Characterising antibody immunity and ageing in a short-lived teleost

Inaugural-Dissertation
zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von

William John Bradshaw
aus Hastings

Köln 2020

Gutachter: Dr. Dario Riccardo Valenzano
Prof. Dr. Andreas Beyer

Tag der mündlichen Prüfung: 6. Juni 2019

Creér, c'est recombiner.

– François Jacob [1]

Kurzzusammenfassung

Alternde Menschen zeigen einen stetigen Rückgang der adaptiven Immunfunktion, mit wichtigen Auswirkungen auf Gesundheit und Lebensdauer. Systemische Veränderungen, die in der Struktur und Diversität des Antikörperrepertoires mit dem Alter beobachtet werden, spielen eine wichtige Rolle in diesem immunoseneszenten Phänotypen; die relativ lange Lebensdauer der meisten Wirbeltier-Modellorganismen macht eine gründliche Untersuchung des alternden Immunrepertoires jedoch schwierig. Als natürlich kurzlebiges Wirbeltier bietet der Türkise Prachtgrundkärpfling (*Nothobranchius furzeri*) eine aufregende neue Gelegenheit, die Alterung des adaptiven Immunsystems im Allgemeinen und des Antikörperrepertoires im Besonderen zu untersuchen.

In dieser Arbeit habe ich eine Kombination aus bestehenden genomischen Assemblierungen und neu generierten Sequenzierungsdaten verwendet, um den Immunoglobulin-H-Ketten-Locus des Türkisen Prachtgrundkärpflings zusammenzusetzen, zu charakterisieren und mit Loci eng verwandter Arten zu vergleichen. Dies zeigt eine dynamische Locus-Entwicklung mit sich wiederholenden Duplizierungen und Verlusten des spezialisierten mukosalen Isotypen *IGHZ*. Diese Ergebnisse unterstützen eine hohe Evolutionsrate der Immunoglobulin-H-Ketten-Loci in den Teleostei und bilden eine solide Grundlage für die Forschung der vergleichenden evolutionären Immunologie in Zahnkärpflingen.

Nachdem ich die Immunoglobulin-H-Ketten-Locus-Sequenz in *N. furzeri* charakterisiert hatte, nutzte ich sie, um Immunoglobulin-Sequenzierung gezielt für diese Spezies zu etablieren, welche eine quantitative Analyse des Antikörperrepertoires ermöglicht. Die Anwendung dieses Protokolls auf Ganzkörper-Fischproben ergab komplexe und individualisierte Antikörperrepertoires, die in der intraindividuellen Diversität mit dem Alter rasch abnehmen und in der interindividuellen Variabilität zunehmen. Dies zeigt, dass die Antikörperrepertoires von Türkisen Prachtgrundkärpflingen, entsprechend ihrer kurzen Lebensdauer schnelle Alterungsprozesse aufweisen. Dieser altersbedingte Diversitätsverlust war besonders stark in Darmproben - ein Phänomen, das mit der konstanten ausgeprägten Antigenexposition an Schleimhautoberflächen zusammenhängen kann und bisher nicht in einem Wirbeltier-Modell untersucht wurde. Zusammenfassend etablieren diese Ergebnisse den Türkisen Prachtgrundkärpfling als neuartiges Modell für die Erforschung von Immunoseneszenz in Wirbeltieren und bilden die Grundlage für zukünftige Analysen von und Interventionsstudien an adaptiver Immunalterung.

Abstract

Ageing individuals exhibit a pervasive decline in adaptive immune function, with important implications for health and lifespan. Systemic changes observed in the structure and diversity of antibody repertoires with age are thought to play an important role in this immunosenescent phenotype; however, the relatively long lifespan of most vertebrate model organisms makes thorough investigation of the ageing repertoire challenging. As a naturally short-lived vertebrate, the turquoise killifish (*Nothobranchius furzeri*) offers an exciting new opportunity to study the ageing of the adaptive immune system in general and antibody repertoires in particular.

In this thesis, I used a combination of existing genomic assemblies and new sequencing data to assemble and characterise the immunoglobulin heavy chain (*IGH*) locus sequence in the turquoise killifish and compare it to those of closely related species, revealing a history of dynamic locus evolution and repeated duplication and loss of the specialised mucosal isotype *IGHZ*. The *N. furzeri* locus itself lacks *IGHZ*, making it one of the few known teleost species not to possess this isotype. These results support a high rate of evolution in teleost *IGH* loci and set a strong foundation for the study of comparative evolutionary immunology in cyprionodontiform fishes.

Having characterised the *IGH* locus sequence in *N. furzeri*, I used it to establish targeted immunoglobulin sequencing in this species, enabling quantitative interrogation of the antibody repertoire. Applying this protocol to whole-body killifish samples revealed complex and individualised antibody repertoires which decline rapidly in within-individual diversity and increase in between-individual variability with age, demonstrating that turquoise killifish exhibit a rapid repertoire-ageing phenotype in line with their short lifespans. This loss of diversity with age was particularly strong in isolated gut samples, a phenomenon that may be related to the constant strong antigenic exposure experienced at mucosal surfaces and has not been previously investigated in a vertebrate model. Taken together, these results establish the turquoise killifish as a novel model for vertebrate immunosenescence and lay the groundwork for future interrogation of – and intervention in – adaptive-immune ageing.

Acknowledgements

A doctoral thesis in the life sciences might have a single author, but almost always has a great many direct and indirect contributors. Mine is no exception, and so I have a number of people to thank.

First, my sincere thanks to Dario Valenzano, who welcomed me into his research group in 2014 and has always been an enthusiastic, attentive and understanding supervisor. None of the work in this thesis would have been possible without him, and I would be a far poorer scientist without his mentorship and support.

Many members of the Valenzano group have actively contributed to this thesis. Patrick Smith performed the microbiota-transfer studies that were the inspiration for the project in the first place and supervised and mentored me through my first and most difficult year. Joanna Dodzian generously bred and provided the fish cohorts I used for the pilot and ageing experiments in Chapter 4. Rongfeng (Ray) Cui performed the genome assembly that laid the foundation for the work in Chapter 3 and provided a great deal of advice on both experimental procedures and computational analysis. My student supervisees, Linda Zirden, Lena Schlautmann, Pascha Hokama, and Davina Patel all made essential contributions to developing and testing the turquoise-killifish immunoglobulin-sequencing protocol; you were all great.

Our tireless and wonderful technician, Aleksandra Placzek, has been an essential part of the IgSeq project in the lab and performed several of the library preps; many of my favourite parts of this thesis would not have been possible without her. Most recently, Michael Poeschla has joined the repertoire-sequencing team and helped prepare the libraries for the final IgSeq experiment in Chapter 4; I look forward to seeing where he takes the project in the future. In additional to Dario and Michael, my fellow students David Willemsen, Jens Seidel, Miriam Popkes and Daniel Davila all helped proofread the manuscript for this thesis; Miriam, David and especially Jens also did the lion's share of the work translating my abstract into German. I am sincerely grateful to all of them, and to all other past and present members of the Valenzano lab, for their help, support, and enthusiasm, and for making the group such a pleasant and friendly place in which to work.

A number of external collaborators and other scholars also made vital contributions to this work. Kathrin Reichwald provided me with essential BAC clones and a substantial amount of important advice. John Beausang, Aleksandra Walczak, Thierry Mora and Susana Magadan helped me get started in the enormous and somewhat intimidating world of immune-repertoire analysis, and Aleksandra in

particular has repeatedly provided invaluable advice, introductions and insight. Jason Vander Heiden has tirelessly answered my many questions about the Immcantation pipeline and even made some changes to it at my request, while Quentin Marcou has been an invaluable help in getting to grips with his fantastic software IGoR. Tsimafei Padvitski and Davi Silva generously provided additional proofreading. Many other researchers have responded to my pestering emails, given advice where I needed it, and generally been friendly, welcoming and generous with their time. I am grateful to you all.

The co-ordination team of the Cologne Graduate School of Ageing Research has consistently been amazing, and I have repeatedly been impressed at the depth of their dedication, sympathy and willingness to help us students with all our big and little problems. My sincere thanks to Daniela Morick, Doris Birker and Jenny Ostermann for all their help and support. Thanks also to Ruth Willmott, for her invaluable soft-skills and careers coaching; Jorge Boucas, Franzi Metge, Daniel Rosskopp and the rest of the bioinformatics core facility for their patient help and advice; and all the other people I have unaccountably failed to mention here whose assistence, friendship and support I have relied on throughout this lengthly process.

My family, of course, have been a constant source of love and support. Mum, Dad, Anna, thank you for everything. I love you. My love and thanks also to the rest of my family, especially Ron and Julie, for their enduring support and belief in me and my research.

Finally, my deepest and sincerest thanks to Simon Woolf, without whom this entire long adventure would have been a great deal more difficult and a great deal less fun.

Table of contents

List of figures

List of tables

List of abbreviations

#	number of
±	plus or minus
×	multiples (e.g. of standard concentration)
%	percent
°	degrees (angle)
°C	degrees Celsius
µM	micromolar concentration (micromoles per litre)
µl	microlitre(s)
µmol	micromole(s)
A. australe	*Aphyosemion australe*
A. limnaeus	*Austrofundulus limnaeus*
a.k.a.	also known as
AA	amino acid(s)
AID	activation-induced cytidine deaminase
BAC	bacterial artificial chromosome
BCR	B-cell receptor
bit	binary digit (unit of information)
bp	base pair(s)
BR	broad-range (Qubit assay)
c.	*circa* (approximately)

C. toddi . . . . . *Callopanchax toddi*

C. variegatus . . . . . *Cyprinodon variegatus*

CD . . . . . IGHD constant-region exon

C_δ . . . . . IGHD constant-region exon

cDNA . . . . . complementary DNA

CDR . . . . . complementarity-determining region

CH . . . . . heavy-chain constant-region exon

chr . . . . . chromosome

CM . . . . . IGHM constant-region exon

C_μ . . . . . IGHM constant-region exon

CSR . . . . . class-switch recombination

CZ . . . . . IGHZ constant-region exon

C_ζ . . . . . IGHZ constant-region exon

dATP . . . . . deoxyadenosine triphosphate

dCTP . . . . . deoxycytidine triphosphate

dGTP . . . . . deoxyguanosine triphosphate

DH . . . . . heavy-chain diversity (gene segment)

DNA . . . . . deoxyribonucleic acid

dNTP . . . . . deoxyribonucleoside triphosphate mix (dATP, dGTP, dCTP and dTTP)

DSB . . . . . double-strand break

DTT . . . . . dithiothreitol (reducing agent)

dTTP . . . . . deoxythymidine triphosphate

E-value . . . . . Expect value (BLAST)

E. coli . . . . . *Escherichia coli*

e.g. . . . . *exempli gratia* (for example)

EB . . . . . elution buffer

EDTA . . . . . ethylenediaminetetraacetic acid

EMBOSS . . . . . European Molecular Biology Open Software Suite

et al. . . . . *et alia* (and others)

etc. . . . . *et cetera* (and so on)

F. heteroclitus . . . . . *Fundulus heteroclitus*

FDC . . . . . follicular dendritic cell

FLI . . . . . Leibniz Institute on Aging – Fritz Lipmann Institute (Jena, Germany)

FR . . . . . framework region

FWR . . . . . framework region

g . . . . . gram(s)

g . . . . . relative centrifugal force

GALT . . . . . gut-associated lymphoid tissue

GLM . . . . . generalised linear model

GRZ . . . . . strain of *N. furzeri* (from Gonarezhou National Park, Zimbabwe)

GRZ-AD . . . . . substrain of GRZ

GRZ-Bellemans . . . . substrain of GRZ

GSP . . . . . gene-specific primer (for reverse transcription)

h . . . . . hour(s)

H. sapiens . . . . . *Homo sapiens*

HiFi . . . . . high-fidelity (PCR)

HMM . . . . . Hidden Markov Model

HS . . . . . high-sensitivity (Qubit assay)

HSC . . . . . haematopoietic stem cell

HSP . . . . . high-scoring segment pair (BLAST)

i.e. . . . . *id est* (that is to say)

ID . . . . . identity

IGH . . . . . immunoglobulin heavy chain

IGHD . . . . . Immunoglobulin D (isotype of IGH)

IGHD-TM . . . . . transmembrane IGHD

IGHM . . . . . Immunoglobulin M (isotype of IGH)

