Genomic analysis of a second rainbow trout line (Arlee) leads to an extended description of the IGH VDJ gene repertoire

Susana Magadan, Stanislas Mondot, Yniv Palti, Guangtu Gao, Marie Paule Lefranc, Pierre Boudinot

Genomic analysis of a second rainbow trout line (Arlee) leads to an extended description of the IGH VDJ gene repertoire

Susana Magadan[1,*], Stanislas Mondot[2], Yniv Palti[3], Guangtu Gao[3], Marie Paule Lefranc[4], Pierre Boudinot[5,*]

[1] Centro de Investigaciones Biomédicas, Universidade de Vigo, Campus Universitario Lagoas Marcosende, 36310 Vigo, España.

[2] Université Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, 78350, Jouy-en-Josas, France

[3] USDA-ARS National Center for Cool and Cold Water Aquaculture, 11861 Leetown Road, Kearneysville, WV 25430, USA

[4] IMGT®, the international ImMunoGeneTics Information System®, Laboratoire d'ImmunoGénétique Moléculaire (LIGM), Institut de Génétique Humaine (IGH), UMR9002 CNRS, Université de Montpellier, Montpellier, France.

[5] Université Paris-Saclay, INRAE, UVSQ, VIM, 78350, Jouy-en-Josas, France.

Correspondence:

Susana Magadan, CINBIO, Centro de Investigaciones Biomédicas, Universidade de Vigo, Immunology Group, Campus Universitario Lagoas Marcosende, 36310 Vigo, España.

Phone: 0034 986130142 smaga@uvigo.es

And

Pierre Boudinot, Université Paris-Saclay, INRAE, UVSQ, VIM, 78350, Jouy-en-Josas, France

Phone : 0033 1 34652585 Pierre.Boudinot@inrae.fr

Keywords: IMGT, immunoglobulin, immune repertoires, Salmonids, rainbow trout

Abstract

High-throughput sequencing technologies brought a renewed interest for immune repertoires. Fish Ab and B cell repertoires are no exception, and their comprehensive analysis can both provide new insights into poorly understood immune mechanisms, and identify markers of protection after vaccination. However, the lack of genomic description and standardized nomenclature of IG genes hampers accurate annotation of Ig mRNA deep sequencing data. Complete genome sequences of Atlantic salmon and rainbow trout (Swanson line) recently allowed us to establish a comprehensive and coherent annotation of Salmonid IGH genes following IMGT standards. Here we analyzed the IGHV, D, and J genes from the newly released genome of a second rainbow trout line (Arlee). We confirmed the validity of salmonid IGHV subgroups, and extended the description of the rainbow trout IGH gene repertoire with novel sequences, while keeping nomenclature continuity. This work provides an important resource for annotation of high-throughput Ab repertoire sequencing data.

Introduction

Immunoglobulins (IG) – also known as antibodies - are the antigen specific receptors expressed by B cells (Lefranc and Lefranc 2001, Lefranc and Lefranc 2020). Germline IG loci contain non contiguous V,(D) and J genes that can be assembled to produce expressible genes encoding IG variable regions ((Tonegawa 1983), reviewed in (Max and Fugmann 2013)). During the B lymphocyte differentiation, IG loci are subjected to genomic rearrangements of variable (V), joining (J), diversity (D) genes, leading to the expression of a unique antigen receptor by each B lymphocyte. The IG heavy (IGH) chains present the highest diversity due to the inclusion of the D gene in a two-step recombination that produces the VH domain (V-D-J-REGION), while the L chains result from a V-J rearrangement that produces the VL domain (V-J-REGION). The V-(D)-J junction recombination process is not exact, and the deletion of nucleotides at the joint sites as well as the insertion of so-called palindromic (P) and non-templated (N) nucleotides by the enzyme terminal deoxynucleotidyltransferase (TdT) encoded by the DNA nucleotidylexotransferase (DNTT) gene, are commonly observed (Lefranc and Lefranc 2020). An IGH V-(D)-J rearrangement leads to an exon encoding an IG variable domain (V-DOMAIN) (Lefranc, Pommie et al. 2003). V-domains are beta-barrels, and their antigen binding site is made of three loops named complementarity determining region (CDR)1, 2 and 3 protruding at the top of the domain. The CDR3 is produced by the recombination of the IGH-V(D)J genes and is the most variable mainly due to imprecise recombination and addition/deletion of nucleotides at the D-J and V-(D)-J gene junctions (Wu and Kabat 1970, Lefranc and Lefranc 2001). CDR1 and CDR2 are encoded by IGHV genes; being structurally unconstrained loops, they are highly variable across these genes and also participate to the IGV domain variation. Thus, the diversity of IG

results from the association of heavy and light chains, from the combination of different V-(D)-J genes and from junctional mechanisms.

Teleost fish have typical B cells expressing a large diversity of IG (reviewed in (Fillatreau, Six et al. 2013)). As in other jawed vertebrates, fish B cell clonal responses are induced by infection or immunization, and antibodies (Abs) are critical for the protection induced by most vaccines. Fish B cells can express three IG heavy chain isotypes (reviewed in (Fillatreau, Six et al. 2013)): μ and δ (for IgM and IgD classes, respectively) which are conserved in all classes of vertebrates with jaws; and τ (for IgT class), which is specific to fish and is specialized to mucosal defense (Salinas, Zhang et al. 2011). Isotypic commutation and switch recombination do not occur in fish, but the structure of the IGH locus and recombination mechanisms ensure an exclusive expression of either IgM/IgD or IgT within a recombination unit. IgM/D and IgT are not co-expressed by the same B cell (Hansen, Landis et al. 2005). Thus, they define two lineages of B cells expressing distinct repertoires built on the same pool of IGHV genes combined with different IGHD and IGHJ genes (Danilova, Bussmann et al. 2005, Hansen, Landis et al. 2005, Zhang, Salinas et al. 2010, Castro, Jouneau et al. 2013).

Many aspects of the biology of fish B cells remain poorly known. The location and mechanisms of fish B lymphopoiesis are still elusive (Zwollo 2011, Liu, Li et al. 2017); the equivalent of VPREB (V-preB) and IGLL5 (lambda-5) chains are unknown in these species, and the main stages of B cell development might be different in fish and mammals. The bases of allelic/loci exclusion are not understood. It is also important to note that the anatomy of lymphoid tissues is very different between fish and well-known mammalian models. For example, the lack of lymph nodes in fish raises the questions of where immune

responses are initiated. Detailed phenotypes and functional specialization of B cell subsets still have to be defined, and memory B cells in fish have not been accurately described (Yamaguchi, Quillet et al. 2018). The existence of somatic hypermutation of fish IG sequences is now well established (Yang, Waldbieser et al. 2006, Jiang, Weinstein et al. 2011, Magor 2015), but its regulation remains poorly understood. Because IG are central to B cell biology, antibody repertoire analysis is essential for the characterization of these mechanisms and populations. Such approaches remain complex in the absence of a complete and comprehensive genomic description of IG loci in many fish species.

