frontiers in Immunology

ORIGINAL RESEARCH
published: 21 February 2018
doi: 10.3389/fimmu.2018.00251

High-Throughput Sequencing of the Expressed Torafugu (*Takifugu rubripes*) Antibody Sequences Distinguishes IgM and IgT Repertoires and Reveals Evidence of Convergent Evolution

*Xi Fu[1,2†], Jianqiang Sun[3†], Engkong Tan[2], Kentaro Shimizu[3], Md Shaheed Reza[4,5], Shugo Watabe[4] and Shuichi Asakawa[2]**

[1] State Key Laboratory of Biotherapy, West China Hospital, Collaborative Innovation Center and Sichuan University, Chengdu, China, [2] Laboratory of Aquatic Molecular Biology and Biotechnology, Department of Aquatic Bioscience, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo, Japan, [3] Bioinformational Engineering Laboratory, Department of Biotechnology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo, Japan, [4] School of Marine Biosciences, Kitasato University, Kanagawa, Japan, [5] Department of Fisheries Technology, Bangladesh Agricultural University, Mymensingh, Bangladesh

OPEN ACCESS

Edited by:
Victoriano Mulero,
Universidad de Murcia, Spain

Reviewed by:
Yuko Ota,
University of Maryland, Baltimore, United States
Pierre Boudinot,
Institut National de la Recherche Agronomique, France

***Correspondence:**
Shuichi Asakawa
asakawa@mail.ecc.u-tokyo.ac.jp

[†]These authors have contributed equally to this work.

Specialty section:
This article was submitted to Comparative Immunology, a section of the journal Frontiers in Immunology

Received: 06 November 2017
Accepted: 29 January 2018
Published: 21 February 2018

Citation:
Fu X, Sun J, Tan E, Shimizu K, Reza MS, Watabe S and Asakawa S (2018) High-Throughput Sequencing of the Expressed Torafugu (Takifugu rubripes) Antibody Sequences Distinguishes IgM and IgT Repertoires and Reveals Evidence of Convergent Evolution. Front. Immunol. 9:251. doi: 10.3389/fimmu.2018.00251

B-cell antigen receptor (BCR) or antibody diversity arises from somatic recombination of immunoglobulin (Ig) gene segments and is concentrated within the Ig heavy (H) chain complementarity-determining region 3 (CDR-H3). We performed high-throughput sequencing of the expressed antibody heavy-chain repertoire from adult torafugu. We found that torafugu use between 70 and 82% of all possible V (variable), D (diversity), and J (joining) gene segment combinations and that they share a similar frequency distribution of these VDJ combinations. The CDR-H3 sequence repertoire observed in individuals is biased with the preferential use of a small number of VDJ, dominated by sequences containing inserted nucleotides. We uncovered the common CDR-H3 amino-acid (aa) sequences shared by individuals. Common CDR-H3 sequences feature highly convergent nucleic-acid recombination compared with private ones. Finally, we observed differences in repertoires between IgM and IgT, including the unequal usage frequencies of V gene segment and the biased number of nucleotide insertion/deletion at VDJ junction regions that leads to distinct distributions of CDR-H3 lengths.

Keywords: antibody repertoire, heavy (H) chain complementarity-determining region 3, teleost fish, convergent evolution, IgM, IgT

INTRODUCTION

The adaptive immune system (AIS) is fundamentally reliant on a highly diverse set of antigen receptors. These receptors are generated through somatic recombination of tandemly arranged variable (V), diversity (D), and joining (J) segments of the B-cell antigen receptor [BCR, the membrane-bound form of antibodies or immunoglobulins (Igs)] and T-cell receptor (TCR) genes, and the insertion and deletion of nucleotides (nts) at the junctions between ligated segments (1). The variety of generated antibody or Ig repertoires is required to recognize and bind various antigens (pathogens).

The majority of diversity in the naïve Ig repertoire lies within the heavy (H) chain complementarity-determining region 3 (CDR-H3), which consists of the VDJ recombination junctions, and thus is the most diverse component and a major determinant of Ig specificity (2, 3).

Estimates of Ig diversity have been previously surveyed in several species (4–9). For the human BCRs, for example, the potential diversity of Ig molecules is estimated to be $>10^{13}$ (9), while this number exceeds the total number of B cells in the human body (approximately $1–2 \times 10^{11}$) (10). This excess of potential Ig diversity leads to the expectation that different individuals could seldom share the same segment rearrangement. Nevertheless, several studies have demonstrated overlap among the Ig repertoires of different individuals occurring, i.e., in the naïve human and zebrafish IgH repertoires (8, 11), and in the B-cell responses to virus infection in the rainbow trout (12).

Teleost fish are the most primitive bony vertebrates that contain Igs (13). Like humans, teleost fish have Ig gene rearrangement, junctional diversity during recombination, and somatic hypermutation (13, 14). It is also well known that smaller model organisms that contain fewer cells in total and obviously fewer immune cells can provide a better starting point from which to obtain a better coverage of the immune repertoire (15). In this regard, we chose to characterize the antibody repertoire of torafugu (*Takifugu rubripes*), a species that possesses the shortest genome of any known vertebrate, yet contains gene characteristics similar to those in the human genome (16).

We developed an analytical tool (PyDAIR, https://github.com/biunit/PyDAIR) to investigate the quiescent state of the torafugu immune system. Using the Illumina Miseq high-throughput sequencing platform allowed 12 million expressed antibody sequences from three healthy adult torafugu. We found evidence that the repertoire of individual fish is highly biased toward sequences with specific VDJ combinations, in addition to the convergent evolution of CDR-H3 sequences in both IgM and IgT. We observed distinct repertoires for IgM and IgT, respectively, which may suggest different mechanisms of diversity creation in fish naïve immune system.

RESULTS

We characterized the expressed IgH repertoire from healthy adult torafugu. To this end, we used a high-throughput sequencing approach that determines CDR-H3 profiles for both IgM and IgT, based on PCRs between Ig V and C domains. More than 8 million CDR-H3 amino-acid (aa) sequences were captured in our data set. We found that on an average, 180,062 and 24,933 distinct CDR-H3 clusters were presented in the naïve IgM and IgT repertoires, respectively. Each consensus (merged) read was assigned to a V and J gene by alignment to the germline references with success rates of ~94% for IgM and of ~98% for IgT (**Table 1**). Identifiable VJ reads were used for subsequent analysis. On an average, Dμ (Dm) segments were assigned to ~14% of the VJm reads and Dτ (Dt) to ~40% of VJt reads (**Table 1**); many of the unidentifiable cases had D regions deleted.

While the use of a PCR step before sequencing could potentially introduce bias in the inferred relative abundance of the sequences, due to differences in the efficiency of PCR amplification using different sets of primers. It is noteworthy that in the present study, the estimation of relative usage of V, D, and J gene segments was based on the hypothesis that there is no significant difference of PCR efficiency between primer combinations.

TABLE 1 | Summary of consensus reads assigned to each V, D, and J gene segment measured for both IgM and IgT groups.

Fish	1	2	3
Total IgM reads	1,798,717	2,337,015	2,029,082
Identifiable VJm	1,684,225	2,215,821	1,915,237
Identifiable VDmJm	251,746	320,638	244,172
VJm assignment rate	0.936	0.948	0.943
VDmJm assignment rate	0.149	0.144	0.127
Total IgT reads	2,160,668	1,516,736	1,905,817
Identifiable VJt	2,096,641	1,483,036	1,876,739
Identifiable VDtJt	1,462,606	545,230	306,943
VJt assignment rate	0.970	0.977	0.984
VDtJt assignment rate	0.697	0.367	0.163

Highly Biased VDJ Usage in the Naïve IgM and IgT Repertoires

IgM

Focused on the naïve IgM repertoire, the results showed that frequencies at which specific V, D, and J gene segments were used were roughly consistent among individuals. Of the 32 potentially functional V genes, we found that 30 V genes from the three known IGHV groups were used by IgM, suggesting that most of the potential IgM repertoire was expressed in healthy fish. A preference for the usage of IGHV1 family sequences was evident, with V1.13, V1.14, and V1.8 occurring most frequently. The top five V genes accounted for over 65% of all VJm reads. Jm usage could vary by more than 100-fold, with Jm1 and Jm2 being favored, ranging from 55% for Jm1 to 0.6% for Jm5 (average values across individuals). Due to substantial base deletion and overall transformation, D gene segments were usually unrecognizable without ambiguity. Sequences where Dm could be identified unambiguously and accurately (with minimum length = 4 nucleotide (nt)) represent 816,556 out of the 5,815,283 VJm sequences. For these 816,556 sequences, a similar trend in Dm usage was observed in individual fish, with Dm1 showing the highest frequency (**Figure 1A**).