IGHM-S . . . . . secretory IGHM

IGHM-TM . . . . . transmembrane IGHM

IGHT . . . . . Immunoglobulin T (equivalent to IGHZ)

IGHT/Z . . . . . Immunoglobulin T/Z (equivalent to IGHZ)

IGHZ . . . . . Immunoglobulin Z (isotype of IGH)

IGHZ/T . . . . . Immunoglobulin Z/T (equivalent to IGHZ)

IGL . . . . . immunoglobulin light chain

IgSeq . . . . . immunoglobulin sequencing

iSB . . . . . incomplete SeraBind (buffer)

JH . . . . . heavy-chain joining (gene segment)

K. marmoratus . . . . *Kryptolebias marmoratus*

kb . . . . . kilobase(s)

KWT . . . . . Kruskal-Wallis test

L . . . . . litre(s)

LB . . . . . lysogeny broth

M . . . . . molar concentration (moles per litre)

MIG . . . . . molecular identifier group

min . . . . . minute(s)

ml . . . . . millilitre(s)

mM . . . . . millimolar concentration (millimoles per litre)

mmol . . . . . millimole(s)

mol . . . . . mole(s) (unit of amount of substance)

mRNA . . . . . messenger RNA

Mya . . . . . million years ago

N. furzeri . . . . . *Nothobranchius furzeri*

N. orthonotus . . . . . *Nothobranchius orthonotus*

n_c . . . . . cycle number (PCR)

nat . . . . . natural unit (unit of information)

ng . . . . . nanogram(s)

NHEJ . . . . . non-homologous end-joining

nM . . . . . nanomolar concentration (nanomoles per litre)

nmol . . . . . nanomole(s)

NT . . . . . nucleotide(s)

nt . . . . . nucleotide(s)

O. latipes . . . . . *Oryzias latipes*

oligo . . . . . oligonucleotide

P. formosa . . . . . *Poecilia formosa*

P. playfairii . . . . . *Pachypanchax playfairii*

P. reticulata . . . . . *Poecilia reticulata*

PCI . . . . . phenol:chloroform:isoamyl alcohol mixture

PCoA . . . . . principal co-ordinate analysis

PCR . . . . . polymerase chain reaction

PEG . . . . . polyethylene glycol

PhiX . . . . . PhiX174 bacteriophage genome

prep . . . . . preparation

RAG . . . . . recombination-activating gene

RDI . . . . . Repertoire Dissimilarity Index

RIN . . . . . RNA integrity number

RNA . . . . . ribonucleic acid

RNA-seq . . . . . RNA sequencing

RSS . . . . . recombination signal sequence

s . . . . . second(s)

SA . . . . . suffix array

scf . . . . . scaffold (genome assembly)

SHM . . . . . somatic hypermutation

SPRI . . . . . solid-phase reversible immobilisation

T_a . . . . . annealing temperature (PCR)

t_{ext} . . . . . extension time (PCR)

TdT . . . . . terminal dideoxy transferase

TET . . . . . Tris:EDTA:Tween (buffer)

TM . . . . . transmembrane exon

TRB . . . . . T-cell receptor beta

TSA . . . . . template-switch adapter

U . . . . . unit(s) of enzyme activity

UDG . . . . . uracil-DNA glycosylase

UMI . . . . . unique molecular identifier

UTR . . . . . untranslated region

v/v . . . . . volume/volume

VH . . . . . heavy-chain variable (gene segment)

vs . . . . . *versus* (against, compared to)

w/v . . . . . weight/volume

X. maculatus . . . . . *Xiphophorus maculatus*

Chapter 1

Introduction

1.1 Overview

All organisms exist in a state of intense competition for resources. For many organisms, among the most dangerous competitors are parasites, which attempt to colonise the host's own body and co-opt its internal resources and systems to their own advantage. The evolutionary arms race between parasites and hosts is ancient, and has led to the development of a stunning variety of complex offensive and defensive adaptations on each side [2]. Immunology has therefore had a profound effect on evolution.

One of the most complex and sophisticated systems developed as part of this ancient host/parasite conflict is the vertebrate adaptive immune system. By dynamically recombining their own DNA within specialised lymphocyte cells (Section 1.2.2), jawed vertebrates are capable of producing an almost unlimited variety of different *de novo* antigen-receptor proteins, and hence of responding effectively to entirely novel immune threats [2]. In addition, by producing long-lived memory cells in response to antigenic stimulation, the adaptive immune system can retain the ability to respond to recurrent threats years or even decades after they were first encountered by an individual [3] (Section 1.2.4). This combination of dynamic adaptability to novel immune threats and persistent immune memory enables vertebrates to progressively improve their protection against predictable aspects of their immune environment [4], while also coping effectively with the rapid evolution of bacterial and viral pathogens.

Among the different branches of vertebrate adaptive immunity, the B-cell-mediated humoral immune system is unique in both the breadth of antigenic compounds it can respond to and its ability to produce secreted antigen-receptor proteins capable of acting independently of the cells that produced them. Whereas the T-lymphocytes of the cellular adaptive immune system can respond only to processed peptide antigens expressed on the surface of antigen-presenting cells, the antibodies produced by B-lymphocytes (Section 1.2.1) are capable of responding to a much wider variety of organic molecules, including proteins, carbohydrates, nucleic acids, and lipids [5]. Secreted antibodies in serum and mucosal secretions play a number of essential roles in vertebrate immunity, including opsonisation (recruitment of phagocytic cells), activation of complement, and inactivation and aggregation of antigens and pathogens [6], while membrane-bound antibodies (also known as B-cell receptors or BCRs [7]) orchestrate B-cell development and response to antigen exposure. An effective humoral immune system is essential to immune function in vertebrates, and mutations that impair antibody production can cause severe immunodeficiency syndromes in humans [8].

Despite all its sophistication, however, the functionality of the vertebrate adaptive immune system declines dramatically with age, leaving older individuals increasingly vulnerable to infectious disease (Section 1.3). In humans and other species, the decline in immune functionality with age manifests as a dramatic increase in deaths from infection in older people, as well as increasing levels of infection-associated morbidity and a decline in the effectiveness of pro-immune interventions such as vaccination [9]. The molecular and physiological changes underlying this immunosenescent

phenotype are wide-ranging, implicating many different parts of the immune system; however, one particularly important contributor is the systemic decline in the effectiveness of the adaptive immune system. For the humoral immune system, well-established immunosenescent phenotypes include a fall in naïve B-cell output from the bone marrow, reduced antibody quality, impaired affinity maturation, and a serious decline in the ability to mount an effective antibody response to vaccination (Section 1.3).

While much has been discovered about the cellular and physiological changes that take place in the humoral adaptive immune system with age in some species, particularly humans and mice, much less is known about how these changes translate to alterations in the high-level diversity, clonal makeup, and other structural aspects of the overall antibody repertoire in older individuals. Such a high-level systemic approach to antibody diversity could potentially reveal a great deal about how ageing and other processes affect adaptive immune functionality in an organism. Specialised, high-throughput, quantitative approaches [10] have been used to assess the structure, diversity and health of adaptive immune repertoires in various contexts, including development, disease, and ageing; in the latter case, this pre-existing work has confirmed a loss of antibody-repertoire diversity with age, accompanied by impaired primary and clonal B-cell selection, an increase in memory-cell clonal expansions, and an increase in between-individual variablity in the repertoires of older individuals (Section 1.3). However, much remains to be discovered about the temporal and spatial progression of repertoire ageing, the effect of anti-ageing interventions, and the comparative effects of ageing on repertoires from species other than humans and mice.

When it comes to investigating the ageing of vertebrate-specific adaptations like the adaptive immune system, research is made more difficult by a relative lack of well-suited model organisms. Most established short-lived model organisms used in ageing research (e.g. yeast, nematode worms, or fruit flies) are not vertebrates, while most established vertebrate models (e.g. zebrafish or *Xenopus laevis*) were selected for properties other than their lifespan (such as rapid development) and are too long-lived for use as ageing models in many contexts [11, 12]. Even mice, the most widely-used vertebrate model organisms for both ageing and other biomedical applications, have a median lifespan of several years in most commonly-used laboratory strains [13], making many ageing studies prohibitively expensive. In this context, the recent emergence of the turquoise killifish (*Nothobranchius furzeri*), the shortest-lived vertebrate species currently bred in captivity, as a model organism for ageing research (Section 1.4) represents a highly promising development for the study of adaptive immunosenescence, providing the opportunity to rapidly investigate the ageing of the antibody repertoire over the entire lifespan of an organism, in different organs, and in response to known anti-ageing interventions.

In this thesis, therefore, I establish the turquoise killifish as a model for the study of comparative immunology and humoral adaptive immunosenescence. Using a combination of existing genomic assemblies and new sequence data, I assembled and characterised the immunoglobulin heavy chain (*IGH*) gene locus of the turquoise killifish and compared it to other newly-assembled loci from

closely-related species, revealing a group of complex and rapidly-evolving loci with a number of surprisingly ideosyncratic features (Chapter 3). Using the sequences from this newly-characterised locus, I established the first working immunoglobulin-sequencing protocol in this species, which I used to investigate the diversity and complexity of heavy-chain immune repertoires in adult killifish and how this diversity changes with age in the whole body and the gut (Chapter 4). The results of these investigations demonstrate that the turquoise killifish possesses a complex, diverse and individualised antibody repertoire, which undergoes a rapid decline in within-individual diversity and increase in between-individual variability with age. This phenomenon is particularly strong in the gut mucosal repertoire, likely as a result of the much greater relative prelavence of large, antigen-experienced clones and the intense antigen exposure of mucosal B-cell populations. Taken together, these results demonstrate the value of the turquoise killifish as an emerging model system in vertebrate immunology.

1.2 The humoral adaptive immune system in jawed vertebrates

1.2.1 Antibody structure and function

The lineage of lymphocytic white blood cells known as B-cells is ancient, with its origins predating any modern jawed-vertebrate lineage [14, 15]. In modern jawed vertebrates, B-lymphocytes are responsible for the production, diversification and secretion of the antigen-receptor proteins known as antibodies [6, 16], as well as diverse other roles that vary between taxa [17]. In almost all species examined to date, antibodies share a common, and highly distinctive, tetrameric structure (Figure 1.1A), with two identical immunoglobulin heavy chains (IGH) and two light chains (IGL) linked by disulfide bonds into a roughly Y-shaped configuration [6, 16]. The sequence and structure of these chains, and the corresponding regions of their underlying gene loci, is divided into an N-terminal (5') variable region and a C-terminal (3') constant region [16] (Figure 1.1A); together, the variable regions of each light/heavy-chain pair determines the antigen-binding specificity of the antibody, while the constant regions, particularly that of the heavy chain, determine its structure, functional properties, and interactions with the rest of the immune system [6, 16]. Unsurprisingly, the sequence diversity of the constant region is far lower than that of the variable region, with most species expressing only a few distinct constant-region classes (or *isotypes*) but an almost unlimited variety of variable-region sequences (or *ideotypes*).

As members of the immunoglobulin superfamily of proteins, the tertiary structure of antibody chains consists primarily of a species of immunoglobulin-fold domains; most light chains consist of two such domains, while the number in heavy chains varies substantially with isotype [6]. In both heavy and light chains, the most N-terminal immunoglobulin fold comprises the variable region, while the rest make up the constant region (Figure 1.1A); in some taxa (but not in teleost fishes),

one of the constant-region immunoglobulin folds is replaced with a flexible hinge domain in some isotypes to increase the flexibility of the heavy chain [6]. The three loops of the protein chain facing the antigen-binding site are labelled H1, H2 and H3 in the heavy chain and L1, L2 and L3 in the light chain (Figure 1.1B) [18] and are principally responsible for determining the antigen-binding specificity of the antibody; their corresponding gene regions are known as complementarity-determining regions (CDRs) and are the main focus of sequence variability among ideotypes, while the other parts of the antibody sequence are known as framework regions (FRs or FWRs) and exhibit less variability [6].