The development of high-throughput sequencing technologies during the last decade led to many studies of the clonal composition of B cell populations. Deep sequencing of human and mouse B cell repertoires has already revolutionized the field, shedding new light on ontogeny of B cells, their importance for autoimmunity and cancer, and their implication in the responses to infections or vaccines (reviewed in (Hou, Chen et al. 2016)). There is also a growing interest for B cell repertoires among fish immunologists, with connections to basic questions about immune mechanisms as well as to applied issues such as markers of protection after vaccination. Salmonids (family Salmonidae) have become important models for such studies (Castro, Navelsaker et al. 2017, Krasnov, Jorgensen et al. 2017) for several reasons: (1) their immune system and, more specifically, their antibody responses to pathogens have been well-studied (2) high-quality genome sequences are available for farmed and wild species including rainbow trout (*Oncorhynchus mykiss*), Atlantic salmon (*Salmo salar*), brown trout (*Salmo trutta*), Coho salmon (*Oncorhynchus kisutch*) and chinook salmon (*Oncorhynchus tshawytscha*). A whole genome duplication (WGD) during the early evolution of the salmonids led to extraordinary complex repertoires of IGHV, D and J genes

clustered at two genomic locations, and of IGL genes belonging to four subtypes (Yasuike, Boer et al. 2010, Magadan, Krasnov et al. 2019, Rego, Hansen et al. 2020) (3) Atlantic salmon and rainbow trout are key species for fish farming globally, and their pathogens have been extensively investigated. IGH loci of salmonids have been studied for thirty years, but complete genome sequences recently allowed us to establish a comprehensive and coherent annotation for this family, based on Atlantic salmon (Yasuike, Boer et al. 2010) and rainbow trout. This led to an IMGT standardized nomenclature of IGH genes in these two species, taking into account the particularities of Salmonid loci, *ie* a large number of IGHV, IGHD and IGHJ genes, a distribution of these genes into two distinct locations on two chromosomes, and a common set of IGHV subgroups (Magadan, Krasnov et al. 2019). A consistent nomenclature of IGHV subgroups and IGHJ genes is particularly important for comparative studies of AIRRseq data (Adaptive Immune Receptor Repertoire sequencing) *ie*, RNA-seq sequencing data of IG transcripts expressed in multiple contexts across Salmonids.

However, the rainbow trout IGH repertoire defined in (Magadan, Krasnov et al. 2019) based on the genome sequence of the Swanson line (Omyk_1.0) was not complete. Indeed, a number of rainbow trout IGH cDNA sequences present in the GenBank did not have any close counterpart in the Swanson genome assembly (hence, in the IMGT gene directory). When annotating our own deep sequencing datasets from isogenic lines (Quillet, Dorson et al. 2007) genetically rather distant from the Swanson line (Palti, Gao et al. 2014), we also realized that a number of IGHV genes were missing in the reference IMGT directories. In fact, such gaps in our annotation could be due to either genome assembly issues or real differences between the IGH loci of the Swanson line and those of other trout lines. For example, the PacBio long read sequencing technology used for the Arlee genome significantly improved the quality of the assembly, compared to the Swanson genome

(based on Illumina sequencing), and hence some differences in the annotation of IGHV genes between the Swanson and Arlee may be caused by this difference in the quality of the two genome assemblies rather than biological differences between the two rainbow trout lines.

In this work, we pursued our effort of annotation and nomenclature standardization using the newly released genome assembly from the Arlee line of rainbow trout (USDA_OmykA_1.1). Both the Swanson and Arlee are homozygous clonal YY male lines developed through androgenesis by the lab of Gary Thorgaard at Washington State University. The Swanson line was derived in 1991 from a semi-wild population from the Swanson River in the Kenai Peninsula of Alaska (Robison, Wheeler et al. 2001). The Arlee clonal line was derived from the Arlee strain, a domesticated hatchery strain that was used by the Montana Department of Fish, Wildlife and Parks (Ristow, Grabowski et al. 1998) and is thought to have originally been collected from Northern California like most farmed rainbow trout stocks that were imported to Europe (Gary Thorgaard, Personal Communication). In addition, fish from the Arlee line were found to have low nonspecific cellular cytotoxicity in the peripheral blood (Ristow, Grabowski et al. 1995). Here we aimed at extending the IMGT rainbow trout IGH annotation to a second line coming from a different region of North America. Testing the salmonid IGH classification and nomenclature established from the Swanson line on the IGH repertoire present in the Arlee line confirmed the validity of the IGHV subgroups, and provided a first picture of the variation of these genes between rainbow trout populations. This update was performed to take into account all new sequences, while keeping nomenclature continuity. We thus produced a non-redundant directory of IGHV sequences to help with annotation of IG repertoire in this

163 species. This resource will significantly extend the diversity of sequences available for the
164 IMGT/HighV-QUEST (Alamyar, Giudicelli et al. 2010, Alamyar, Giudicelli et al. 2012, Li,
165 Lefranc et al. 2013) or other annotation tools.
166
167 ## Materials and methods
168 *Gene annotation*
169 Chromosome 12 (CM023230.2) and 13 (CM023231.2) from the recently released rainbow
170 trout genome assembly (USDA_OmykA_1.1) which was derived from the Arlee line were
171 examined to locate IGH loci. IGHV, IGHD and IGHJ gene sequences were previously
172 identified by Magadan *et al.* (Magadan, Krasnov et al. 2019) in the rainbow trout genome
173 assembly (Omyk_1.0) which was derived from the Swanson line (Pearse, Barson et al. 2019)
174 and were used as queries to identify the chromosomal regions containing the IGH loci in the
175 Arlee genome assembly. All these IGH genes identified in the Swanson genome assembly
176 had been previously submitted to the Immunoglobulins (IG), T cell receptors (TR) and major
177 histocompatibility (MH) Nomenclature Sub-Committee (IMGT-NC) of the International
178 Union of Immunological Societies (IUIS) Nomenclature Committee. The obtained IMGT gene
179 names had been included in the IMGT-NC Report #2019-10-0402 (www.imgt.org).
180
181 *Extension of the IMGT nomenclature to the Arlee line IGH repertoire*
182 IGHV sequences from the two lines were aligned pairwise, and clusters of sequences >= 98%
183 identical were identified. This first step produced a list of sequences for which a
184 straightforward 1:1 correspondence exist between genes from each genetic background.
185 We then subjected the other IGHV sequences from the Arlee line to IMGT/V-QUEST
186 [*Oncorhynchus mykiss* and *Salmo salar*; IGH] and IMGT/BlastSearch (IMGT/GENE-DB

nucleotide sequences (F+ORF+inframeP)) analysis. For functional genes or for ORF/pseudogenes with lower degree of mutation, IMGT/V-QUEST identified the subgroup to which assign the IGHV gene. When a higher mutation/indel number impaired the IMGT/V-QUEST analysis, subgroup assignation was operated from IMGT/BlastSearch results. All IGHV sequences from the Arlee line were assigned to an IGHV subgroup defined from the Atlantic salmon repertoire.

Comparison of the IMGT gene sets from Swanson and Arlee lines with a collection of IGHV cDNA available in GenBank

Rainbow trout IGHV cDNA sequences were collected from GenBank using Blastn with all sequences from Swanson and Arlee. cDNA sequences (301 bp upstream the YxC pattern corresponding to the end of the V gene) were clustered using VSEARCH (Rognes, Flouri et al. 2016) (threshold 98% identity). An identity matrix was inferred from pairwise local sequence alignments computed with ClustalW.