Theoretically, there are 960 possible VDJ combinations in torafugu ($32 \text{ V} \times 6 \text{ Dm} \times 5 \text{ Jm} = 960$ VDJ). The collection of all VDJ combinations is deemed as the VDJ repertoire. In the present study, a total of 788 VDJ combinations were captured. Individual torafugu VDJ repertoires could be visualized in 3D graphical form (**Figure 2**). The combination coverage in any captured torafugu ranges from 70 to 82%, showing similar tendencies compared with the previous observation in zebrafish (8). Most VDJ combinations showed low abundance; however, a similarly small fraction—albeit showing different ranking for specific combinations in individual fish—were observed at high frequencies. Consistently, there was a strong overlap among the top 10 (accounting for an average of over 45% of frequency in all captured combinations) VDJ combinations in

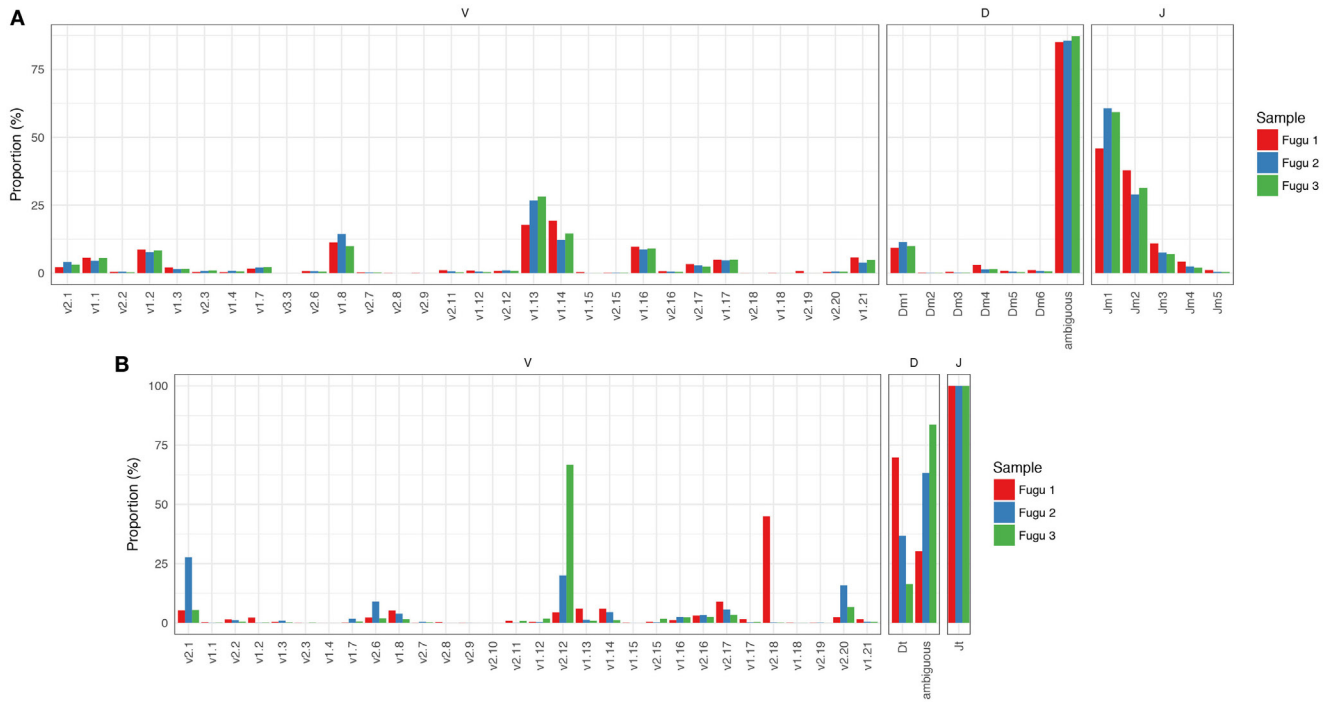

FIGURE 1 | Overall IgM **(A)** and IgT **(B)** repertoires across the three torafugu. Non-uniform usages of V, D, and J gene segments were observed. VDJ gene segments are presented in 5′ to 3′ order as in genomic arrangement.

individual fish, whereas most combinations were identified only once. We also demonstrated that the sampling of the VDJ repertoire was directed toward saturation by performing rarefaction studies (**Figure 3**).

IgT

IgT rearrangements are independent of IgM and are carried out using distinct sets of D and J gene segments. The observed usage pattern of V gene segments suggests that there is a commonality to the frequency distributions of V in IgM, which is not observed in IgT. As can be noted in **Figure 1B**, although the expression of IGHV2 family V sequences in IgT was commonly observed, the usage of individual IGHV2 gene segment was much more diverse across fish than those in IgM.

Differential expression of IgM and IgT has been documented in torafugu (17), and there is some knowledge about their tissue-specific expression as well as their developmental progression; IgM expresses as early as 4 days' post fertilization whereas secretory IgT is observed 4 days later. The finding that V(DJ)-segmental profiles of IgM and IgT are distinct from each other may suggest that the repertoires of IgM- and IgT-secreting B cells are different, and that the two types of B cells bear different distributions of variable regions for antigen binding.

Sequence Diversity of CDR-H3

Here, we defined the position of CDR-H3 as the region between the last conserved cysteine of the V segment and the first conserved tryptophan of the J segment in the conserved motif, WGxG. In the pooled data set, the standard defined subsets of 3,196,578 and 4,345,245 CDR-H3 aa sequences of IgM and IgT, respectively.

IgM

It was observed that most captured CDR-H3 sequences fall in a length range from 8 to 16 aa with a bell-shaped profile (**Figure 4A**), which has been documented in other teleosts (i.e., rainbow trout) (12) and is suggested as a representative feature of the immune repertoire expressed by naïve lymphocytes in mammals (18). In addition to the combinatorial variety, most rearrangements involved non-template nt deletion and insertion, the extent of change within CDR-H3 could be substantial. Here, we investigated random deletion of nts and insertion of non-template nts at the V–D and D–J junctions. As a result, much of the observed diversity in the torafugu CDR-H3 repertoire was generated by insertion of non-template nts: (i) more than 80% of the observed sequences had zero or less than three deletions at the V–D (**Figure 4B**) and D–J junctions for each fish (**Figure 4C**); (ii) in contrast, most sequences with ten or more nt insertions were observed in the region between 3′ V and 5′ J (refers to regions that contain the additional non-template nts along with the remnants of the D gene segment used which cannot be distinguished based on the current method) in individuals (**Figure 4D**).