Most antibody isotypes can be expressed in secreted or membrane-bound form; in the latter case they are also known as B-cell receptors (BCRs) [7]. The choice between secreted and membrane-bound forms of an antibody is usually made through alternative splicing [7, 19]; typically, the transmembrane domain and cytosolic tail of the BCR are expressed via two additional exons (TM1 and TM2). Other constant-region exons are referred to collectively as CH exons, and are numbered by their occurence in the protein chain from the variable region to the C-terminus. In secreted antibodies, the standard four-chain configuration described above is known as an antibody monomer [16], while multiple four-chain antibodies bound together are referred to by the number of four-chain monomer subunits they contain: dimers for two connected subunits, tetramers for four, etc. [6, 16] (Figure 1.1D). These multimeric antibody supercomplexes have increased overall avidity for antigen and so can respond more strongly to low levels of low-specificity antigen [16]; they can be bound together by covalent disulfide bonds between subunits [6] or, more rarely, by noncovalent intermolecular interactions [20].

Of the two types of chain comprising the antibody protein, the heavy chain has by far the larger effect on antigen specificity [18], as well as determining the antibody's effector function [6]; this thesis therefore focuses exclusively on the heavy chain. The number and variety of antibody heavy-chain classes available to B-cells in an organism, and the mechanism by which the isotype of an antibody is determined, vary substantially by species; in tetrapods, the isotype of the antibodies produced by a given B-cell can be modified by a specialised class-switch recombination (CSR) process which is absent in teleost fishes [21]. Different isotypes vary in their length, flexibility, multimerisation behaviour, and effector functions [6, 21]. In teleost fishes, three constant-region classes have been observed to date [7, 19, 22]:

- **Immunoglobulin M** (*IGHM*) was the first *IGH* isotype to be identified in teleosts [22], and is homologous to the isotype of the same name found in mammals and other jawed vertebrates [19]. It is expressed in both secreted and transmembrane form [22]; in most teleosts, the transcript of the secreted form comprises four CH exons (C_μ1-4), while the transmembrane form comprises three CH exons and two TM exons [7, 22]. In contrast to mammals, in which secreted *IGHM* is primarily found as a pentamer, in teleosts it is typically found as a tetramer [22] connected by disulfide bonds between heavy chains [19]. In those fish species which have been tested, secreted *IGHM* is the main form of antibody found in serum [7, 19, 22].

- Like *IGHM*, **immunoglobulin D** (*IGHD*) is a primitive isoform present in most lineages of jawed vertebrates [19], including teleost fishes. The size and structure of *IGHD* varies dramatically between teleost species, with the number of CH exons varying more than twofold, from roughly seven (C_δ1-7) in some species to seventeen in zebrafish [19, 22]. All teleost *IGHD* transcripts to date have possessed a chimeric C_μ1 exon from *IGHM*, a configuration almost unknown in mammals [19, 22]; teleost *IGHD* also lacks the flexible hinge region present in mammalian *IGHD* [22]. A minority of teleost species are known to possess secretory forms of *IGHD*, though the mechanism of producing them varies: in channel catfish, one dedicated sublocus has a dedicated *IGHD* secretory exon in place of the transmembrane exons [23] (Figure 1.3B), while in rainbow trout (and possibly some other species like Atlantic salmon and cod) a run-on event at the end of C_δ7 results in the production of a secretory tail in a manner similar to secretory *IGHZ* [24]. In other species, only transmembrane isoforms have been observed. In teleosts as in mammals, transmembrane *IGHD* is usually co-expressed with *IGHM*; however, its role in the adaptive immune system remains unclear [19].

- Unlike *IGHM* and *IGHD*, **immunoglobulin Z** (*IGHZ*, also known as *IGHT*, *IGHZ/T* and *IGHZ/T*) is unique to teleost fishes [22]. Also unlike *IGHM* and *IGHD*, *IGHZ* is not found universally among teleost loci – of those *IGH* loci characterised to date, *IGHZ* is missing in those of medaka and channel catfish (Figure 1.3B) [22, 25], having apparently been lost independently in these species (Figure 1.3B). In those species in which it is present, *IGHZ* appears to act as a specialised mucosal antibody class, with elevated levels observed in mucosal secretions compared to serum [20, 22, 26]. Unlike *IGHM*, secretory *IGHZ* in serum is predominantly monomeric, while in mucosal secretions it is found primarily as a tetramer held together by noncovalent intermolecular bonds between heavy chains [20]. In most species, *IGHZ* comprises four CH exons (C_ζ1-4) and two TM exons [19], though some species have fewer – for example, stickleback *IGHZ* has only three CH exons [27, 28], while fugu *IGHZ* has only two [22, 29].

1.2.2 Antibody sequence diversification and primary repertoire diversity

The immune environment encountered by a vertebrate organism contains an enormous number of different potential pathogenic threats, each of which has its own antigenic signatures and many of which are capable of evolving much more rapidly than the vertebrate host [2]. In order for the adaptive immune system to cope with this huge diversity of different threats, it must be able to produce antibodies with a correspondingly large diversity of different antigen specificities. The greater the potential diversity of antibody sequences available to the adaptive immune system, the greater its capacity to respond effectively to novel immune threats. In fact, the mechanisms employed by the vertebrate adaptive immune system to diversify its antigen receptors enable an almost unlimited

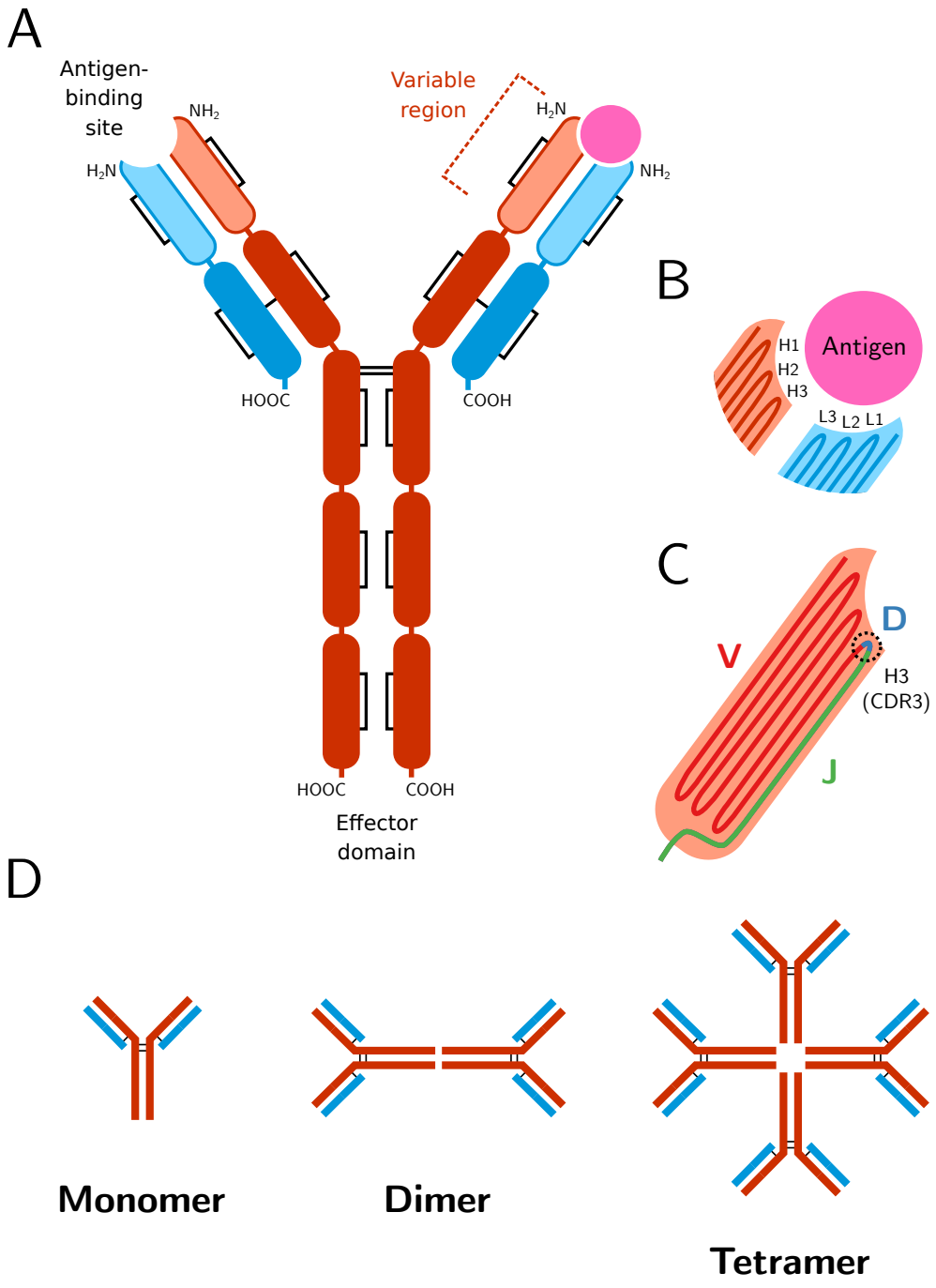


Figure 1.1: Antibody structure and function: (A) Schematic of a hingeless secreted antibody monomer, with heavy chains depicted in red, light chains in blue, and bound antigen in pink. Immunoglobulin fold domains are depicted as rounded rectangles, with variable-region domains shown with light shading and constant-region domains with dark. Black lines indicate disulfide bonds. (B) Schematic close-up of an immunoglobulin antigen-binding site, indicating the six loop regions (H1-3 and L1-3) making up the antigen-binding region. (C) Schematic close-up of the heavy-chain variable region, indicating the regions corresponding to the variable, diversity and joining gene segments from the unrecombined *IGH* locus. Note the chimeric nature of the third antigen-binding loop (H3, marked), which is formed by parts of all three gene segments. (D) In the context of antibodies, "monomer" refers to a single four-chain antibody protein, while "dimer", "tetramer" etc. refer to multiple four-chain monomers linked together through covalent or noncovalent interactions.

diversity of potential antibody sequences, with a correspondingly vast array of potential antigen specificities [30].

The mechanisms by which the adaptive immune system produces this diversity are dramatic, and rely on a highly unusual underlying gene structure [31] and a very high level of cellular wastage [32]. In the humoral immune system, antibody diversification takes place during B-cell development in the primary lymphoid organs (bone marrow in mammals, anterior kidney in teleosts [17]). Prior to this process, the native antibody loci in B-progenitor cells (and other cell types) are highly fragmented [31], with numerous variable-region sequence segments present in series on the chromosome upstream of the constant-region exons. In the heavy chain locus, these variable-region gene segments can be divided into three categories:

- **Variable (VH) segments** are the longest class of gene segment, at roughly 300 bp in length [33]. Each VH segment codes for the majority of the variable region of an antibody, including the entirety of the first three framework regions (FWR1-3) and the first two complementarity-determining regions (CDR1-2) as well as the 5' part of CDR3 [6, 31] (Figure 1.1C). They are therefore highly structured, and include several highly-conserved positions present in virtually all functional V-segments in nearly all species, including two conserved cysteine residues (which give rise to an intra-domain disulfide bond) and a conserved tryptophan residue following CDR1 [33]. Each V-segment is also associated with its own promoter sequence and 5'-UTR, as well as a 5'/C-terminal leader peptide between the translation start site and the start of the functional V-sequence.
- **Diversity (DH) segments** are the shortest class of segment, typically on the order of 10-20 bp, and are the least structured [34]. They form the middle part of CDR3 (Figure 1.1C) [6].
- **Joining (JH) segments** are of intermediate length, typically 50-60 bp [34]. They form the 3'/C-terminal part of heavy chain CDR3 and the 5'/N-terminal part of FR4 (Figure 1.1C) [6]. Each J-segment is succeeded on the chromosome by a splice donor site [25], which is used to join the variable region of the antibody sequence to the constant region via RNA splicing following transcription of *IGH* mRNA (Figure 1.2Aiii). Like VH segments, JH segments can be identified from their conserved structure, particularly the conserved tryptophan residue marking the end of CDR3 [34].