Results and discussion

1. *Comparison of the structure of IGHV loci between Swanson and Arlee rainbow trout lines*

The newly released genome assembly from the Arlee clonal line of rainbow trout (USDA_OmykA_1.1) resulted in elucidation of complex loci such as the IGH locus. In the Arlee genome assembly, the IGH genes are within two regions named IGHA and IGHB, that are located on chromosomes 13 (CM023231.2) and 12 (CM023230.2), respectively. This is similar to the rainbow trout reference genome Omyk_1.0 (GCA_002163495.1) (Pearse, Barson et al. 2019), which was obtained from the Swanson homozygous line (Magadan,

211 Krasnov et al. 2019). IGHA and IGHB loci span a total length of approximately 5 and 4.3 Mbp,
212 respectively. Each locus comprises a number of IGHV genes upstream of a few D-J-C-clusters
213 comprising IGHD and IGHJ genes associated to an IGHC gene, either IGHT or, for the most 3'
214 cluster, IGHM and IGHD) (see Figure 1, and figure S1 for comparison with the map of the
215 IGH loci from Swanson).
216 The annotation of IGH loci in the rainbow trout reference genome Omyk_1.0 had provided a
217 first rainbow trout IMGT reference directory for IGH sequences
218 (http://www.imgt.org/vquest/refseqh.html). In the Omyk_1.0 genome (Swanson line), a
219 total of 129 IGHV genes were identified, of which 57 were considered fully functional or
220 with an open reading frame (ORF) without stop codon. The annotation of USDA_OmykA_1.1
221 genome (Arlee line) resulted in a lower number of IGHV genes, but a higher number of
222 functional genes: a total of 118 IGHV genes were identified, of which 68 seem to be
223 functional or with ORF. The distribution of these IGHV genes on chromosomes is as follows:
224 61 IGHV genes on chromosome 12, of which 31 are functional or with ORF, and 57 IGHV
225 genes on chromosome 13, of which 37 are functional or with ORF.
226 A comparison of the genomic location of IGHV genes between Arlee (Figure 1) and Swanson
227 (Figure S1) genome sequences reveals a substantial level of variation (see genes encircled in
228 red and blue in Figure 1 and Figure S1, which are found in only one assembly) especially in
229 the locus B on chromosome 12. However, the order of functional genes – as identified based
230 on both location and sequence similarity - was globally conserved, allowing an upgrade of
231 the nomenclature without major discontinuity.
232
233 ***2. Creation of an extended nomenclature of rainbow trout IGHV genes***

We have previously established a standardized IMGT nomenclature of salmonid IGH genes based on genome assemblies for Atlantic salmon and the Swanson line of rainbow trout. Differences between the IGH loci of the Swanson and Arlee lines of rainbow trout led us to update this nomenclature for this species.

2.1 Extending IGH nomenclature to the Arlee strain genome

While it was important to maintain nomenclature continuity as much as possible, a gene name could be transferred only when there was convincing evidence of the gene equivalence. We therefore combined sequence similarity and positional information to extend the IMGT nomenclature to the repertoire of the Arlee line. Finding the best compromise can be challenging especially for IGHV genes in the large duplicated loci found in Salmonids. We therefore used at best the flexibility offered by the IMGT names. IGHV names are constituted as follows: first, IGHV, then the subgroup number, the letter D if the gene is located in the duplicated locus (IGH B, on chromosome 12), then a first dash and a number (N_1) defining the gene rank in the locus of reference (for Salmonids, *Salmo salar*, see (Magadan, Krasnov et al. 2019)) and finally, when necessary, a second dash followed by a number(N_1) denoting the identity of a gene within a gene microcluster. For example, the name IGHV16D-69-4 denote a gene belonging to the subgroup 16, located in a microcluster at the reference rank 69 within locus IGH B; the final number "4" does not refer to a position information, but is only an ID usually based on the date of the gene description in the microcluster. While this name format may appear to be unusually complicated, it provides an exquisite flexibility to integrate additional genes, or describe gaps while keeping a nomenclature consistent across haplotypes or genomes updates.

2.2 IGHV genes

We first postulated that IGHV (V exon) nucleotide sequence of the same gene should generally not differ by more than 2% between the two lines. We therefore performed a clustering based on this criterion, and found a Swanson counterpart for 87 genes found in the Arlee line, out of 118. Overall, 73% of the Arlee IGHV genes, belonging to subgroups 1, 2, 3, 4, 6, 8, 9, 10, 11, 12, 14, 15, and 16, could be easily matched to a unique counterpart in Swanson (*ie*, with a nucleotide sequence >98% similar). Within this subset, we could verify that all Arlee IGHV located within the locus A on the chromosome 13 (and, respectively, within the locus B on chromosome 12) matched Swanson genes from the corresponding locus.

The other Arlee IGHV sequences were then analyzed individually. IMGT/BlastSearch on IMGT/GENE-DB nucleotide sequences (F+ORF+inframeP) and IMGT/V-QUEST analysis on Atlantic Salmon and Swanson rainbow trout showed that all these sequences could be classified into one of the 16 IGHV subgroups identified in our previous work (Table S1).

These sequences were attributed novel names following the format described above, and extending the member list of each subgroup previously defined in (Magadan, Krasnov et al. 2019). Table S1 presents the IGHV repertoires from Swanson and Arlee lines and the updated gene nomenclature.

We did not detect obvious imbalance of variation between subgroups between Swanson and Arlee genomes. However, we found 3 functional IGHV7 and 3 IGHV14 in the Arlee line while in the Swanson assembly these subgroups count only pseudogenes (or ORF). In both lines, neither IGHV5 nor IGHV16 subgroup contains any functional gene.

Importantly, our data support the validity of the IGHV salmonid subgroups previously defined in (Magadan, Krasnov et al. 2019) based on *Salmo salar* and *Oncorhynchus mykiss*. Future work will be necessary to determine if other *Oncorhynchus* (such as Coho, Chinook,

Sockeye salmon) or *Salmo* (such as the Brown trout) species possess IG genes that require the creation of additional subgroups.

Rainbow trout Swanson and Arlee lines come from distinct populations: the Swanson line is from Alaska, while the Arlee line is from the southern interior region (Montana). The differences between the subgroup content of these lines illustrate that IGHV sequences are subjected to relatively fast variation. Indeed, germline insertions and deletions involving genes from nearly every IGHV subgroup has been described in humans (Pramanik, Cui et al. 2011, Watson and Breden 2012), suggesting that particular regions of IGH locus may be subject to frequent rearrangements. Significant difference between the levels of divergence might imply difference in selection pressure on regions of IGH locus and the functional importance of these IGHV genes. Thus, a wider dataset collecting the full IGHV repertoire of multiple lines would allow an in-depth analysis to investigate patterns of geographic variation. In particular, it would be interesting to test whether these variations may be adaptive, possibly under selection pressure exerted by pathogens.

2.3 IGHD and IGHJ genes

In fish, IGH rearrangements encoding the three immunoglobulin classes (IgM, IgD and IgT/Z) can use the same set of IGHV genes when expressed from a given IGH locus (i.e. IGHA or IGHB in Salmonids). In contrast, distinct IGHD-IGHJ gene clusters are associated either to IGHT or to IGHM-IGHD constant genes (Hansen, Landis et al. 2005). Our nomenclature of salmonid IGHD and IGHJ genes reflects this particular feature of fish IGH (Tables 2 and 3). D and J gene names are constituted as follows (Magadan, Krasnov et al. 2019): first, IGHD or IGHJ then a number X, followed, if the gene is part of DJC cluster encoding IgT, by “T” and a second number Y, then by “D” if the gene belongs to the duplicated locus (ie locus B), and

finally by a star ("*") and the allele number. X is the rank of the gene within a DJC cluster, and Y is the rank of the DJ cluster, both counted from 5' to 3' end of the locus. For example, IGHD1T2D*01 would be the name of the first allele (*01) of the first D gene (D1) within the second IGHD-IGHJ cluster encoding IgT, in the IGHB locus. Our annotation of the IGHD and J genes of the Arlee strain is largely based on the annotation established for the Swanson strain of rainbow trout and for the Atlantic salmon.

Based on the conserved pattern of their RS sequences, characterized by 12 bp spacer and typical heptamer and nonamer sequences (Ramsden, Baetz et al. 1994), 26 IGHD genes were annotated in the Arlee genome assembly, 14 on Chromosome 12 (locus IGHB) and 12 on Chromosome 13 (locus IGHA) (Figure 1). The previous annotation in Swanson genome assembly described 19 IGHD genes on Chr 12, and 11 on Chr 13 (Magadan, Krasnov et al. 2019) and Figure S1. Analysis of nucleotide similarity of Arlee IGHD coding regions indicated that most of them have identical counterparts in the same or different locus of this trout line (Figure 2). In general, they also present a counterpart in Swanson with 98-100% of nucleotide sequence identity. There are 5 exceptions in the IGHB locus: IGHD2T2D, IGHD1T3D, IGHD2T3D, IGHD3T3D and IGHD4T3D (see Table 1 and Figure S2A), either due to real differences between the IGHD genes of the Swanson and Arlee lines, or to assembly issues likely in Swanson. As previously described in Swanson genome assembly (Magadan, Krasnov et al. 2019), only one "long" D gene (37 bp) is found in Arlee rainbow trout genome assembly within the IGHA locus (Figure 2).