In our dataset, there are examples of independent recombination events that have produced the same CDR-H3 aa sequence. An average of 411,984 distinctive CDR-H3 aa sequences for each fish were captured. Although most captured CDR-H3 aa sequences were encoded by only 1 nt sequence, we found on

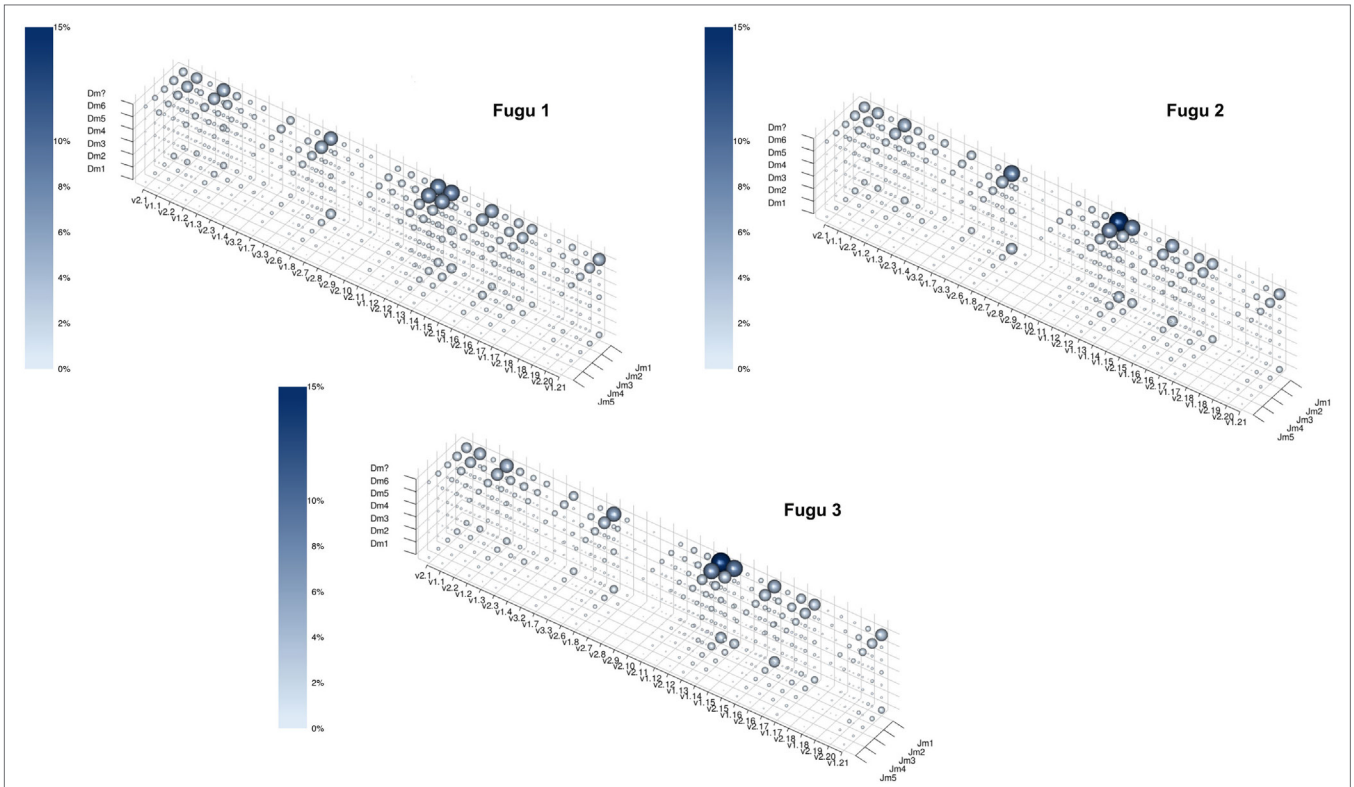

FIGURE 2 | The captured VDJ repertoires for individual torafugu. Each point in the three-dimensional space represents a unique VDJ combination. Both the size of the sphere and the intensity of the color correspond to the percentage of the number of reads that match the VDJ combination relative to all captured reads. To prevent a sphere overlapping with others, the size of the sphere was transformed by $n^{(1/4)}$, where n is the percentage of the corresponding VDJ combination. "Dm?" represents ambiguous D segment.

an average 37,190 examples of rearrangements, constituting approximately 10% of the total, that the same CDR-H3 aa sequence could be encoded by divergent nt sequences (**Table 2**). Additionally, in the mean 43,411 instances, the same CDR-H3 nt sequence was associated with different VJ pairing sequences.

IgT

Compared with IgM, IgT had a narrower distribution of CDR-H3 region lengths, assuming a fewer number of peaks (**Figure 5A**). Specifically, the CDR-H3 length of IgT varied from 7 to 17 aa with a peak at 10 aa. Like IgM, most CDR-H3 diversity in IgT was generated by insertion of nts. However, the distribution frequency of junctional diversity in IgT differed from that in IgM. For IgM, the non-random base addition/deletion at the V–D and D–J junctions was apparent: insertion/deletion was observed at a roughly equal frequency in the three individuals. However, no clear trend was seen for IgT; junction size showed few localized peaks (**Figures 5B–D**); a skewed and restricted insertion/deletion dominating each of the three individuals, i.e., fugu 1 had the highest frequency (45%) of 18 insertions at the two junctions whereas fugu 3 had 15 insertions with the highest frequency (56%) (**Figure 5D**).

The captured distinct CDR-H3 aa sequences in IgT were much fewer than in IgM and similar nt–aa usage trend was observed for IgT. About 15% of the total CDR-H3 aa sequences were from more than 1 nt sequence, whereas most of them were exclusively encoded by 1 nt sequence (**Table 2**). On an average, there were 14,577 instances wherein each CDR-H3 nt sequence could be associated with different (more than one) VJ pairings.

Estimating CDR-H3 Repertoire Size

The vast diversity of CDR-H3 is essential for maintaining functional immune responses. To explore this, we performed a capture–recapture analysis to estimate the potential size of the CDR-H3 repertoire from our sequence data. This approach is commonly used for estimating population sizes and diversity in general (19) or in immunological studies (20, 21). We first combined similar yet slightly distinct CDR-H3 aa sequences differing by at most 20% due to mutation and sequencing/PCR errors, to a single CDR-H3 cluster. We then applied the sample and resample technique on the three individuals to estimate the size of the torafugu CDR-H3 repertoires. This analysis yielded an estimate of, 503,559 IgM and 76,344 IgT lineages with pooled data (**Table 3**). However, since the pooled data has a large number of clusters that contain only one unique sequence (Figure S1 in Supplementary Material), the number of torafugu Ig lineages might be overestimated, as suggested by Haegeman et al. (22). Comparison of the actual size of observed CDR-H3 clusters with the estimated ones revealed that 74% IgM and 78% IgT of the total CDR-H3 clusters were captured in our dataset (**Table 3**). Because the effect of VDJ combination

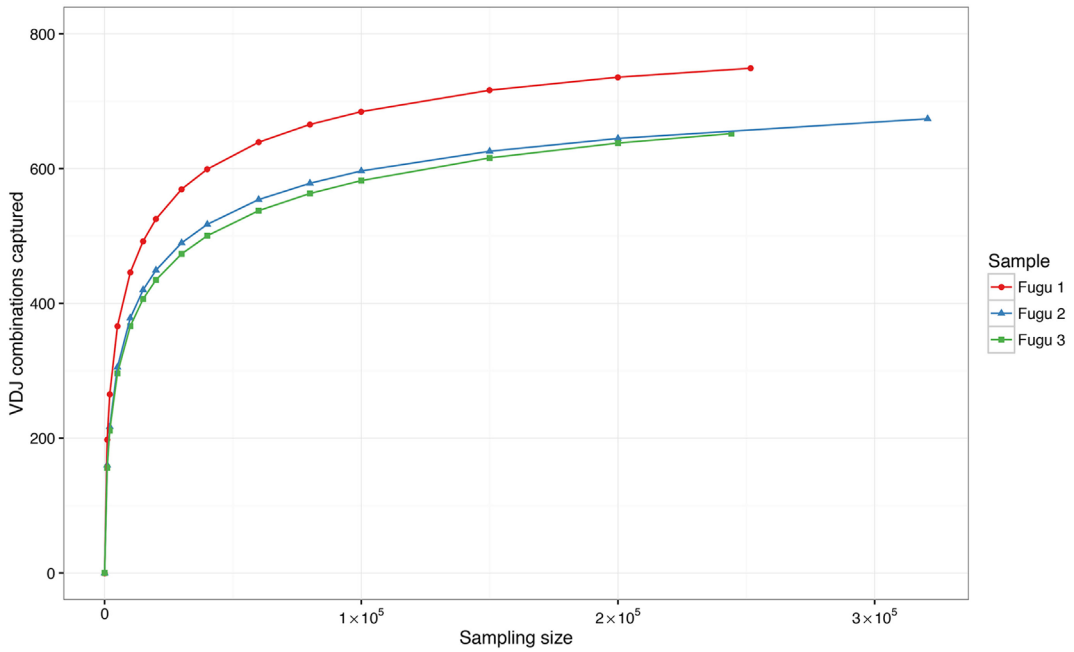

FIGURE 3 | Rarefaction studies of VDJ diversity in IgM. It demonstrates that with deeper sequencing in an individual, the number of novel VDJ combinations saturates. The sampling–resampling process was performed 1,000 times for each data point.

