



HAL
open science

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

► To cite this version:

Susana Magadan, Stanislas Mondot, Yniv Palti, Guangtu Gao, Marie Paule Lefranc, et al.. Genomic analysis of a second rainbow trout line (Arlee) leads to an extended description of the IGH VDJ gene repertoire. *Developmental and Comparative Immunology*, 2021, 118, pp.103998. 10.1016/j.dci.2021.103998 . hal-03230900

HAL Id: hal-03230900

<https://hal.science/hal-03230900>

Submitted on 13 Feb 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

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

3
4 Susana Magadan^{1,*}, Stanislas Mondot², Yniv Palti³, Guangtu Gao³,
5 Marie Paule Lefranc⁴, Pierre Boudinot^{5,*}

6
7 ¹ Centro de Investigaciones Biomédicas, Universidade de Vigo, Campus Universitario Lagoas
8 Marcosende, 36310 Vigo, España.

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

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

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

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

16

17

18 Correspondence:

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

21 Phone: 0034 986130142 smaga@uvigo.es

22 And

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

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

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

26 **Abstract**

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

40

41

42

43

44 Introduction

45 Immunoglobulins (IG) – also known as antibodies - are the antigen specific receptors
46 expressed by B cells (Lefranc and Lefranc 2001, Lefranc and Lefranc 2020). **Germline IG loci**
47 **contain non contiguous V,(D) and J genes that can be assembled to produce expressible**
48 **genes encoding IG variable regions** ((Tonegawa 1983), reviewed in (Max and Fugmann
49 2013)). During the B lymphocyte differentiation, IG loci are subjected to genomic
50 rearrangements of variable (V), joining (J), diversity (D) genes, leading to the expression of a
51 unique antigen receptor by each B lymphocyte. The IG heavy (IGH) chains present the
52 highest diversity due to the inclusion of the D gene in a two-step recombination that
53 produces the VH domain (V-D-J-REGION), while the L chains result from a V-J rearrangement
54 that produces the VL domain (V-J-REGION). The V-(D)-J junction **recombination** process is
55 not exact, and the deletion of nucleotides at the joint sites as well as the insertion of so-
56 called palindromic (P) and non-templated (N) nucleotides by the enzyme **terminal**
57 **deoxynucleotidyltransferase (TdT) encoded by the DNA nucleotidylexotransferase (DNNT)**
58 **gene**, are commonly observed (Lefranc and Lefranc 2020). An IGH V-(D)-J rearrangement
59 leads to an exon encoding an IG variable domain (V-DOMAIN) (Lefranc, Pommie et al. 2003).
60 V-domains are beta-barrels, and their antigen binding site is made of three loops named
61 complementarity determining region (CDR)1, 2 and 3 protruding at the top of the domain.
62 The CDR3 is produced by the recombination of the IGH-V(D)J genes and is the most variable
63 mainly due to imprecise recombination and addition/deletion of nucleotides at the D-J and
64 V-(D)-J gene junctions (Wu and Kabat 1970, Lefranc and Lefranc 2001). CDR1 and CDR2 are
65 encoded by IGHV genes; being structurally unconstrained loops, they are highly variable
66 across these genes and also participate to the IGV domain variation. Thus, the diversity of IG

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

69

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

83

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

91 responses are initiated. Detailed phenotypes and functional specialization of B cell subsets
92 still have to be defined, and memory B cells in fish have not been accurately described
93 (Yamaguchi, Quillet et al. 2018). The existence of somatic hypermutation of fish IG
94 sequences is now well established (Yang, Waldbieser et al. 2006, Jiang, Weinstein et al.
95 2011, Magor 2015), but its regulation remains poorly understood. Because IG are central to
96 B cell biology, antibody repertoire analysis is essential for the characterization of these
97 mechanisms and populations. Such approaches remain complex in the absence of a
98 complete and comprehensive genomic description of IG loci in many fish species.