Twenty IGHJ genes were identified in Arlee genome assembly, 11 on Chromosome 12 (IGHB locus) and 9 on Chromosome 13 (IGHA locus). In the Swanson genome assembly 9 IGHJ genes were identified on Chr 12, and 10 on Chr 13 (Magadan, Krasnov et al. 2019). Two Arlee IGHJ genes (IGHJ1T2D and IGHJ2T2D) located on Chr12 are pseudogenes, while the

other IGHJ genes appear to be functional. They present a conserved 5'J-RS, with the nonamer and heptamer separated by 22/23bp. As described for IGHD genes, the functional Arlee IGHJ genes have similar counterparts in the same or different locus (Figure 3 and Figure S2B). They also have highly similar or identical counterpart in Swanson (with 98-100% of nucleotide sequence identity), with the exception of IGHJ pseudogenes, IGHJ1T3D, and IGHJ2T3D, all of them belonging to IGHB locus (Table 2).

It is important to note that the impact of sequencing and assembly errors is particularly problematic for the comparative annotation of short genes similar to each other and organized in multiple clusters as IGHD and IGHJ. Therefore, it will be important to integrate new information from genomic and transcriptome/repertoire data to reach a definitive consensus.

3. *Combined sequences from Swanson and Arlee lines cover most of the rainbow trout repertoire of functional IGHV genes*

In salmonids, several research groups have started using repertoire sequencing to monitor IG responses to infection or vaccination (Castro, Jouneau et al. 2013, Krasnov, Jorgensen et al. 2017, Magadan, Jouneau et al. 2018, Magadan, Jouneau et al. 2019, Navelsaker, Magadan et al. 2019). However, the annotation of such data has been hampered by the lack of standardized sequence databases. The analysis of somatic hypermutations in IG sequences was especially problematic in absence of a genomic reference resource for IGH genes (Abos, Estensoro et al. 2018). Another issue was that original names did not reflect phylogenetic relationships between genes from rainbow trout and Atlantic salmon, making inter-specific comparison very complicated.

This encouraged us to use our IMGT standardized nomenclature and database of IGH genes from Atlantic salmon and rainbow trout, to construct an IMGT IGH gene directory for annotation of RepSeq data using IMGT/HighV-QUEST (Alamyar, Giudicelli et al. 2010, Alamyar, Giudicelli et al. 2012, Li, Lefranc et al. 2013). For rainbow trout, this directory comprises all IGHV genes detected in the genome of the Swanson line. However, we realized that several IGHV genes in experimental datasets produced from French farmed trout were missing in the Swanson-based database (R Castro, personal communication; S Magadan and P Boudinot, unpublished data; these sequences were from the "Synthetic" strain maintained at INRA (D'Ambrosio, Phocas et al. 2019)). These sequences were actually highly similar to genes found in the Arlee genome. We therefore believe that the annotation of the Arlee IGH loci presented in this work contributes to fill the gaps noted in the Swanson IGH repertoire.

To further test if the combination of Swanson and Arlee IGH loci provides a broad description of the IGHV repertoire of rainbow trout, we mined the GenBank for such sequences. We found a large collection of rainbow trout cDNA (835 entries) sequenced by Steve Kaattari and his colleagues, using fish from a Troutlodge stock (Brown, Kaattari et al. 2006). This stock was presumably coming from Washington state, hence from a third trout population different from Swanson and Arlee. To compare this cDNA collection to the Arlee and Swanson repertoires, we built the complete similarity matrix based on pairwise alignment of all pairs of sequences, and selected all cDNA which were less than 98% similar to any Arlee or Swanson sequence. We thus identified 53 sequences (6% of the whole collection). These sequences clustered into 10 sets of sequences more than 98% similar to each other (Table 3). All these sequences were then analyzed using IMGT/V-QUEST and

IMGT/BlastSearch (IMGT/GENE-DB nucleotide sequences F+ORF+allP) to check how divergent they were from previously annotated germline gene sequences (Table 3).

Six clusters (1-6; 40 sequences) were 95%-98% similar to their best hit among Swanson or Arlee sequences, and two other clusters (7-8; 8 sequences) were 95-96% similar to sequences present in Arlee only. Differences between these sequences and their known best match do not show bias in the CDR1-IMGT and CDR2-IMGT (Figure S3). These observations are not in favor of somatic hypermutation, however we cannot exclude that this mechanism may explain (at least some of) these differences.

Two clusters were without >=95% match either in Arlee or Swanson (clusters 9-10, see Table 3). For cluster 9, differences between cDNA sequences and their best batch IGHV1-41*01 were distributed along the V domain (2 in CDR1, 4 in FR2, 3 in CDR2 and 3 in FR3, see Figure S3), while for cluster 10 they were rather concentrated in FR1 (10 positions in FR1, one in CDR1, and 2 in FR2, see Figure S3). These sequences may represent two novel IGHV genes, which were not present in the IGH loci of Arlee or Swanson. Overall, these observations indicate that the sequence database constructed from Swanson and Arlee genomes covers a large fraction of the entire rainbow trout repertoire of functional IGHV genes. Locally, populations certainly have additional variants but our data suggest that they likely not represent more than a small proportion of their functional genes. This point will be clarified by future analyses of Ig repertoire sequencing data from trout of various origins. There is apparently a substantial level of structural variation across trout populations, which affects the number and composition of the set of pseudogenes in the IGH loci. Future analyses will show if the locus B is particularly prone to such variations.

Conclusions

This work extended the IMGT classification and nomenclature of rainbow trout IGH genes from the Swanson line to the Arlee line. Our results confirmed the validity of the IGHV subgroups, which were all found in both lines, and the global structure of both loci (ie, the order of IGHV functional genes). It also unveiled differences at the gene level, especially for IGHV pseudogenes. It is difficult at this stage to determine the respective contributions of the sequencing technology used and of true genetic divergence to these differences. A comparison of our extended IGH sequence database (Arlee+Swanson) with all cDNA sequences available in the GenBank strongly suggested that it provides a broad description of the rainbow trout repertoire of functional genes. Hence, this database provides a basis for robust annotation of repertoire sequencing (AIRR) datasets. Importantly, it makes it possible to get sequences annotated at the gene level rather than at the subgroup level, which will be crucial for polymorphism / hypermutation analyses. Future data combining genome and transcript sequencing will help to determine the respective contribution of polymorphism and somatic hypermutation in the variation of IGHV mRNA sequences across various fish stocks, organs, and immune status.

Acknowledgements

This work was supported by the Institut National de la Recherche Agronomique, by the ANR-16-CE20-0002-01 (FishRNAVax). Xunta de Galicia “Grupo Referencia Competitiva 2020” (ED431C 2020/02). SM also acknowledges the contract from Retención de Talento Investigador- Universidade de Vigo. We are grateful to Ben Koop for discussions and for sharing information about salmonid genomics. The assembly of the Arlee line rainbow trout

genome was supported by funds from the USDA Agricultural Research Service in house project number 8082-31000-012. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

Figure legends.