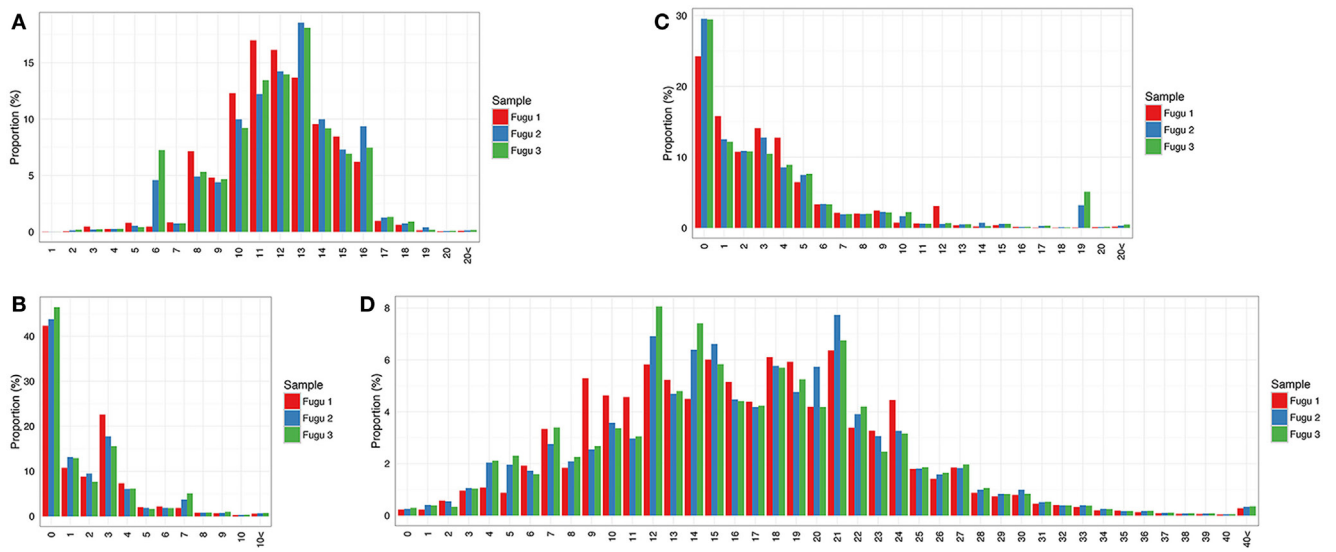

FIGURE 4 | Sequence diversity of heavy (H) chain complementarity-determining region 3 (CDR-H3) repertoire in IgM. **(A)** Distribution of CDR-H3 amino-acid length. **(B,C)** Distribution of the total number of nucleotide deletions and **(D)** V–J insertion.

diversity on CDR-H3 aa diversity is not trivial, we also calculated the average ratio between the number of CDR-H3 aa clusters and the number of captured VDJ combinations in IgM (503,559 clusters/788 VDJ = 639) and IgT (76,344 clusters/30 VDJ = 2,545), respectively. Such an observation might implicate a higher degree of repertoire diversity in naïve IgT$^+$ B cells than in IgM$^+$ B cells.

Commonly Used CDR-H3 Sequences among Individuals

Next, we explored the CDR-H3 characteristics based on the sharing level (**Figures 6A,B**). A CDR-H3 cluster was formed with a group of CDR-H3 aa sequences with at least 80% similarities. Our analysis revealed that common CDR-H3 aa clusters (i.e., found in more than one individual) manifest a higher level of "convergent

recombination" (the same CDR3 aa sequence could be generated from different nt recombinations) (23, 24); increased sharing was involved with a gradual increase in the mean degree of convergent recombination. CDR-H3 aa clusters found in one individual (termed as "private CDR-H3 cluster") were encoded by one (the median value) nt sequence for both IgM and IgT, whereas the common CDR-H3 clusters were encoded by 14 (for IgT; **Figure 7A**) and 8 (for IgM; **Figure 7B**) nt sequences on average; similar behavior was shown for T cells (25–27). We also compared the contents of sequence abundance, average length, and mutation levels between private and common CDR-H3 aa clusters; however, no significant difference was observed, with similar distribution trend appearing in groups (Figure S2 in Supplementary Material). This may demonstrate that shared CDR-H3 aa sequences (clusters) differ from private ones in a detectable degree of convergent recombination, irrespective of their intrinsic characteristics.

Finally, although CDR-H3 clusters were common among individuals at a low level (8% for IgM and 4% for IgT), 41,491 IgM and 2,743 IgT shared clusters were identified in our dataset, indicating a potential for public responses in this species.

TABLE 2 | Summary of the convergent recombination of nucleotide (nt) sequences for heavy (H) chain complementarity-determining region 3 (CDR-H3) amino-acid (aa) sequences in IgM and IgT.

	Fish	1	2	3
IgM	Unique CDR-H3 aa sequences	363,728	479,043	393,180
	CDR-H3 aa with 1 coding sequence	330,174	435,994	358,213
	CDR-H3 aa with >1 coding sequences	33,554	43,049	34,967
	Ratio of CDR-H3 aa with >1 representative nt sequences	0.092	0.089	0.088
IgT	Unique CDR-H3 aa sequences	93,683	72,364	90,292
	CDR-H3 aa with 1 coding sequence	78,573	60,121	76,454
	CDR-H3 aa with >1 coding sequences	15,110	12,243	13,838
	Ratio of CDR-H3 aa with >1 representative nt sequences	0.161	0.169	0.153

MATERIALS AND METHODS

Torafugu and Total RNA Preparation

Adult torafugu ($n = 3$) weighing between 800 and 900 g were obtained from Fish Interior (Tokyo, Japan). All fish were

