGTA	12	ACAAAAACT	sca483 381317-381384 (chr5:5165694-5165761)
DH6μ	AGTTTTTGT	12	CATTGTG	GACTATAGCTGGAAC	CACAGTG	12	ACAAAAACC	sca483 381002-381091 (chr5: 5165927-5165997)
DH7μ	TGTTTTTGT	12	CACTGTG	TATACGGGGGTGGGG	CACAGTG	12	GCAAAAACC	sca483 380452-380522 (chr5: 5166464-5166534)

Novel DH segments are in bold.

2.3.11 Constant domains

We used the sequence of cDNA encoding torafugu μ, δ, and τ to identify corresponding genomic sequences in the current genome assembly and the scaffolds from our data. We identified a region of approximately 45 kb on chromosome 5 that contains the CH segments. The torafugu τ was identified in both chromosome 5 and scaffold 483, consistent with the structure of the previously characterized torafugu $C\tau$ (Savan, et al., 2005). The torafugu $C\mu$ gene locates in a 3.5 kb-region of scaffold 483 and show a conserved 4-CH structure (Flajnik, 2005) as well as 2 exons encoding membrane spanning regions. However, we did not identify any nucleotide sequence that resembles published $C\mu$ gene sequence in chromosome 5 because a large gap region (approximately 27 kb) is present downstream of the 4 reported $D\mu$ segments. In addition, we found the $C\delta$ domains in chromosome 5 immediately

downstream of this large gap region and the transmembrane regions of torafugu C*δ* in both chromosome 5 and scaffold 483. We therefore speculate that the nucleotide sequences in scaffold 483 can be assigned to chromosome 5 from the most 3′ *IGHV* gene of the C-proximal cluster to the Cδ region (Fig. 2-1d). Interestingly, the torafugu *δ* gene was found to contain at least 14 intact and 1 truncated Cδ exons on chromosome 5. This finding differs from an earlier report, in which the CH region of torafugu *δ* gene was reported to contain 13 Cδ exons with a tandem Cδ1-Cδ6 duplication followed by a single Cδ7 domain (Saha, et al., 2004). Our finding based on chromosome 5 in the v5 assembly shows a structural variation in the IGHδ locus (Fig. 2-8). Five exons (Cδ1–Cδ3, Cδ5, and Cδ6, in accordance with the reported Cδ nomenclature) are repeated twice, while one exon (Cδ4) is repeated three times with 98–100% identity in amino acid sequence. We identified one additional Cδ5 exon that contains a 3′ truncation due to gaps as well as a single Cδ7 exon and two membrane exons. On the other hand, scaffold 483 was more incomplete in this region. The difference of gene organization between the previous report and the data based on chromosome 5 might reflect structural variation in this region. Alternatively, it might reflect on misassembling of sequence in this region of v5 assembly as often found in such duplicated or multiplied genomic regions.

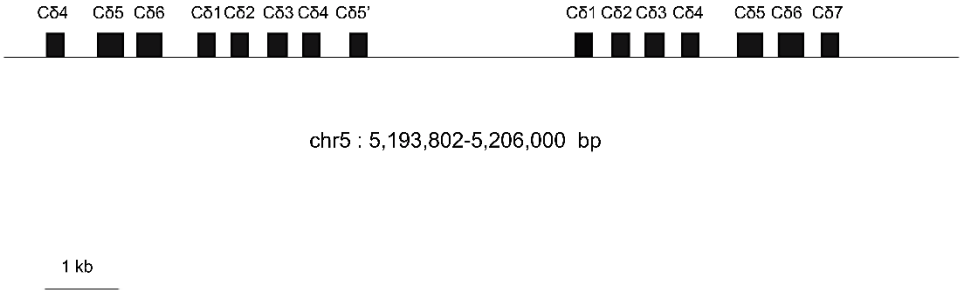


Fig. 2-8 Genomic organization of the constant region of torafugu IgD

A novel repeat pattern of Cδ domains is present on chromosome 5. Cδ exons are numbered in accordance with the reported Cδ nomenclature. The Cδ5′ represents the truncated Cδ exon.

2.4 Discussion

Our data provide updated information regarding the torafugu IGH locus, in which at least 48 *IGHV* genes (Table 2-3), 7 *IGHD* genes (Table 2-4), and 6 *IGHJ* genes (Table 2-5) are spread out over the variable region. This locus is located in a 115 kb region of chromosome 5 and is organized in a translocon organization with multiple *IGHV* genes followed by tandem DH-JH-CH cassettes. The structure of the IGH loci is diverse among teleosts because of successive episodes of genome duplications and gene loss. For example, rainbow trout possesses two IGH isoloci while catfish does not contain IgT. The overall organization of the IGH locus in torafugu is very similar to the corresponding region in zebrafish, despite the fact that these two species are not evolutionarily close among fish families (Fig. 2-9). This

organization can also be found in mouse Tcrα-Tcrδ locus (Krangel, et al., 2004), suggesting the conservation of such arrangement from mammals to teleosts.

Table 2-5 Nucleotide sequences of JH gene segments

J-segment	Nonamer		Heptamer	Nucleotide sequence& translation	location
					chr5
Jτ1	AGTTTTTGC	23	CACTGTG	Y F D V W G N G T K V T V S S attattttgacgtctggggtaatggaactaaagtcacagtatcatcag	5159133-5159219
					sca483
Jμ1	CTTTCTTGT	23	CACGGTG	Y Y A Y F D Y W G K G T T V T V T S ttactacgcatactttgactactgggggaaaggaacaacagttacagtaacatctg	378104-378198 R
					sca483
Jμ2	AGTTTTGGT	22	CACTGTG	Y Y F D Y W G K G T M V T V T S attactactttgactactgggggaaaggaaccatggtgaccgtcacatcag	377856-377944 R
					sca483
Jμ3	GGTTTTTGT	23	CACTGTG	Y A L D Y W G K G T K V T V T S actatgctctggactactggggaaaaggcacaaaggtcacagtgacatcag	377682-377771 R
					sca483
Jμ4	GGTTTTTGT	23	AACTGTG	Y G F D Y W G K G T T V T V S S actacggcttcgactactggggaaaaggcaccacagtaaccgtcagctcag	377393-377482 R
					sca483
Jμ5	GGATTTCAT	23	CAATGTG	Y Y E F D Y W G K G T S V T V S S ctactatgaatttgactactggggaaagggcacatcggtgactgtttcgtccg	376941-377032 R

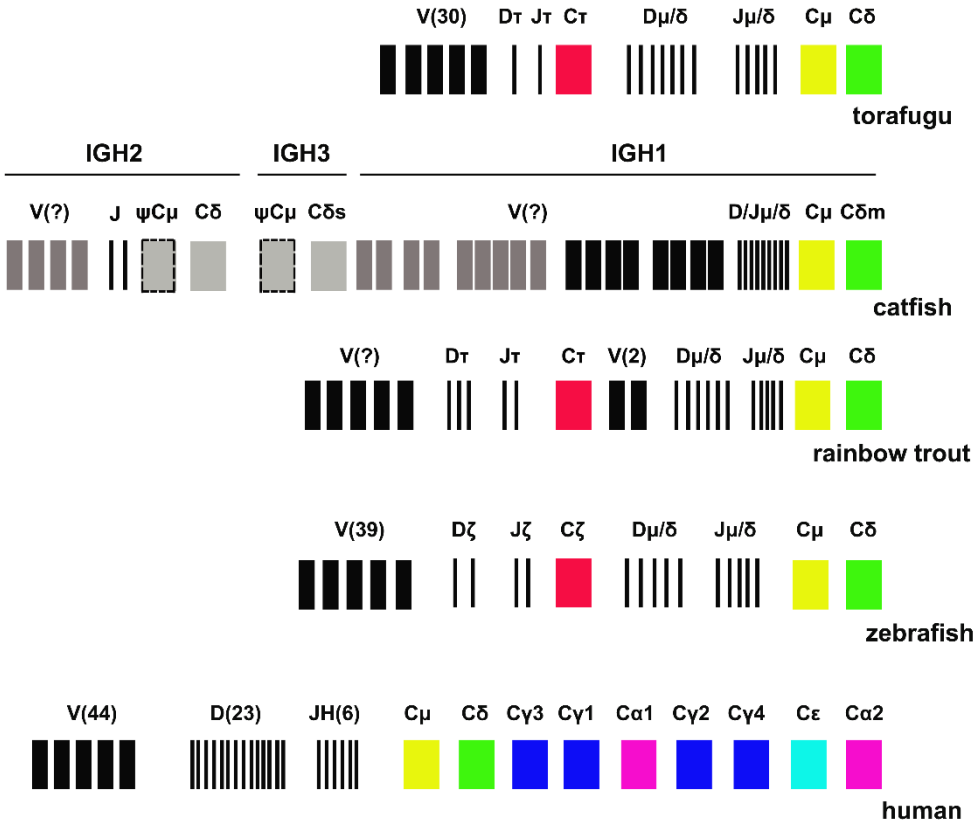


Fig. 2-9 Schematic structures of torafugu IGH loci and those of other vertebrates

The torafugu IGH locus was sequenced and characterized in this study. The diagrams of the IGH locus of catfish (725 kb), rainbow trout (125 kb), and zebrafish (175 kb) were modified from earlier reports (Bengtén, et al., 2006; Danilova, et al., 2005; Hansen, et al., 2005). The IGH locus of human (1,250 kb) was retrieved from IMGT at http://www.imgt.org. The number of potential functional *IGHV* genes is indicated in brackets. In the catfish locus, the complete sequences of IGH2 and IGH3 and the upstream of IGH1 VH region have not been reported. Therefore, these regions are shown in gray boxes. The "?" symbol indicates lack of data. Dashed-line boxes indicate the IGHC pseudogenes (ψ). Cδs and Cδm correspond to the secreted and membrane IgD coding genes, respectively.

A total of 48 *IGHV* genes have been characterized, including those revealed by sequences in scaffold 287. Of the 48 *IGHV* genes, 30 (62.5%) are potential functional germline *IGHV* genes. The remaining genes are either ORFs or pseudogenes because they contain one or more defects. The number of *IGHV* genes among teleosts has been proved to be more or less uniform (Das, et al., 2008). In other teleosts, for example, the total number of functional *IGHV* genes is 37, 38, and 41 for the zebrafish, medaka, and stickleback, respectively, while the number of nonfunctional *IGHV* genes is 10, 15, and 11, respectively (Das, et al., 2008). In contrast, the number of *IGHV* genes varies widely among mammalian species. The rat genome possesses the highest number of functional *IGHV* genes (at least 131) of all mammalian species studied so far (Hendricks, et al., 2010), while the numbers of both functional and nonfunctional *IGHV* genes in sheep and cow are much smaller than those in other placental mammals (Das, et al., 2008). The number of *IGHV* genes in torafugu, identified in this study (30 functional and 18 nonfunctional genes), was equivalent to those in other teleosts. Similar numbers of *IGHV* genes and similar ratio between functional and nonfunctional *IGHV* genes among teleosts may suggest that the distribution pattern of *IGHV* genes was already present in a teleost ancestor and was conserved in the different lineages. Moreover, RT-PCR identified IGHV1 and IGHV2 segments in high to moderate abundance, indicating a meaningful contribution of these segments to the available germline repertoire in torafugu.

There is a positive correlation between the number of functional *IGHV* genes and the number of nonfunctional *IGHV* genes (Das, et al., 2008). In this study, more pseudogenes were identified in scaffold 287 than in chromosome 5. Given that the number of functional *IGHV* genes is higher in scaffold 287 than in chromosome 5, we suggest that torafugu IGHV locus has evolved following the birth-and-death evolution model. It has also been proposed that gene duplication and functional elimination are involved in the presence of nonfunctional genes (Nei and Rooney, 2005). Here, we grouped two pseudogenes into

an additional IGHV4 family. This family branches off from other IGHV families at an early stage (Fig. 2-4). The existence of this family may be partially explained by the theory of functional elimination after deleterious mutations. *IGHV* genes from this family may go through unsuccessful gene duplication and are functionally eliminated by deleterious mutations.

We have shown that the torafugu VH region contains a much lower proportion of interspersed repeats than those in humans and mouse. A prevalent hypothesis of large scale changes in genomic sequences, including the IGHV loci, suggests that such alterations may be the result of the content of different repetitive elements. Taking into account that LTR elements constitute the biggest part of torafugu IGHV locus which differs that in humans and mouse, we speculate that the differences in the content of repetitive element of the IGHV locus may explain in part the gene content heterogeneity of the IGHV locus in mammals as compared with torafugu.

Identification of a more expanded IGHδ locus in this study may suggest a high degree of variation in the number of CH domain exons in the δ gene. It has been shown that torafugu δ was transcribed as a chimeric molecule with Cμ1 spliced to Cδ1 (Saha, et al., 2004). The presence of three exons (Cδ4– Cδ6) upstream of the Cδ1 exon raises questions regarding the expression of torafugu IgD. Alternative splicing patterns of IgD may therefore exist, as revealed in the zebrafish (Zimmerman, et al., 2011).

The structure of IGHδ varies among teleosts because of the extent of CH domain duplications. For example, Cδ2–Cδ3–Cδ4 domains are repeated three times in the Atlantic salmon IgHA and catfish, and four times in the zebrafish and Atlantic salmon IgHB (Fillatreau, et al., 2013). It is clear that most, if not all, teleost δ genes have seven unique CH exons (Cδ1–Cδ7) and it seems that torafugu δ retains these CH exons even after structural alteration. However, it is noteworthy that the IGHδ locus, identified here, contains neither an upstream μ gene sequence nor sequence information between the truncated Cδ5 exon

and the second block of Cδ1–Cδ6 domains. Whether this organization of IGHδ locus occurred by random genomic drift (Nei, 2007) or was the result of an error in the assembly of the torafugu v5 assembly remains undetermined. Thus, our observation must be taken as a trial and future studies on more complete genome assembly may help to answer these questions.

chapter 3:

Analysis of the immunoglobulin light chain genes in torafugu

3.1 Introduction

IgL genes have been found in all groups of jawed vertebrates studied so far, although differences in the number of IgL isotypes, expression patterns and genomic organization are commonly observed depending on the species. It has long been known that mammals express two IgL isotypes, κ and λ (Das, et al., 2008; Wu and Kabat, 1970). As IgL chain sequences were identified in additional vertebrate groups, problems occurred when attempts were made to classify the ectothermic IgL sequences based on mammalian κ and λ distinctions. This was not only due to the lack of data from key phylogenetic groups, but also to the intrinsic rapid evolutionary rate of antigen receptors (Litman, 1999). For example, three IgL isotypes are present in the amphibian Xenopus: κ, λ, and a unique isotype σ (Qin, et al., 2008); four IgL isotypes has been categorized in elasmobranchs: type I (σ-cart-type) is restricted to elasmobranchs, type II (λ-type) and type III (κ-type) are present in all jawed vertebrates except for birds, and type IV (σ-type) (Hikima, et al., 2011).