Figure 1. Rainbow trout (*Oncorhynchus mykiss*, line Arlee) IGH locus (A) on chromosome 13 (CM023231.2) (IGHA) and (B) on chromosome 12 (CM023230.2)(IGHB). The orientation of both rainbow trout IGH loci are forward (FWD). IGHV genes are in blue, IGHD in black, IGHJ in red, and IGHC in green. Pseudogenes are in grey. The boxes representing the genes are not to scale. Exons are not shown. Distances are indicated in Mbp (million base pair) or in kb (kilobase). IGH gene names are according to IMGT nomenclature. The location of IGHC exon clusters is schematically indicated and is analyzed in details in Gao *et al.* (Gao, Magadan et al. 2020). This map can be compared with the map of IGH A and B loci based on the Swanson genome (see http://www.imgt.org/IMGTrepertoire/index.php?section=LocusGenes&repertoire=locus&species=rainbow_trout&group=IGH).

IGHV genes encircled in red dotted lines were not found in the Swanson assembly.

Figure 2. Alignment of coding nucleotide IGHD sequences annotated from the Arlee (OmykA) genome assembly (USDA_OmykA_1.1). Sequences from the locus A (respectively B) are represented in blue (respectively in black).

Figure 3. Alignment of coding nucleotide IGHJ sequences annotated from the Arlee (OmykA) genome assembly (USDA_OmykA_1.1). Pseudogene sequences are denoted by a (ψ).

Tables

Table 1. IGHD genes annotated in Arlee genome (USDA_OmykA_1.1) on chromosome 13 (CM023231.2) and chromosome 12 (CM023230.2) and the corresponding counterpart in Swanson genome (Omyk_1.0)

Locus	***O. mykiss Arlee***			***O. mykiss Swanson***	
	IGHD gene	Chromosomal position[1] Start	End	IGHD gene	IGHC associated gene
IGHA (Chr 13)	IGHD1T1*01	51936009	51936021	IGHD1T1*01	IGHT
	IGHD2T1*01	51936280	51936294	IGHD2T1*01	
	IGHD3T1*01	51943169	51943181	IGHD3T1*01	
	IGHD1T2*02	51987496	51987532	IGHD1T2*01	
	IGHD3T2*01	51988273	51988285	IGHD3T2*01	
	IGHD2T2*01	51988548	51988562	IGHD2T2*01	
	IGHD1*02	52049134	52049146	IGHD1*01	IGHM-IGHD
	IGHD2*01	52049585	52049598	IGHD2*01	
	IGHD3*01	52050027	52050038	IGHD3*01	
	IGHD4*01	52050814	52050825	IGHD4*01	
	IGHD5*01	52051256	52051267	IGHD5*01	
	IGHD6*01	52051711	52051722	IGHD6*01	
IGHB (Chr 12)	IGHD2T1D*01	84581653	84581667	IGHD2T1D*01	IGHTD
	IGHD3T1D*01	84588394	84588406	IGHD3T1D*01	
	IGHD4T1D*01	84589031	84589042	IGHD4T1D*01	
	IGHD2T2D*01	84644559	84644573		
	IGHD1T3D*01	84681104	84681116		
	IGHD2T3D*01	84681518	84681537		
	IGHD3T3D*01	84688239	84688251		
	IGHD4T3D*01	84688876	84688887		
	IGHD1D*02	84769576	84769588	IGHD1D*01	IGHMD-IGHDD
	IGHD2D*01	84770031	84770043	IGHD2D*01	
	IGHD3D*01	84770461	84770472	IGHD3D*01	
	IGHD4D*01	84771248	84771259	IGHD4D*01	
	IGHD5D*01	84771692	84771703	IGHD5D*01	
	IGHD6D*01	84772148	84772159	IGHD6D*01	

[1] Chromosomal position of the coding region, without RS sequences.

458

Table 2. IGHJ genes annotated in Arlee genome (USDA_OmykA_1.1) on chromosome 13 (CM023231.2) and chromosome 12 (CM023230.2) and the corresponding counterpart in Swanson genome (Omyk_1.0)

Locus	*O. mykiss Arlee*			*O. mykiss Swanson*	
	IGHJ gene	Chromosomal position[1] Start	End	IGHJ gene	IGHC associated gene
IGHA (Chr 13)	IGHJ1T1*01	51936627	51936674	IGHJ1T1*01	IGHT
	IGHJ2T1*01	51937103	51937154	IGHJ2T1*01	
	IGHJ1T2*01	51988942	51988989	IGHJ1T2*01	
	IGHJ2T2*02	51989455	51989506	IGHJ2T2*01	
	IGHJ1*01	52065893	52065943	IGHJ1*01	IGHM-IGHD
	IGHJ2*01	52066062	52066112	IGHJ2*01	
	IGHJ3*01	52066267	52066317	IGHJ3*01	
	IGHJ4*02	52066602	52066652	IGHJ4*01	
	IGHJ5*02	52066987	52067036	IGHJ5*01	
IGHB (Chr 12)	IGHJ1T1D*01	84582048	84582095	IGHJ1T1D*01	IGHTD
	IGHJ2T1D*01	84582556	84582607	IGHJ2T1D*01	
	IGHJ1T2D [2]	84645351	84645407		
	IGHJ2T2D [2]	84646319	84646392		
	IGHJ1T3D*01[3]	84681919	84681966		
	IGHJ2T3D*01[3]	84682434	84682485		
	IGHJ3D*02	84789478	84789528	IGHJ3D*01	IGHMD-IGHDD
	IGHJ4D*01	84789647	84789697	IGHJ4D*01	
	IGHJ5D *01	84789852	84789902	IGHJ5D*01	
	IGHJ6D*02	84790211	84790261	IGHJ6D*01	
	IGHJ7D*01	84790599	84790648	IGHJ7D*01	

[1] Chromosomal position of the coding region, without RS sequences. [2] Pseudogenes are highlighted in grey. [3]Arlee IGHJ1T3D and IGHJ2T3D nucleotide sequences present 100% identity with Swanson IGHJ1T1D*01 and IGHJ2T1D*01, respectively.

Table 3. Clusters of IGHV sequences from the cDNA collection with no match (>98% similarity) in Swanson and/or Arlee genomes. When a match was found in Arlee only, sequences are underlined and in italic.

	Cluster composition	IGHV subgroup	% of identity with the best match	Best match (swanson)
1	EF442475	8	96.81%	Swanson IGHV8-30*01 F, or Swanson IGHV8-40*01 F
2	EF438713.1; EF438707.1		97.54%; 97.54%	Swanson IGHV9-23*01 F
	EF438675.1; EF438627.1		97.19%; 97.19%	
	EF438621.1; EF438612.1		97.54%; 97.54%	
	EF438604.1; EF438574.1		97.54%; 97.54%	
	EF438570.1; EF438569.1		97.54%; 97.54%	
	EF438517.1; EF438516.1		97.54%; 97.54%	
	EF438499.1	9	97.19%	
3	EF438527.1; EF438508.1; EF438505.1	9	96.49%; 96.14% 96.49%	Swanson IGHV9-15*01 F

4	EF438463.1; EF442512.1; EF442511.1.1		96.81%; 96.10% 96.81%	Swanson IGHV1-13*01 F
5	EF438762.1; EF438736.1; EF438439.1		95.39%; 95.74% 95.39%	Swanson IGHV8-11*01 P
6	EF467934.1; EF438660.1 EF438653.1; EF438651.1 EF438640.1; EF438563.1 EF438560.1; EF438525.1 EF438500.1; EF438491.1 EF438465.1; EF438450.1 EF438449.1; EF438421.1 EF438407.1; EF438399.1 EF438387.1		97.59%; 97.59% 97.94%; 97.94% 97.25%; 97.94% 97.94%; 97.59% 97.59%; 97.94% 97.94%; 97.94% 96.91%; 97.94% 97.59%; 97.94% 97.94%	Swanson IGHV3D-30*01 ORF
7	EF442660.1		91.40%	Swanson IGHV9D-2*01 F
		9	*95%*	*Arlee IGHV9-22-2*
8	EF442518.1; EF442482.1 EF442442.1; EF442438.1 EF442441.1; EF442440.1 EF442439.1		87.72%; 87.72% 87.72%; 87.72% 86.67%; 87.37% 87.37%	Swanson IGHV1-42*01 F
		1	*96%*	*Arlee IGHV1D-14-3 F*
9	EF438784.1;EF438709.1; EF438666.1;EF442641.1		93.06%; 93.06%	Swanson IGHV1-41*01 P
		4	92.71%; 93.06%	
10	EF442503.1	1	92.01%	Swanson IGHV1-18*01 F

References

Abos, B., I. Estensoro, P. Perdiguero, M. Faber, Y. Hu, P. Diaz Rosales, A. G. Granja, C. J. Secombes, J. W. Holland and C. Tafalla (2018). "Dysregulation of B Cell Activity During Proliferative Kidney Disease in Rainbow Trout." Front Immunol **9**: 1203.