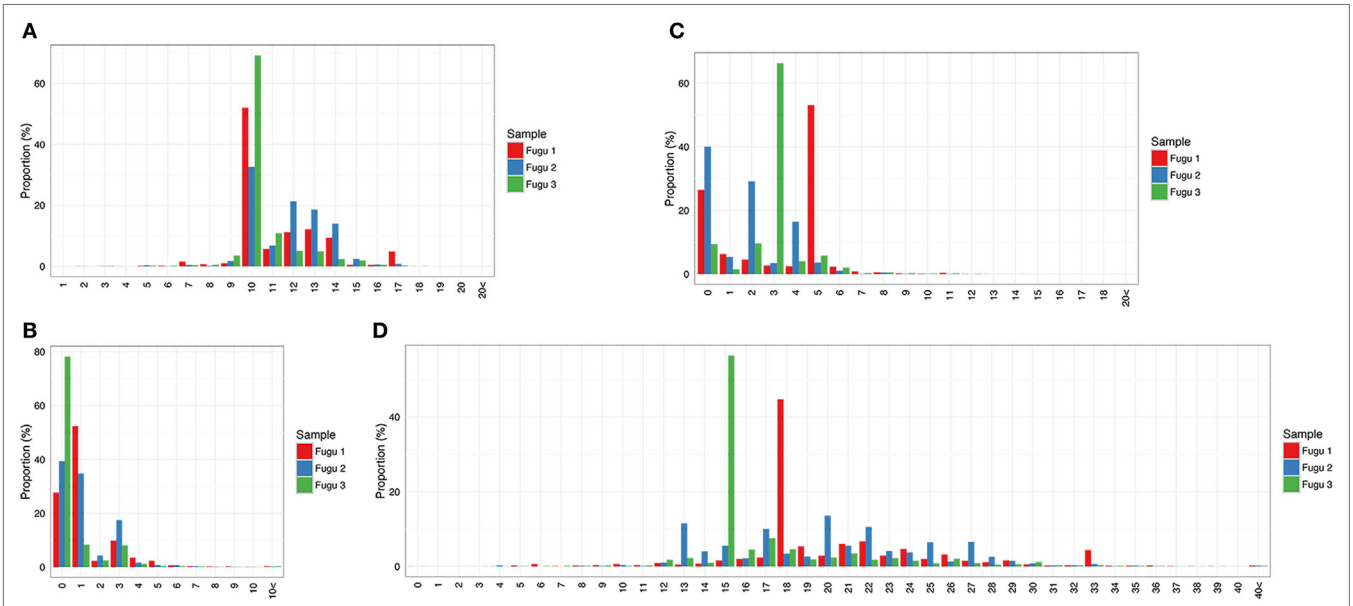

FIGURE 5 | Sequence diversity of heavy (H) chain complementarity-determining region 3 (CDR-H3) repertoire in IgT. **(A)** Distribution of CDR-H3 amino-acid length. **(B,C)** Distribution of the total number of nucleotide deletions and **(D)** V–J insertion.

TABLE 3 | Heavy (H) chain complementarity-determining region 3 (CDR-H3) amino-acid (aa) clustering statistics.

	IgM			IgT		
Fish	CDR-H3 aa	CDR-H3 aa clusters[a]	Population size[b]	CDR-H3 aa	CDR-H3 aa clusters	Population size
1	929,522	156,730	230,079	1,663,460	25,931	35,513
2	1,225,559	201,461	292,175	1,177,589	20,920	28,655
3	1,041,497	181,996	280,449	1,504,196	27,949	38,128
Pooled	3,196,578	372,933	503,559	4,345,245	59,619	76,344

[a]*Number of CDR-H3 clusters computed by CD-HIT.*

[b]*Number of population clusters estimated by abundance-based coverage estimator.*

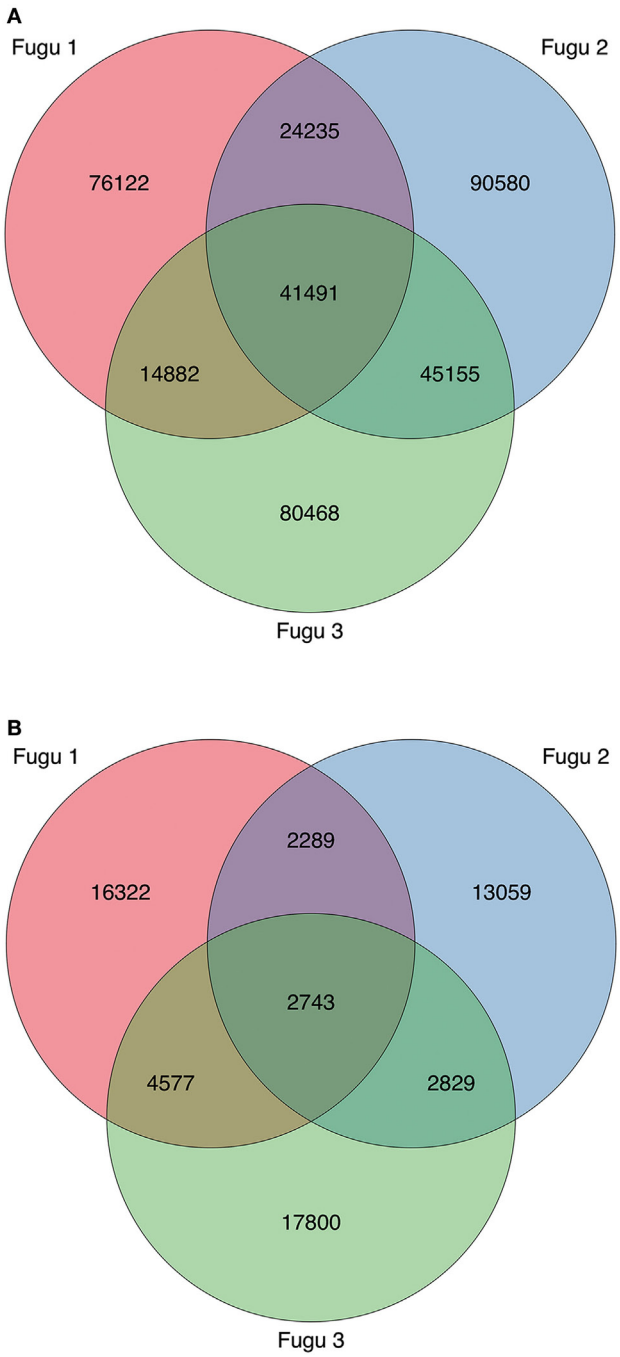

FIGURE 6 | Profile and overlap of heavy (H) chain complementarity-determining region 3 (CDR-H3) cluster abundance across the three torafugu. **(A)** IgM and **(B)** IgT. The Venn diagram corresponds to **Table 3** showing the cluster sequence statistics.

maintained in tanks with aerated seawater at 20°C. The fish were euthanized, followed by rapid dissection of tissues. Spleens and trunk kidneys from each torafugu were collected and directly fixed in RNA*later*® (Ambion, Austin, TX, USA). Total RNA was extracted using the RNeasy Lipid Tissue Mini Kit (Qiagen, Valencia, CA, USA). The mRNA was further purified using Dynabeads® mRNA DIRECT™ Purification Kit (Ambion, Austin, TX, USA). Manufacturer's protocols were followed during these processes and the concentrations of both the total RNA and mRNA were measured using a Qubit3.0 fluorometer.

Primer Design

The torafugu IgH locus was described in our previous study (28) (Figure S3 in Supplementary Material). Here, we used the IgM and IgT classes referring to the protein products of their isotypes, μ and τ, respectively, which correspond to their associated constant region genes. The consensus leader sequences for 32 potentially functional V genes were used to design five forward primers (as part of a family) (Table S1 in Supplementary Material). The reverse primers were derived from the first exon of Cμ and the second exon of Cτ, respectively (Figure S4 in Supplementary Material). To obtain cleaner products, a second independent primer set (Table S1 in Supplementary Material) 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 the nested Cμ and Cτ sequences; these (nVhCm1 and nVhCt1) were located 3-nt upstream from the first round Cμ-specific PCR primer and in the first exon of Cτ. Gene-specific primers (GSP-μ and GSP-τ) were also designed for reverse transcription.

cDNA Synthesis and Multiplex PCR Amplification

First-strand cDNA was synthesized using SuperScript® III reverse transcriptase (Invitrogen, Carlsbad, CA, USA). Total mRNA purified from each fish was split into four cDNA synthesis reactions with both the primers for IgM and IgT constant regions. RNase H (Invitrogen, Carlsbad, CA, USA) was added to each reaction to remove RNA at the end of the cDNA synthesis step.