99

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

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

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

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

143

144 In this work, we pursued our effort of annotation and nomenclature standardization using
145 the newly released genome assembly from the Arlee line of rainbow trout
146 (USDA_OmykA_1.1). Both the Swanson and Arlee are homozygous clonal YY male lines
147 developed through androgenesis by the lab of Gary Thorgaard at Washington State
148 University. The Swanson line was derived in 1991 from a semi-wild population from the
149 Swanson River in the Kenai Peninsula of Alaska (Robison, Wheeler et al. 2001). The Arlee
150 clonal line was derived from the Arlee strain, a domesticated hatchery strain that was used
151 by the Montana Department of Fish, Wildlife and Parks (Ristow, Grabowski et al. 1998) and
152 is thought to have originally been collected from Northern California like most farmed
153 rainbow trout stocks that were imported to Europe (Gary Thorgaard, Personal
154 Communication). In addition, fish from the Arlee line were found to have low nonspecific
155 cellular cytotoxicity in the peripheral blood (Ristow, Grabowski et al. 1995). Here we aimed
156 at extending the IMGT rainbow trout IGH annotation to a second line coming from a
157 different region of North America. Testing the salmonid IGH classification and nomenclature
158 established from the Swanson line on the IGH repertoire present in the Arlee line confirmed
159 the validity of the IGHV subgroups, and provided a first picture of the variation of these
160 genes between rainbow trout populations. This update was performed to take into account
161 all new sequences, while keeping nomenclature continuity. **We thus** produced a non-
162 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 $\geq 98\%$
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

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

193

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

196 Rainbow trout IGHV cDNA sequences were collected from GenBank using Blastn with all
197 sequences from Swanson and Arlee. cDNA sequences (301 bp upstream the YxC pattern
198 corresponding to the end of the V gene) were clustered using VSEARCH (Rognes, Flouri et al.
199 2016) (threshold 98% identity). An identity matrix was inferred from pairwise local sequence
200 alignments computed with ClustalW.

201

202 **Results and discussion**

203 **1. Comparison of the structure of IGHV loci between Swanson and Arlee rainbow** 204 **trout lines**

205 The newly released genome assembly from the Arlee clonal line of rainbow trout
206 (USDA_OmykA_1.1) resulted in elucidation of complex loci such as the IGH locus. In the
207 Arlee genome assembly, the IGH genes are within two regions named IGHA and IGHB, that
208 are located on chromosomes 13 (CM023231.2) and 12 (CM023230.2), respectively. This is
209 similar to the rainbow trout reference genome Omyk_1.0 (GCA_002163495.1) (Pearse,
210 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 IGJ 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**

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

238 **2.1 Extending IGH nomenclature to the Arlee strain genome**

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

256

257 **2.2 IGHV genes**

258 We first postulated that IGHV (V exon) nucleotide sequence of the same gene should
259 generally not differ by more than 2% between the two lines. We therefore performed a
260 clustering based on this criterion, and found a Swanson counterpart for 87 genes found in
261 the Arlee line, out of 118. Overall, 73% of the Arlee IGHV genes, belonging to subgroups 1, 2,
262 3, 4, 6, 8, 9, 10, 11, 12, 14, 15, and 16, could be easily matched to a unique counterpart in
263 Swanson (*ie*, with a nucleotide sequence >98% similar). Within this subset, we could verify
264 that all Arlee IGHV located within the locus A on the chromosome 13 (and, respectively,
265 within the locus B on chromosome 12) matched Swanson genes from the corresponding
266 locus.

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

271 These sequences were attributed novel names following the format described above, and
272 extending the member list of each subgroup previously defined in (Magadan, Krasnov et al.
273 2019). [Table S1](#) presents the IGHV repertoires from Swanson and Arlee lines and the
274 updated gene nomenclature.

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

279 Importantly, our data support the validity of the IGHV salmonid subgroups previously
280 defined in (Magadan, Krasnov et al. 2019) based on *Salmo salar* and *Oncorhynchus mykiss*.

281 Future work will be necessary to determine if other *Oncorhynchus* (such as Coho, Chinook,

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

284 Rainbow trout Swanson and Arlee lines come from distinct populations: the Swanson line is
285 from Alaska, while the Arlee line is from the southern interior region (Montana). The
286 differences between the subgroup content of these lines illustrate that IGHV sequences are
287 subjected to relatively fast variation. Indeed, germline insertions and deletions involving
288 genes from nearly every IGHV subgroup has been described in humans (Pramanik, Cui et al.
289 2011, Watson and Breden 2012), suggesting that particular regions of IGH locus may be
290 subject to frequent rearrangements. Significant difference between the levels of divergence
291 might imply difference in selection pressure on regions of IGH locus and the functional
292 importance of these IGHV genes. Thus, a wider dataset collecting the full IGHV repertoire of
293 multiple lines would allow an in-depth analysis to investigate patterns of geographic
294 variation. In particular, it would be interesting to test whether these variations may be
295 adaptive, possibly under selection pressure exerted by pathogens.