Alamyar, E., V. Giudicelli, P. Duroux and M.-P. Lefranc (2010). IMGT/HighV-QUEST: A high-throughput system and web portal for the analysis of rearranged nucleotide sequences of antigen receptors—High-throughput version of IMGT/V-QUEST. 11èmes Journées Ouvertes de Biologie, Informatique et Mathématiques (JOBIM), Montpellier, France.

Alamyar, E., V. Giudicelli, S. Li, P. Duroux and M.-P. Lefranc (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**: 1-2.

Brown, G. D., I. M. Kaattari and S. L. Kaattari (2006). "Two new Ig VH gene families in Oncorhynchus mykiss." Immunogenetics **58**: 933-936.

Castro, R., L. Jouneau, H. P. Pham, O. Bouchez, V. Giudicelli, M. P. Lefranc, E. Quillet, A. Benmansour, F. Cazals, A. Six, S. Fillatreau, O. Sunyer and P. Boudinot (2013). "Teleost fish mount complex clonal IgM and IgT responses in spleen upon systemic viral infection." PLoS Pathog **9**(1): e1003098.

Castro, R., S. Navelsaker, A. Krasnov, L. Du Pasquier and P. Boudinot (2017). "Describing the diversity of Ag specific receptors in vertebrates: Contribution of repertoire deep sequencing." Dev Comp Immunol **75**: 28-37.

D'Ambrosio, J., F. Phocas, P. Haffray, A. Bestin, S. Brard-Fudulea, C. Poncet, E. Quillet, N. Dechamp, C. Fraslin, M. Charles and M. Dupont-Nivet (2019). "Genome-wide estimates of genetic diversity, inbreeding and effective size of experimental and commercial rainbow trout lines undergoing selective breeding." Genet Sel Evol **51**(1): 26.

Danilova, N., J. Bussmann, K. Jekosch and L. A. Steiner (2005). "The immunoglobulin heavy-chain locus in zebrafish: identification and expression of a previously unknown isotype, immunoglobulin Z." Nature immunology **6**: 295-302.

Fillatreau, S., A. Six, S. Magadan, R. Castro, J. O. Sunyer and P. Boudinot (2013). "The astonishing diversity of Ig classes and B cell repertoires in teleost fish." Front Immunol **4**: 28.

Gao, G., S. Magadan, G. C. Waldbieser, R. C. Youngblood, P. A. Wheeler, B. E. Scheffler, G. H. Thorgaard and Y. Palti (2020). "A long reads-based de-novo assembly of the genome of the Arlee homozygous line reveals structural genome variation in rainbow trout." bioRxiv **12.28.424581**.

Hansen, J., E. Landis and R. Phillips (2005). "Discovery of a unique Ig heavy-chain isotype (IgT) in rainbow trout: Implications for a distinctive B cell developmental pathway in teleost fish." Proceedings of the National Academy of Sciences **102**: 6919-6924.

Hou, D., C. Chen, E. J. Seely, S. Chen and Y. Song (2016). "High-Throughput Sequencing-Based Immune Repertoire Study during Infectious Disease." Front Immunol **7**: 336.

Jiang, N., J. A. Weinstein, L. Penland, R. A. White, D. S. Fisher and S. R. Quake (2011). "Determinism and stochasticity during maturation of the zebrafish antibody repertoire." Proceedings of the National Academy of Sciences **108**: 5348-5353.

Krasnov, A., S. M. Jorgensen and S. Afanasyev (2017). "Ig-seq: Deep sequencing of the variable region of Atlantic salmon IgM heavy chain transcripts." Mol Immunol **88**: 99-105.

Lefranc, M.-P. and G. Lefranc (2001). The immunoglobulin factsbook. San Diego, Academic Press.

Lefranc, M. P. and G. Lefranc (2020). "Immunoglobulins or Antibodies: IMGT((R)) Bridging Genes, Structures and Functions." Biomedicines **8**(9).

Lefranc, M. P., C. Pommie, M. Ruiz, V. Giudicelli, E. Foulquier, L. Truong, V. Thouvenin-Contet and G. Lefranc (2003). "IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains." Dev Comp Immunol **27**(1): 55-77.

Li, S., M. P. Lefranc, J. J. Miles, E. Alamyar, V. Giudicelli, P. Duroux, J. D. Freeman, V. D. Corbin, J. P. Scheerlinck, M. A. Frohman, P. U. Cameron, M. Plebanski, B. Loveland, S. R. Burrows, A. T. Papenfuss and E. J. Gowans (2013). "IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling." Nat Commun **4**: 2333.

Liu, X., Y. S. Li, S. A. Shinton, J. Rhodes, L. Tang, H. Feng, C. A. Jette, A. T. Look, K. Hayakawa and R. R. Hardy (2017). "Zebrafish B Cell Development without a Pre-B Cell Stage, Revealed by CD79 Fluorescence Reporter Transgenes." J Immunol **199**(5): 1706-1715.

Magadan, S., L. Jouneau, P. Boudinot and I. Salinas (2019). "Nasal Vaccination Drives Modifications of Nasal and Systemic Antibody Repertoires in Rainbow Trout." J Immunol **203**(6): 1480-1492.

Magadan, S., L. Jouneau, M. Puelma Touzel, S. Marillet, W. Chara, A. Six, E. Quillet, T. Mora, A. M. Walczak, F. Cazals, O. Sunyer, S. Fillatreau and P. Boudinot (2018). "Origin of Public Memory B Cell Clones in Fish After Antiviral Vaccination." Front Immunol **9**: 2115.

Magadan, S., A. Krasnov, S. Hadi-Saljoqi, S. Afanasyev, S. Mondot, D. Lallias, R. Castro, I. Salinas, O. Sunyer, J. Hansen, B. F. Koop, M. P. Lefranc and P. Boudinot (2019). "Standardized IMGT® Nomenclature of Salmonidae IGH Genes, the Paradigm of Atlantic Salmon and Rainbow Trout: From Genomics to Repertoires." Front Immunol **10**: 2541.

Magor, B. G. (2015). "Antibody Affinity Maturation in Fishes-Our Current Understanding." Biology (Basel) **4**(3): 512-524.

Max, E. and S. Fugmann (2013). Immunoglobulins: molecular genetics. Fundamental Immunology. W. Paul. New York Wolkers Kluver and Lippincott.

Navelsaker, S., S. Magadan, L. Jouneau, E. Quillet, N. J. Olesen, H. M. Munang'andu, P. Boudinot and Ø. Evensen (2019). "Sequential Immunization With Heterologous Viruses Does Not Result in Attrition of the B Cell Memory in Rainbow Trout." Front Immunol **10**: 2687.

Palti, Y., G. Gao, M. R. Miller, R. L. Vallejo, P. A. Wheeler, E. Quillet, J. Yao, G. H. Thorgaard, M. Salem and C. E. Rexroad, 3rd (2014). "A resource of single-nucleotide polymorphisms for rainbow trout generated by restriction-site associated DNA sequencing of doubled haploids." Mol Ecol Resour **14**(3): 588-596.