Each 20-μL cDNA synthesis reaction was split into two PCRs, and a total of eight PCRs were set up for individual fish. Each of the five forward primers was added to represent each V segment at a final concentration of 300 nM. Some primers covered multiple V segments, and their concentration was proportionate with the number of V segments. Both reverse primers were added at a concentration of 10 μM. The reverse transcription reaction (2 μL) subsequently served as a template for PCR amplification using Platinum® *Taq* DNA Polymerase High Fidelity (Invitrogen, Carlsbad, CA, USA). The thermal cycling conditions were as follows: 94°C for 2 min, followed by 30 cycles of 94°C for 30 s, 55°C for 30 s, and 68°C for 1 min, and a final extension at 68°C for 5 min. PCR products were purified using QIAquick PCR Purification Kit (Qiagen, Valencia, CA, USA). A second-round PCR was performed on 2 μL of the first-round reaction with proportionate nested primers. Reaction conditions were as follows: 94°C for 2 min, 28 cycles of 94°C for 30 s, 55°C for 30 s, and 68°C for 1 min, and a final 5-min extension at 68°C. The second-round PCR products were purified as described above, and their concentration was measured using a Qubit® 3.0 fluorometer. The size distribution of the PCR products was determined using the Agilent 2200 Tape Station.

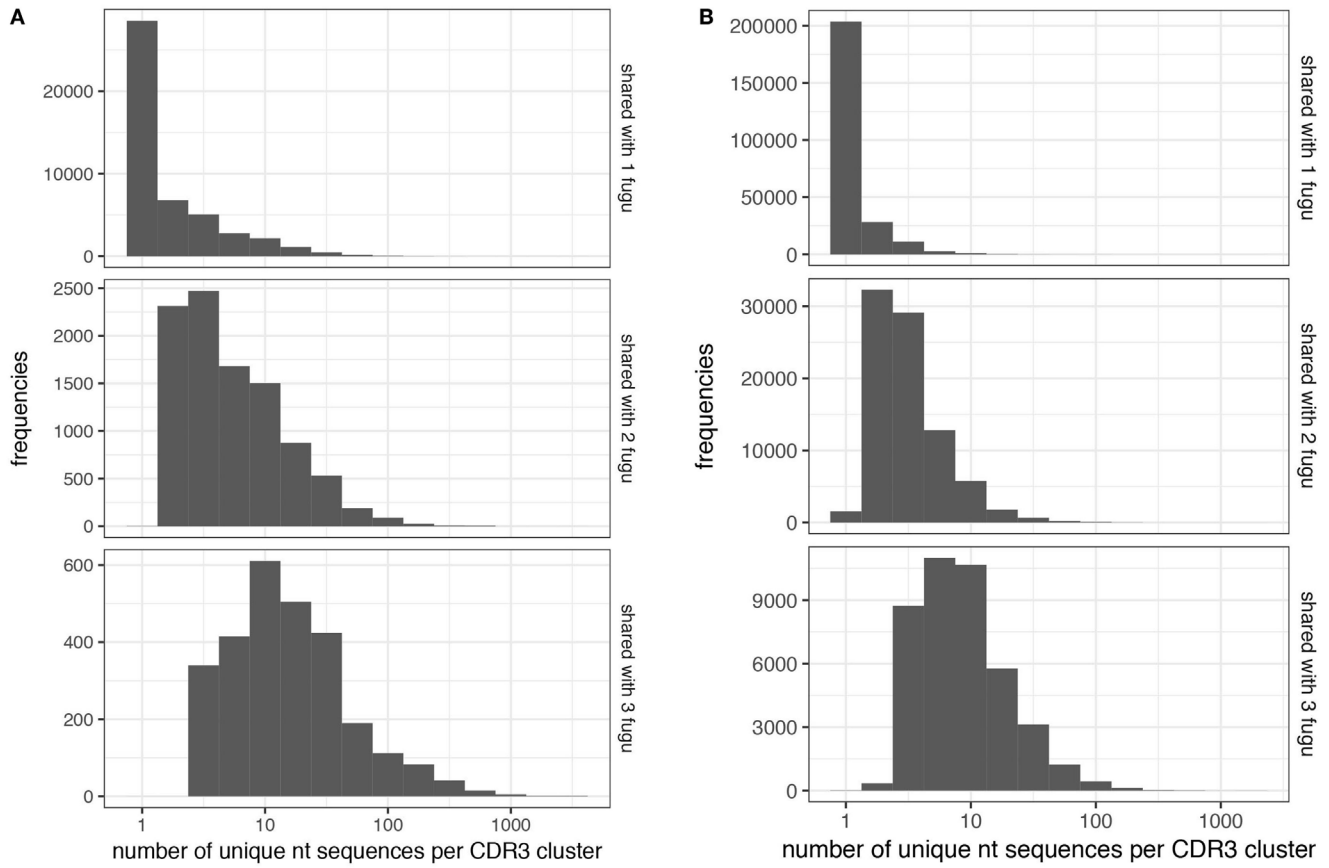

FIGURE 7 | Convergent recombination analysis of heavy (H) chain complementarity-determining region 3 (CDR-H3) clusters. Convergent recombination gradually increases as a function of sharing in **(A)** IgT and **(B)** IgM CDR-H3 clusters.

Amplicon Library Construction

About 1 µg of QIAquick cleaned PCR product for each fish was used to initiate the Illumina library preparation process. The Illumina TruSeq® DNA PCR-free sample preparation protocol was followed with slight modification. Briefly, double-stranded DNA was end-repaired and ligated to indexing adaptors for hybridization onto a flow cell. The DNA library templates were quantified by qPCR using a 7300 real-time PCR cycler (Applied Biosystems, Waltham, MA, USA) and the Library Quant Illumina Kit (KAPA Biosystems, Boston, MA, USA) with standards in a range from 0.2 fM to 20 pM. Finally, equal amounts of quantified cDNA (10 ng) from each of the three libraries, corresponding to each fish, were pooled to obtain the final amplicon library, which represented the complete collection of IgH transcripts from the three fishes.

Library Sequencing and Pre-Processing of Data

Libraries were sequenced on the Illumina MiSeq platform using the V3 (2 × 300 base) kit. The sequencing runs yielded approximately 6 million raw reads per sample. Raw reads were sorted into isotypes IgM/IgT according to their primer sequences using cutadapt (29) (parameter: −O 10, −e 0.2), and bases with a Phred quality score $Q < 20$ were trimmed from the 3'-end and the 5'-end by Trimmomatic (30) (parameter: LEADING:20, TRAILING:20, MINLEN:30). The remaining paired reads were merged using PEAR (31) (parameter: −v 10, −p 0.05, −n 300).

Data Analysis

Most existing tools [e.g., IMGT/V-Quest (32), IgBLAST (33), and JoinSolver (34)] for Ig repertoire analysis are restricted to organisms registered with International ImMunoGeneTics (IMGT). Till now, sequence annotation for torafugu Ig sequences remains unavailable on IMGT. Here, we have developed PyDAIR (https://github.com/biunit/PyDAIR) as a more flexible tool for analyzing Ig sequences. This tool can be applied to any organism without database restriction.

Identification of the VDJ usage and definition of the CDR-H3 sequence were performed using an in-house developed pipeline as illustrated in Figure S5 in Supplementary Material. In brief, merged (consensus) reads were first aligned to each V- and J-germline sequence using the Basic Local Alignment Search Tool (BLAST) to determine the optimal alignment (parameters used are shown in Table S2 in Supplementary Material). The

CDR-H3 sequences were then extracted by locating the two conserved motifs (Table S3 in Supplementary Material) between V (the 2nd CYS) (35) and J (the tryptophan in the conserved WGxG motif characteristic of J) (17) using regular expression matching. Finally, parts of the CDR-H3 sequences were aligned to D-germline sequences for complete D identification (Figure S5 in Supplementary Material). We counted the read number of each V, D, and J gene segment and calculated the frequency distribution. The computational strategy used for indel detection within the VDJ junctions was adapted from Decombinator (36).