In the simplest "translocon" configuration of the *IGH* locus, blocks of repeated VH, DH and JH segments are present in series on the chromosome in contiguous V-, D- and J-regions (Figure 1.2Ai) [6, 31]. During B-cell development, a single VH, DH and JH segment are selected from these segment blocks, and the intervening genomic regions are permanently excised from the genome to produce a single, contiguous VDJ sequence coding for the complete variable region of an antibody (Figure 1.2Aii). The mechanism by which this excision occurs is called VDJ recombination, and

relies on a specialised recombinase complex containing lymphocyte-specific recombination-activating genes 1 and 2 (*RAG1* and *RAG2*) [31, 35]. This complex recognises specialised recombination signal sequences (RSSs) flanking each variable gene segment, composed of highly conserved heptamer and nonamer sequences separated by a spacer sequence of conserved length (either 12 or 23 bp, corresponding respectively to one or two turns of the DNA helix) [36] (Figure 1.2B). Each functional VH segment is succeeded by a 23 bp-spacer RSS in 5'-3' orientation, while each JH segment is preceded by a 23 bp-spacer RSS in 3'-5' orientation; each DH segment, meanwhile, is flanked by 12 bp-spacer RSSs in 5'-3' and 3'-5' orientation, respectively (Figure 1.2C) [35].

In VDJ recombination, recombinase complexes bind two of these RSSs and associate with each other to bring the two corresponding gene segments into close proximity (Figure 1.2Di-iii). A single-strand DNA nick is introduced between each RSS and its gene segment, and the exposed 3'-hydroxyl group of the break attacks the 5'-phosphate on the other strand to produce an asymmetric double-strand break, with the coding sequence terminating in a hairpin loop and the cleaved RSS in a blunt end [6, 35] (Figure 1.2Div-v). These breaks are then resolved by the ubiquitous non-homologous end-joining (NHEJ) machinery of the DNA damage response: the blunt ends of the cleaved RSSs can be directly ligated together, while for the coding sequences the hairpin loops must first be cleaved and the resulting 3'-overhangs resolved [6, 35] (Figure 1.2Dvi-vii). The net result of this process is a contiguous V/D or D/J join in the coding sequence and a ligated DNA circle containing the excised RSSs and intervening sequence, which is subsequently degraded. Via unknown mechanisms [35], the recombination machinery exhibits a strong preference for RSS sequences of different lengths (the so-called "12/23 rule"), encouraging D-J and V-D joins while largely preventing V-J joins.

The simple joining of gene segments during VDJ recombination provides a basic combinatorial sequence diversity, with a number of possible sequences in the simplest case equal to the product of the numbers of VH, DH and JH segments in the *IGH* gene locus. The number of potential *IGH* variable-region sequences in most species, however, is vastly higher than this: in humans, for example, there are roughly 8000 possible functional VDJ combinations [37], but the number of different possible nucleotide sequences exceeds 10^{21} [38]. This huge difference arises from the inexactness with which the coding ends produced by the recombinase complex are joined following RSS excision (Figure 1.2E). Before the two gene segments brought together in VDJ recombination can be joined by NHEJ, the hairpin loops formed by the recombinase complex must be resolved by re-nicking the DNA [6, 35]. This re-nicking process is not exact, and often occurs within the coding sequence, resulting in a palindromic 3'-overhang that can be filled in (resulting in palindromic *P-insertions*) or trimmed (resulting in *deletions*). The resulting blunt ends then serve as substrates for the lymphocyte-specific terminal dideoxy transferase enzyme (TdT), which adds a variable number of untemplated nucleotides (nonpalindromic *N-insertions*) prior to sequence ligation [6, 35]. As a result, the final recombined sequence is frequently not a clean ligation of the two original gene segments, but an inexact join involving multiple missing terminal positions and/or *de novo* inserted nucleotides, a phenomenon collectively known as *junctional diversity* [39].

The three processes contributing to junctional diversity (P-insertion, N-insertion, and deletion) together vastly increase the potential antibody sequence diversity available to a developing B-cell at the heavy-chain CDR3 region, which as a consequence is by far the most important single region on either antibody chain in determining antigen specificity [18]. This junctional diversity, however, comes at a substantial cost. The indel mutations introduced by VDJ recombination are not constrained to occur in multiples of three over both junctions, resulting in a high rate of recombinations in which the VH and JH segments are out of frame with each other; still more loci are rendered nonfunctional through the introduction of *de novo* stop codons. As the sequence changes introduced by VDJ recombination are irreversible, many developing B-cells are left with permanently disrupted *IGH* loci on both chromosomes, and are left with no recourse but programmed cell death. In addition, the huge and untemplated sequence diversity introduced in this way means that many B-cells whose *IGH* loci do successfully undergo VDJ recombination are left with BCRs that either cannot effectively bind any antigen or strongly bind self-antigen, resulting in useless antibodies in the first case and a dangerous risk of autoimmunity in the second; these B-cells are eliminated through a process of antigenic selection before naïve B-cells are permitted to exit the primary lymphoid organs, resulting in still more wastage as cells with nonfunctional or self-binding antigens undergo programmed cell death. Overall, as many as 90 % of developing B-cells may be eliminated by one or other of these processes before successfully exiting to the periphery [32].

Taken together, the processes of VDJ recombination, junctional diversity and primary selection will produce a population of naïve B-cells with an extremely high diversity of heavy-chain ideotypes, with virtually every naïve B-cell expressing a unique variable-region sequence. Taken together, the resulting population of sequences represents the *primary antibody heavy-chain repertoire* of the organism, and determines the range of different antigen sequences the humoral immune system of that organism can potentially respond to. The diversity of this primary repertoire depends on the native structure of the *IGH* locus in that species (which determines the number of segment choices available to the VDJ-recombination process), the distribution of possible numbers of insertions or deletions contributing to junctional diversity, and the stringency of the primary-selection process.

1.2.3 *IGH* locus structure in teleost fishes

Section 1.2.2 describes the process of *IGH* locus maturation in terms of an idealised translocon locus with a simple V-D-J-C structure (Figure 1.2A). In some species, including humans and mice, *IGH* loci roughly correspond to this layout [31]; however, in many species, including all teleosts, this idealised structure is a significant oversimplification of the actual layout of their *IGH* loci [22]. Due to their size, repetitiveness and complexity, comprehensive assembly of *IGH* loci is often difficult, and full elucidation of their structure often requires a focused characterisation effort. Nevertheless, a number of teleost loci have been characterised to date, primarily from species used widely in research (e.g.

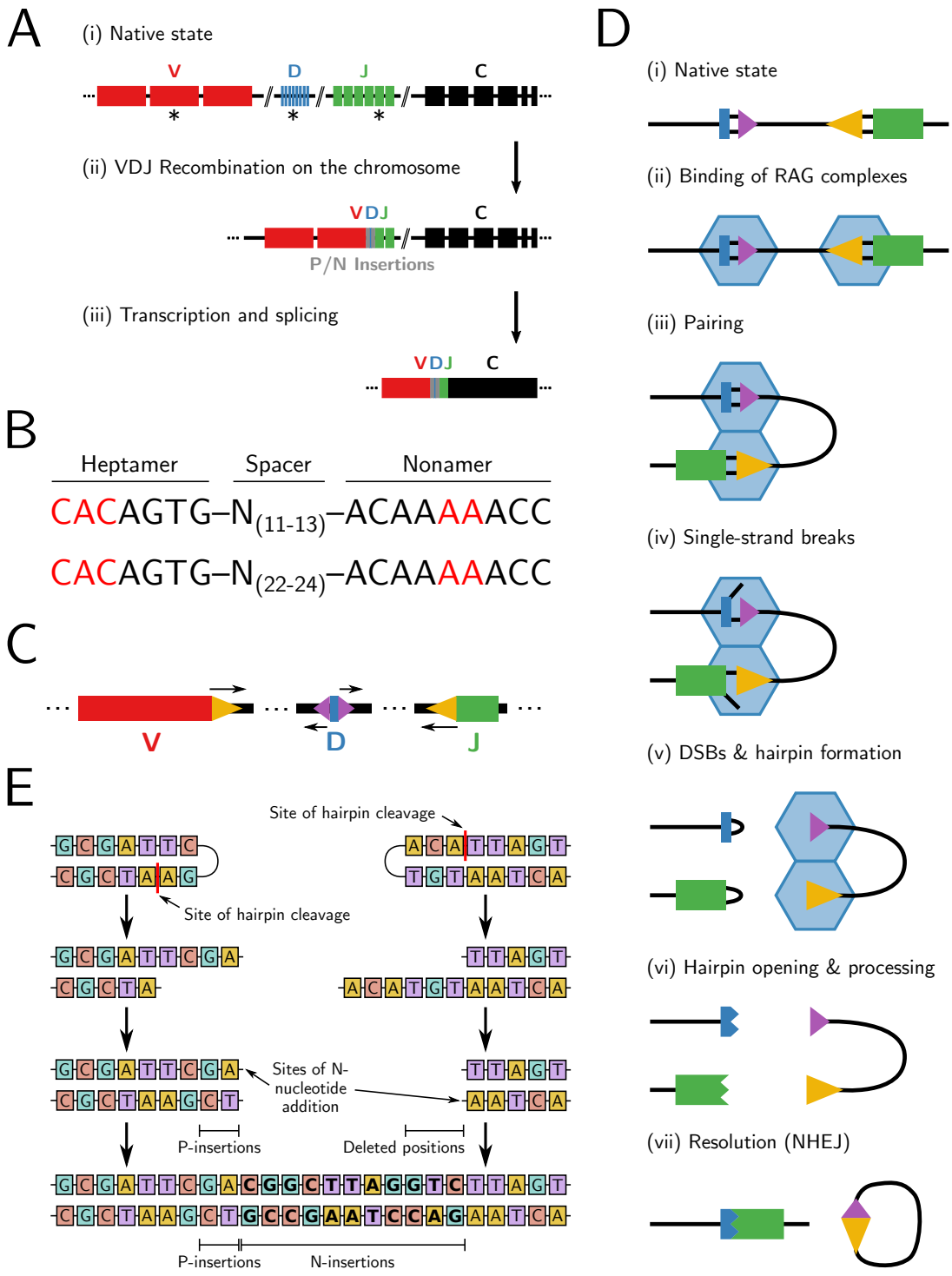


Figure 1.2: Primary sequence diversification in the immunoglobulin heavy chain: (A) Schematic of antibody sequence formation at the locus level. Black asterisks indicate the segments to be recombined. (B) Consensus sequence of short (above) and long (below) RSSs in jawed vertebrates. Nucleotides marked in red are the most strongly conserved and most required for efficient recombination [36]. (C) Schematic of position and orientation of long (yellow) and short (purple) RSSs relative to variable-region gene segments. (D) Schematic of recombination mechanism for a D/J segment pair, illustrating hairpin formation and imprecise end-joining of coding sequences; adapted from [35], Figure 2. (E) Example of the inexact hairpin resolution mechanisms leading to junctional diversity in (D-vi); adapted from [39], Figure 10-5.

zebrafish, medaka, stickleback), agriculture (e.g. salmon, grass carp, channel catfish) or both (e.g. rainbow trout) [7, 22].

Among teleost fishes, the simplest *IGH* locus structures known to date are exhibited by species such as zebrafish, grasscarp, and fugu (Figure 1.3A) [22]. In these species, the *IGH* locus adopts a V-D-J-C_ζ-D-J-C_μ-C_δ structure, with a single shared V-region followed by D- and J-regions specific to the *IGHZ* and *IGHM/D* constant regions, respectively. The rainbow trout locus has a similar organisation, but with two additional VH segments following the *IGHZ* region [40]. During VDJ recombination in these loci, the choice of variable gene segments also determines the choice of constant region: if a pair of DH and JH segments upstream of the *IGHZ* constant region is selected, the cell will express *IGHZ*, while if the segments chosen are downstream of *IGHZ*, that constant region will be excised during VDJ recombination and *IGHM* and/or *IGHD* will be used instead. As a result, VDJ recombination in these species determines both the ideotype of a developing B-cell and its isotype [22].

While some teleost species possess such relatively simple *IGH* loci, many species exhibit much more complicated, and often much larger, locus structures. In many species, multiple distinct subloci (comprising V-, D- and J-regions and least one constant region) are present in tandem on the chromosome, often with distinct combinations of variable gene segments and constant regions (Figure 1.3B); in a few cases, such as medaka and Atlantic salmon, one or more of these subloci is in inverted orientation relative to the rest of the locus. In some species, such as Atlantic salmon, whole-genome duplication has led to the existence of multiple distinct *IGH* loci on different chromosomes, each of which has its own complement of subloci, gene segments, and constant regions [41]. In this milieu, pseudogenisation of gene segments or constant-region exons is common, resulting in loci with large numbers of pseudogenised V-segments and constant regions.