Pearse, D. E., N. J. Barson, T. Nome, G. Gao, M. A. Campbell, A. Abadia-Cardoso, E. C. Anderson, D. E. Rundio, T. H. Williams, K. A. Naish, T. Moen, S. Liu, M. Kent, M. Moser, D. R. Minkley, E. B. Rondeau, M. S. O. Brieuc, S. R. Sandve, M. R. Miller, L. Cedillo, K. Baruch, A. G. Hernandez, G. Ben-Zvi, D. Shem-Tov, O. Barad, K. Kuzishchin, J. C. Garza, S. T. Lindley, B. F. Koop, G. H. Thorgaard, Y. Palti and S. Lien (2019). "Sex-dependent dominance maintains migration supergene in rainbow trout." Nat Ecol Evol **3**(12): 1731-1742.

Pramanik, S., X. Cui, H. Y. Wang, N. O. Chimge, G. Hu, L. Shen, R. Gao and H. Li (2011). "Segmental duplication as one of the driving forces underlying the diversity of the human immunoglobulin heavy chain variable gene region. ." BMC Genomics. **12**: 78.

Quillet, E., M. Dorson, S. Le Guillou, A. Benmansour and P. Boudinot (2007). "Wide range of susceptibility to rhabdoviruses in homozygous clones of rainbow trout." Fish Shellfish Immunol **22**(5): 510-519.

Ramsden, D. A., K. Baetz and G. E. Wu (1994). "Conservation of sequence in recombination signal sequence spacers." Nucleic Acids Res **22**(10): 1785-1796.

Rego, K., J. D. Hansen and E. S. Bromage (2020). "Genomic architecture and repertoire of the rainbow trout immunoglobulin light chain genes." Dev Comp Immunol **113**: 103776.

Ristow, S. S., L. D. Grabowski, C. Ostberg, B. Robison and G. H. Thorgaard (1998). "Development of Long-Term Cell Lines from Homozygous Clones of Rainbow Trout." Journal of Aquatic Animal Health **10**: 75-82.

Ristow, S. S., L. D. Grabowski, P. A. Wheeler, D. J. Prieur and G. H. Thorgaard (1995). "Arlee line of rainbow trout (Oncorhynchus mykiss) exhibits a low level of nonspecific cytotoxic cell activity." Dev Comp Immunol **19**(6): 497-505.

583 Robison, B. D., P. A. Wheeler, K. Sundin, P. Sikka and G. H. Thorgaard (2001). "Composite
584 interval mapping reveals a major locus influencing embryonic development rate in
585 rainbow trout (Oncorhynchus mykiss)." J Hered **92**(1): 16-22.
586 Rognes, T., T. Flouri, B. Nichols, C. Quince and F. Mahe (2016). "VSEARCH: a versatile open
587 source tool for metagenomics." PeerJ **4**: e2584.
588 Salinas, I., Y.-A. Zhang and J. O. Sunyer (2011). "Mucosal immunoglobulins and B cells of
589 teleost fish." Developmental and comparative immunology **35**: 1346-1365.
590 Tonegawa, S. (1983). "Somatic generation of antibody diversity." Nature **302**(5909): 575-
591 581.
592 Watson, C. T. and F. Breden (2012). "The immunoglobulin heavy chain locus: genetic
593 variation, missing data, and implications for human disease." Genes Immun. **13**: 363-
594 373.
595 Wu, T. T. and E. A. Kabat (1970). "An analysis of the sequences of the variable regions of
596 Bence Jones proteins and myeloma light chains and their implications for antibody
597 complementarity." J Exp Med **132**(2): 211-250.
598 Yamaguchi, T., E. Quillet, P. Boudinot and U. Fischer (2018). "What could be the mechanisms
599 of immunological memory in fish?" Fish Shellfish Immunol.
600 Yang, F., G. C. G. C. Waldbieser and C. J. C. J. Lobb (2006). "The Nucleotide Targets of
601 Somatic Mutation and the Role of Selection in Immunoglobulin Heavy Chains of a
602 Teleost Fish." Journal of immunology **176**: 1655.
603 Yasuike, M., J. D. Boer, K. R. V. Schalburg, G. A. Cooper, L. Mckinnel, A. Messmer, S. So, W. S.
604 Davidson and B. F. Koop (2010). "Evolution of duplicated IgH loci in Atlantic salmon ,
605 Salmo salar." BMC genomics **11**: 486.
606 Zhang, Y.-a., I. Salinas, J. Li, D. Parra, S. Bjork, Z. Xu, S. E. Lapatra, J. Bartholomew and J. O.
607 Sunyer (2010). "IgT , a primitive immunoglobulin class specialized in mucosal
608 immunity." Nature Immunology **11**: 827-835.
609 Zwollo, P. (2011). "Dissecting teleost B cell differentiation using transcription factors."
610 Developmental and comparative immunology **35**: 898-905.
611

Figure 1

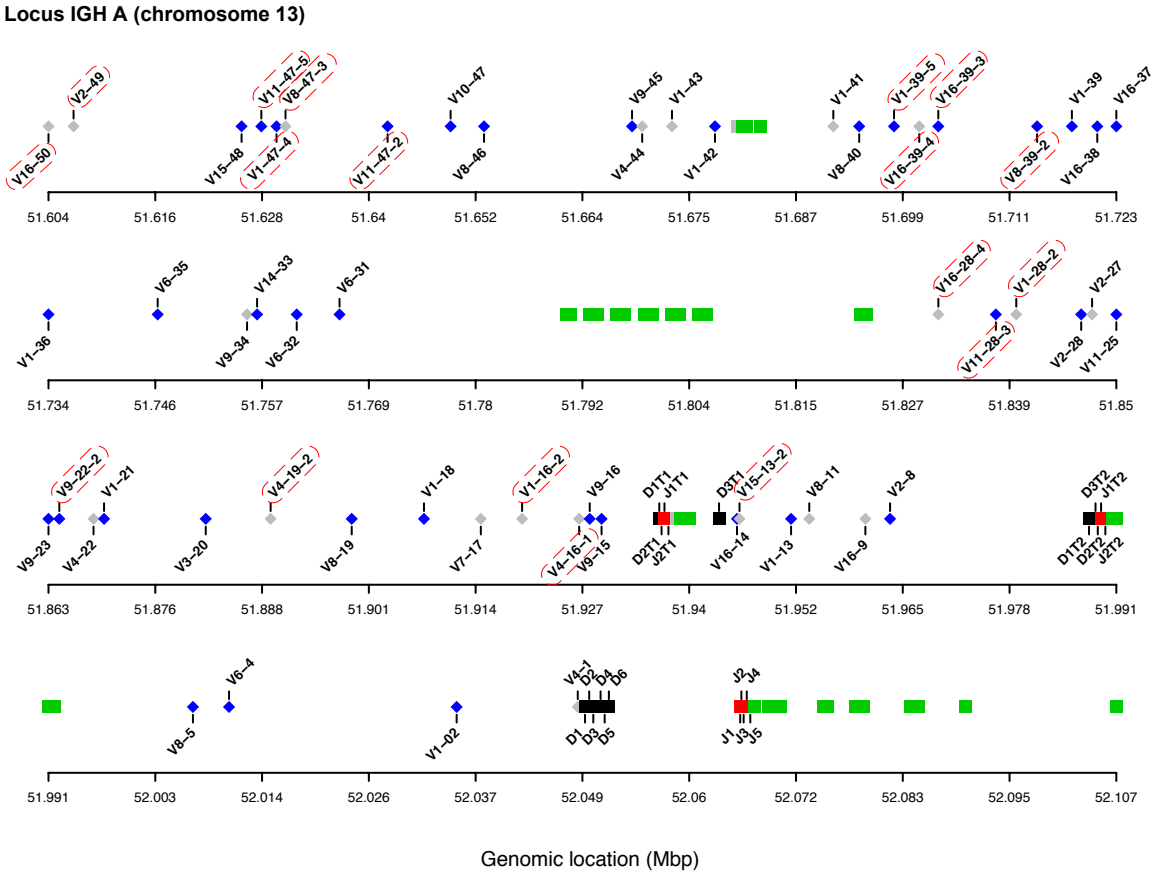
Locus IGH A (chromosome 13)
Genomic location (Mbp)
Locus IGH B (chromosome 12)
Genomic location (Mbp)