In teleosts, there are four groups containing one λ-type, two κ-types (teleost type 1 (L1) and type 3 (L3)), and one σ-type (type 2 (L2)) as summarized in Table 3-1. It can be suggested that all teleost species possessing the λ-type L chain are more ancient diverged lineages of teleost fish such as two types of catfish (*b*, Siluriformes), rainbow trout (*c*, Salmoniformes), and Atlantic cod (*d*, Gadiformes). On the other hand, most of the other fishes are modern types in teleost evolution. It seems that the λ-type L chains in the Acanthopterygii superorder (including the orders from *e* to *h*) have been lost in their divergence, and that the other early divergent orders, like cypriniformes (*i.e.* carp and zebrafish) and salmoniformes (*i.e.* Atlantic salmon), could also have the λ type isotype (Hikima, et al., 2011).

Table 3-1 IgL isotypes reported in teleost fish

Teleost species		Orders	Isotypes
Common carp	*Cyprinus carpio*	*a*	κ (L1A, L1B, L3), σ (L2)
Zebrafish	*Danio rerio*	*a*	κ (L1, L3), σ (L2)
Catfish	*Ictalurus punctatus*	*b*	κ (F, L1; G, L3), σ (L2), λ
Atrantic Salmon	*Salmo salar*	*c*	κ (L1, L3), σ (L2)
Rainbow trout	*Oncorhinchus mykiss*	*c*	κ (L1, L3), σ (L2), λ
Atlantic cod	*Gadus morhua*	*d*	κ (L1), λ (L2)
Three-spined stickleback	*Gasterosteus aculeatus*	*e*	κ (L1a, L1b), σ (L2)
Torafugu	*Takifugu rubripes*	*f*	κ (L1), σ (L2)
Japanese flounder	*Paralichthys olivaceus*	*g*	κ (L1a, L3)
Spotted wolffish	*Anarhichas minor*	*h*	κ (L1, L3), σ (L2)
Yellowtail	*Seriola quinqueradiata*	*h*	L1

Adapt from (Hikima, et al., 2011)

Orders for teleosts accordong to the divergence: (a) Cypriniformes, (b) Siluriformes, (c) Salmoniformes, (d) Gadiformes, (e)Gasterosteiformes, (f)Tetraodoniformes, (g) Pleuronectiformes, and (h) Perciformes.

It was clear that many ectothermic vertebrates expressed more than two IgL chain isotypes including some isotypes not orthologous to κ or λ. Criscitiello and Flajnik (Criscitiello and Flajnik, 2007) thus proposed the grouping of all vertebrate IgL chains into four main ancestral branches: κ, λ, σ, and σ-cart. This classification system is based on criteria of sequence homology, and the spacing of heptamer and nonamers of RSS sequence. In addition, the genomic configuration of IgL gene segments and the length of the CDR of corresponding VL gene segments support a syntenic approach to the IgL classification. Several studies have also used molecular sequence markers to distinguish IgL isotypes (Das, et al., 2008; Edholm, et al., 2009). For the purpose of this thesis the original IgL nomenclature is used together with the more recent classification into κ (mammalian κ, elasmobranch type III, teleost L1, and L3, Xenopus ρ), λ (mammalian λ, elasmobranch type II), σ (Xenopus σ, teleost L2, elasmobranch type IV), and σ-cart (elasmobranch type I/NS5).

Teleosts IgL genes are usually arranged in multiple clusters of VL, JL, and CL region segments: (VL-JL-CL)n or (VL-VL-JL-CL)n (Edholm, et al., 2011). Thus, teleost fish possess a chimeric gene organization for the H and L chain genes (Flajnik, 2002). For the L1 and L3 loci, the VL segments are in opposite transcriptional orientation to the JL and CL segments, while in L2 and λ clusters VL, JL, and CL genes are in the same transcriptional orientation (Daggfeldt, et al., 1993; Ghaffari and Lobb, 1997; Timmusk, et al., 2000). The only exception was found in the stickleback, where there are eight σ clusters consisting of (V2-JL-CL-V1) and the VL 3′ of the CL is in the opposite transcription orientation (Bao, et al., 2010). To date, IgL isotypes found in teleosts have been found to exist on different chromosomes in a cluster assemblage. In a genome-wide study on zebrafish, such clusters have been found in four different chromosomes (Bao, et al., 2010; Daggfeldt, et al., 1993; Edholm, et al., 2009; Zimmerman, et al., 2011).

Here, we report a genome-wide analysis of *IgL* genes in the torafugu, revealing (VL-JL-CL) clusters spanning at least 3 separate chromosomes and multiple scaffolds, as well as the identification of a third IgL isotype in this species.

3.2 Materials and methods

3.2.1 Retrieval of *IgL* genes from torafugu genome

Genome builds of torafugu (assembly v4, October 2004 and assembly v5, January 2010) available from Fugu Genome Project (http://www.fugu-sg.org/) (Kai, et al., 2011) were searched to locate *IgL* genes. Published IgL sequences from torafugu (Hsu and Criscitiello, 2006; Saha, et al., 2004) and other teleost fishes (Bao, et al., 2010; Edholm, et al., 2009; Zimmerman, et al., 2011) were used as queries the encoded amino acid sequences in TBLASTN alignments (cutoff *E*-value of 10^{-15}) to retrieve IgL-gene-containing scaffolds and chromosomes. Genomic sequences that contain matches for both IGLV and IGLC were downloaded and analyzed further.

The identified genomic sequences were then used as queries in BLASTN searches against the EST database at NCBI to retrieve any expression data. In this study, expression of *IGLV* genes was determined using BLAST hits using a 95% threshold identity and a 10^{-15} *E*-value threshold, while ESTs were assigned to concordant *IGLC* when a ≥99% identity was met.

3.2.2 Annotation of torafugu IGL

The Artemis software (Carver, et al., 2008) was used to annotate the IGL locus including the

transcriptional polarity and relative positions of *IGLV* and *IGLC* in genomic sequences. CL exons were discerned by comparing resultant genomic sequences with published IgL mRNAs. *IGLV* genes were determined by the presence of canonical RSS (allowing 2 nucleotide mismatches), by ORFs that match for Ig signature sequences using IgBlAST of NCBI (www.ncbi.nlm.nih.gov/projects/igblast) and IMGT/V-QUEST software (the Teleostei unit) (Brochet, et al., 2008), and pattern searches for 23RSS or 12RSS flanking ends of gene segments. To identify the *IGLJ* genes, which are too short to be detected by BlAST searches, we performed pattern searches to find JL-specific RSS among the initial genomic sequences identified to contain *IGLV* and *IGLC*. The pattern is a consensus RSS heptamer and nonamer with a 22-24 bp spacer (CACAGTG-$N_{22\text{-}24}$-ACAAAAACC) region. Splice sites between leader and VL exons were discerned by FSPLICE (http://linux1.softberry.com/). Exon boundaries of IGLV, IGLJ, and IGLC were refined by aligning with known VJ-C cDNA sequence (Saha, et al., 2004) and torafugu EST sequences (from Fugu Genome Project) (Clark, et al., 2003).

3.2.3 Determination of functionality of *IGL* genes

Identified gene segments were determined according to the IMGT® nomenclature (Lefranc, 2007). For the *IGLV* genes, any retrieved sequence without a truncation, frameshift mutation, premature stop codons in the leader exon and the VL exon, had conserved residues (1st-CYS, conserved-TRP, and 2nd-CYS) in FR1, FR2, and FR3 regions, respectively, and possessed a proper RSS was deemed functional gene. For the *IGLC* and *IGLJ* genes, retrieved sequences without any frameshift mutations and internal stop codons were regarded as potentially functional genes. In addition, the examination of RSS was implemented to determine putative functionality of *IGLJ* genes.

3.2.4 Comparative phylogenetic studies

Phylogenetic studies were carried out using the MEGA6 program (Tamura, et al., 2013). Multiple sequence alignments were performed using MAFFT (Katoh and Standley, 2013). The neighbor-joining method was used to construct the phylogenetic trees (pair-wise deletion, Jones-Taylor-Thornton matrix) and enter range activated sites by gamma parameter 2.5. Evaluation of the veracity of these trees was done by executing a bootstrap procedure of 1000 replicates.

3.3 Results

3.3.1 Torafugu IGL genomic organization

There are a total of 76 IGL gene segments were identified to be localized in multiple clusters to three different chromosomes (chromosome 2, 3, and 5) and 38 different genomic scaffolds. Of the scaffolds, four were assigned to different chromosomes. Collectively, 46 IGLV, 16 IGLC, and 14 IGLJ torafugu gene segments were identified (Table 3-2 and Table 3-3).

Table 3-2 Genomic features of the torafugu IGLV genes

IGLV family	*IGLV* gene	Fct	Gene structure			Location on scaffolds	RSS		
			L-PART1 (nt)	Intron (nt)	V-eoxn (nt)		7mer	Spacer (nt)	9mer
IGLV1	sca2115_v3	F	52	84	303	936-1374	CACAGTG	12	ACAAACCCT
	sca2488_v2	F	52	85	314	6109-6559 R	CACAGTG	12	ACAAAAACT

	sca2488_v3	F	52	85	317	13008-13461 R	CACAGTG	12	ACAAAAACC
	sca3401_v2	F	52	84	315	3276-3726 R	CACAGTG	12	ACAAAAACC
	sca6557_v2	P[a]	-	-	306	2716-3021 R	CACAGTG	12	ACAAAAACC
	sca9808_v2	P[a]	-	-	315	2521-2835 R	CACAGTG	12	ACAAAAACC
IGLV2	sca54_v1	P[b]	40	92	230	360582-360943	-	-	-
	sca139_v1	P[c]	40	92	329	478-938 R	CACAGTG	12	ACAAAAACC
	sca2352_v1	P[a]	-	-	329	7618-7946	CACAGTG	12	ACAAAAACC
	sca2352_v2	F	40	92	326	9663-10120	CACAGTG	12	ACAAAAACC
	sca2681_v1	P[b]	40	92	215	2-348 R	-	-	-
	sca3001_v1	F	40	92	329	5112-5572 R	CACAGTG	12	ACAAAAACC
	sca3330_v1	P[d]	40	195	225	4995-5454 R	CACAGTG	12	ACAAAAACC
	sca4312_v1	P[e]	40	92	326	4083-4540	CACAGTG	12	-
	sca4520_v1	P[f]	40	92	326	4390-4847	CACAGTG	12	ACAAAAACC
	sca4852_v1	P[g]	40	92	326	3504-3961 R	CACAGTG	12	ACAAAAACC
	sca4988_v1	P[b]	40	92	206	833-1170	-	-	-
	sca5604_v1	P[h]	40	92	323	308-762	CACAGTG	12	ACAAAAACC
	sca5821_v1	F	40	217	182	2961-3399 R	CACAGTG	12	ACAAAAACC
	sca6813_v1	P[i]	40	92	326	1660-2117	CACAGTG	12	ACAAAAACC
	sca7335_v1	P[b]	40	92	302	4267-4700 R	-	-	-
	sca7335_v2	P[j]	40	92	329	275-735 R	CACAGTG	12	ACAAAAACC
	sca7665_v1	P[a]	-	92	326	304-629 R	CACAGTG	12	ACAAAACCT

	sca7989_v1	F	40	92	302	1561-1994	CACAGTG	12	ACAAAAACC
	sca8603_v1	P[k]	40	92	326	1186-1643	CACAGTG	12	ACAAAAACC
	sca9980_v1	F	40	89	323	1721-2172	CACAGTG	12	ACAAAAACC
	sca11246_v1	P[a]	-	-	329	1497-1825	CACAGTG	12	ACAAAAACC
	sca11893_v1	F	40	92	329	1964-2424 R	CACAGTG	12	ACAAAAACC
IGLV3	sca10_v1	ORF[l]	40	137	300	2769863-2770162	CACAGTG	12	ACAAAAACC
	sca10_v2	P[b]	40	133	168	2772729-2772896	-	-	-
	sca10_v3	ORF[l]	40	137	291	2781060-2781527	CACAGTG	12	ACAAAAACC
	sca158_v1	ORF[l]	40	137	291	690626-691093	CACAGTG	12	ACAAAAACC
	sca2115_v1	P[a]	-	-	114	20505-20618 R	CACAGTG	12	ACAAAAACC
	sca2115_v2	P[b]	40	137	249	17729-18154 R	-	-	-
	sca2422_v1	ORF[l]	40	137	285	4065-4526	CACAGTG	12	ACAAAAACC
	sca2422_v2	P[a]	-	-	267	13647-13913	CACAGTG	12	ACAAAAACC
	sca2488_v1	ORF[l]	40	137	291	1139-1606 R	CACAGTG	12	ACAAAAACC
	sca3326_v1	ORF[l]	40	137	300	301-777 R	CACAGTG	12	ACAAAAACC
	sca3401_v1	P[a]	-	-	300	8198-8497 R	CACAGTG	12	ACAAAAACC
	sca4248_v1	ORF[l]	40	137	300	5574-6050	CACAGTG	12	ACAAAAACC
	sca4557_v1	P[a]	-	-	300	2247-2546 R	CACAGTG	12	ACAAAAACC
	sca6105_v1	ORF[l]	40	137	285	4890-5351	CACAGTG	12	ACAAAAACC
	sca6723_v1	ORF[l]	40	137	291	5284-5751 R	CACAGTG	12	ACAAAAACC
	sca6723_v2	ORF[l]	40	133	291	2419-2882 R	CACAGTG	12	ACAAAAACC

sca7391_v1	ORF[l]	40	140	285	2666-3130 R	CACAGTG	12	ACAAAAACC
sca9808_v1	ORF[l]	40	137	300	61-537 R	CACAGTG	12	ACAAAAACC
sca9855_v1	ORF[l]	40	137	285	2234-2695	CACAGTG	12	ACAAAAACC
sca11195_v1	ORF[l]	40	137	288	2011-2475 R	CACAGTG	12	ACAAAAACC

Fct functionality, *F* functional, *P* pseudogene, *ORF* open reading frame, *R* reverse strand