The diversity in locus size and organisation among teleost fishes is likely to have important consequences for humoral adaptive immunity in these species. The native locus constitutes the raw substrate for the VDJ recombination process, and the number of different VH, DH and JH gene segments available for recombination defines the baseline sequence diversity of antibodies in that species. It is not clear to what extent VDJ recombination can take place across different subloci in the large tandem loci described in Figure 1.3B; if recombination between subloci is restricted, this will also have important effects on the kinetics of VDJ recombination in a species. The division of an *IGH* locus into tandem subloci may also have effects on the statistics of B-cell maturation, with more subloci conceivably presenting the opportunity for a greater number of recombination attempts before the *IGH* loci of a developing B-cell are exhausted and so reducing the rate at which cells fail to mature successfully; however, to my knowledge little or nothing is known about the effects of locus structure on B-cell developmental processes at present. Finally, variations in the constant regions present in different species are likely to have very important effects on humoral immunity; most fundamentally, whereas *IGHZ/T* appears to be specialised for mucosal adaptive immunity in those

teleost species that possess it [20, 22, 26], it is not known how mucosal immune responses manifest in teleost species lacking this isotype. Given all these important (and potentially-important) effects of *IGH* locus structure on humoral adaptive immunity, characterising the sequence and organisation of this locus is an essential step in understanding the adaptive immune system of any given vertebrate species.

1.2.4 Affinity maturation and secondary repertoire diversity

Following VDJ recombination and primary selection, developed naïve B-cells emerge from the primary lymphoid organs and circulate in the periphery. At some point, a subset of these naïve cells will make contact with their cognate antigen. If this contact occurs in the correct signalling context (e.g. in the presence of T-cell help for T-dependent responses [5]), the B-cell becomes activated and begins to proliferate [43]. During this clonal expansion process, the B-cell's antibody genes undergo a very high rate of somatic mutation (known as *somatic hypermutation* or SHM), with mutations focused in the complementarity-determining regions coding for the antigen-binding loops of the variable region; in mammals, the rate can be as high as 10^{-3} per nucleotide per cell division [44]. This SHM process is orchestrated by the activation-induced cytidine deaminase enzyme (AID), which preferentially targets particular hotspot motifs and deaminates cytidine residues to uracil, which can either be corrected to thymine on DNA replication (resulting in a C-to-T transition mutation) or removed via base-excision repair (resulting in a variety of other mutations) [45]. As a result, the original antibody sequence of the ancestral naïve cell is diversified into a cluster of related sequences, with widely varying affinity for the cognate antigen.

In order for the combination of clonal expansion and somatic hypermutation to improve the adaptive immune system's ability to respond to the stimulating antigen, the resulting clone needs to undergo a selection process, to identify and favour cells expressing antibodies with improved antigen affinity. This selection is effected through a competitive process in which clonally-expanded B-cells attempt to bind cognate antigen trapped on the surface of helper cells: those cells which successfully bind antigen receive growth and differentiation signals, while those which do not undergo programmed cell death. In mammals, this process takes place primarily in histologically-distinct germinal centres within specialised secondary lymphoid organs such as the spleen [46], in which B-cells that have encountered antigen first undergo clonal expansion and SHM, then compete for antigen on the surface of follicular dendritic cells (FDCs) [46, 47]. In teleosts, which lack specialised, histologically-differentiated germinal centres, B-cell proliferation takes place near clusters of melanomacrophages surrounded by reticular cells, both of which have been attributed an antigen-trapping and -presentation role analogous to that performed by FDCs in mammals [45].

While clonal expansion, somatic hypermutation and clonal selection, collectively known as *affinity maturation*, are present in both mammals and teleosts, the increase in antibody affinity resulting from this process in teleosts is much weaker than in mammals (e.g. 3-to-10-fold in rainbow trout, compared

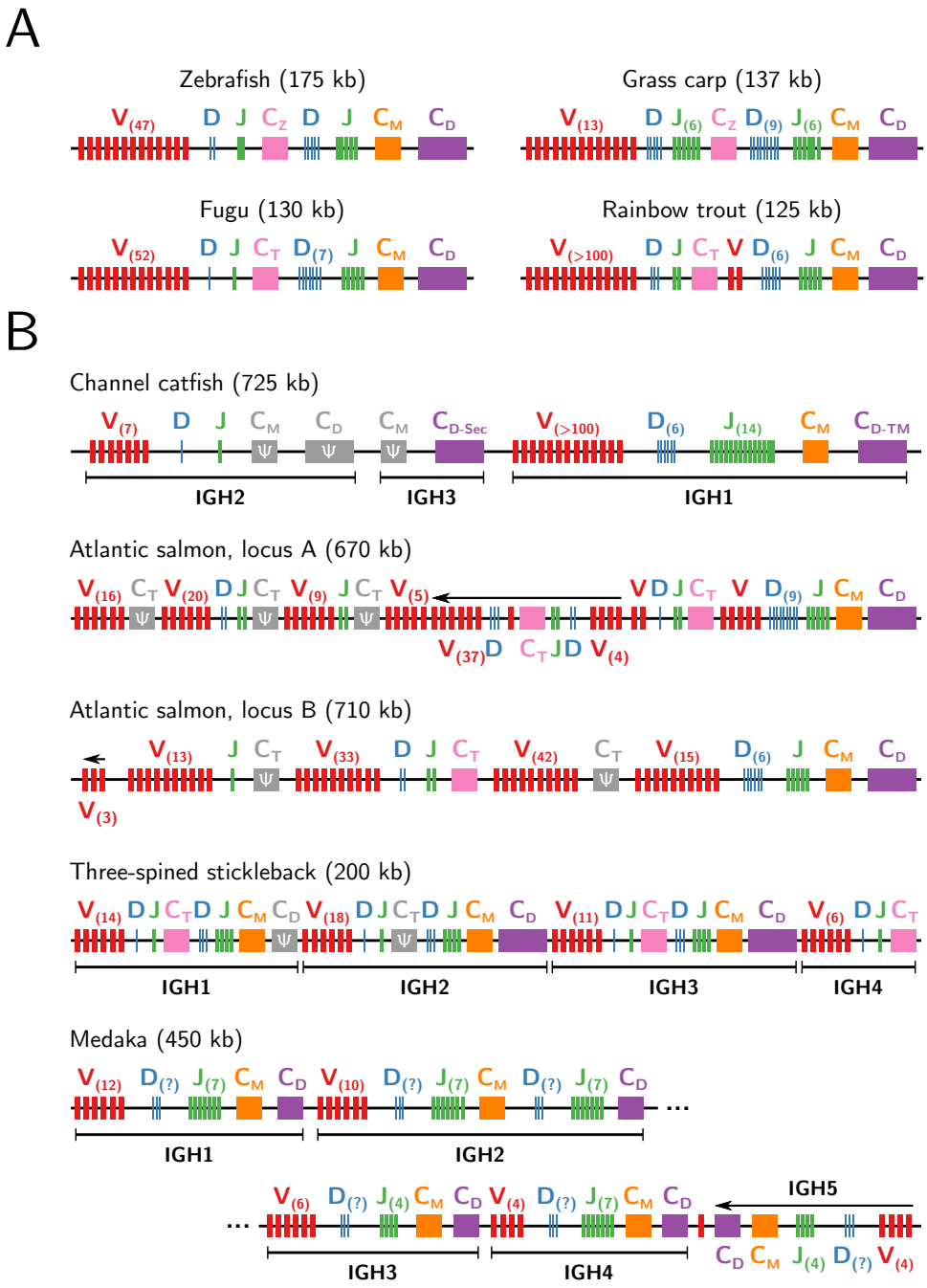


Figure 1.3: *IGH* locus structure in teleost fishes: Simplified schematics of *IGH* loci from nine teleost species. (A) In the simplest teleost *IGH* loci, single *IGHZ* and *IGHM/D* constant regions share a common V-region, but are preceded by separate D/J regions. In rainbow trout, a small number of V-segments also separate *IGHZ* from the pre-*IGHM* D-region. (B) In many teleosts, the *IGH* loci are much larger and more complex, with repeated tandem subloci (labelled regions), pseudogenised constant regions (grey, ψ), inverted segments (leftward arrows), and other deviations from the classic locus organisation. Loci are not to scale; blocks of more than five segments are condensed and labelled with their segment number, as are some smaller segment blocks for clarity. Adapted from [22], Figure 2 and [7], Figure 4, with additional information from [25, 27, 28, 41, 42]. The number of D-regions in the medaka locus is not provided in the sources.

to as much as 1000-fold in mammals [45]). Since SHM appears to be fully functional in those fish so far investigated [45], this difference seems likely to arise from differences in clonal-selection dynamics between teleosts and mammals. The reasons for such a difference are still not entirely clear, but may involve histological differences in where and how affinity maturation takes place in these taxa. In mammalian germinal centres, the ratio between expanded B-cells and FDCs is such that the amount of presenting antigen is limiting, forcing competition for antigen among B-cells and favouring those cells which can bind antigen most strongly [45]. In teleosts, conversely, proliferating B-cells are far outnumbered by antigen-trapping melanomacrophages and reticular cells, resulting in an oversupply of antigen relative to B-cell demand; this antigen surplus may result in faster overall proliferation, but at the cost of much weaker selection for high-affinity antibodies [45].

Following affinity maturation, activated B-cells undergo differentiation into memory cells (which persist in the bloodstream for long periods and provide secondary immune memory) [4, 46] and plasmablasts/plasma cells (which secrete large amounts of antigen to mount a powerful immune response) [43, 46, 47]. Affinity-matured cells can also re-enter germinal centres (or their less-developed teleost equivalents) to undergo additional rounds of proliferation, hypermutation and selection [46, 47]; memory cells can also undergo further rounds of affinity maturation, resulting in even larger and more diverse clones and still-higher levels of antigenic affinity. In tetrapods, class switching between different constant-region isotypes also occurs as part of affinity maturation, and is also orchestrated by AID; however, this process appears to be absent in teleosts [45].

The combination of the primary (naïve) repertoire described in Section 1.2.2 with the clonal expansions and additional sequence diversity arising from affinity maturation constitutes the *secondary heavy-chain antibody repertoire* of the organism (Figure 1.4). This is the repertoire actually encountered by incoming pathogens and available to relatively direct experimental interrogation. The structure and diversity of this secondary repertoire depends on the makeup of the primary repertoire, the degree of clonal expansion and hypermutation during affinity maturation, the strength of clonal selection, and the relative abundance of different B-cell subtypes. Inferring the composition of the primary repertoire from that of the secondary is therefore non-trivial, and requires an attempt to distinguish sequences arising from naïve vs activated B-cells or infer the former from the latter; once obtained, the naïve sequences can be used to infer the parameters of the generative process [38, 48].