Figure 2

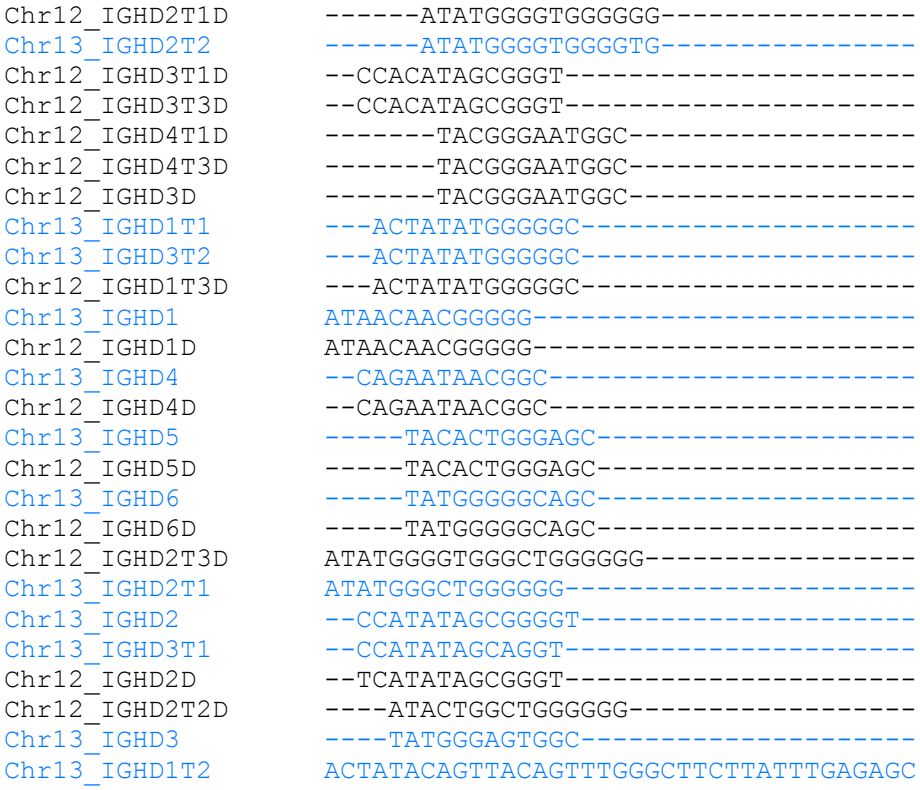
Chr12_IGHD2T1D ------ATATGGGGTGGGGGG----------
Chr13_IGHD2T2 ------ATATGGGGTGGGGTG----------
Chr12_IGHD3T1D --CCACATAGCGGGT----------
Chr12_IGHD3T3D --CCACATAGCGGGT----------
Chr12_IGHD4T1D -------TACGGGAATGGC----------
Chr12_IGHD4T3D -------TACGGGAATGGC----------
Chr12_IGHD3D -------TACGGGAATGGC----------
Chr13_IGHD1T1 ---ACTATATGGGGGC----------
Chr13_IGHD3T2 ---ACTATATGGGGGC----------
Chr12_IGHD1T3D ---ACTATATGGGGGC----------
Chr13_IGHD1 ATAACAACGGGGG----------
Chr12_IGHD1D ATAACAACGGGGG----------
Chr13_IGHD4 --CAGAATAACGGC----------
Chr12_IGHD4D --CAGAATAACGGC----------
Chr13_IGHD5 -----TACACTGGGAGC----------
Chr12_IGHD5D -----TACACTGGGAGC----------
Chr13_IGHD6 -----TATGGGGGCAGC----------
Chr12_IGHD6D -----TATGGGGGCAGC----------
Chr12_IGHD2T3D ATATGGGGTGGGCTGGGGGG----------
Chr13_IGHD2T1 ATATGGGCTGGGGGG----------
Chr13_IGHD2 --CCATATAGCGGGGT----------
Chr13_IGHD3T1 --CCATATAGCAGGT----------
Chr12_IGHD2D --TCATATAGCGGGT----------
Chr12_IGHD2T2D ----ATACTGGCTGGGGGG----------
Chr13_IGHD3 ----TATGGGAGTGGC----------
Chr13_IGHD1T2 ACTATACAGTTACAGTTTGGGCTTCTTATTTGAGAGC

Figure 3.


Chr13_IGHJ1T1 ------ATGCTTTTGACTACTGGGGTAAAGGGACACAAGTCACCGTCTCAACAG------
Chr12_IGHJ1T1D ------ATGCTTTTGACTACTGGGGTAAAGGGACACAAGTCACCGTCTCAACAG------
Chr12_IGHJ1T3D ------ATGCTTTTGACTACTGGGGTAAAGGGACACAAGTCACCGTCTCAACAG------
Chr13_IGHJ1T2 ------ATGCTTTTGACTACTGGGGTAAAGGGACACAAGTCACCGTCTCAGCAG------
Chr12_IGHJ1T2D(ψ) ------ATGCTTTTGACTACTGGGGTAATGTTGGGGAATAATCAGAATTAGTTGGTAACATAG
Chr13_IGHJ2T1 --TATAGCTACTTTGACTATTGGGGGAAAGGAACAATGGTCACAGTCTCATCAG------
Chr12_IGHJ2T1D --TATAGCTACTTTGACTATTGGGGGAAAGGAACAATGGTCACAGTCTCATCAG------
Chr12_IGHJ2T3D --TATAGCTACTTTGACTATTGGGGGAAAGGAACAATGGTCACAGTCTCATCAG------
Chr13_IGHJ2T2 --TATGGATACTTTGACTATTGGGGGAAAGGAACATTGATCACAGTATCATCAG------
Chr12_IGHJ2T2D(ψ) TACACTGATACTTTGACTATTGGGGGAAAGGACCAATGGTCACAGTCTCATCGGGTAAGCCATTGTTTTTGTA
Chr13_IGHJ1 ---ATTACTATTTCGACTACTGGGGGAAAGGGACCCAAGTCACGATAACTTCTG------
Chr12_IGHJ3D ---ATTACTATTTCGACTACTGGGGGAAGGGGACCCAAGTCACGATAACTTCTG------
Chr13_IGHJ2 ---ACAACTACTTTGACTACTGGGGGAAAGGGACCATGGTGACCGTGTCCACAG------
Chr12_IGHJ4D ---ACAACTACTTTGACTACTGGGGGAAAGGGACCATGGTGACCGTGTCCACAG------
Chr13_IGHJ3 ---ACAATGCTTTTGACTACTGGGGGAAAGGCACAATGGTTACCGTTTCATCAG------
Chr12_IGHJ5D ---ACAATGCTTTTGACTACTGGGGGAAAGGCACAATGGTTACCGTTTCATCAG------
Chr13_IGHJ5 -CTACGCTGCTTTTGACTCCTGGGGTCAGGGTACAATAGTAACCGTTTCAG---------
Chr12_IGHJ7D -CTACGCTGCTTTTGACTATTGGGGTCAGGGTACAATAGTAACCGTTTCAG---------
Chr13_IGHJ4 ---ACTACAGTTTTGACTACTGGGGGAAAGGGACAATGGTTACTGTTTCATCAG------
Chr12_IGHJ6D ---ACTACCGTTTTGACTACTGGGGGAAAGGGACAATGGTTACTGTTTCATCAG------