296

297 **2.3 IGHD and IGHJ genes**

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

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

312 Based on the conserved pattern of their RS sequences, characterized by 12 bp spacer and
313 typical heptamer and nonamer sequences (Ramsden, Baetz et al. 1994), 26 IGHD genes
314 were annotated in the Arlee genome assembly, 14 on Chromosome 12 (locus IGHB) and 12
315 on Chromosome 13 (locus IGHA) (Figure 1). The previous annotation in Swanson genome
316 assembly described 19 IGHD genes on Chr 12, and 11 on Chr 13 (Magadan, Krasnov et al.
317 2019) and Figure S1. Analysis of nucleotide similarity of Arlee IGHD coding regions
318 indicated that most of them have identical counterparts in the same or different locus of
319 this trout line (Figure 2). In general, they also present a counterpart in Swanson with 98-
320 100% of nucleotide sequence identity. There are 5 exceptions in the IGHB locus:
321 IGHD2T2D, IGHD1T3D, IGHD2T3D, IGHD3T3D and IGHD4T3D (see Table 1 and Figure S2A),
322 either due to real differences between the IGHD genes of the Swanson and Arlee lines, or
323 to assembly issues likely in Swanson. As previously described in Swanson genome
324 assembly (Magadan, Krasnov et al. 2019), only one "long" D gene (37 bp) is found in Arlee
325 rainbow trout genome assembly within the IGHA locus (Figure 2).

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

330 other IGHJ genes appear to be functional. They present a conserved 5'J-RS, with the
331 nonamer and heptamer separated by 22/23bp. As described for IGHD genes, the functional
332 Arlee IGHJ genes have similar counterparts in the same or different locus (Figure 3 and
333 Figure S2B). They also have highly similar or identical counterpart in Swanson (with 98-
334 100% of nucleotide sequence identity), with the exception of IGHJ pseudogenes,
335 IGHJ1T3D, and IGHJ2T3D, all of them belonging to IGHB locus (Table 2).

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

341

342 **3. Combined sequences from Swanson and Arlee lines cover most of the rainbow** 343 **trout repertoire of functional IGHV genes**

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

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

365

366 To further test if the combination of Swanson and Arlee IGH loci provides a **broad**
367 description of the IGHV repertoire of rainbow trout, we mined the GenBank for such
368 sequences. We found a large collection of rainbow trout cDNA (835 entries) sequenced by
369 Steve Kaattari and his colleagues, using fish from a Troutlodge stock (Brown, Kaattari et al.
370 2006). This stock was presumably coming from Washington state, hence from a third trout
371 population different from Swanson and Arlee. To compare this cDNA collection to the Arlee
372 and Swanson repertoires, we built the complete similarity matrix based on pairwise
373 alignment of all pairs of sequences, and selected all cDNA which were less than 98% similar
374 to any Arlee or Swanson sequence. We thus identified 53 sequences (6% of the whole
375 collection). These sequences clustered into 10 sets of sequences more than 98% similar to
376 each other (Table 3). All these sequences were then analyzed using IMGT/V-QUEST and

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

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

385

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

400

401 **Conclusions**

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

417

418 **Acknowledgements**

419 This work was supported by the Institut National de la Recherche Agronomique, by the ANR-
420 16-CE20-0002-01 (FishRNAVax). Xunta de Galicia “Grupo Referencia Competitiva 2020”
421 (ED431C 2020/02). SM also acknowledges the contract from Retención de Talento
422 Investigador- Universidade de Vigo. We are grateful to Ben Koop for discussions and for
423 sharing information about salmonid genomics. The assembly of the Arlee line rainbow trout

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

429

430 **Figure legends.**

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

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

443

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

447

448 **Figure 3.** Alignment of coding nucleotide IGJH sequences annotated from the Arlee (OmykA)449 genome assembly (USDA_OmykA_1.1). Pseudogene sequences are denoted by a (ψ).