[a] L-PART1 is missing; [b] 3′ truncation; [c] 1 nt deletion and frameshift at position 659 R; 2 nt deletion and frameshift from 637 R; [d] 1 nt deletion and frameshift at position 5176 R; 2 nt deletion and frameshift from 5154 R; [e] 1 nt deletion and frameshift at position 4359; 2 nt deletion and frameshift from 4381; [f] 1 nt deletion and frameshift at position 4666; 2 nt deletion and frameshift from 4678; [g] 1 nt deletion and frameshift at position 3685 R; 2 nt deletion and frameshift from 3673 R; [h] 1 nt insertion and frameshift at position 540; 1 nt deletion and frameshift at position 586; 1 nt deletion and frameshift at position 608; [i] 6 nt deletion and frameshift from 1896; 1 nt deletion and frameshift at position 1936; 2 nt deletion and frameshift from 1955; [j] 1 nt insertion and frameshift at position 439 R; 4 nt deletion and frameshift from 456 R; [k] 2 nt deletions in CDR1-IMGT and CDR2-IMGT regions and frameshift mutations at 1418 and 1487; 4 nt deletion and frameshift from 1429; 1 nt deletion and frameshift at position 1462; [l] 1st-CYS replaced by Ala

Table 3-3 Torafugu IGLJ nucleotide and amino acid sequences with associated RSS

IGLJ gene	Fct	J-Nonamer	Spacer	J-Heptamer	J region nt and AA sequences
sca2422_J	F	GGTTTTTGT	ACGACCACTTGA TGAGTTTGTAT	CACTGTG	TGGACGTTTGGTGGAGGAACCAAACTCATCATATTC W T F G G G T K L I I F
sca3698_J	F	GGTTTTTGT	ACGACCACTTGA	CACTGTG	TGGACGTTTGGTGGAGGAACCAAACTCATCATATTC W T F G G G T K L I I F

			TGAGTTTGTAT		
sca4248_J	F	GGTTTTTGT	ACGACCACTTGA	CACTGTG	TGGACGTTTGGTGGAGGAACCAAACTCATCGTTTTC W T F G G G T K L I V F
			TGAGTTTGTAT		
sca6782_J	F	GGTTTTTGT	ACGACCACTTGA	CACTGTG	TTGACGTTTGGTGGAGGAACCAAACTCATCGTTGAC L T F G G G T K L I V D
			TGAGTTTGTAT		
sca2115_J	F	GGTTTTTGT	ACGACCACTTGA	CACTGTG	TTCACGTTTGGTGGAGGAACCAAACTCATCGTATTCTGTAAG F T F G G G T K L I V F C K
			TGAGTTTGTAT		
sca4520_J	F	GGTTTTTGT	ACAGCTGTGTGT	CACTGTG	GTATTCGGACCAGGAACCAAGCTGATTGTCACCAGT V F G P G T K L I V T S
			ACAAACTGAAT		
sca4988_J	F	GGTTTTTGT	ACAGCTGTGTGT	CACTGTG	GTATTCGGACCAGGAACCAAGCTGATTGTCACCAGT V F G P G T K L I V T S
			ACAAACTGAAT		
sca7989_J	F	GGTTTTTGT	ACAGCTGTGTGT	CACTGTG	GTATTCGGACCAGGAACCAAGCTGATTGTCGCCAGT V F G P G T K L I V A S
			ACAATCTGAAT		
sca8603_J	P	-	-	CACTGTG	GTATTCGGACCAGGAACCAAGCTGATTGTCACCAGT V F G P G T K L I V T S
sca5604_J	F	GGTTTTTGT	ACAGCTGTGTGT	CACTGTG	GTATTCGGACCAGGAACCAAGCTGATTGTCACCAGT V F G P G T K L I V T S
			ACAAACTGAAT		
sca2126_J	F	GGTTTTTGT	ACAGCTGTGTGT	CACTGTG	GTATTCGGACCAGGAACCAAGCTGATTGTCACCAGT V F G P G T K L I V T S
			ACAAACTGAAT		
sca2352_J	F	GGTTTTTGT	ACAGCTGTGTGT	CACTGTG	GTATTCGGACCAGGAACCAAGCTGATTGTCACCAGT V F G P G T K L I V T S
			ACAAACTGAAT		

sca5821_J	F	GGTTTTTGT	ACAGCTGTGTGT ACAAACTGAAT	CACTGTG	GTATTCGGACCAGGAACCAAGCTGATTGTCACCAGT V F G P G T K L I V T S

3.3.2 Identification of a third IGL isotype in torafugu

The classification of a teleost fish IgL chain is traditionally established through (1) sequence homology/identities, (2) spacing of heptamer and nonamer sequences of VL-RSS and JL-RSS, and (3) gene organization. Among these approaches, CL region homology is the most reliable one. As mentioned, two IGL isotypes have been reported in torafugu: L1 and L2. Here, we used the published IgL sequences from various teleost species to search the torafugu database (http://www.fugu-sg.org/). Three scaffolds (scaffold 2422, scaffold 2488, and scaffold 3698) were found to carry IGLC exons that showed matches (47-53% amino acid identities) from L3 C domains of zebrafish, carp, and catfish; this degree of shared sequences in CL region exceeds what was used (35-37%) to distinguish mammalian κ and λ C regions, thus further strengthens the identification of a torafugu L3. BlAST searches with the IGLV sequences on the three scaffolds revealed similarities to L1/3 V from other teleost fishes. It has been clearly demonstrated that teleost L3 chain is κ ortholog and consistent with the designation, the torafugu L3 RSS adheres to the κ-like RSS with 12-bp spacer of VL-RSS and 23-bp spacer of JL-RSS.

3.3.3 Type 3 IGL organization

The L3 IGL gene organization is shown in Fig. 3-1. There are 5 IGLV gene sequences, 2 IGLJ gene segments, 3 partial and complete CL exons. Because scaffold 2422 has a functional V, a V without leader

sequence, one C and J segment, scaffold 2488 two functional V sequences, one V designated as ORF sequence, one partial C, scaffold 3698 contains one J segment and one C, the heterogeneity suggests an organization of multiple clusters. If a region harboring one C exon is considered as a cluster, three clusters may exist at the L3 locus. The L3 C exons share 48-75% identity with each other at the amino acid level, which, suggesting their divergence from each other but nonetheless is distinguishable from L1/L2 C sequences (10-31% identity in all inter-type pair-wise comparisons). The functional IGLV sequences fall into two groups: V3d, V3e and V3a, V3b, V3c.Within a group, they are 88-91.5% at amino acid level over the IGLV coding sequences; between the two groups, they are 34-42% identical. Of the five IGLV sequences, four are in opposite transcriptional polarity to IGLC and IGLJ segments on individual scaffold, similar to other teleost fish. The V3e segment is in a different transcriptional orientation, since the coding sequences is not outstanding, its position may resulted from a meiotic inversion event.

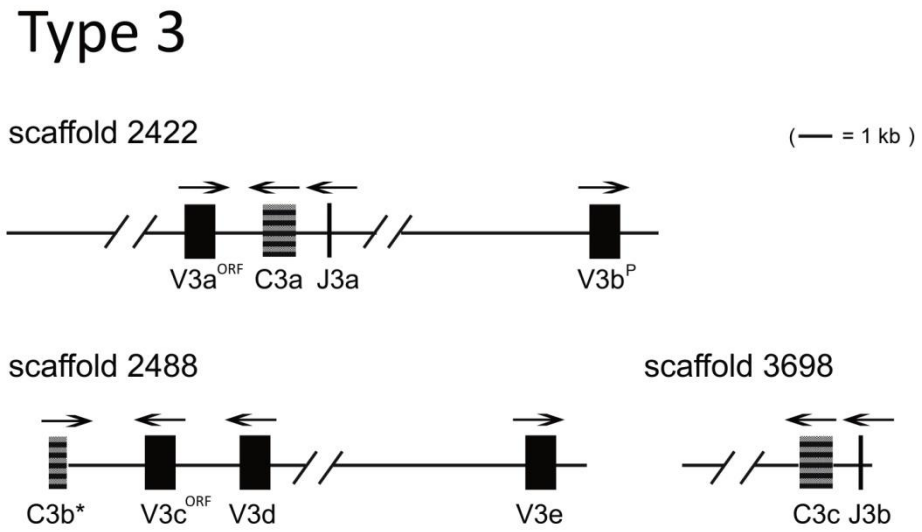


Fig. 3-1 Overall organization of representative *IGL* genes from type 3. Scaffold 2422 of 14,667 bp, 2488 of 13,611 bp, and 3698 of 3784 bp, to scale, with exon size exaggerated. The transcriptional polarity is indicated by overhead arrows. Each gene is labeled, and an asterisk denotes incomplete coding sequence. $V^{P/ORF}$ denotes pseudogene (P) or open reading frame (ORF) sequence.

The leader sequence of V3b is not available from the current database due to regions with gaps, however, it contains a functional RSS and a VL region exon that may potentially able to recombine with the IGLJ with subsequent genome builds additional IGL sequence inserted. IGLV segments on both sides of the J/C will likely undergo rearrangement with the C3a and J3a segments through inversion as in other teleost fishes, the V3b downstream will itself invert to join J3a, while the V3a upstream can recombine by inversion of the J3a and C3a (Fig. 3-2).

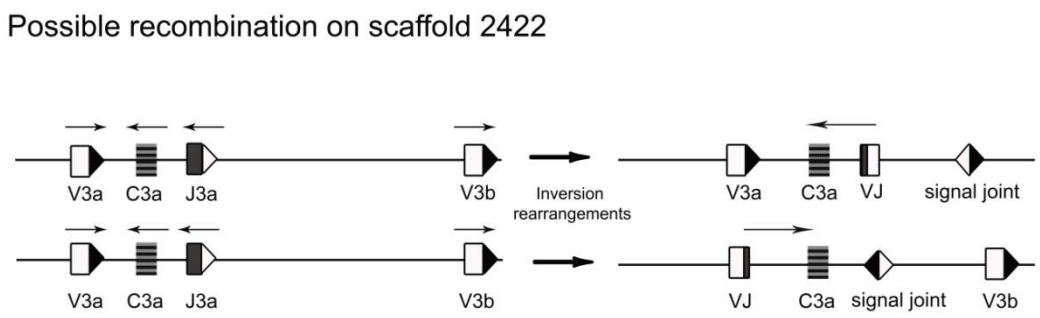


Fig. 3-2 Inversion rearrangements on scaffold 2422. The transcription polarity of the rearranged VJ, at the *right*, is indicated by arrowhead on the top of VJ-C. The J-RSS is indicated as a white triangle, the V-RSS is indicated as a black triangle.

3.3.4 Type 2 IGL organization

A search with L2 C sequences from various teleost fishes showed good matches with ten scaffolds (scaffold 4520, 4988, 5604, 7989, 8603, 2126, 2352, 2681, 3001, and 3330) in the v4 torafugu assembly; other scaffolds were found to contain either L2 IGLV gene segments or IGLJ sequences (not depicted in Fig. XX). The torafugu L2 locus includes 22 sequences matching IGLV gene sequences, 8 IGLJ, and 11 IGLC segments. All the 22 IGLV-matching sequences (some are found only as fragments due to sequences missing in gaps) were characterized and named as included in Table 3-2.

Type 2 IGL locus is depicted in Fig. 3-3. The C2a, C2c, and C2i are identical with the published L2 torafugu C region (Hsu and Criscitiello, 2006). Other L2 C segments (those with complete coding sequences) are 92-99% identical with the C2a in the derived amino acid sequences and share 15-35% identity with L1/L3 C sequences, showing that they duplicated among themselves and diverged long ago from other types. The transcriptional orientations of the V segments within each scaffold are either the

same or opposite to the J and C segments, which is topologically similar to the IGL gene organizations of zebrafish and three-spined stickleback on chromosome 25 and chromosome 11, respectively (Bao, et al., 2010; Zimmerman, et al., 2011; Zimmerman, et al., 2008). However, since all the scaffolds with V segment in the opposite orientation as C and J are missing sequence information between V to J-C (sequences in scaffold 4988, 2352, 2681, and 3001), the possibility that they all contain the *IGL* genes in the same transcriptional direction cannot be ruled out. For example, V2f and V2g in scaffold 2352 are in the opposite orientation to C2h and J2g, but missing segments in gaps, if any, may present additional C and J segments that are in the same direction to V2f and V2g; or with efforts to join scaffolds together, additional V segments may be downstream and in the same orientation of C2h and J2g. The L2 locus is most likely occupied by eleven clusters, and on average one V segment resides in each cluster. Conventional recombination at the L2 locus would occur. For example, rearrangement between V2d and J2d in scaffold 7989 will delete the intervening DNA to form a VJ.

On scaffold 54 and 139, assigned to chromosome 3 and chromosome 5, respectively, only one V segment on each scaffold were detected, which circumvents exact delineation of the L2 gene organization. However, based on the statistically robust phylogenetic groupings of torafugu L2 IGLV sequences, we speculate that scaffolds that habor L2 *IGL* genes, like zebrafish, may also grouped by chromosome and thus should be assigned to either chromosome 3 or chromosome 5.

