論 文 の 内 容 の 要 旨

水圏生物科学専攻

平成２４年度博士課程入学

氏　　名　付　希

指導教員名　浅川　修一

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）

Adaptive immune system (AIS) is fascinating with its specific yet incredibly diverse ability to fight pathogens and has a memory. Potent AIS is fundamentally reliant on the generation of a diverse repertoire of B-cell antigen receptors (BCRs, or immunoglobulins (Igs)). A classic Ig comprises two identical heavy (H) chains and two identical light (L) chains, each of which containing one variable (V) domain and one (in the L chain) or more (in the H chain) constant (C) domains. Antigen binding is performed by using both the H and L chain V domains. 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 have annotated both the torafugu Ig H and L chain genes, based on available torafugu genome assemblies. The present study also took advantage of next-generation sequencing (NGS) capturing relevant Ig coding region sequences to characterize the Ig repertoire of torafugu.

1. Organization of the torafugu immunoglobulin heavy chain gene locus

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 against all the genome databases. This search resulted in the identification of chromosome 5 from the current assembly and three genomic scaffolds (scaffold 287, 483, and 1358) from our database. Some of the gaps in chromosome 5 were closed by the identified scaffolds. The remaining gaps in the IGH gene locus were resolved by chain-termination sequencing using primers specific for flanking region of each gap for polymerase chain reaction (PCR) amplification. 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), three new IGHV gene families (IGHV3–IGHV5) were discovered. We observed the interspersion of IGHV1 and IGHV2 family members and that they often intermingled with each other, while other family members were further interspersed. 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. *IGHC* genes have also been mapped in this study, with 3 genes encoding immunoglobulin classes: IgT, IgM, and IgD.

The rearrangements of *IGH* genes were also investigated here. Briefly, total RNA was extracted and reverse transcribed. Equal amounts of each cDNA were combined and the mixture used as PCR template. The consensus leader sequences for functional IGHV gene segments were used to design the forward primers (as part of a family), and the reverse primers were derived from the C region. PCR products were cloned using TOPO® TA Cloning Kit for sequencing and twelve clones from each positive PCR product were picked and plasmids were sequenced. The cDNAs sequences were BLAST against the IGHV and IGHJ gene sequences to identify their presence in the clones. As a result, we confirmed the expression of newly identified IGHV3 family sequences in adult torafugu. A favorable IGHV segment usage by IgM and IgT was observed, that is, the IGHV1 family genes were tend to be used by Cμ2 and Cτ2, ~70% of the time. Possible structural variation in the IGHδ locus was observed based on the current torafugu assembly, wherein a more expanded number of Cδ exons compared to previous report were found. 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 or was the result of an error in the assembly of the torafugu v5 assembly remains undetermined.

2. Analysis of the immunoglobulin light chain genes in torafugu

Genome builds of torafugu (assembly v4, October 2004 and assembly v5, January 2010) available from Fugu Genome Project were searched to locate *IGL* genes. Published IgL sequences from torafugu and other teleost fishes 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. In total, 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.

Comparisons of the torafugu V segments revealed three distinct groups (designated IGLV1, IGV2, and IGLV3). The type 2 (L2) V sequences grouped strongly together and are distinct from the κ group (teleost type 1 (L1) /type 3 (L3)), which seem to be mingled. The phylogenetic tree of the torafugu C sequences and those from other vertebrates showed 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).

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. L1 and L2 have been reported in torafugu. Here, three scaffolds were found to carry IGLC exons that showed matches (47-53% amino acid identities) with L3 C domains of zebrafish, carp, and catfish; this degree of shared sequences in CL region exceeds limit used (35-37%) to distinguish mammalian κ and λ C regions, and further strengthens the identification of a torafugu L3. BlAST searches with the IGLV sequences on the three scaffolds revealed similarities to the type1/3 V from other teleost fishes. The L3 RSS identified here adheres to the κ-like RSS with 12-bp spacer of VL-RSS and 23-bp spacer of JL-RSS.

Torafugu IGL had previously been showed to locate to 4 scaffolds in L1 locus and 2 clones in L2 locus, respectively. Here, we provide an extended maps of torafugu L1 and L2 loci. A BLAST search with L2 C sequences from various teleost fishes showed high homology with 10 scaffolds in the v4 torafugu assembly. Collectively, the torafugu L2 locus includes 22 sequences matching IGLV segments, 8 IGLJ, and 11 IGLC segments scattered through 21 scaffolds. The transcriptional orientations of the V segments within each scaffold are either the same or opposite to the J and C segments. The L1 genes are located on at least 7 genomic scaffolds (scaffolds with L1 C sequences) and they might operate as seven loci. The overall picture is that the transcriptional orientation in L1 locus is V opposite to nearby J and C.

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. In total, 15 torafugu EST sequences were identified. Alignment of these ESTs to concordant genomic IGLV sequences revealed that all possibly functional IGLV1 group genes were expressed, while only one IGLV2 group gene was shown to express. Additionally, expression of all the IGLV3 group genes was observed although they are missing the key amino acid 1st-CYS in FR1 region. In detail, 9 ESTs and 6 ESTs were found to be concordant to L2 locus and L1/L3 loci, respectively.

3. Profiling the IgH repertoire in torafugu by massively parallel sequencing

Based on the characterized IGH locus, we designed specific primers to isolate the relevant IgH sequences to profile the Ig repertoire of torafugu. In the present study, we focused on analyzing the complimentarity-determining region (CDR3) of the H chain, because it is the most diverse component in terms of length and sequence of the H chain repertoire and is a major determinant of Ig specificity.

We performed deep sequencing of amplicon library pooled from 3 torafugu, which represents the complete collection of IGH transcripts of the 3 samples using Illumina NGS platform (Miseq). Paired end sequencing was performed with 301 cycles. The Miseq generated ~18 million paired end reads (5,881,846 - 6,199,423 reads per sample). Reads entered into the analysis after being identically matched to the corresponding forward and reverse primers (IgM or IgT). 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. Each consensus sequence represents an original IGH molecule. Consensus reads were divided into different C-region groups (IgM or IgT) by matching the portion of the CH1 upstream of the reverse primers. 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 last conserved Cysteine (C) in YYC motif at the 3′ of *IGHV* gene and the first conserved Tryptophan (W) in WGxG motif at the 5′ of *IGHJ* gene (in some cases, sequences that are without the conserved C or W and with identifiable CDR3 region were also included since we are trying to provide a broader picture of the variability of CDR3 sequence). 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 (Jμ1>Jμ2> Jμ3> Jμ4> Jμ5). The length of CDR3 varies from 19 to 69 nt with a peak at 43 nt, which is shorter than that in humans.

4. Conclusion

The present study provides detailed annotations of both the Ig H and L chain gene locus in torafugu. The characterization of IGH locus showed expansion of *IGHV* genes and *IGHD* genes than demonstrated in previous studies. The expression of newly identified IGHV3 family sequences was confirmed in rearrangement study. The torafugu *IGL* genes are organized in multiple clusters and located in at least three chromosomes. Sequence and phylogenetic analyses revealed three IGL isotypes, L1, L2, and L3. We used high-throughput sequencing of the V domain of the Ig H chain from 3 torafugu to analyze VDJ usage and Ig sequence variability. Torafugu were found to use preferentially between different IGHV segments by certain Ig isotypes. A precise measurement of CDR3 length diversity was also performed. This study provides insights into the genetic basis of diversity in the antibody response of torafugu as well as the breadth of expressed Ig repertoire.