1.3 Humoral adaptive immunosenescence in jawed vertebrates

The immune system of aged vertebrates has long been known to undergo a severe and systemic decline in functionality with age [49]. As a result of this decline, older individuals exhibit increased susceptibility to a wide range of bacterial, viral, and fungal infectious diseases, as well as higher rates of complications and mortality from those infections when they occur [9]. In addition, the effectiveness of vaccination against these infections often declines dramatically with age, with older

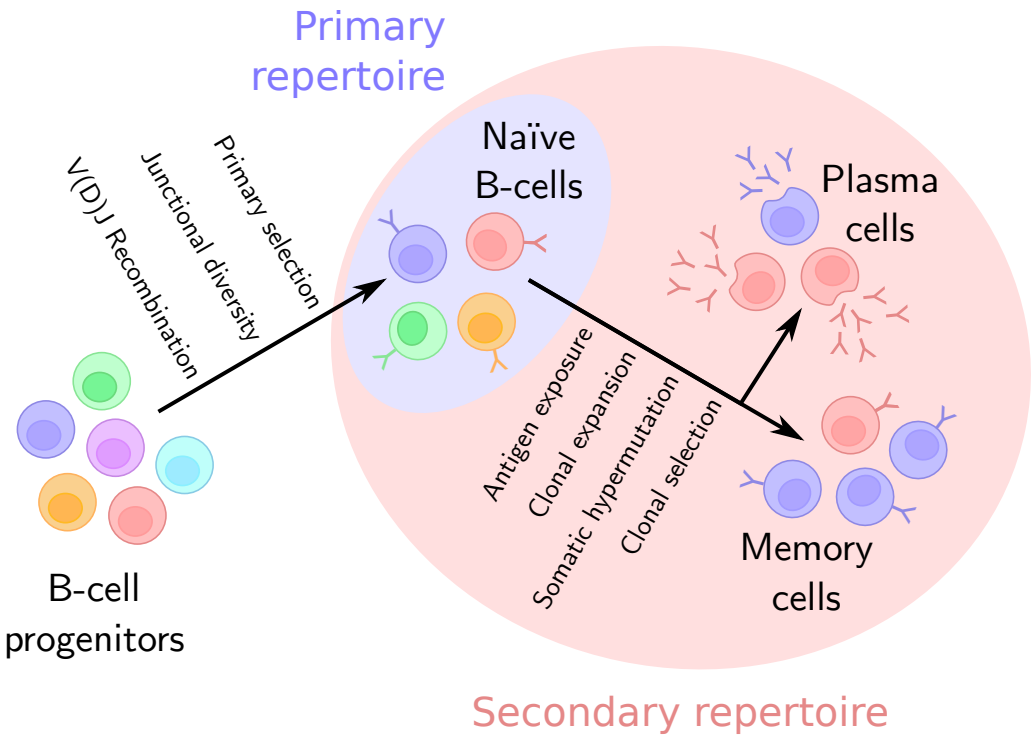


Figure 1.4: Primary and secondary antibody repertoires: Schematic of processes giving rise to the primary (naïve) and secondary (total) B-cell repertoires in the vertebrate immune system. Cell colours indicate clonal membership, with a loss of clonal diversity during primary development (left) and affinity maturation (right). Primary B-cell development (left) takes place in primary lymphoid organs (bone marrow in mammals, anterior kidney in teleosts [17]), while affinity maturation (right) takes place in germinal centres (in mammals) or proto-germinal clusters (in teleosts) in secondary lymphoid organs such as the gut-associated lymphoid tissue (GALT) or spleen [45, 46].

individuals often receiving half or less of the immune protection from vaccination exhibited by younger recipients [9]. Similar declines in immune functionality with age are also observed in other model organisms, and have been especially well-studied in mice, which share many details of their immune system with humans while being substantially easier to study in depth. While many different parts of the immune system are implicated in this immunosenescent phenotype [32], the decline in the humoral adaptive immune system has emerged as one particularly important contributor [50].

The changes in the humoral immune system underlying its impaired functionality in old age begin at the level of B-cell development. In mice, the population of haematopoietic stem cells (HSCs) resident in the bone marrow increases in number but exhibits a decreased propensity to regenerate the pool of developing B-cells [32, 50]; factors contributing to this change include a deteriorating bone marrow niche, changes in blood-borne factors affecting HSC differentiation, and the well-known bias of older mammalian HSCs towards myeloid (rather than lymphoid) cell lineages [32, 43]. As a result of this change in HSC behaviour, the number of new B-cell progenitor and precursor cells in the bone marrow declines with age in mice; furthermore, these precursors exhibit decreased proliferation

capacity, reduced rates of VDJ recombination and an increased rate of apoptosis [32, 51, 52]. The rate of B-cell output from the bone marrow therefore declines significantly in aged mice, falling as low as 10 % of its level in young adults [32]. As these naïve cells also have a shorter lifespan than antigen-experienced memory- and long-lived-plasma-cell populations, the net result of this decline in output from the bone marrow is a decrease in the number of naïve cells in the periphery and a progressive increase in the relative prevalence of antigen-experienced memory and plasma cells [32, 53]; the pool of circulating immunoglobulins, meanwhile, becomes progressively dominated by hypermutated antibodies specific to previously-encountered antigens [32]. As the naïve-cell pool is essential for responding to pathogenic threats not previously encountered by the immune system, this change is thought to lead to a progressive reduction in the capacity of the humoral immune system to respond effectively to these novel threats [32].

Despite the decline in new naïve B-cells produced by the bone marrow, the total number of B-cells in aged mice does not appear to change significantly, reflecting an increase in the absolute number, as well as the relative prevalence, of antigen-experienced B-cell subpopulations [50]. In humans, conversely, the absolute number of total B-cells appears to decline with age [50, 51, 54], with most B-cell subsets showing a significant decline in absolute abundance [55]. Sources differ as to the age-related changes in relative abundance of different B-cell subtypes in humans, with some reporting an increase in the relative prevalence of naïve cells and decrease in that of memory cells [55, 56] and others reporting a more mouse-like shift in favour of the memory compartment [50]. These differences may arise from differences in the methods and cell markers used to quantify abundance of different B-cell populations [50], genuine differences between populations, or the limitations of peripheral blood as a method of sampling the whole-body B-cell population of an individual [57, 58].

Whatever changes in cellular composition occur in ageing humans, it is clear that both humans and mice undergo a severe age-related decline in B-cell functionality. In humans, the best-understood aspect of this decline is a deterioration in the responsiveness of older patients to vaccination. Older humans frequently exhibit a substantially smaller [9] and slower [32] increase in the titre of antigen-specific serum antibodies in response to vaccination [50, 54, 59], with defects in humoral response observed to vaccines for diseases as diverse as influenza, hepititis A and B, tetanus, diphtheria, pneumococcus and tick-borne encephalitis [43]. At least in the case of influenza, the observed decline in specific antibody production is primarily due to a decrease in the number of plasmablasts activated in response to vaccination [51, 54, 59].

In addition to being less abundant, the antibodies produced by elderly humans in response to vaccination also appear to be of lower quality: anti-influenza antibodies produced by older human individuals demonstrate reduced ability to effectively neutralise viral haemagglutinin [32, 59], while *IGHM* antibodies produced by elderly subjects in response to pneumococcal polysaccharide vaccine demonstrate reduced opsonisation ability [32]. While basal serum antibody titres actually increase with age in humans and mice [60], suggesting that antibody production *per se* is not impaired, these

antibodies also decline in quality [51], with measurable increases in the rate of polyspecific and self-reactive antibodies produced in older individuals [32]. This decline in baseline antibody quality in older individuals suggests a breakdown in primary B-cell selection in the bone marrow, enabling a greater number B-cells with low-specificity or self-reactive antibody sequences to emerge into the periphery [50].

Various aspects of affinity maturation are also impaired in aged mammals. While memory-cell clones established early in life often persist into old age, naïve B-cells in aged humans demonstrate a severely reduced ability to give rise to new antigen-specific memory B-cells following antigenic stimulation, impairing the establishment of functional immune memory for novel threats [54]. B-cells from older humans and mice also exhibit impaired class-switch recombination capacity, possibly as a result of decreased AID expression [51, 55, 56], impairing the generation of antibodies with the same specificity but different effector functions. The decreased ability of older individuals to produce high-affinity antigen-specific antibodies [55] also suggests a defect in affinity maturation, though this could also be attributable to the known ageing-related defects in the T-helper and follicular dendritic cells involved in secondary B-cell selection [50, 51, 54] rather than changes in the B-cells themselves. Conversely, the effect of ageing on the somatic hypermutation process remains controversial [47, 50, 60, 61], with different groups reporting an increase, a decrease, or no change in the rate and level of SHM accumulation of different B-cell subtypes with age.

The various cellular and population-level changes that occur in the humoral immune system with age naturally have important effects on the antibody repertoires of ageing individuals. Aged humans [57] and mice [43] exhibit increased rates of non-malignant clonal expansions in the memory-cell compartment, which when combined with the apparent maintenance or decline in total B-cell number would lead to a reduction in overall diversity [43]. Similarly, early techniques that investigated the repertoire by investigating the distribution of CDR3 lengths (CDR3 spetratyping) indicated that older humans frequently exhibit more distorted distributions dominated by CDR3 regions of a particular length, indicative of clonal expansion, and found that older individuals with more distorted spectratypes exhibited greater frailty and worse health and lifespan outcomes than those with more young-like spectratypes [62]. On the basis of these and similar findings, the antibody repertoire has long been thought to decline in diversity in older individuals, impairing the adaptability of the ageing adaptive immune system.

More recently, the development of specialised high-throughput-sequencing-based techniques for interrogating antibody repertoires (immunoglobulin sequencing, or IgSeq [10, 63]) has enabled a few studies to investigate the ageing of these repertoires in greater detail. These studies, primarily on human peripheral blood, have confirmed many observations made using older lower-throughput methods: a reduced number of clonal lineages in older repertoires supports a decrease in naïve-cell output from the bone marrow [64], while an increase in mean CDR3 length [65] and the rate of premature stop codons [66] support the finding that primary and clonal selection are impaired with

age. An increase in clonal expansion with age is also observed in sequencing data: an oligoclonal phenotype, in which one or a few large clones dominate the repertoire, is often seen in peripheral blood from elderly humans but very rarely in the young [65, 66]. These expanded clones are persistent before and after vaccination [66] and even across multiple years [65]; in contrast, large clones observed in young human blood repertoires are typically transient responses to recent antigen challenge [65].

The reduction in clonal richness and increase in clonal expansions seen in older individuals contribute to an overall reduction in the estimated number of unique sequences present in the peripheral repertoire [66], a change observed for both naïve and mutated sequences. The same study also found a drop in the percentage of unique sequences identified as coming from naïve B-cells, suggesting a mouse-like shift in repertoire prevalence towards antigen-experienced cell types [66]. However, while the alpha (within-individual) sequence diversity of the human peripheral repertoire appears to decline with age, the beta (between-individual) diversity of the repertoire actually increases, with repertoires from older individuals differing more from one another than those from young individuals do [66]. Elderly repertoires also showed decreased flexibility in response to immune challenge, with repertoires taken from the same individual pre- and post-vaccination being significantly more similar in elderly than in young samples [66]. Taken together, these findings suggest a pattern in which human peripheral antibody repertoires progressively lose diversity and flexibility over the course of the human lifespan, while also becoming increasingly individualised and distinct.

In conclusion, there is strong evidence for a variety of cellular and physiological changes in the B-cell immune system with age in both mice and humans, giving rise to a severe decline in functional performance and immune protection. Many of these changes affect the composition of the antibody repertoire, with a reduction in clonal and sequence diversity and an increase in between-individual variability with age in human blood. However, despite this abundance of data, there remain some serious limitations in our knowledge of the ageing antibody repertoire. From an evolutionary and comparative perspective, almost nothing is known about adaptive immunosenescence in species other than humans and mice; while several IgSeq studies have been performed on teleost fish [63, 67–70], for example, none to my knowledge have investigated B-cell immunosenescence in this taxon. Many of the findings reported above are specific to *IGHG* or *IGHA* antibodies in mice and humans [32], and may not generalise to species lacking these isotypes. Even in humans, the majority of immunosenescence studies, including virtually all repertoire studies, are limited to peripheral blood, and therefore primarily sample the minority of B-cells in transition between tissues; the majority of B-cells, which are resident in some immune organ or tissue, are systematically underrepresented in these samples [57, 58]. Relatively little is therefore known about how the ageing of the B-cell repertoire differs between organs; the only paper I know of which investigated changes in antibody-repertoire diversity with age in biopsies from multiple human tissues failed to find any significant changes in alpha diversity, except for a significant *increase* in repertoire sequence entropy in old spleen [58]. To my knowledge, no publication has yet investigated the ageing of the antibody repertoire at mucosal

surfaces, where secreted antibodies play a particularly important role in regulating microbiotal communities and defending the body from pathogenic invasion [71].

In addition to this lack of spatial resolution in our knowledge of antibody repertoire ageing, there is a serious lack of temporal resolution; most studies of antibody repertoire ageing simply compare a "young" group (say 20-30 years of age) with one or two "old" groups (say $\geq$ 70 years of age), with little or no information about the intervening progression of any observed phenotypes [58, 66]. The effect of known lifespan-increasing interventions on immune-repertoire ageing also remains largely unstudied. Due to their very long lifespans and restrictions on experimental manipulation, humans are unsuitable subjects for these sorts of experiments, as are long-lived vertebrate model organisms such as zebrafish (median lifespan c. 3.5 years [11]) and *Xenopus* (median lifespan c. 9 years [12]). Even mice, though widely used in biogerontology, are inconveniently long-lived for many ageing experiments (median lifespan c. 2 to 2.5 years for the most common laboratory strains [13]). Conversely, many major model organisms in ageing research (such as fruit flies and nematode worms) are invertebrates, and so lack a mammal-like adaptive immune system. The development of a short-lived vertebrate species as a model organism for antibody-repertoire experiments would therefore be highly valuable for research into adaptive immunosenescence.