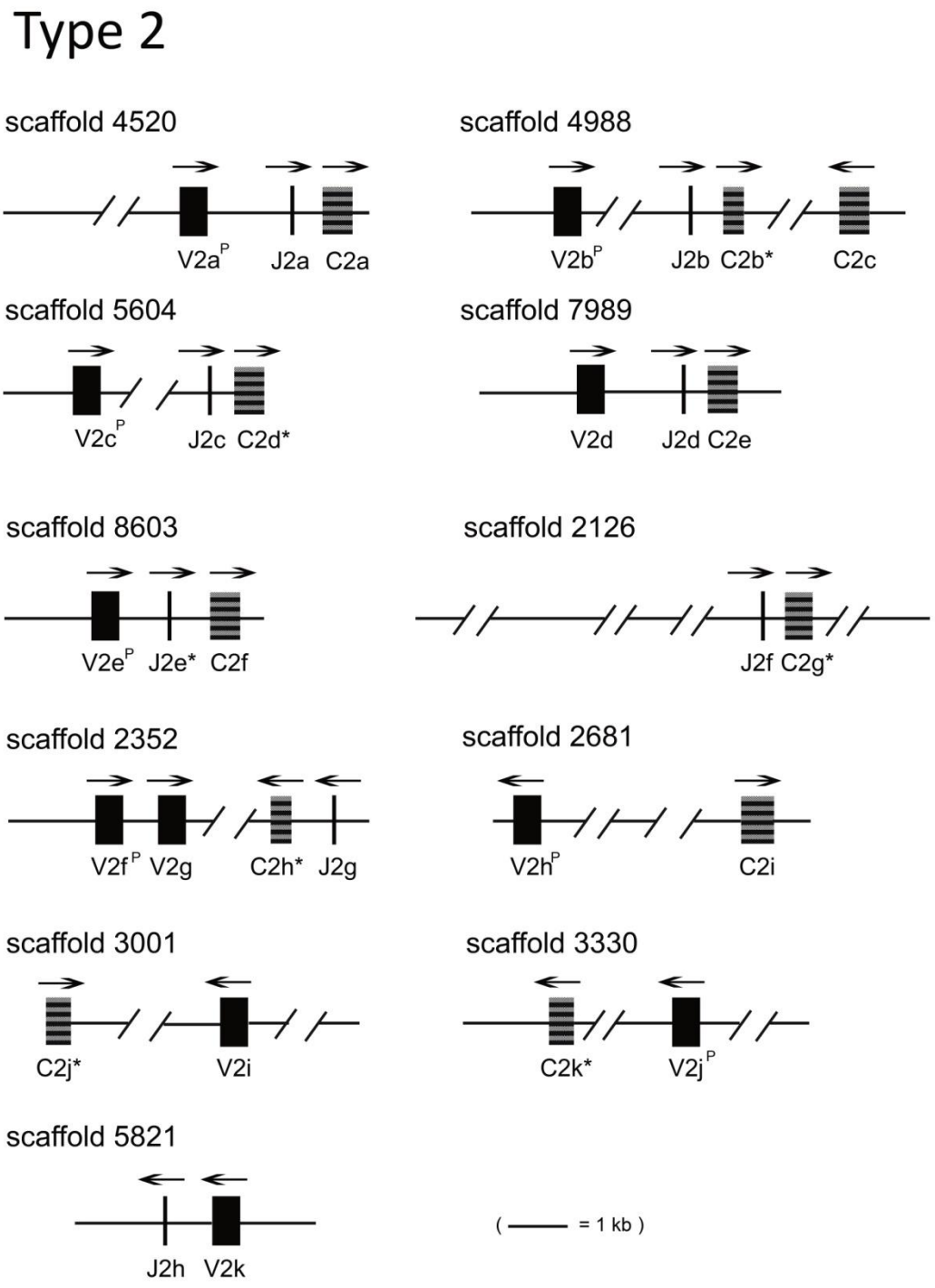


Fig. 3-3 Overall organization of representative type 2 *IGL* genes to scale, with exon size exaggerated. The transcriptional polarity is indicated by overhead arrows. An asterisk denotes incomplete coding sequence. $V^{P/ORF}$ denotes pseudogene or ORF sequence.

3.3.5 Type 1 IGL organization

L1 and L3 V sequences appear to be intermixed (discussed below), we described the L1 genes as on at least seven genomic scaffolds (scaffolds with L1 C sequences), thus the L1 genes might operate as seven loci. As expected, L1 C sequences possess high amino acid identity (≥96%) and the divergence from other types was evident (15-35% identity compared to L2/L3). As depicted in Fig. 3-4, transcriptional polarity pattern in the type 1 locus presents as V in both orientations to J and C. In fact, in all but one instance (chromosome 2), the overall impression is that the transcriptional orientation in type 1 locus is V opposite to nearby J and C. On chromosome 2, four V sequences were identified, with three placed in the same transcriptional orientation to the C and another one in the opposite direction. However, scaffold 158 carries C and V in the opposite orientation, and scaffold 10 has three V sequences in the same orientation. We thus speculate an error in the assembly of chromosome 2 in regards to this region.

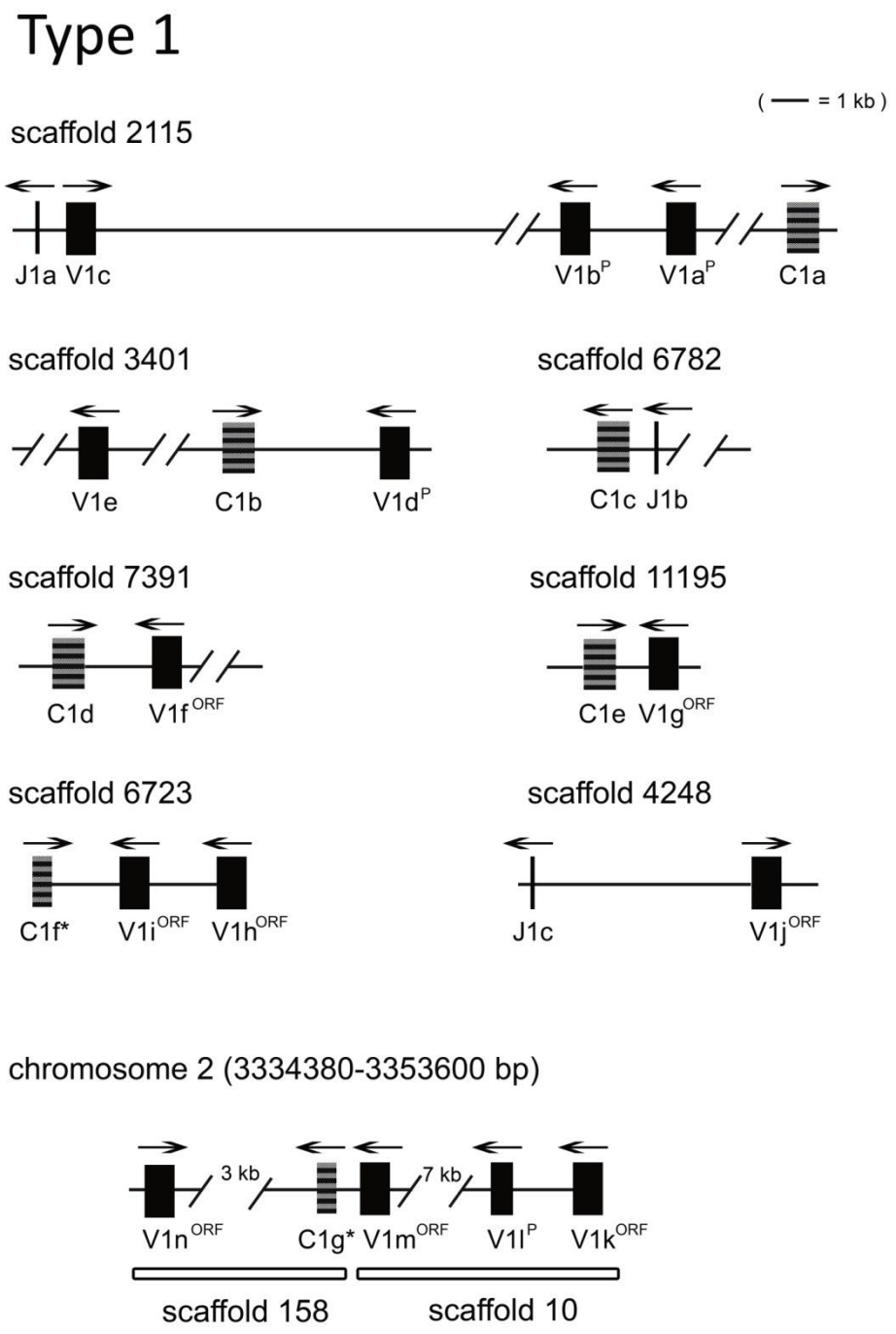


Fig. 3-4 Overall organization of representative type 1 *IGL* genes to scale, with exon size exaggerated. The transcriptional polarity is indicated by overhead arrows. An asterisk denotes incomplete coding sequence. $V^{P/ORF}$ denotes pseudogene or ORF sequence. Scaffold 158 and 10 were assigned to chromosome 2.

3.3.6 Torafugu IGLV segments

Comparisons of the torafugu V segments revealed three distinct groups (designated IGLV1, IGV2, and IGLV3) (Fig. 3-5 and Fig. 3-6). As previously suggested, the alignment of torafugu V sequences showed a conservation of long CDR2 regions in IGLV2 sequences (L2/σ type) compared to other isotypes, while a long CDR1 and a short CDR2 (relative to L2/σ type) in IGLV1 sequences was observed. The one exception is the IGLV3 group, in which the V sequences are with both short CDR1 and short CDR2 regions. We speculate that the missing cysteine which is supposed to be conserved within the FR1 region may contribute to the structure deviation. A phylogenetic tree based on the alignment of V amino acid sequences from various vertebrates was constructed (Fig.3-7). The L2/σ V sequences grouped strongly together and are distinct from the κ group (including teleost L1 and L3 isotypes), which seemed to be mingled. Importantly, torafugu IGLV1 and IGLV3 sequences (κ group) clustered in individual groups, presumably they are associated with different isotypes, as is the case in the stickleback (Bao, et al., 2010).

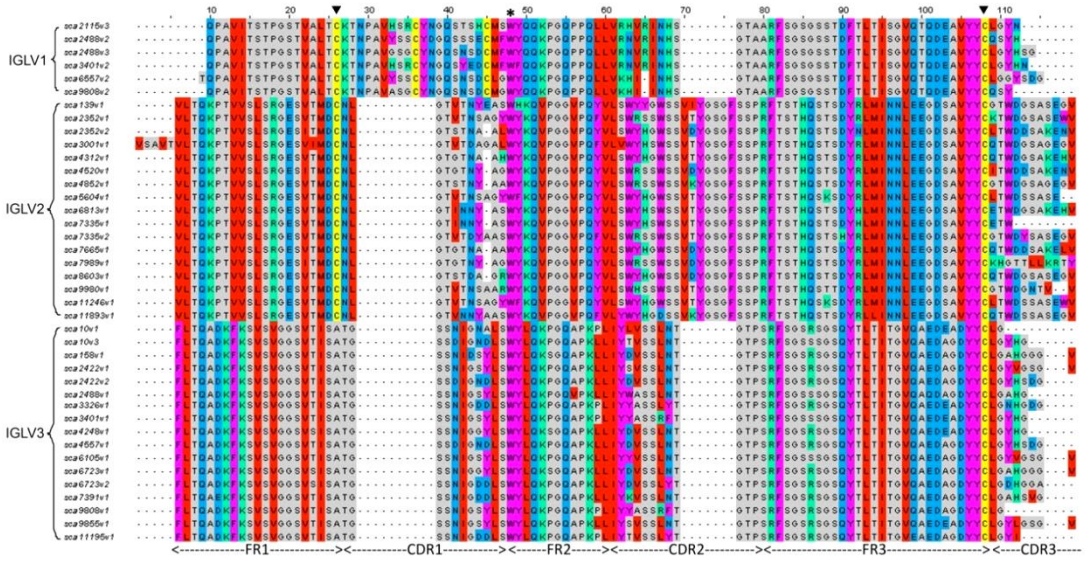


Fig. 3-5 Overview window from Jalview of alignment of torafugu IGLV representative amino acid sequences as determined by MAFFT. Hyphens denote gaps. Framework regions (FR) and complementarity determining regions (CDR) are labeled according to Kabat delineation (Kabat EA, et al., 2001). Conserved Tryptophan in FR2 region is indicated by asterisk symbol. Cysteines that are expected to form the intra-chain disulfide bridges are indicated by solid black triangles, with the exception of IGLV3 family (replaced by Ala).

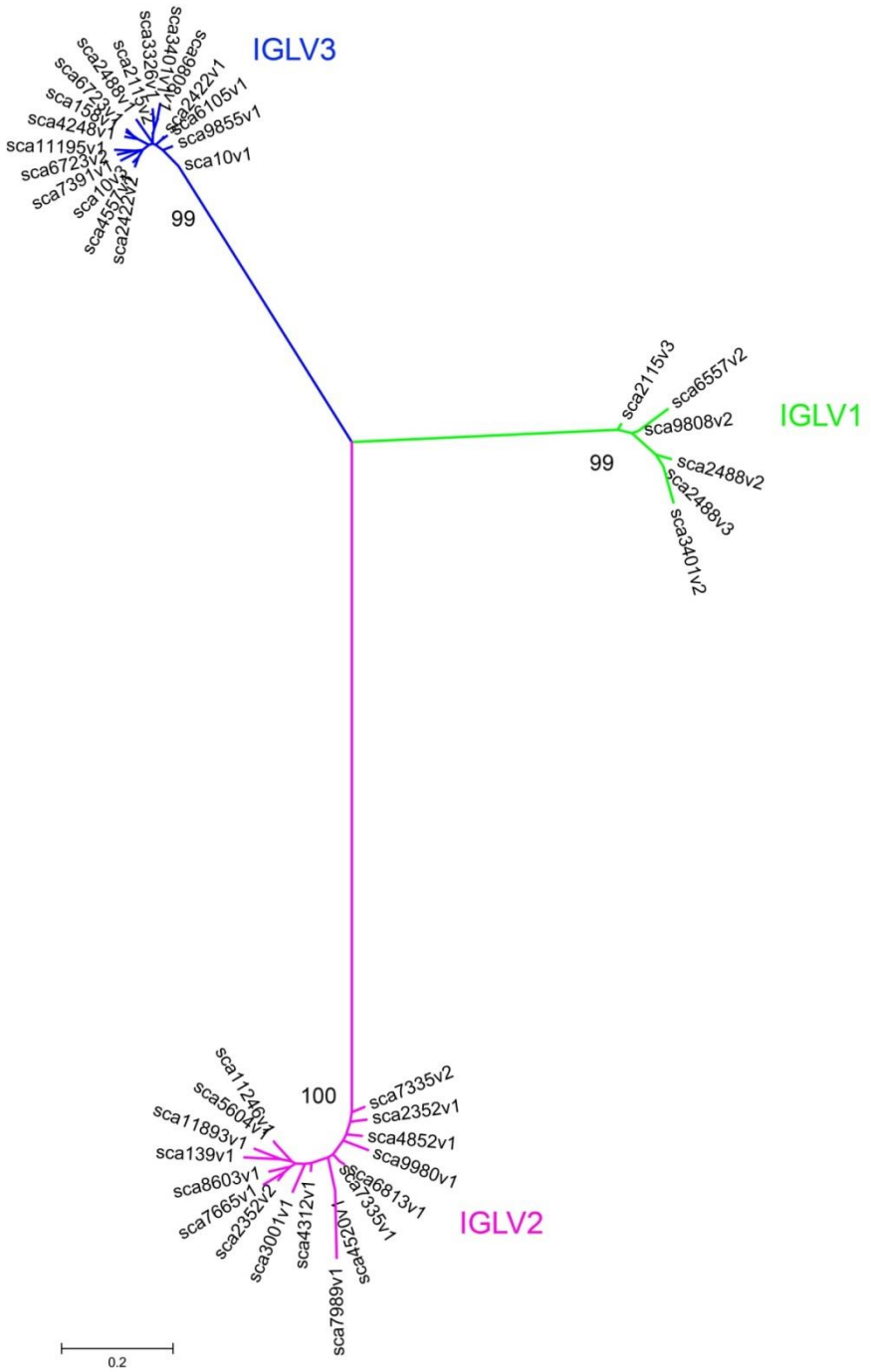


Fig. 3-6 Comparison analysis of torafugu genomic IGLV segments, revealing three distinct IGLV groups. The phylogenetic tree was constructed by MEGA 6 from representative amino acid sequences aligned in Fig. 3-5.