CDR-H3 Diversity Estimation

Capture–recapture analysis was performed to assess the CDR-H3 diversity. CDR-H3 aa sequences were clustered into lineages according to sequence similarity using CD-HIT (37, 38). Clusters were created according to the following steps.

First, CDR-H3 sequences from the three samples were pooled. Then, each CDR-H3 sequence in the pool was compared with all other sequences using CD-HIT (parameter: −c 0.8, −n 4, −s 0.8, −M 2000, −l 5, −d 200). Input CDR-H3 sequences were added to the same cluster if they shared at least 80% similarity and that the shorter one matched at least 80% of the length of the longer one. The number of CDR-H3 clusters for each sample and for the pooled data are summarized in Table S6 in Supplementary Material. The population sizes of CDR-H3 repertoires (i.e., CDR-H3 sequence clusters) were then estimated with abundance-based coverage estimator (19).

DISCUSSION

Here, we characterized the expressed μ and τ IgH repertoires (IgM and IgT) in adult torafugu by massively parallel sequencing of IgH amplicons. The approach developed in this study allowed us to identify Ig sequences and determine the abundance and isotype of IgH mRNA as well as the CDR-H3 diversity. Consistent with previous observations from other groups (8, 11, 39), we observed that certain V, D, and J gene segments were preferentially used in torafugu. This observation highlighted the usage bias of V, D, and J gene segments in the quiescent immune system of torafugu, which reflects bias in the VDJ recombination mechanisms. The reasons for such bias are not well understood but are likely due to a combination of proximity effects that influence BCR development. It is also possible that preferential PCR amplification of certain V sequences over others skewed the usage frequencies presented here.

Notably, there is an obvious connection between J gene segments in the VDJ combinations of the IgM repertoire and their relative positions to V and D gene segments on the H chain locus. J gene segments that are close to the V and D gene segments are preferentially used by adult torafugu. In detail, Jm1 and Jm2 accounted for 55 and 31%, respectively, of the Jm usage compared with an expected frequency of 20% assuming unbiased gene usage. This observation is reminiscent of reports in humans that (*i*) V gene segment that are close to the D and J segments are preferentially used (40, 41) and (*ii*) D genes positioned in the J proximal locus are preferentially used in sterile DJ rearrangements (42). These results suggest that the antibody variable region repertoire is partly determined by the chromosomal position of these gene segments. One explanation for such restriction could be chromatin conformations that render certain gene segments more accessible than others (43). In addition, recent research has suggested that chromatin structure governs a large part of the biases in TCR Jβ gene usage in both mice and humans (44). Since it is a general approach, we speculate that the observed location-related Jm biases could be further explained by this theory.

When studying the IgM repertoire, a high stereotypy, i.e., a biased yet common use of a small number of VDJ combinations, was observed. Similarly, in the large-scale sequencing of the naïve zebrafish IgM repertoire, convergences have been found in the normal repertoire (8). This result suggests the presence of a common pool of pre-existing B cells in which the structured IgM repertoire could be recruited. It is also possible that the source of the naïve IgM repertoire is convergent evolution, and fish living similar environments show that selection in the quiescent immune systems converges to certain VDJ combinations. Finally, the observed public IgM component is consistent with the paradigm of the B-cell clonal selection theory (specific antigen only activates (i.e., selection) its counter-specific cell to produce its clones for antibody production) as noted in mammals (45, 46), suggesting that general features are already in place in the common ancestors of fish and tetrapod.

We also implemented VDJ junction and mutation analysis as a function of CDR-H3 diversity among different individuals. Our results suggest that most of the junction mutations in adult fish are generated from insertions of nts rather than deletions. The high frequency of nt insertions may correspond to an ordered expression of terminal deoxyribonucleotidyl transferase, as previously noted (47). Interestingly, our analysis of the spread of insertion diversities indicates that the distribution of nt insertion in IgM is much wider than that in IgT. Perhaps this diversity is related to the compartmentalization of the two Ig isotypes.

To observe how often repertoires converged to the same core region of the antibody, we searched for CDR-H3 clusters that are shared between fish. As a result, there were a number of public components consisting of slightly varying CDR-H3 aa sequences that were shared by different individuals. Importantly, this represents a large potential for public responses. Additionally, the same CDR-H3 aa sequence could be encoded by divergent nt sequences. Taken together, these data suggest the powerful forces of antigen-driven clonal selection.

By combining high-throughput sequencing with informatics analysis, we revealed the naïve μ and τ IgH repertoires in torafugu. We discovered that VDJ usage is not uniform and that the same handful VDJ combinations often common to different individuals in the naïve IgM repertoire. We also found that the mutation content of IgM and IgT repertoires is diverse, and that convergent evolution of CDR-H3 sequences is common. Collectively, these data provide a window into the mechanism of Ig diversity creation and allow us to better understand B-cell clonal selection.

DATA ACCESS

The sequencing data from this study have been submitted to DDBJ (http://www.ddbj.nig.ac.jp/) under the accession numbers DRA004021 to DRA004023.

ETHICS STATEMENT

This study was conducted in accordance with the recommendations of "Guidelines for Proper Conduct of Animal Experiments" released by the Science Council of Japan. The protocol was approved by the Subcommittee on Institutional Animal Care and Use of the Graduate School of Agricultural and Life Sciences, The University of Tokyo (Protocol no. P14-952).

AUTHOR CONTRIBUTIONS

XF performed experiments, analyzed data, and wrote the manuscript. JS contributed analytic tools and analyzed data. ET performed experiments. KS contributed analytic tools. MR assisted in experiments. SW conceived ideas and oversaw the research. SA conceived ideas, oversaw the research, and cowrote the manuscript. All authors contributed to the manuscript preparation.

FUNDING

This work was funded by the Japan Society for the Promotion of Science (grant number 24248034 and grant number 15H02461 to SA, and grant number 15J10504 to JS).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/articles/10.3389/fimmu.2018.00251/full#supplementary-material.