1.4 The African turquoise killifish as a model for vertebrate ageing

The genus *Nothobranchius* comprises a broad group of annual freshwater fishes distributed across equatorial and subequatorial Africa [72], with species diversity concentrated in the south-east of the continent [73]. Members of this genus share a suite of adaptations to life in ephemeral pools and rivers, most notably the production of desiccation-resistant embryos capable of surviving through the dry season in a diapause state [73] (Figure 1.5B). Fish from this genus have been known for several decades to exhibit very rapid growth and short lifespans, consistent with their evolving under conditions of very high extrinsic mortality [72], with many species exhibiting a median lifespan of less than one year. Nevertheless, there is wide variation within the genus in body size, growth rate and lifespan, with species from less arid regions tending to show slower growth and longer median lifespans [73].

Like other *Nothobranchius* species, the turquoise killifish (*Nothobranchius furzeri*, Figure 1.5A) is a medium-sized annual fish first isolated from ephemeral freshwater pools – in this case, from a relatively arid region of southeastern Zimbabwe and Mozambique [73, 74]. Even by the standards of the *Nothobranchius* genus, *N. furzeri* exhibits extremely rapid growth, maturation, and ageing, with the most widely-used laboratory strain (GRZ) exhibiting a median lifespan of just 9-16 weeks [72, 73, 75–78] (Figure 1.5C) – the shortest lifespan of any captive-bred vertebrate, and a dramatic outlier on the distribution of vertebrate lifespans (Figure 1.5D). Combined with their possession of several important vertebrate-specific adaptations – including an adaptive immune system – this extremely

short lifespan makes the turquoise killifish a highly promising model organism for ageing research [79].

Despite its very short lifespan, *N. furzeri* has been found to show a wide range of senescent phenotypes in even the shortest-lived strains, including lipofuscin deposition [73]; accumulation of senescence markers [73]; increased neurodegeneration [80, 81]; impaired learning and behavioural phenotypes [73, 80]; and a high incidence of degenerative and neoplastic lesions [82]. These diverse phenotypes indicate that the short lifespan of the turquoise killifish is the result of an accelerated general ageing process, rather than the specific failure of a particular organ or system. Moreover, established anti-ageing interventions such as resveratrol treatment [80], reduction in ambient temperature [83] and dietary restriction [84] also extend lifespan in the turquoise killifish, indicating a strong analogy with the ageing phenotypes observed in canonical model systems.

Due primarily to its potential as a model organism for ageing research, the turquoise killifish has also seen rapid development as a genetic model. The short-lived GRZ strain has been bred in captivity for fifty years and at least a hundred generations [85] and exibits a very high degree of homozygosity [76, 86, 87], providing a uniform genetic background for experimental interventions. A number of important genetic resources are now available, including several assemblies of the nuclear genome [77, 88, 89], the most recent of which [89] is of very high quality. These assemblies have yielded a genome with an estimated size of roughly 1.5 gigabases [89]; karyotyping [86] and sequencing analysis [88] both indicate a chromosome number of $2n = 38$, corresponding to 19 distinct linkage groups in the haploid genome.

Prior to the work contained in this thesis, almost nothing was known about the adaptive immune system of the turquoise killifish. However, as a teleost, and therefore as a jawed vertebrate, it could be strongly expected to have a roughly mammal-like adaptive immune system, and a number of genes related to B-cell adaptive immunity (including *RAG1*, various B-cell developmental markers, and fragments of antibody genes) were identified in one or more genome annotations. Phylogenetically, the genus *Nothobranchius* falls within the Cyprinodontiformes, and the turquoise killifish is therefore relatively closely related to several species with previously-characterised *IGH* loci, including fugu, stickleback and medaka [85, 90]. Despite the relatively unknown state of adaptive immunity in the turquoise killifish, therefore, there was good reason to believe it possessed a functional B-cell immune system similar to those of other teleosts, and therefore comparable in many respects with those of mammalian species, including humans. When combined with its short lifespan and rapid ageing phenotype, this made the turquoise killifish a potentially highly-valuable model for studying the forms and mechanisms of humoral adaptive immunosenescence in jawed vertebrates.

In this thesis, therefore, I characterised the *IGH* locus of the turquoise killifish (Chapter 3), established a working immunoglobulin sequencing protocol for killifish samples, and performed the first experiments investigating adaptive immunosenescence in this species (Chapter 4). The results demonstrated that turquoise killifish possess a complex and sophisticated adaptive immune system

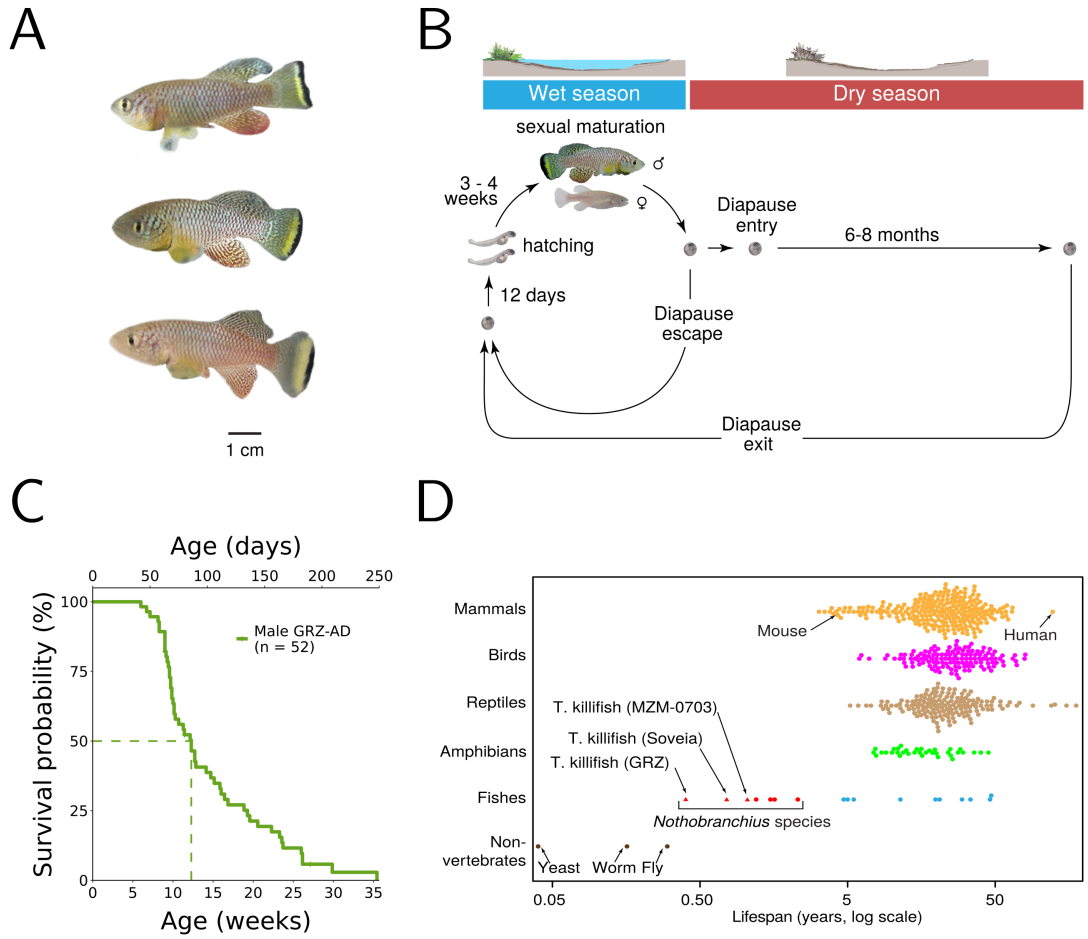


Figure 1.5: The turquoise killifish (*Nothobranchius furzeri*) as a model for vertebrate ageing: (A) Photographs of adult male yellow-tailed turquoise killifish. (B) Life cycle of the turquoise killifish in its natural environment, showing hatching and breeding during the rainy season and embryo survival in diapause during the dry season. (C) Example lifespan curve of short-lived GRZ-AD turquoise-killifish males, hatched in December 2016, with a median lifespan of roughly 12 weeks. (D) Comparison of turquoise-killifish maximum lifespan (red triangles) to other vertebrate species, demonstrating the extremely short lifespan of GRZ-strain *N. furzeri* compared to other vertebrates. (A), (B) and (D) adapted from [77], Figure 1.

with diverse and individualised antibody repertoires, which nevertheless undergoes a rapid loss of within-individual diversity with age. This loss of diversity was especially pronounced in isolated gut samples, possibly due to the intense antigenic exposure experienced at mucosal surfaces. Taken together, these results confirm the value of the turquoise killifish as a model system for investigating adaptive immunosenescence in vertebrates.

Chapter 2

Materials and methods

2.1 Killifish husbandry and sample-preparation methods

Male turquoise killifish (*Nothobranchius furzeri*, GRZ-AD strain) from a single hatching were raised under standard husbandry conditions [91] and housed from four weeks post-hatching in individual 2.8 L tanks connected to a water-recirculation system. Fish received 12 h of light per day on a regular light/dark cycle, and were fed blood worm larvae and brine shrimp nauplii twice a day during the week and once a day during the weekend [78, 91].

Sacrificed fish (Table E.24) were killed by anaesthetisation in $1.5\,\mathrm{g\,L^{-1}}$ Tricaine solution in room-temperature tank water [92], then flash-frozen in liquid nitrogen and ground to a homogenous powder with a pestle in a liquid-nitrogen-filled mortar. The powder was mixed thoroughly and stored at $-80\,^{\circ}\mathrm{C}$ prior to RNA isolation.

2.2 Biochemistry and molecular-biology methods

2.2.1 Standard laboratory techniques

2.2.1.1 PCR

The polymerase chain reaction is a well-established method for rapid amplification of a DNA sequence through repeated cycles of denaturation, primer-annealing and replication by a high-temperature-tolerant DNA polymerase enzyme [93]. Unless otherwise specified, all PCRs in this chapter were performed using $2\times$ Kapa HiFi HotStart ReadyMix PCR Kit (Appendix A.1) according to the manufacturer's instructions. Briefly, for a 25 µl reaction, 12.5 µl Kapa ReadyMix was combined with 12.5 µl total of template, nuclease-free water, and 10 µM primers; these volumes were scaled linearly for reactions of different volumes. The mixture was then heated in a thermocycler as described in Table 2.1.

Table 2.1: Thermocycler protocol for Kapa high-fidelity hot-start PCR

Step	**Temperature** (°C)	**Duration** (s)	**# Cycles**
Initial denaturation	95	180	1
Denaturation	98	20	
Annealing	T_a [a]	15	n_c [a]
Extension	72	t_{ext} [a]	
Final extension	72	$t_{ext} \times 4$ [a]	1

[a] Annealing temperature (T_a), extension time (t_{ext}) and cycle number (n_c) determined separately for each reaction.

2.2.1.2 Nucleic-acid purification with SeraSure magnetic beads

Nucleic-acid isolation, size-selection and concentration in the IgSeq library-preparation protocol (and elsewhere where necessary) were performed using SeraSure SPRI (solid-phase reversible immobilisation) bead preparations [94–97]. In SPRI, carboxyl-coated paramagnetic beads bind DNA in the presence of polyethylene glycol (PEG), with the affinity of the beads for DNA depending on the concentration of PEG in the binding buffer. As a result, the range of nucleic-acid sequence lengths retained by SPRI bead purification depends primarily on the concentration of PEG, which in turn depends on the relative volume of SeraSure bead suspension added to a sample; the higher the concentration, the shorter the minimum fragment length retained during the purification process. In combination with a magnetic rack to remove the DNA-bound beads from suspension, this allows DNA of the desired size range to be isolated from a solution and resuspended in the desired volume of fresh buffer.

To prepare 50 ml of SeraSure bead suspension for DNA (or DNA:RNA heteroduplex) isolation, a stock of SeraMag beads (Appendix A.2) was vortexed thoroughly, and 1 ml was transferred to a new tube. This tube was then transferred to a magnetic rack and incubated at room temperature for 1 min, then the supernatant was removed and replaced with 1 ml TET buffer (Appendix A.3) and the tube was removed from the rack and vortexed thoroughly. This washing process was repeated twice more, for a total of three washes in TET. A fourth cycle was used to replace the TET with incomplete SeraBind buffer (iSB, Appendix A.3). The vortexed 1 ml aliquot of beads in iSB was then transferred to a conical tube containing 28 ml iSB and mixed by inversion. To add the PEG, 20 ml 50 % (w/v) PEG 8000 solution was dispensed slowly down the side of the conical tube, bringing the total volume to 49 ml. Finally, this was brought to 50 ml by adding 250 µl 10 % (w/v) Tween 20 solution and 750 µl autoclaved water to complete the SeraSure bead suspension.