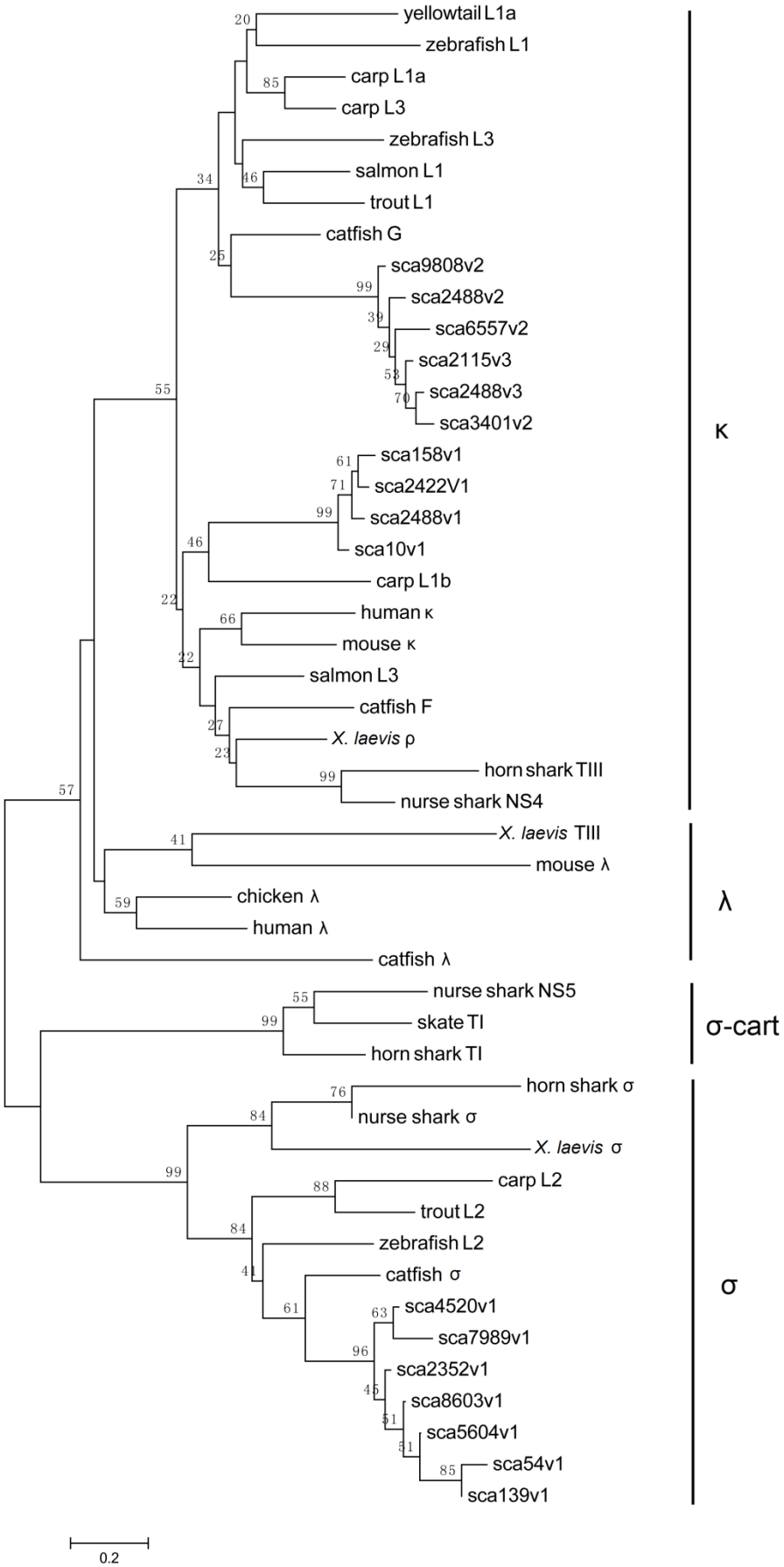

yellowtail L1a
zebrafish L1
carp L1a
carp L3
zebrafish L3
salmon L1
trout L1
catfish G
sca9808v2
sca2488v2
sca6557v2
sca2115v3
sca2488v3
sca3401v2
sca158v1
sca2422V1
sca2488v1
sca10v1
carp L1b
human κ
mouse κ
salmon L3
catfish F
X. laevis ρ
horn shark TIII
nurse shark NS4
X. laevis TIII
mouse λ
chicken λ
human λ
catfish λ
nurse shark NS5
skate TI
horn shark TI
horn shark σ
nurse shark σ
X. laevis σ
carp L2
trout L2
zebrafish L2
catfish σ
sca4520v1
sca7989v1
sca2352v1
sca8603v1
sca5604v1
sca54v1
sca139v1
κ
λ
σ-cart
σ
0.2

Fig. 3-7 Phylogenetic analysis of representative IGLV sequences from various vertebrates. The NJ tree was constructed using MEGA 6 with 1000 bootstrap replications. GenBank accession numbers are: yellowtail L1a (AB062619); zebrafish L1 (AF246185); carp L1a (AB073328); carp L3 (AB073335); zebrafish L3 (AF246193); salmon L1 (AF273012); trout L1 (X65260); catfish G (L25533); carp L1b (AB073332); human κ (S46371); mouse κ (MUSIGKACN); salmon L3 (AF406956); catfish F (U25705); *X. laevis* ρ (XELIGLVAA); horn shark TIII (L25561); nurse shark NS4 (A49633); *X. laevis* TIII (L76575); mouse λ (AY648665); chicken λ (M24403); human λ (AAA59013); catfish λ (EU925383); nurse shark NS5 (AAV34678); skate (*Leucoraja erinacea*) TI (L25568); horn shark TI (X15315); horn shark σ (EF114760); nurse shark σ (EF114765); *X. laevis* σ (S78544); carp L2 (AB091113); trout L2 (AAB41310); zebrafish L2 (AF246162); catfish σ (EU872021).

3.3.7 Torafugu IGLC segments

The classification of a teleost L chain is largely dependent on the C region. The phylogenetic trees of the torafugu C sequences and those from other vertebrates are shown in Fig. 3-8 and Fig. 3-9, respectively. It can be seen that none of the torafugu C cluster with mammalian κ or λ isotypes. They do, however, strongly group in branches where sequences belonging to the same teleost isotype (L1, L2, and L3, respectively), suggesting common derivation among these teleosts and that three or more IGLC may have been present in a teleost ancestor. A close relationship between torafugu (belonging to the Tetradontiformes order, Acanthopterygii superorder), and other species from the Perciformes order (Acanthopterygii), such as seabass (*Dicentrarchus labrax*), rockcod (*Trematomus bernacchii*), and wolffish (*Anarhichas minor*), was also evident in the tree. In addition, the result of phylogenetic analysis

shows that IGLC tend to cluster with taxonomic group than isotype, as suggested in previous studies (Criscitiello and Flajnik, 2007; Rast, et al., 1994).

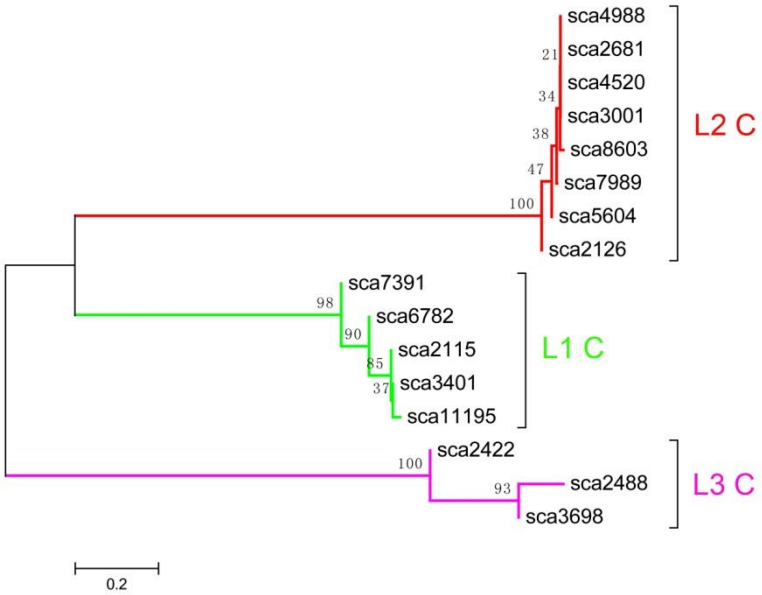


Fig. 3-8 The tree of CL amino acid sequences revealing three distinct groups that correspond to each isotype, respectively.

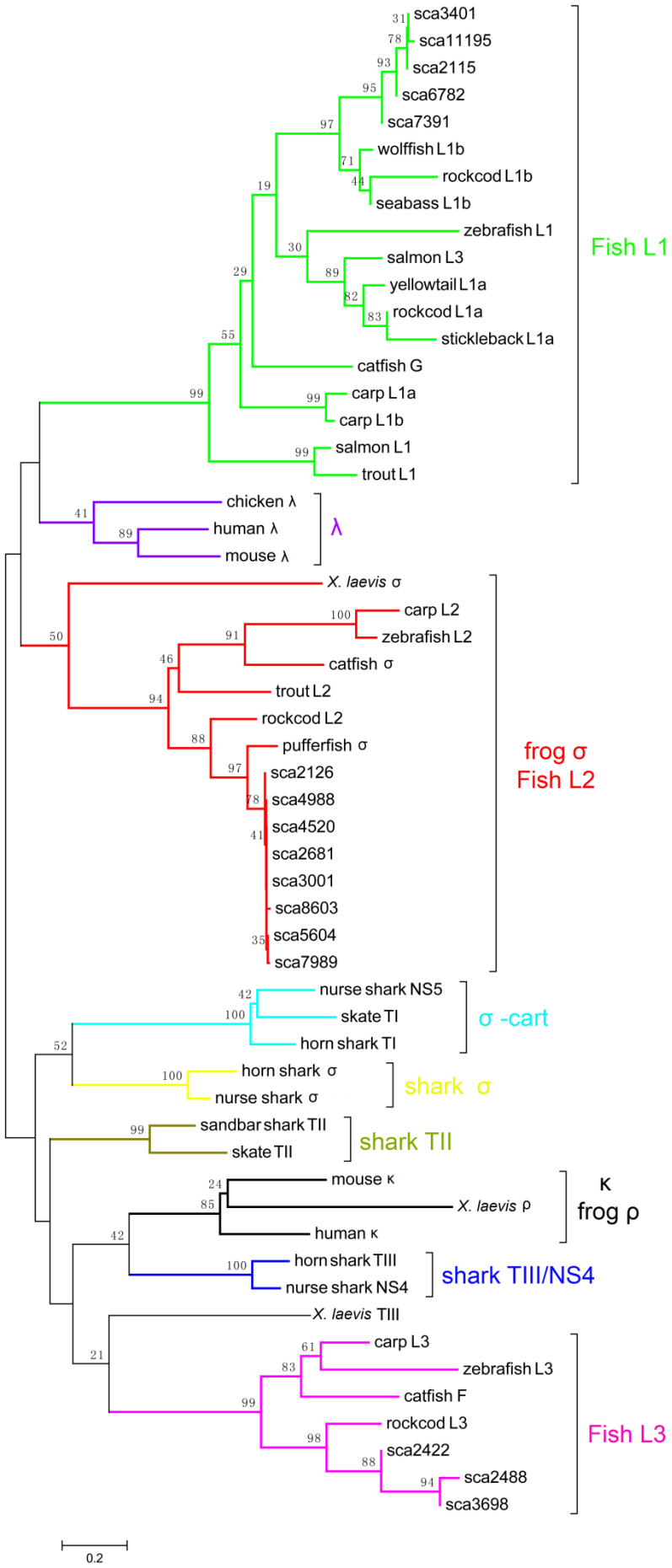

sca3401
sca11195
sca2115
sca6782
sca7391
wolffish L1b
rockcod L1b
seabass L1b
zebrafish L1
salmon L3
yellowtail L1a
rockcod L1a
stickleback L1a
catfish G
carp L1a
carp L1b
salmon L1
trout L1
Fish L1
chicken λ
human λ
mouse λ
λ
X. laevis σ
carp L2
zebrafish L2
catfish σ
trout L2
rockcod L2
pufferfish σ
sca2126
sca4988
sca4520
sca2681
sca3001
sca8603
sca5604
sca7989
frog σ
Fish L2
nurse shark NS5
skate TI
horn shark TI
σ -cart
horn shark σ
nurse shark σ
shark σ
sandbar shark TII
skate TII
shark TII
mouse κ
X. laevis ρ
human κ
κ
frog ρ
horn shark TIII
nurse shark NS4
shark TIII/NS4
X. laevis TIII
carp L3
zebrafish L3
catfish F
rockcod L3
sca2422
sca2488
sca3698
Fish L3
0.2

Fig. 3-9 Phylogenetic analysis of IGLC sequences from various vertebrates. GenBank accession numbers for sequences are as follows: wolffish L1b (AF137398); rockcod L1b (DQ842622); seabass L1b (AJ400216); zebrafish L1 (AF246185); salmon L3 (AF406956); yellowtail L1a (AB062619); rockcod L1a (EF114784); stickleback L1a (AY278356); catfish G (L25533); carp L1a (AB035728); carp L1b (AB035729); salmon L1 (AF273012); trout L1 (X65260); chicken λ (M24403); human λ (AAH07782); mouse λ (J00592); *X. laevis* (*Xenopus laevis*) σ (S78544); carp L2 (AB103558); zebrafish L2 (AF246162); catfish σ (EU872021); trout L2 (AAB41310); rockcod L2 (EF114785); pufferfish (*Tetraodon nigroviridis*) σ (AJ575637); horn shark (*Heterodontus francisci*) TI (X15315); nurse shark (*Ginglymostoma cirratum*) NS5 (AAV34681); skate (*Leucoraja erinacea*) TI (L25568); horn shark σ (EF114760); nurse shark σ (EF114765); sandbar shark (*Carcharhinus plumbeus*) TII (M81314); skate TII (L25566); mouse κ (AB048524); *X. laevis* ρ (XELIGLVAA); human κ (M11937); carp L3 (AB035730); zebrafish L3 (AF246193); catfish F (U25705); rockcod L3(DQ842626).