450

451

452 **Tables**453 **Table 1.** IGHD genes annotated in Arlee genome (USDA_OmykA_1.1) on chromosome 13

454 (CM023231.2) and chromosome 12 (CM023230.2) and the corresponding counterpart in

455 Swanson genome (Omyk_1.0)

456

Locus	<i>O. mykiss Arlee</i>			<i>O. mykiss Swanson</i>	
	IGHD gene	Chromosomal position ¹		IGHD gene	IGHC associated gene
		Start	End		
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	IGHM-IGHD
	IGHD1*02	52049134	52049146	IGHD1*01	
	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		IGHMD-IGHDD
	IGHD1D*02	84769576	84769588	IGHD1D*01	
	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		

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

459 **Table 2.** IGHJ genes annotated in Arlee genome (USDA_OmykA_1.1) on chromosome 13
 460 (CM023231.2) and chromosome 12 (CM023230.2) and the corresponding counterpart in
 461 Swanson genome (Omyk_1.0)
 462

Locus	<i>O. mykiss</i> Arlee			<i>O. mykiss</i> Swanson	
	IGHJ gene	Chromosomal position ¹		IGHJ gene	IGHC associated gene
		Start	End		
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	IGHM-IGHD
	IGHJ1*01	52065893	52065943	IGHJ1*01	
	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		IGHMD-IGHDD
	IGHJ2T3D*01 ³	84682434	84682485		
	IGHJ3D*02	84789478	84789528	IGHJ3D*01	
	IGHJ4D*01	84789647	84789697	IGHJ4D*01	
	IGHJ5D*01	84789852	84789902	IGHJ5D*01	
	IGHJ6D*02	84790211	84790261	IGHJ6D*01	
	IGHJ7D*01	84790599	84790648	IGHJ7D*01	

463 ¹ Chromosomal position of the coding region, without RS sequences. ² Pseudogenes are highlighted in
 464 grey. ³Arlee IGHJ1T3D and IGHJ2T3D nucleotide sequences present 100% identity with Swanson
 465 IGHJ1T1D*01 and IGHJ2T1D*01, respectively.
 466

467 **Table 3.** Clusters of IGHV sequences from the cDNA collection with no match (>98%
 468 similarity) in Swanson and/or Arlee genomes. When a match was found in Arlee only,
 469 sequences are underlined and in italic.
 470

	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 EF438675.1; EF438627.1 EF438621.1; EF438612.1 EF438604.1; EF438574.1 EF438570.1; EF438569.1 EF438517.1; EF438516.1 EF438499.1	9	97.54%; 97.54% 97.19%; 97.19% 97.54%; 97.54% 97.54%; 97.54% 97.54%; 97.54% 97.54%; 97.54%	Swanson IGHV9-23*01 F
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	1	96.81%; 96.10% 96.81%	Swanson IGHV1-13*01 F
5	EF438762.1; EF438736.1; EF438439.1	8	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	3	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
8	EF442518.1; EF442482.1 EF442442.1; EF442438.1 EF442441.1; EF442440.1 EF442439.1	9	<u>95%</u> 87.72%; 87.72% 87.72%; 87.72% 86.67%; 87.37% 87.37%	<u>Arlee IGHV9-22-2</u> Swanson IGHV1-42*01 F
		1	<u>96%</u>	<u>Arlee IGHV1D-14-3 F</u>
9	EF438784.1; EF438709.1; EF438666.1; EF442641.1	4	93.06%; 93.06% 92.71%; 93.06%	Swanson IGHV1-41*01 P
10	EF442503.1	1	92.01%	Swanson IGHV1-18*01 F

471

472

473 References

- 474 Abos, B., I. Estensoro, P. Perdiguero, M. Faber, Y. Hu, P. Diaz Rosales, A. G. Granja, C. J.
475 Secombes, J. W. Holland and C. Tafalla (2018). "Dysregulation of B Cell Activity During
476 Proliferative Kidney Disease in Rainbow Trout." *Front Immunol* **9**: 1203.
- 477 Alamyar, E., V. Giudicelli, P. Duroux and M.-P. Lefranc (2010). IMGT/HighV-QUEST: A high-
478 throughput system and web portal for the analysis of rearranged nucleotide
479 sequences of antigen receptors—High-throughput version of IMGT/V-QUEST. 11èmes
480 Journées Ouvertes de Biologie, Informatique et Mathématiques (JOBIM), Montpellier,
481 France.
- 482 Alamyar, E., V. Giudicelli, S. Li, P. Duroux and M.-P. Lefranc (2012). "IMGT/HighV-QUEST: The
483 IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR)
484 analysis from NGS high throughput and deep sequencing." *Immunome Res* **8**: 1-2.
- 485 Brown, G. D., I. M. Kaattari and S. L. Kaattari (2006). "Two new Ig VH gene families in
486 *Oncorhynchus mykiss*." *Immunogenetics* **58**: 933-936.
- 487 Castro, R., L. Jouneau, H. P. Pham, O. Bouchez, V. Giudicelli, M. P. Lefranc, E. Quillet, A.
488 Benmansour, F. Cazals, A. Six, S. Fillatreau, O. Sunyer and P. Boudinot (2013). "Teleost
489 fish mount complex clonal IgM and IgT responses in spleen upon systemic viral
490 infection." *PLoS Pathog* **9**(1): e1003098.