REFERENCES

1. Schatz DG. Antigen receptor genes and the evolution of a recombinase. *Semin Immunol* (2004) 16:245–56. doi:10.1016/j.smim.2004.08.004
2. Xu JL, Davis MM. Diversity in the CDR3 region of VH is sufficient for most antibody specificities. *Immunity* (2000) 13:37–45. doi:10.1016/S1074-7613(00)00006-6
3. Ippolito GC, Schelonka RL, Zemlin M, Ivanov II, Kobayashi R, Zemlin C, et al. Forced usage of positively charged amino acids in immunoglobulin CDR-H3 impairs B cell development and antibody production. *J Exp Med* (2006) 203:1567–78. doi:10.1084/jem.20052217
4. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel V-D-J pyrosequencing. *Sci Transl Med* (2009) 1(12):12ra23. doi:10.1126/scitranslmed.3000540
5. Choi NM, Loguercio S, Verma-Gaur J, Degner SC, Torkamani A, Su AI, et al. Deep sequencing of the murine IgH repertoire reveals complex regulation of nonrandom V gene rearrangement frequencies. *J Immunol* (2013) 191:2393–402. doi:10.4049/jimmunol.1301279
6. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc Natl Acad Sci U S A* (2009) 106:20216–21. doi:10.1073/pnas.0909775106
7. Lavinder JJ, Hoi KH, Reddy ST, Wine Y, Georgiou G. Systematic characterization and comparative analysis of the rabbit immunoglobulin repertoire. *PLoS One* (2014) 9:e101322. doi:10.1371/journal.pone.0101322
8. Weinstein JA, Jiang N, White RA III, Fisher DS, Quake SR. High-throughput sequencing of the Zebrafish antibody repertoire. *Science* (2009) 324:807–10. doi:10.1126/science.1170020
9. Schroeder HW. Similarity and divergence in the development and expression of the mouse and human antibody repertoires. *Dev Comp Immunol* (2006) 30:119–35. doi:10.1016/j.dci.2005.06.006
10. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat Biotechnol* (2014) 32:158–68. doi:10.1038/nbt.2782
11. Arnaout R, Lee W, Cahill P, Honan T, Sparrow T, Weiand M, et al. High-resolution description of antibody heavy-chain repertoires in humans. *PLoS One* (2011) 6:e22365. doi:10.1371/journal.pone.0022365
12. Castro R, Jouneau L, Pham H-P, Bouchez O, Giudicelli V, Lefranc M-P, et al. Teleost fish mount complex clonal IgM and IgT responses in spleen upon systemic viral infection. *PLoS Pathog* (2013) 9:e1003098. doi:10.1371/journal.ppat.1003098
13. Flajnik MF, Kasahara M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat Rev Genet* (2009) 11:1–14. doi:10.1038/nrg2703
14. Bengtén E, Wilson M. Antibody repertoires in fish. In: Hsu E, Du Pasquier L, editors. *Pathogen-Host Interactions: Antigenic Variation v. Somatic Adaptations*. New York: Springer International Publishing (2015). p. 193–234.
15. Benichou J, Ben-Hamo R, Louzoun Y, Efroni S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* (2012) 135:183–91. doi:10.1111/j.1365-2567.2011.03527.x
16. Aparicio S, Chapman J, Stupka E, Putnam N, Chia J, Dehal P, et al. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* (2002) 297:1301–10. doi:10.1126/science.1072104
17. Savan R, Aman A, Sato K, Yamaguchi R, Sakai M. Discovery of a new class of immunoglobulin heavy chain from fugu. *Eur J Immunol* (2005) 35:3320–31. doi:10.1002/eji.200535248
18. Pannetier C, Even J, Kourilsky P. T-cell repertoire diversity and clonal expansions in normal and clinical samples. *Immunol Today* (1995) 16:176–81. doi:10.1016/0167-5699(95)80117-0
19. Chao A, Lee S-M. Estimating the number of classes via sample coverage. *J Am Stat Assoc* (1992) 87:210–7. doi:10.1080/01621459.1992.10475194
20. Laydon DJ, Bangham CRM, Asquith B. Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach. *Philos Trans R Soc Lond B Biol Sci* (2015) 370:20140291. doi:10.1098/rstb.2014.0291
21. Qi Q, Liu Y, Cheng Y, Glanville J, Zhang D, Lee J-Y, et al. Diversity and clonal selection in the human T-cell repertoire. *Proc Natl Acad Sci U S A* (2014) 111:13139–44. doi:10.1073/pnas.1409155111
22. Haegeman B, Hamelin J, Moriarty J, Neal P, Dushoff J, Weitz JS. Robust estimation of microbial diversity in theory and in practice. *ISME J* (2013) 7:1092–101. doi:10.1038/ismej.2013.10
23. Venturi V, Kedzierska K, Price DA, Doherty PC, Douek DC, Turner SJ, et al. Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proc Natl Acad Sci U S A* (2006) 103:18691–6. doi:10.1073/pnas.0608907103
24. Venturi V, Price DA, Douek DC, Davenport MP. The molecular basis for public T-cell responses? *Nat Rev Immunol* (2008) 8:231–8. doi:10.1038/nri2260
25. Li H, Ye C, Ji G, Wu X, Xiang Z, Li Y, et al. Recombinatorial biases and convergent recombination determine interindividual TCRβ sharing in murine thymocytes. *J Immunol* (2012) 189:2404–13. doi:10.4049/jimmunol.1102087
26. Quigley MF, Greenaway HY, Venturi V, Lindsay R, Quinn KM, Seder RA, et al. Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire. *Proc Natl Acad Sci U S A* (2010) 107:19414–9. doi:10.1073/pnas.1010586107
27. Madi A, Shifrut E, Reich-Zeliger S, Gal H, Best K, Ndifon W, et al. T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Res* (2014) 24:1603–12. doi:10.1101/gr.170753.113
28. Fu X, Zhang H, Tan E, Watabe S, Asakawa S. Characterization of the torafugu (*Takifugu rubripes*) immunoglobulin heavy chain gene locus. *Immunogenetics* (2015) 67:179–93. doi:10.1007/s00251-014-0824-z
29. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.J* (2011) 17:10–2. doi:10.14806/ej.17.1.200
30. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* (2014) 30:2114–20. doi:10.1093/bioinformatics/btu170

31. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina paired-end reAd mergeR. *Bioinformatics* (2014) 30:614–20. doi:10.1093/bioinformatics/btt593
32. Giudicelli V, Brochet X, Lefranc M-P. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb Protoc* (2011) 2011:695–715. doi:10.1101/pdb.prot5633
33. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* (2013) 41:W34–40. doi:10.1093/nar/gkt382
34. Souto-Carneiro MM, Longo NS, Russ DE, Sun H, Lipsky PE. Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER. *J Immunol* (2004) 172:6790–802. doi:10.4049/jimmunol.172.11.6790
35. Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, et al. IMGT/LIGM-DB, the IMGT® comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Res* (2006) 34:D781–4. doi:10.1093/nar/gkj088
36. Thomas N, Heather J, Ndifon W, Shawe-Taylor J, Chain B. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics* (2013) 29:542–50. doi:10.1093/bioinformatics/btt004
37. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* (2006) 22:1658–9. doi:10.1093/bioinformatics/btl158
38. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* (2012) 28:3150–2. doi:10.1093/bioinformatics/bts565
39. Ivanov II, Schelonka RL, Zhuang Y, Gartland GL, Zemlin M, Schroeder HW. Development of the expressed Ig CDR-H3 repertoire is marked by focusing of constraints in length, amino acid use, and charge that are first established in early B cell progenitors. *J Immunol* (2005) 174:7773–80. doi:10.4049/jimmunol.174.12.7773
40. Yancopoulos GD, Desiderio SV, Paskind M, Kearney JF, Baltimore D, Alt FW. Preferential utilization of the most JH-proximal VH gene segments in pre-B-cell lines. *Nature* (1984) 311:727–33. doi:10.1038/311727a0
41. Schroeder HW, Walter MA, Hofker MH, Ebens A, Willems van Dijk K, Liao LC, et al. Physical linkage of a human immunoglobulin heavy chain variable region gene segment to diversity and joining region elements. *Proc Natl Acad Sci U S A* (1988) 85:8196–200. doi:10.1073/pnas.85.21.8196
42. Hansen TØ, Lange AB, Barington T. Sterile DJ H rearrangements reveal that distance between gene segments on the human Ig H chain locus influences their ability to rearrange. *J Immunol* (2015) 194:973–82. doi:10.4049/jimmunol.1401443
43. Jhunjhunwala S, van Zelm MC, Peak MM, Cutchin S, Riblet R, van Dongen JJM, et al. The 3D Structure of the immunoglobulin heavy-chain locus: implications for long-range genomic interactions. *Cell* (2008) 133:265–79. doi:10.1016/j.cell.2008.03.024
44. Ndifon W, Gal H, Shifrut E, Aharoni R, Yissachar N, Waysbort N, et al. Chromatin conformation governs T-cell receptor J gene segment usage. *Proc Natl Acad Sci U S A* (2012) 109:15865–70. doi:10.1073/pnas.1203916109
45. Magadan S, Sunyer OJ, Boudinot P. Unique features of fish immune repertoires: Particularities of adaptive immunity within the largest group of vertebrates. In: Hsu E, Du Pasquier L, editors. *Pathogen-Host Interactions: Antigenic Variation v. Somatic Adaptations*. New York: Springer International Publishing (2015). p. 235–64.
46. Rajewsky K. Clonal selection and learning in the antibody system. *Nature* (1996) 381:751–8. doi:10.1038/381751a0
47. Schroeder HW, Zhang L, Philips JB. Slow, programmed maturation of the immunoglobulin HCDR3 repertoire during the third trimester of fetal life. *Blood* (2001) 98:2745–51. doi:10.1182/blood.V98.9.2745

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.