3.3.8 Functionality of IGL loci

In total, 15 torafugu EST sequences were identified from the NCBI EST database. Alignment of torafugu ESTs to concordant genomic IGLV sequences revealed that all functional IGLV1 group genes are expressed, while only one IGLV2 group gene (*V2k* gene on scaffold 5821) is expressed. Additionally, expression of all the IGLV3 group genes was observed despite they are missing the key amino acid 1st-CYS in FR1 region. Expression of all the complete IGLC segments was also observed with only one exception of the C1d segment on scaffold 7391. In detail, 9 ESTs and 6 ESTs were found to be concordant to L2 locus and L1/L3 loci, respectively. 80% of these ESTs were found to either lack or only

contain a V segment while VJC segments were only present in 3 of the ESTs. The identity of all the retrieved ESTs to genomic V and C segments is 95%-100%, suggesting the feasibility of using this method to assign ESTs to concordant genomic sequences.

3.4 Discussion

In the present study, we have characterized the torafugu IGL locus based on available genome data. Three torafugu L chain isotypes, designated L1, L2, and L3, were identified to array on at least three different chromosomes (v5 assembly) and multiple scaffolds (v4 assembly). During vertebrate phylogeny, *IGL* genes have undergone major evolutionary transitions involving genomic organizations. One extreme example is the presence of a single IGL isotype (λ) in bird species, such as chicken and zebra finch, where a single set of functional VL, JL, and CL that can perform primary rearrangement (Das, et al., 2010; Ruti Parvari, et al., 1987). The mammalian κ and λ loci are often organized in a translocon fasion, for example, human κ and λ contain a large numbers of IGLV in a single cluster per locus, and the mouse λ locus has a small number of IGLV organized in $(VL\text{-}VL\text{-}JL\text{-}C)_2)\text{-}(VL\text{-}(JL\text{-}C)_2)$. Herein, we show that torafugu have a totally different configuration with multiple (VL-JL-C) clusters similar to that found in teleost fishes, suggesting a conservation of the IGL genomic organization among teleost species.

Phylogenetic study of CL sequences showed that torafugu CL shared the same cluster with teleost L1, L2, and L3, respectively. Moreover, a sister-group relationship in the superorder Acanthopterygii between torafugu L1 C sequences and those of the L1b subgroup (wolffish L1b, seabass L1b, and rockcod L1b) supported by high bootstraps values was observed. At this time, we did not find a L1a C

sequence homolog in torafugu, but if in the future such sequences are found, this would further support the hypothesis that L1a and L1b subtypes exist in the Acanthopterygii L1 isotype (Coscia, et al., 2008). Teleost L1 and L3 were previously suggested to share a common origin. The identification of a L3 in torafugu (Acanthopterygii), together with the presence of L3 in rockcod (Acanthopterygii), and in Ostariophysi (catfish, zebrafish, and carp), suggesting the divergence between L1 and L3 took place at or before the emergence of Euteleosts (Hsu and Criscitiello, 2006).

chapter 4:
Profiling the IgH repertoire in torafugu by massively parallel sequencing

4.1 Introduction

NGS has established itself as a highly useful platform in studying several aspects of genomics research. Previously intractable problems can be solved because of its high efficiency, accuracy and cost-effectiveness. One such area is the expressed Ig repertoire. Compared with Sanger sequencing, Ig-seq can provide invaluable information on the diversity and composition of Ig repertoire. For example, prior to NGS, four decades of research provided < 30,000 human VH, Vκ, and Vλ sequences, less than a single NGS run, as reported in human (Glanville, et al., 2009) and zebrafish (Weinstein, et al., 2009). When we try and get the best coverage of the Ig repertoire, we actually aim to sequence as many Ig sequences as possible. Therefore, smaller model organisms that contain fewer cells in total and, of course, fewer immune cells, can provide a better starting point from which to maximize sequenced Ig sequences . Torafugu have the earliest recognizable AIS and more importantly one of the smallest genomes (~400 Mb) among vertebrates, which, makes them an ideal model system for studying the Ig repertoire.

4.2 Materials and methods

4.2.1 Torafugu and Total RNA preparation

Three adult torafugu weighing between 800 and 900g were obtained from Fish Interior (Tokyo, Japan). Fish were maintained in tanks with aerated seawater and maintained at 20°C. Adult fish were euthanized, followed by rapid dissection of tissues. Spleen and trunk kidney were collected and directly fixed in RNAlater (Applied Biosystems). Total RNA was extracted from individual spleens and trunk kidneys using RNeasy Lipid Tissue Mini Kit (Qiagen).

4.2.2 Primer design

The torafugu IGH locus was described in our previous study (Fu, et al., 2015). The consensus leader sequences for 32 potentially functional IGHV gene segments of the torafugu were used to design the 5 forward primers (as part of a family). The reverse primers were derived from the first exon of Cμ and the second exon of Cτ. In order to obtain cleaner products, a second, independent primer set (set number 2) was designed. The forward primers of set 2 utilized the consensus frame region 1(FR1) sequences for each IGHV family. And the reverse primers were based on nested Cμ and Cτ sequences; these were located 3-nucleotides upstream from the first round Cμ-specific PCR primer and in the first exon of Cτ. Gene specific primers were also designed for the reverse transcription step (Table 4-1).

Table 4-1 Primer sets for repertoire study

Primer	Name	V gene segment	Sequence
set1	FVH1	1S1,1S2,1S3,1S4,1S7,1S8,1S12,1S13orf, 1S14,1S15,1S16,1S17,1S18,1S21	GGACAGGACTGCTGCTTCTAAC
	FVH2-1	2S1,2S2,2S6,2S7,2S8,2S9,2S10,2S11, 2S12,2S15,2S16,2S17,2S18,2S20	AGCTCTGCTGCTGCTGTTG
	FVH2-2	2S19	TTCTCTGCAGCTGTGGTGC
	FVH3-1	3S2	CAGAGGTTTACTGATCATTGTC
	FVH3-2	3S3orf	TCTTCAGTGTCTGGTGGACG

	GSP-μ	C-μ	AGGGCTACCGTCCCAGTCCTGT
	GSP-τ	C-τ	GTGATCAGACACACAAGAGTGACG
	VhCm1	C-μ	TGCCGTTCATGGTTGGAGGGT
	VhCt2	C-τ	GCTGATCATGTCTTTCTCTGGCG
set2	nFVH1	1S1,1S2,1S3,1S4,1S7,1S8,1S12,1S13orf, 1S14,1S15,1S16,1S17,1S18,1S21	CTGACCCAGTCTGAACCAGT
	nFVH2	2S1,2S2,2S6,2S7,2S8,2S9,2S10,2S11, 2S12,2S15,2S16,2S17,2S18,2S19,2S20	TGAACAGTTGACACAGCCAGC
	nFVH3	3S2,3S3orf	GCCTGAAGTAAAAAGACCTGGA
	nVhCm1	C-μ	CGTTCATGGTTGGAGGGTAC
	nVhCt1	C-τ	TCTGGGAAGAAGTCGAGAGC

4.2.3 cDNA synthesis and multiplex PCR amplification

First-strand cDNA was synthesized using SuperScript® III reverse transcriptase (Invitrogen). Total RNA purified from each fish was split into 4 cDNA synthesis reactions with both the primers for IgM and IgT constant regions. RNase H (Invitrogen) was added to each reaction to remove RNA at the end of the cDNA synthesis step. All enzyme concentrations, reaction volumes and incubation temperature were according to the manufacturer's protocol. Each 20-μl cDNA synthesis reaction was split into two PCR reactions, and a total of 8 PCR reactions were set up for each fish. Each of the 5 forward primers was added to represent each IGHV segment in a final concentration of 300nM primer. Some primers cover multiple IGHV segments, thus their concentration was in proportion to the number of IGHV segments.

Both reverse primers were added at a final concentration of 10μM. Two microliters of the RT reaction subsequently served as template for the PCR amplification by Platinum® Taq DNA Polymerase High Fidelity (Invitrogen) with an initial denaturation of 2 min at 94°C, followed by 30 cycles of denaturation at 94°C for 30 s, annealing of primer to DNA at 55°C for 30 s, and extension at 68°C for 1 min, plus a final extension for 5 min at 68°C. The reaction products were purified using QIAquick PCR Purification Kit (Qiagen). A second-round PCR was performed on 2-μl of the first-round reaction with proportional nested primers as described in the previous PCR. Reaction conditions were as follows: 94°C for 2 min, followed by 28 cycles of 94°C for 30 s, 55°C for 30 s, 68°C for 1 min, and a final 5 min extension at 68°C. The product of the second-round PCR was purified as described above, and the concentration was measured using a Qubit® 2.0 fluorometer (Life Technologies). The size distribution of PCR products was determined using the Agilent 2200 Tape Station (Agilent Technologies).

4.2.4 Amplicon library construction

About 1μg of QIAquick cleaned PCR product for each fish was used to start the Illumina library preparation process. Illumina TruSeq® DNA PCR-free sample preparation protocol was followed for all samples without fragmentation. Briefly, double stranded DNA was end repaired and ligated to sequencing adaptors for hybridization onto a flow cell. The DNA library templates were then quantified by Library Quant Illumina Kit (KAPA Biosystems) with standards in a range from 0.2fM to 20pM using a 7300 real-time PCR cycler (Applied Biosystems) (Fig. 4-1). Finally, an equal amount of quantified cDNA (10ng) from each of the 3 libraries, corresponding to the 3 fish samples, was pooled to obtain the final amplicon library, which represents the complete collection of IGH transcripts from the 3 torafugu.

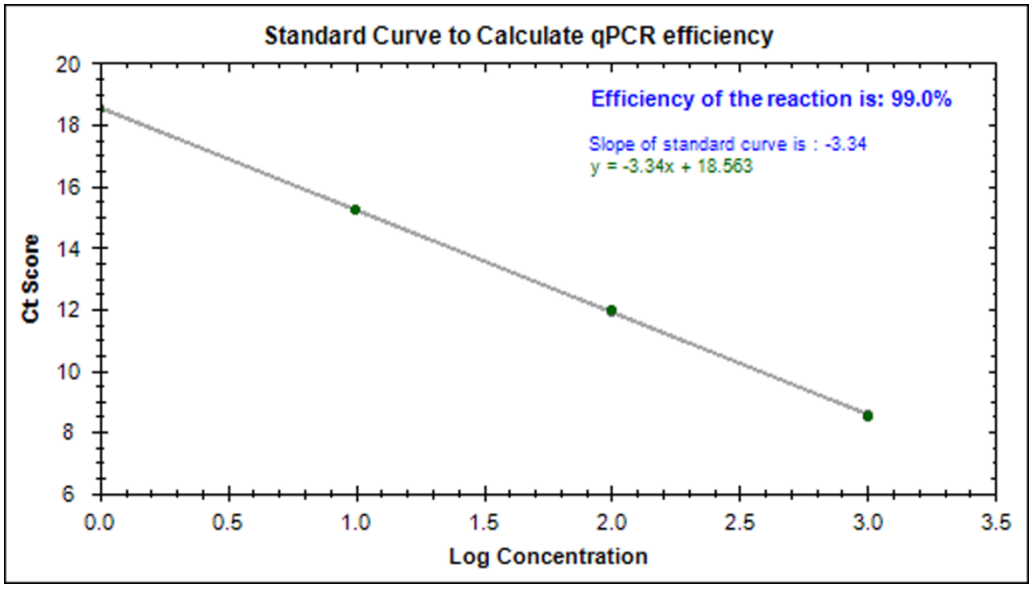


Fig. 4-1 Determine the concentration of ligated DNA library templates by KAPA Biosystems

4.2.5 Miseq run and data analysis

The library was prepared as recommended by Illumina (Miseq Reagent Kit v3) and was loaded at 12pM on one Miseq flowcell. Paired end sequencing was performed with 301 cycles. For the generation of consensus reads, raw reads with Illumina quality scores less than 20 and length below 150bp were filtered using the Trimmomatic tool (Bolger, et al., 2014). The Paired-End reAd mergeR (PEAR) tool (http://sco.h-its.org/exelixis/web/software/pear/doc.html) (Zhang, et al., 2014) was then used in the generation of consensus reads (with parameter -v 15, other parameters with default setting).

Consensus reads were split into different C-region groups (IgM or IgT) by matching the portion of the CH1 upstream of the reverse primers. These reads were then filtered for a minimum length of 300bp. All the resultant consensus sequences were aligned first to germline V segment to determine the optimal alignment and then aligned to all J segments to generate corresponds to genomic V-J (Fu, et al., 2015)

with match gain of +3, mismatch cost of -3, gap-open cost of -8, and gap-extend cost of -8. IMGT/HighV-QUEST software (Alamyar, et al., 2012) was used to filt sequences with stop codons. All the resultant consensus sequences were aligned first to germline V segment to find the optimal alignment and then aligned to all J segments to determine corresponding genomic V-J. The results were summarized and were subsequently used for D assignment. Briefly, the VJ-positive sequences were first translated into amino acid sequences, aligned by MUSCLE, and finally CDR3 sequences were manually determined by the presence of the second conserved cysteine (C) in YYC motif by the 3′ portion of the VH gene segment and the conserved tryptophan (W) in WGxG motif by the 5′ portion of the JH gene segment. The number of nucleotides between these codons determines the length and thus the frame of the CDR3 region. In some cases, sequences that are not with the conserved C or W amino acids but have identifiable CDR3 region were still selected since we are trying to provide a broad picture of the variability of CDR3 sequences.

4.3 Results

4.3.1 Sequences

The Miseq run generated a total of 18 million paired end (PE) reads (5,881,846 - 6,199,423 reads per sample), that passed initial sequence filters. Reads entered into the analysis after being identically matched to the corresponding forward and reverse primers (IgM or IgT). After the merger process which generated consensus reads and final quality filtering which discarded low-quality consensus reads, ~1,000,000 high-quality consensus sequences (947,833-1,188,573 sequences) per sample were obtained (Table 4-2). Each consensus sequence represents an original IGH molecule.

Table 4-2 Summary of Miseq sequence reads assigned for the primer sets and 3 torafugu samples

Torafugu	Total reads	Primer containing reads	Consensus reads	Total IgM reads	Identifiable VJm	Total IgT reads
T1	5,881,846	2,640,816	947,833	277,261	22,924	446,502
T2	6,199,423	2,715,689	955,601	346,194	34,071	400,515
T3	5,936,417	2,816,232	1,188,573	334,734	77,409	587,831

4.3.2 IGHV and IGHJ usage

Between 947,833 and 1,188,573 useful consensus reads were obtained per fish, and we focused our analysis on IgM, which is the most abundant species and CDR3 sequences, which is the most diverse component in the antigen binding region; IgT data were analyzed for the usage of VH gene segment. Consensus reads were grouped into IgM (277,261-346,194) or IgT (400,515-587,831), each read was then assigned VH (IgM and IgT groups) and JH (only IgM group) by alignment to a local reference. The DH gene segment is too short to reliably assign and is contained in CDR3 variation.