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

494 D'Ambrosio, J., F. Phocas, P. Haffray, A. Bestin, S. Brard-Fudulea, C. Poncet, E. Quillet, N.
495 Dechamp, C. Fraslin, M. Charles and M. Dupont-Nivet (2019). "Genome-wide
496 estimates of genetic diversity, inbreeding and effective size of experimental and
497 commercial rainbow trout lines undergoing selective breeding." Genet Sel Evol **51**(1):
498 26.

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

502 Fillatreau, S., A. Six, S. Magadan, R. Castro, J. O. Sunyer and P. Boudinot (2013). "The
503 astonishing diversity of Ig classes and B cell repertoires in teleost fish." Front Immunol
504 **4**: 28.

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

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

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

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

517 Krasnov, A., S. M. Jorgensen and S. Afanasyev (2017). "Ig-seq: Deep sequencing of the
518 variable region of Atlantic salmon IgM heavy chain transcripts." Mol Immunol **88**: 99-
519 105.

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

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

524 Lefranc, M. P., C. Pommie, M. Ruiz, V. Giudicelli, E. Foulquier, L. Truong, V. Thouvenin-
525 Contet and G. Lefranc (2003). "IMGT unique numbering for immunoglobulin and T cell
526 receptor variable domains and Ig superfamily V-like domains." Dev Comp Immunol
527 **27**(1): 55-77.

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

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

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

539 Magadan, S., L. Jouneau, M. Puelma Touzel, S. Marillet, W. Chara, A. Six, E. Quillet, T. Mora,
540 A. M. Walczak, F. Cazals, O. Sunyer, S. Fillatreau and P. Boudinot (2018). "Origin of
541 Public Memory B Cell Clones in Fish After Antiviral Vaccination." Front Immunol **9**:
542 2115.

543 Magadan, S., A. Krasnov, S. Hadi-Saljoqi, S. Afanasyev, S. Mondot, D. Lallias, R. Castro, I.
544 Salinas, O. Sunyer, J. Hansen, B. F. Koop, M. P. Lefranc and P. Boudinot (2019).
545 "Standardized IMGT® Nomenclature of Salmonidae IGH Genes, the Paradigm of
546 Atlantic Salmon and Rainbow Trout: From Genomics to Repertoires." Front Immunol
547 **10**: 2541.

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

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

552 Navelsaker, S., S. Magadan, L. Jouneau, E. Quillet, N. J. Olesen, H. M. Munang'andu, P.
553 Boudinot and Ø. Evensen (2019). "Sequential Immunization With Heterologous Viruses
554 Does Not Result in Attrition of the B Cell Memory in Rainbow Trout." Front Immunol
555 **10**: 2687.

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

560 Pearse, D. E., N. J. Barson, T. Nome, G. Gao, M. A. Campbell, A. Abadia-Cardoso, E. C.
561 Anderson, D. E. Rundio, T. H. Williams, K. A. Naish, T. Moen, S. Liu, M. Kent, M. Moser,
562 D. R. Minkley, E. B. Rondeau, M. S. O. Briec, S. R. Sandve, M. R. Miller, L. Cedillo, K.
563 Baruch, A. G. Hernandez, G. Ben-Zvi, D. Shem-Tov, O. Barad, K. Kuzishchin, J. C. Garza,
564 S. T. Lindley, B. F. Koop, G. H. Thorgaard, Y. Palti and S. Lien (2019). "Sex-dependent
565 dominance maintains migration supergene in rainbow trout." Nat Ecol Evol **3**(12):
566 1731-1742.

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

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

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

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

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

580 Ristow, S. S., L. D. Grabowski, P. A. Wheeler, D. J. Prieur and G. H. Thorgaard (1995). "Arlee
581 line of rainbow trout (*Oncorhynchus mykiss*) exhibits a low level of nonspecific
582 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.