The profiles of the frequencies of sequences associated with each IGHV and IGHJ were compared across the three samples (Fig. 4-2). We focused on sequences that had identifiable V and J. For IgM, this subset corresponded to 8-23% of the total IgM sequences across different torafugu. Of 134,404 VJ sequences, the VH sequences are not equally used; a preference for the IGHV1 family sequences is

evident, with the VH1.13 occurs most frequently. The data also shows that IgM barely use the IGHV2 sequences with only 5 instances observed.

All known, functional Jμ (Jm) gene segments are represented in the subset of 134,404 identifiable VJm sequences. Usage ranged from 63% for Jm1 to 0.5% for Jm5 (Fig. 4-3).

IGHV gene segments associated with Cτ (IgT) occasionally use IGHV1 family sequences; the preference for the usage within IGHV2 family sequences shows an individually specific way (Fig. 4-4). For example, three instances were observed where rearrangement took place differentially across fish samples: VH2.12 are used most frequently in T3 sample (63%), VH2.1 with 33% usage in T2 sample (6% and 9% in T1 and T3, respectively), VH2.18 with 41% usage in T1 sample (0.23% and 0.16% in T2 and T3, respectively).

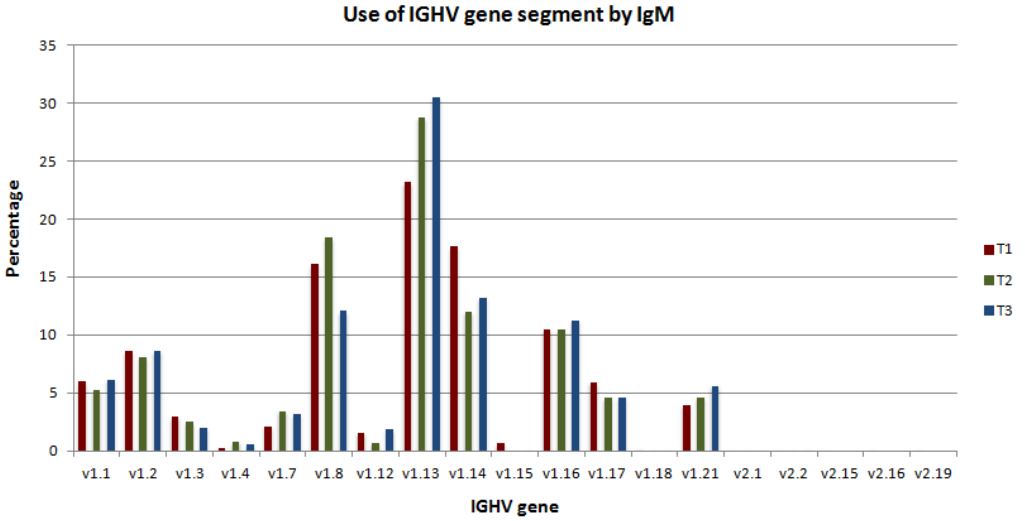


Fig. 4-2 Profiles for IGHV gene usage of IgM across torafugu RNA samples. The percentage occurrence of sequences derived from the indicated *IGHV* genes relative to the total identifiable VJm sequences, is shown for the three torafugu (T1, T2, and T3).

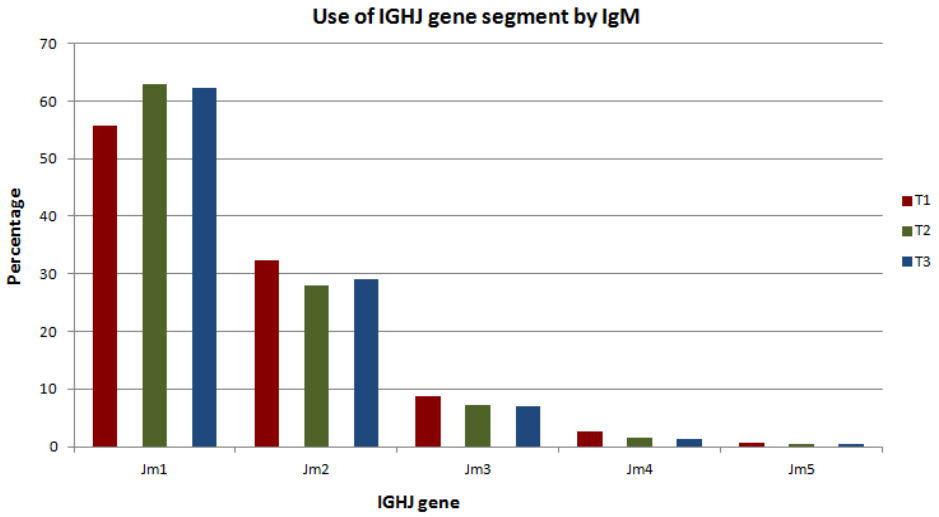


Fig. 4-3 Jμ (Jm) gene segment usage by IgM across the three torafugu (T1, T2, and T3).

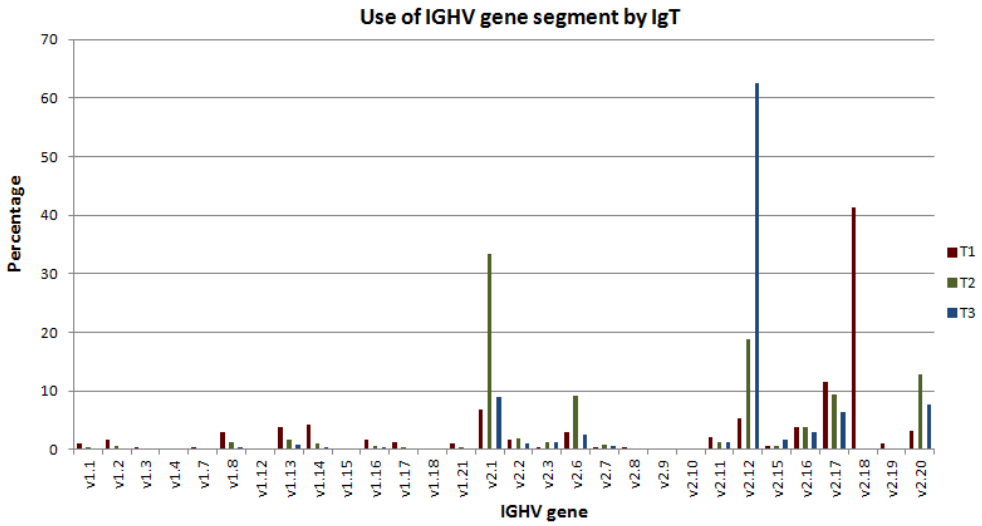


Fig. 4-4 Profiles for IGHV gene usage of IgT across torafugu RNA samples. The percentage occurrence of sequences derived from the indicated *IGHV* genes relative to the total identifiable VJm sequences, is shown for the three torafugu (T1, T2, and T3).

4.3.3 Sequence diversity in CDR3 region

In order to compare sequences, we manually defined Ig H chain CDR3 coordinates as starting at the codon for the last cysteine of IGHV and ending at the tryptophan in the conserved IGHJ segment motif WGxG. The length of CDR3 varies from 19 to 69 nt with a peak at 43 nt (Fig. 4-5, *top*). We find no evidence of overrepresented sequence in the center nucleic acid logo of CDR3 (Fig. 4-5, *bottom*), which indicates the diverse nature and coding potential of IGHD gene segments. The left and right ends of the logos are apparently conserved, suggesting the contributions of IGHV and IGHJ sequences, respectively.

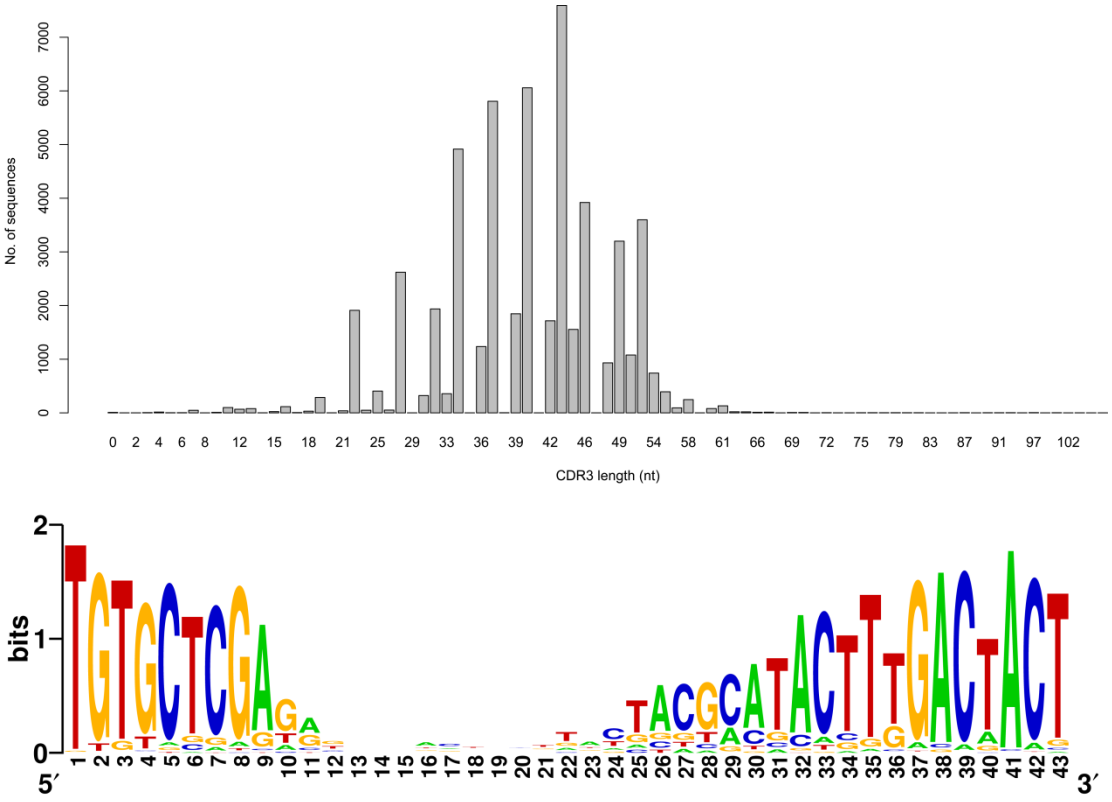


Fig. 4-5 CDR3 nucleotide length distribution (*top*) and sequence composition of the most abundant CDR3 length (*bottom*). The most frequently observed length was 43 nt. For the subset of identifiable

VJ sequences with 43 nt CDR3 sequences, we created logo for the nucleotide composition, using WebLogo (Crooks GE, et al., 2004).

4.4 Discussion

To evaluate the diverse nature of abundant Igs in the torafugu, we applied NGS to H chain that was PCR amplified from RNA of enriched spleen and kidney from non-immunized torafugu. We have identified 134,404 torafugu CDR3 sequences. Analysis of the data from this study has provided information concerning the extent of CDR3 length diversity and on many fundamental IgH repertoires such as preferences for gene segment usage. As expected from previous studies, we see that certain *IGHV* and *IGHJ* genes are commonly used while others are quite rare. For the *IGHV* genes, sequences related to IgM and IgT, are found associated with each specific C region in different ways: a preference for the IGHV1 family genes used by IgM, while IGHV2 family genes are used by IgT, in most cases. The pairing of IGHJ and IGHV is not random: Jm gene segments are represented in all the subset of identifiable VJm sequences, with a clearly bias utilization reaching to 63% (Jm1). The reasons for such bias are not well understood but are likely due to a preferential combination of certain IGHV and IGHJ and RSS compatibilities that affect initial BCR development.

chapter 5:
General Discussion

The ability of the AIS to respond to any of the vast number of potential foreign antigens relies largely on the highly diverse polymorphic receptors expressed by B cells (Igs). Igs are formed by a mixture of recombination among gene segments, sequence diversification in the junction part of these gene segments, and adjusted point mutations throughout the gene. Sequencing of genomes from almost every major class of vertebrate has greatly furthered the understanding of the diversity and evolutionary origins of Igs. Torafugu is a good model organism for comparative genome studies. The organization of mammalian Ig H and L chain gene loci are well characterized, but information about the Ig loci in lower vertebrate such as torafugu is fragmentary. The sequencing of the torafugu genome provides an opportunity for analyzing its Ig loci. Moreover, elementary questions concerning the Ig diversity have remained open: It is unclear what fraction of the repertoire is expressed in an individual and how similar repertoires are shared by individuals live in similar environments. Torafugu have the earliest recognizable AIS and more importantly one of the smallest genomes (~400 Mb) among vertebrates, which, makes them an ideal model system for studying the Ig repertoire. In this study, a step-wise approach was followed to investigate the genomic features of torafugu Ig genes.

A detailed analysis of the complete torafugu IGH locus (Chapter 2) showed an expansion of *IGH* genes compared to previous studies. In total, 48 *IGHV* genes, 7 *IGHD* genes, and 6 *IGHJ* genes are spread out over the IGH variable region. The IGH locus is located in a 115 kb region of a single chromosome (chromosome 5) and possesses over 40 VH segments which could be classified into 5 families. Among these VH segment families, three were newly identified (IGHV3–IGHV5) and the expression of IGHV3 family VH sequences (functional group) was confirmed by reverse transcription experiment. Our results indicate that the expansion version of the IGH locus in torafugu contributes heavily to the increases in diversity of the Ig repertoire.

Identification of a more expanded IGHδ locus in this study may suggest a high degree of variation in the number of CH domain exons in the *δ* gene. It is clear that most, if not all, teleost *δ* genes have seven unique CH exons (Cδ1–Cδ7) and it seems that torafugu *δ* retains these CH exons even after structural alteration. Torafugu *δ* was reported to be transcribed as a chimeric molecule with Cμ1 spliced to Cδ1. The presence of three exons (Cδ4– Cδ6) upstream of the Cδ1 exon raises questions regarding the expression of torafugu IgD. Alternative splicing patterns of IgD may therefore exist, as revealed in the zebrafish.

The torafugu IGL gene loci were annotated in the present study (Chapter 3). Three torafugu L chain isotypes, designated L1, L2, and L3, were identified to array on at least three different chromosomes (v5 assembly) and multiple scaffolds (v4 assembly). We show that torafugu IGL has a configuration with multiple (VL-JL-C) clusters similar to that found in other teleost fishes, suggesting a conservation of the IGL genomic organization among teleost species. Identification of a third IgL isotype has provided new evidence for the divergence of teleost L1 and L3, which further improved knowledge on the origin of teleost L chain types.

In the Ig repertoire study (Chapter 4), we applied NGS to characterize the torafugu Ig H chain and focused our analysis on the CDR3 sequences, the most diverse component responsible for antigen binding. We have identified 134,404 torafugu CDR3 sequences. Analysis of the data from this study has provided information concerning the extent of CDR3 length diversity and on many fundamental IgH repertoires such as preferences for gene segment usage. We observed that certain *IGHV* and *IGHJ* genes are commonly used while others are quite rare. The reasons for such bias are not well understood but are likely due to a preferential combination of certain IGHV and IGHJ and RSS compatibilities that affect initial BCR development.

References

Alamyar, E., *et al.* (2012) IMGT/HighV-QUEST: the IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing., *Immunome Res.*, **8**, 26.

Alberts B, Johnson A and J, L. (2002) The Adaptive Immune System. *Molecular Biology of the Cell. 4th edition.* Garland Science, New York

Allen, E., *et al.* (2003) High concentrations of long interspersed nuclear element sequence distinguish monoallelically expressed genes, *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 9940-9945.

Aparicio, S., *et al.* (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes, *Science*, **297**, 1301-1310.

Bao, Y., *et al.* (2010) The immunoglobulin gene loci in the teleost Gasterosteus aculeatus, *Fish Shellfish Immunol.*, **28**, 40-48.

Bengtén, E., *et al.* (2006) Structure of the catfish IGH locus: analysis of the region including the single functional IGHM gene, *Immunogenetics*, **58**, 831-844.

Bengtén, E., *et al.* (2002) The IgH Locus of the Channel Catfish, Ictalurus punctatus, Contains Multiple Constant Region Gene Sequences: Different Genes Encode Heavy Chains of Membrane and Secreted IgD, *The Journal of Immunology*, **169**, 2488-2497.

Bengtén, E., *et al.* (2000) Immunoglobulin Isotypes: Structure, Function, and Genetics. In Du Pasquier, L. and Litman, G. (eds), *Origin and Evolution of the Vertebrate Immune System*. Springer Berlin Heidelberg, pp. 189-219.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics*.

Brochet, X., Lefranc, M.P. and Giudicelli, V. (2008) IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis, *Nucleic Acids Res.*, **36**, W503-508.

Brodeur, P.H. and Riblet, R. (1984) The immunoglobulin heavy chain variable region (Igh-V) locus in the mouse. I. One hundred Igh-V genes comprise seven families of homologous genes, *Eur. J. Immunol.*, **14**, 922-930.

Carver, T., *et al.* (2008) Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database, *Bioinformatics*, **24**, 2672-2676.

Chen, T. and Manuelidis, L. (1989) SINEs and LINEs cluster in distinct DNA fragments of Giemsa band size, *Chromosoma*, **98**, 309-316.

Clark, M.S., *et al.* (2003) Fugu ESTs: new resources for transcription analysis and genome annotation, *Genome Res.*, **13**, 2747-2753.

Coscia, M.R., *et al.* (2008) Immunoglobulin light chain isotypes in the teleost Trematomus bernacchii, *Mol. Immunol.*, **45**, 3096-3106.

Criscitiello, M.F. and Flajnik, M.F. (2007) Four primordial immunoglobulin light chain isotypes, including lambda and kappa, identified in the most primitive living jawed vertebrates, *Eur. J. Immunol.*, **37**, 2683-2694.

Crooks GE, *et al.* (2004) WebLogo: A sequence logo generator., *Genome Res.*, **14**, 1188-1190.

Daggfeldt, A., Bengtén, E. and Pilström, L. (1993) A cluster type organization of the loci of the immunoglobulin light chain in Atlantic cod (Gadus morhua L.) and rainbowttrout (Oncorhynchus mykiss Walbaum) indicated by nucleotide sequences of cDNAs and hybridization analysis, *Immunogenetics*, **38**, 199-209.

Danilova, N., *et al.* (2005) The immunoglobulin heavy-chain locus in zebrafish: identification and expression of a previously unknown isotype, immunoglobulin Z, *Nat. Immunol.*, **6**, 295-302.

Das, S., *et al.* (2010) Analysis of the immunoglobulin light chain genes in zebra finch: evolutionary implications, *Mol. Biol. Evol.*, **27**, 113-120.

Das, S., *et al.* (2008) Evolutionary redefinition of immunoglobulin light chain isotypes in tetrapods using molecular markers, *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 16647-16652.

Das, S., *et al.* (2008) Evolutionary dynamics of the immunoglobulin heavy chain variable region genes in vertebrates, *Immunogenetics*, **60**, 47-55.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.*, **32**, 1792-1797.

Edholm, E.S., Wilson, M. and Bengten, E. (2011) Immunoglobulin light (IgL) chains in ectothermic vertebrates, *Dev. Comp. Immunol.*, **35**, 906-915.

Edholm, E.S., *et al.* (2009) Identification of Igsigma and Iglambda in channel catfish, Ictalurus

punctatus, and Iglambda in Atlantic cod, Gadus morhua, *Immunogenetics*, **61**, 353-370.

Fillatreau, S., *et al.* (2013) The astonishing diversity of Ig classes and B cell repertoires in teleost fish, *Front. Immunol.*, **4**, 28.

Flajnik, M.F. (2002) Comparative analyses of immunoglobulin genes: surprises and portents, *Nat. Rev. Immunol.*, **2**, 688-698.

Flajnik, M.F. (2005) The last flag unfurled? A new immunoglobulin isotype in fish expressed in early development, *Nat. Immunol.*, **6**, 229-230.

Flajnik, M.F. and Kasahara, M. (2010) Origin and evolution of the adaptive immune system: genetic events and selective pressures, *Nature reviews. Genetics*, **11**, 47-59.

Fu, X., *et al.* (2015) Characterization of the torafugu (Takifugu rubripes) immunoglobulin heavy chain gene locus, *Immunogenetics*, **67**, 179-193.

Gambon-Deza, F., Sanchez-Espinel, C. and Magadan-Mompo, S. (2010) 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, *Dev. Comp. Immunol.*, **34**, 114-122.

Georgiou, G., *et al.* (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire, *Nat. Biotechnol.*, **32**, 158-168.

Ghaffari, S.H. and Lobb, C.J. (1997) Structure and genomic organization of a second class of immunoglobulin light chain genes in the channel catfish, *J. Immunol.*, **159**.

Glanville, J., *et al.* (2009) Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire, *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 20216-20221.

Hansen, J.D., Landis, E.D. and Phillips, R.B. (2005) Discovery of a unique Ig heavy-chain isotype (IgT) in rainbow trout: Implications for a distinctive B cell developmental pathway in teleost fish, *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 6919-6924.

Hendricks, J., *et al.* (2010) Organization of the variable region of the immunoglobulin heavy-chain gene locus of the rat, *Immunogenetics*, **62**, 479-486.

Hikima, J., Jung, T.S. and Aoki, T. (2011) Immunoglobulin genes and their transcriptional control in teleosts, *Dev. Comp. Immunol.*, **35**, 924-936.

Hsu, E. and Criscitiello, M.F. (2006) Diverse Immunoglobulin Light Chain Organizations in Fish

Retain Potential to Revise B Cell Receptor Specificities, *The Journal of Immunology*, **177**, 2452-2462.

Ippolito, G.C., *et al.* (2006) Forced usage of positively charged amino acids in immunoglobulin CDR-H3 impairs B cell development and antibody production, *The Journal of Experimental Medicine*, **203**, 1567-1578.

Jung, D., *et al.* (2006) Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus, *Annu. Rev. Immunol.*, **24**, 541-570.

Kabat EA, *et al.* (2001) *Sequences of proteins of immunological interest.* National Institutes of Health, Bethesda, MD.

Kai, W., *et al.* (2011) Integration of the genetic map and genome assembly of fugu facilitates insights into distinct features of genome evolution in teleosts and mammals, *Genome Biol. Evol.*, **3**, 424-442.

Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability, *Mol. Biol. Evol.*, **30**, 772-780.

Kielbasa, S.M., *et al.* (2011) Adaptive seeds tame genomic sequence comparison, *Genome Res.*, **21**, 487-493.

Krangel, M.S., *et al.* (2004) Enforcing order within a complex locus: current perspectives on the control of V(D)J recombination at the murine T-cell receptor α/δ locus, *Immunol. Rev.*, **200**, 224-232.

Lefranc, M.-P. (2007) WHO-IUIS Nomenclature Subcommittee for immunoglobulins and T cell receptors report, *Immunogenetics*, **59**, 899-902.

Lefranc, M.-P., *et al.* (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains, *Dev. Comp. Immunol.*, **27**, 55-77.

Litman, G.W. (1999) Evolution of antigen binding receptors, *Annu. Rev. Immunol.*, **17**, 109-147.

Mansilla-Soto, J. and Cortes, P. (2003) VDJ Recombination: Artemis and Its In Vivo Role in Hairpin Opening, *J. Exp. Med.*, **197**, 543-547.

Mutoloki, S., Jørgensen, J.B. and Evensen, Ø. (2014) The Adaptive Immune Response in Fish. In, *Fish Vaccination*. John Wiley & Sons, Ltd, pp. 104-115.

Nei, M. (2007) The new mutation theory of phenotypic evolution, *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 12235-12242.

Nei, M. and Rooney, A.P. (2005) Concerted and Birth-and-Death Evolution of Multigene Families, *Annu. Rev. Genet.*, **39**, 121-152.

Ota, T. and Nei, M. (1994) Divergent Evolution and Evolution by the Birth-and-Death Process in the Immunoglobulin VH Gene Family, *Mol. Biol. Evol.*, **11**, 469-482.

Peixoto, B.R. and Brenner, S. (2000) Characterization of approximately 50 kb of the immunoglobulin *VH* locus of the Japanese pufferfish, *Fugu rubripes*, *Immunogenetics*, **51**, 443-451.

Qin, T., *et al.* (2008) Genomic organization of the immunoglobulin light chain gene loci in Xenopus tropicalis: evolutionary implications, *Dev. Comp. Immunol.*, **32**, 156-165.

Rast, J.P., *et al.* (1994) Immunoglobulin light chain class multiplicity and alternative organizational forms in early vertebrate phylogeny, *Immunogenetics*, **40**, 83-99.

Ruti Parvari, *et al.* (1987) Analyses of chicken immunoglobulin light chain cDNA clones indicate a few germline Vλ genes and allotypes of the Cλ locus, *The EMBO Journal* **6**, 97-102.

S.Brenner, *et al.* (1993) Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome, *Nature*, **336**, 265-268.

Saha, N.R., *et al.* (2004) Fugu immunoglobulin D: a highly unusual gene with unprecedented duplications in its constant region, *Immunogenetics*, **56**, 438-447.

Saha, N.R., Suetake, H. and Suzuki, Y. (2004) Characterization and expression of the immunoglobulin light chain in the fugu: evidence of a solitaire type, *Immunogenetics*, **56**, 47-55.

Saha, N.R., Suetake, H. and Suzuki, Y. (2005) Analysis and characterization of the expression of the secretory and membrane forms of IgM heavy chains in the pufferfish, Takifugu rubripes, *Mol. Immunol.*, **42**, 113-124.

Savan, R., *et al.* (2005) Discovery of a novel immunoglobulin heavy chain gene chimera from common carp (Cyprinus carpio L.), *Immunogenetics*, **57**, 458-463.

Savan, R., *et al.* (2005) Discovery of a new class of immunoglobulin heavy chain from fugu, *Eur. J. Immunol.*, **35**, 3320-3331.

Schatz, D.G. and Ji, Y. (2011) Recombination centres and the orchestration of V(D)J recombination, *Nat. Rev. Immunol.*, **11**, 251-263.

Schroeder Jr, H.W. and Cavacini, L. (2010) Structure and function of immunoglobulins, *J. Allergy Clin. Immunol.*, **125**, S41-S52.

Stavnezer, J. and Amemiya, C.T. (2004) Evolution of isotype switching, *Semin. Immunol.*, **16**, 257-275.

Tamura, K., *et al.* (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods, *Mol. Biol. Evol.*, **28**, 2731-2739.

Tamura, K., *et al.* (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0, *Mol. Biol. Evol.*, **30**, 2725-2729.

Tarailo-Graovac, M. and Chen, N. (2009) Using RepeatMasker to identify repetitive elements in genomic sequences, *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, **Chapter 4**, Unit 4 10.

Timmusk, S., Partula, S. and Pilström, L. (2000) Different genomic organization and expression of immunoglobulin light-chain isotypes in the rainbow trout, *Immunogenetics*, **51**, 905-914.

Warr, G.W. (1997) The adaptive immune system of fish, *Dev. Biol. Stand.*, **90**, 15-21.

Weinstein, J.A., *et al.* (2009) High-throughput sequencing of the zebrafish antibody repertoire, *Science*, **324**, 807-810.

Wu, T.T. and Kabat, E.A. (1970) An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity., *J. Exp. Med.*, **132**, 211-250.

Xu, J.L. and Davis, M.M. (2000) Diversity in the CDR3 Region of VH Is Sufficient for Most Antibody Specificities, *Immunity*, **13**, 37-45.

Yasuike, M., *et al.* (2010) Evolution of duplicated IgH loci in Atlantic salmon, Salmo salar, *BMC Genomics*, **11**, 486.

Zhang, H., *et al.* (2013) Assessment of homozygosity levels in the mito-gynogenetic torafugu (Takifugu rubripes) by genome-wide SNP analyses, *Aquaculture*, **380-383**, 114-119.

Zhang, H., *et al.* (2014) Dramatic improvement in genome assembly achieved using doubled-haploid genomes, *Sci. Rep.*, **4**, 6780.

Zhang, J., *et al.* (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR, *Bioinformatics*, **30**, 614-620.

Zhang, Y.A., *et al.* (2010) IgT, a primitive immunoglobulin class specialized in mucosal immunity, *Nat. Immunol.*, **11**, 827-835.

Zimmerman, A.M., *et al.* (2011) Zebrafish immunoglobulin IgD: unusual exon usage and quantitative expression profiles with IgM and IgZ/T heavy chain isotypes, *Mol. Immunol.*, **48**, 2220-2223.

Zimmerman, A.M., Romanowski, K.E. and Maddox, B.J. (2011) Targeted annotation of immunoglobulin light chain (IgL) genes in zebrafish from BAC clones reveals kappa-like recombining/deleting elements within IgL constant regions, *Fish Shellfish Immunol.*, **31**, 697-703.

Zimmerman, A.M., *et al.* (2008) Immunoglobulin light chain (IgL) genes in zebrafish: Genomic configurations and inversional rearrangements between (V(L)-J(L)-C(L)) gene clusters, *Dev. Comp. Immunol.*, **32**, 421-434.