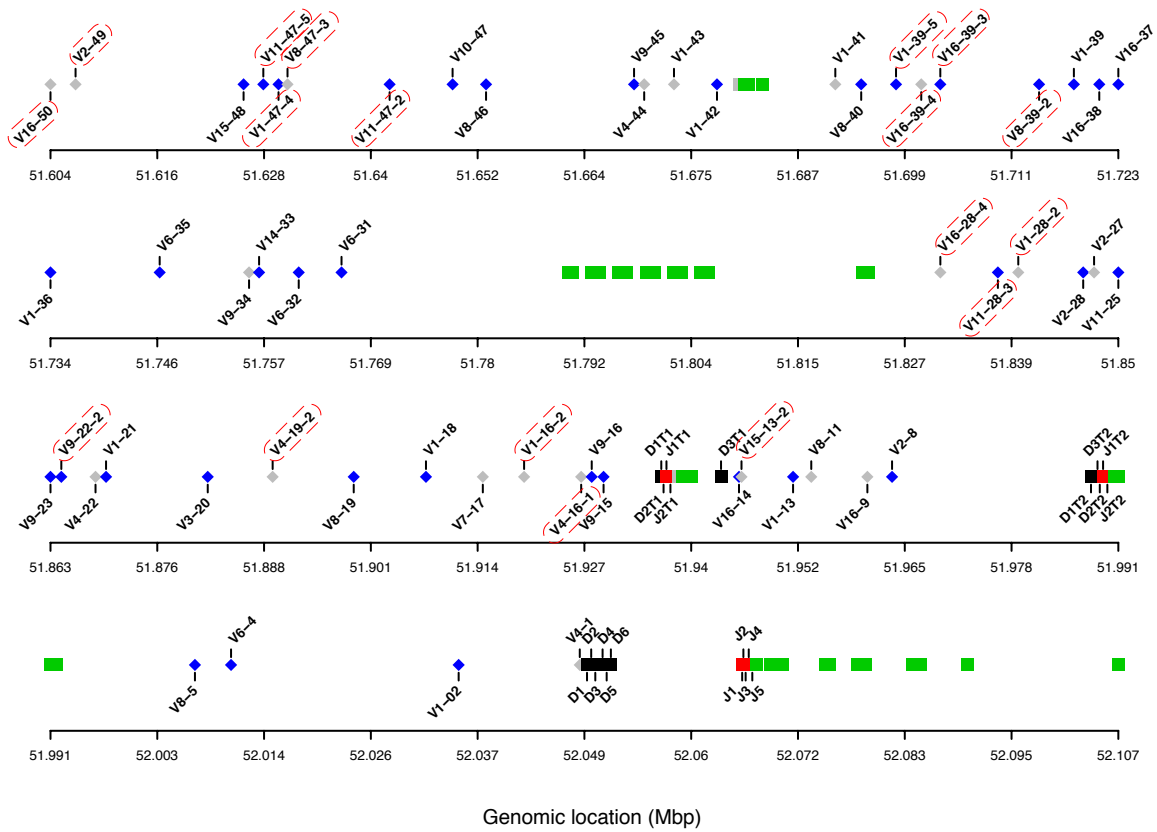
606 Zhang, Y.-a., I. Salinas, J. Li, D. Parra, S. Bjork, Z. Xu, S. E. Lapatra, J. Bartholomew and J. O.
607 Sunyer (2010). "IgT , a primitive immunoglobulin class specialized in mucosal
608 immunity." Nature Immunology **11**: 827-835.

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)



Locus IGH B (chromosome 12)

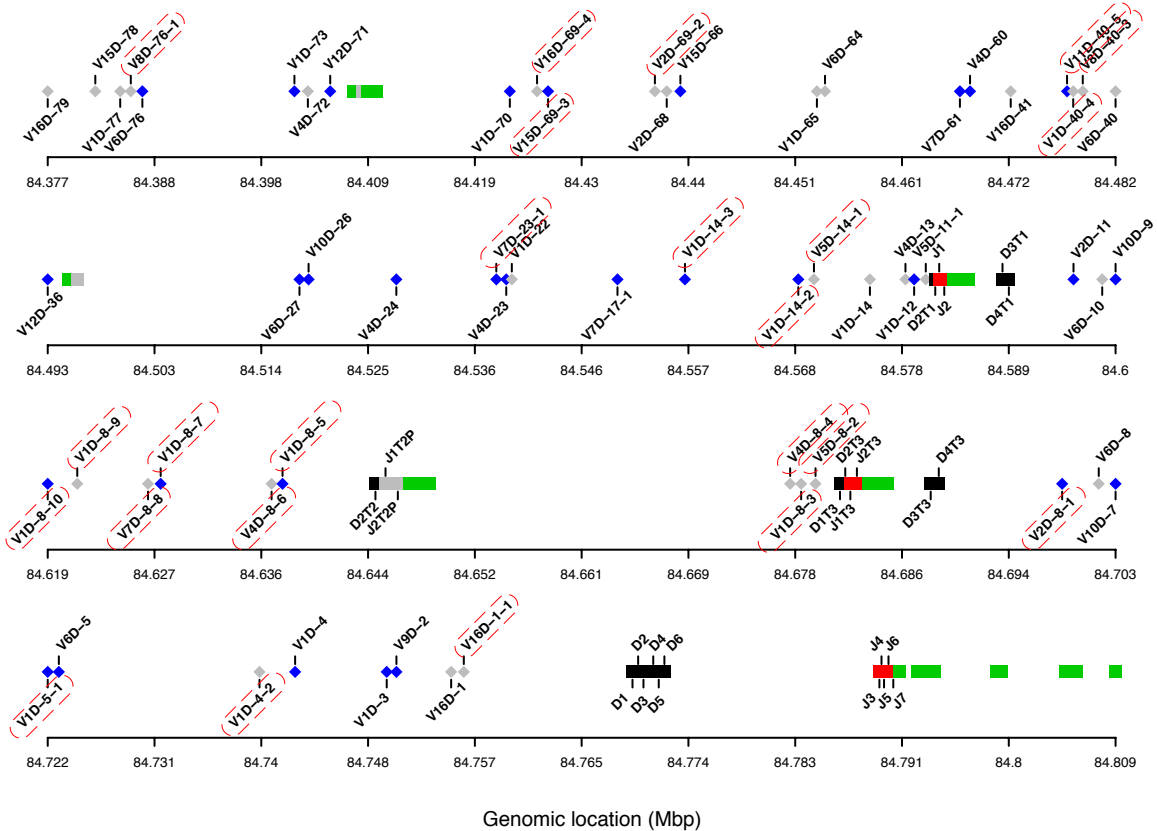


Figure 2

Chr12_IGHD2T1D	-----ATATGGGGTGGGGG-----
Chr13_IGHD2T2	-----ATATGGGGTGGGGT-----
Chr12_IGHD3T1D	--CCACATAGCGGGT-----
Chr12_IGHD3T3D	--CCACATAGCGGGT-----
Chr12_IGHD4T1D	-----TACGGGAATGGC-----
Chr12_IGHD4T3D	-----TACGGGAATGGC-----
Chr12_IGHD3D	-----TACGGGAATGGC-----
Chr13_IGHD1T1	---ACTATATGGGGC-----
Chr13_IGHD3T2	---ACTATATGGGGC-----
Chr12_IGHD1T3D	---ACTATATGGGGC-----
Chr13_IGHD1	ATAACAACGGGG-----
Chr12_IGHD1D	ATAACAACGGGG-----
Chr13_IGHD4	--CAGAATAACGGC-----
Chr12_IGHD4D	--CAGAATAACGGC-----
Chr13_IGHD5	-----TACACTGGGAGC-----
Chr12_IGHD5D	-----TACACTGGGAGC-----
Chr13_IGHD6	-----TATGGGGCAGC-----
Chr12_IGHD6D	-----TATGGGGCAGC-----
Chr12_IGHD2T3D	ATATGGGGTGGGCTGGGGG-----
Chr13_IGHD2T1	ATATGGGCTGGGGG-----
Chr13_IGHD2	--CCATATAGCGGGT-----
Chr13_IGHD3T1	--CCATATAGCAGGT-----
Chr12_IGHD2D	--TCATATAGCGGGT-----
Chr12_IGHD2T2D	---A-TACTGGCTGGGGG-----
Chr13_IGHD3	----TATGGGAGTGGC-----
Chr13_IGHD1T2	ACTATACAGTTACAGTTTGGGCTTCTTATTTGAGAGC