To perform a bead cleanup, an aliquot of prepared SeraSure suspension was vortexed thoroughly to completely resuspend the beads, then the appropriate relative volume of SeraSure suspension was added to a sample, mixing thoroughly by gentle pipetting. The sample was incubated at room temperature for 5 min to allow the beads to bind the DNA, then transferred to a magnetic rack and incubated for a further 5 min to draw as many beads as possible out of suspension. The supernatant was removed and discarded and replaced with 80 % ethanol, to a volume sufficient to completely submerge the bead pellet. The sample was incubated for 0.5-1 min, then the ethanol was replaced and incubated for a further 0.5-1 min. The second ethanol wash was removed, and the tube left on the rack until the bead pellet was almost, but not completely, dry, after which it was removed from the rack. The bead pellet was resuspended in a suitable volume of elution buffer (EB, Appendix A.3) then incubated at room temperature for at least 5 minutes to allow the nucleic-acid molecules to elute from the beads.

Unless otherwise specified, the beads from a cleanup were left in a sample during subsequent applications. To remove beads from a sample, the sample was mixed gently but thoroughly to resuspend the beads, incubated for an extended time period (at least 10 min) to maximise nucleic-acid elution, then transferred to a magnetic rack and incubated for 2-5 min to remove the beads from suspension. The supernatant (containing the eluted nucleic-acid molecules) was then transferred to a new tube, and the beads discarded.

2.2.1.3 Phenol-chloroform extraction and ethanol precipitation of DNA

Phenol-chloroform extraction is a well-established method for removing proteins and other hydrophobic or amphipathic contaminants from nucleic-acid solutions [98]. By thoroughly mixing an aqueous solution of nucleic acid with an organic solvent (phenol), hydrophobic contaminants are dissolved into the organic phase while proteins are denatured at the organic/aqueous boundary. The addition of chloroform further encourages the denaturation of proteins, as well as improving separation of the aqueous and organic phases following centrifugation. The effect on nucleic acids depends upon the pH of the phenol: under acidic conditions, the negatively-charged DNA phosphate backbone is neutralised and so primarily dissolves in the organic phase, while RNA is kept in the aqueous phase through hydrogen bonding between water and exposed bases; under neutral or basic conditions, both DNA and RNA are retained in the aqueous phase [98].

To remove protein from isolated DNA samples, each sample was diluted to 500 µl in nuclease-free water and mixed with 500 µl of equilibrated (non-acidic) phenol:chloroform:isoamyl alcohol (PCI) mixture (Appendix A.2) in a fume hood. The sample/PCI mixture was shaken vigorously by hand for 15 s to thoroughly mix the different components, then centrifuged in a benchtop centrifuge (5 min, room temperature, top speed). Again in a fume hood, the mixed sample was held at an angle, and the upper aqueous phase containing the DNA was removed and transferred to a new tube while the lower organic phase was discarded. A second aliquot of 500 µl PCI was added and the sample was mixed, centrifuged and separated as before. Finally, in order to remove residual phenol in the sample, 500 µl pure chloroform was added to the newly-separated aqueous phase and the sample was once again mixed, centrifuged and separated.

Following this final round of separation, the DNA in the aqueous phase was re-isolated using ethanol precipitation, another widely-used protocol exploiting the ability of alcohols to precipitate nucleic acids in the presence of monovalant cations [99]. 0.1 volumes of 3 M sodium acetate solution was added to each sample, followed by 2.5 volumes of fresh 100 % ethanol. The mixture was mixed gently by inversion, then incubated for 1-3 h at $-80\,^{\circ}$C or at $-20\,^{\circ}$C overnight. The suspension of precipitated DNA was pelleted through centrifugation in a benchtop centrifuge (30 min, 4 °C, top speed). The supernatant was discarded and replaced with 500 µl chilled 70 % ethanol, and the sample centrifuged again (5 min, 4 °C, top speed). After this, the supernatant was again discarded, and the samples allowed to air-dry before being resuspended in 30-50 µl EB (Appendix A.3).

2.2.1.4 Guanidinium thiocyanate-phenol-chloroform extraction of RNA

Guanidinium thiocyanate-phenol-chloroform extraction is a technique for purifying RNA samples closely related to the phenol-chloroform-extraction method described in Section 2.2.1.3 [98]. By using acid rather than equilibrated phenol, DNA in the sample is dissolved in the organic rather than the aqueous phase and so is removed from the aqueous solution along with proteins and other contaminants [98, 100]. Meanwhile, the addition of guanidinium thiocyanate, a chaotropic agent which disrupts hydrogen bonds in aqueous solution, strongly encourages the rapid denaturation of proteins and so helps protect the RNA from degradation by RNase enzymes [98, 100].

To isolate total RNA from homogenised killifish tissues, 1 ml of QIAzol lysis reagent (Appendix A.2, containing acid phenol and guanidinium thiocyanate) was added to 0.1 g of tissue, mixed gently but thoroughly by inversion, then incubated at room temperature for 5 min to allow the QIAzol to penetrate the tissue. 0.2 volumes of chloroform was added and the mixture was shaken vigorously for 15 s, then incubated at room temperature for 3 min. The mixture was then centrifuged (15 min, 4 °C, 12 000 *g*). Holding the tube at an angle, the upper aqueous phase containing the RNA was removed and transferred to a new tube, while the lower organic phase was discarded.

Following phase separation, the RNA was precipitated using isopropanol precipitation, which works on the same principles as ethanol precipitation (Section 2.2.1.3) but requires shorter incubation times and lower relative volumes of alcohol [99]. 0.5 volumes of room-temperature isopropanol was added to each sample, mixing gently by inversion and incubating for 5 min at room temperature. The suspension was centrifuged (10 min, 4 °C, 12 000 *g*) and the supernatant discarded. 1 volume of freshly prepared 75 % ethanol was added and the tube was vortexed briefly and centrifuged again (5 min, 4 °C, 7500 *g*). The supernatant was discarded and the RNA pellet allowed to air-dry for 5-10 min, then resuspended in 50 µl EB (Appendix A.3). The concentration and quality of the resulting total-RNA solution were assayed with the Qubit 2.0 flourometer (Thermo Fisher, RNA BR assay kit) and TapeStation 4200 (Agilent, RNA tape), respectively, according to the manufacturer's instructions.

2.2.2 Library size-selection with the BluePippin

The BluePippin (Sage Science) is a DNA size-selection system based on agarose gel electrophoresis, which uses timed switching between positively-charged electrodes at a forked gel channel in an agarose cassette to redirect DNA of a desired size range into a separate lane from the rest of the sample [101]. The timing of the switch is determined based on the size range input by the user and calibrated using flourescent internal standards, which are added to the sample during the preparation process and designed to run well ahead of the possible size ranges for that cassette type. The combination of the choice of cassette and the choice of standards determines which fragment lengths can be effectively isolated using the machine.

For the experiments described in this thesis, a 1.5 % cassette with R2 markers was used, enabling size selection of targets in the range of 250–1500 bp [101]. Machine calibration and testing, cassette preparation, and protocol design were performed in accordance with the BluePippin documentation and instructions given by the machine software. During this process, the elution wells of the lanes to be used in the size-selection run were emptied and refilled with 40 µl of electrophoresis buffer (Appendix A.2), then sealed for the duration of the run, and a broad-range size-selection protocol with a target range of 400 to 800 bp was specified. 30 µl of sample was then combined with 10 µl of loading solution (Appendix A.2) and vortexed to mix, then 40 µl of buffer was removed from the appropriate loading well and replaced, slowly, with the prepared sample mixture. The protocol was started and run until the final elution was complete. Finally, the eluted samples were removed from the elution wells of the appropriate lanes, and the unused lanes of the cassette were re-sealed for future use.

2.2.3 Isolation and sequencing of bacterial artificial chromosomes

All BAC clones that were sequenced for this research were provided by the FLI in Jena as plate or stab cultures of transformed *E. coli*, which were replated and stored at 4 °C. Prior to isolation, the clones of interest were cultured overnight in at least 100 ml LB medium. The resulting liquid cultures were transferred to 50 ml conical tubes and centrifuged (10-25 min, 4 °C, 3500 *g*) to pellet the cells. The supernatant was carefully discarded and the cells were resuspended in 18 ml buffer P1 (Appendix A.3).

After resuspension, the cultures underwent alkaline lysis [102] to release the BAC DNA and precipitate genomic DNA and cellular debris. 18 ml buffer P2 (Appendix A.3) was added to each tube, which was then mixed gently but thoroughly by inversion and incubated at room temperature for 5 min. 10 ml ice-chilled neutralisation buffer P3 (Appendix A.3) was added to precipitate genomic DNA and cellular debris, and each tube was mixed gently but thoroughly by inversion and incubated on ice for 15 min. The tubes were then centrifuged (20-30 min, 4 °C, 12 000 *g*) to pellet cellular debris and the supernatant was transferred to new conical tubes. This process was repeated at least two more times, until no more debris was visible in any tube; this repeated pelleting was necessary to minimise contamination in each sample, as the normal column- or paper-based filtering steps used during alkaline lysis resulted in the loss of the BAC DNA.

Following alkaline lysis, the DNA in each sample underwent isopropanol precipitation: 0.6 volumes of room-temperature isopropanol were added to the clean supernatant in each tube, followed by 0.1 volumes of 3 M sodium acetate solution. Each tube was mixed well by inversion, incubated for 10-15 min at room temperature, then centrifuged (30 min, 4 °C, 12 000 *g*) to pellet the DNA. The supernatant was discarded and the resulting DNA smear was "resuspended" in 1 ml 100 % ethanol and transferred to a 1.5 ml tube, which was re-centrifuged (5 min, 4 °C, top speed) to obtain a concentrated pellet. Finally, the pelleted samples were resuspended in EB (Appendix A.3) and purified of proteins and RNA using phenol-chloroform extraction and ethanol precipitation (Section 2.2.1.3).

The resuspended BAC isolates were sent to the Cologne Center for Genomics, where they underwent Illumina Nextera XT library preparation and were sequenced on an Illumina MiSeq sequencing machine (MiSeq Reagent Kit v3, 2×300 bp reads).

2.2.4 Immunoglobulin sequencing of killifish samples

2.2.4.1 RNA template quantification and quality control

Total RNA from whole-body killifish samples was isolated as described in Section 2.2.1.4; gut RNA from microbiota-transfer experiments [78] was already prepared and available. Quantification of RNA samples was performed with the Qubit 2.0 flourometer (Thermo Fisher, RNA BR assay kit), while quality control and integrity measurement was performed using the TapeStation 4200 (Agilent, RNA tape), both according to the manufacturer's instructions.

2.2.4.2 Reverse transcription and template switching

Reverse transcription of total RNA and template switching for IgSeq library preparation was performed using SMARTScribe Reverse Transcriptase (Appendix A.1), in line with the protocol specified in Turchaninova *et al.* [103]. Briefly, 750 ng total RNA from a killifish sample was combined with 2 µl 10 µM gene-specific primer (GSP), homologous with the second CH exon of *N. furzeri IGHM* (Appendix B.2, designed using Primer3 [104]). The reaction volume was brought to a total of 8 µl with nuclease-free water, and the resulting mixture was incubated for 2 minutes at 70 °C to denature the RNA, then cooled to 42 °C to anneal the GSP [103].

Following annealing, the RNA-primer mixture was combined with 12 µl of reverse-transcription master-mix (Table 2.2), including the reverse-transcriptase enzyme and template-switch adapter (Appendix B.1, sequence provided in [103]). The complete reaction mixture was incubated for 1 h at 42 °C for the reverse-transcription reaction, then mixed with 1 µl of uracil DNA glycosylase (UDG, Appendix A.1) and incubated for a further 40 min at 37 °C to digest the template-switch adapter. Finally, the reaction product was purified using SeraSure beads (Section 2.2.1.2, Table 2.4).

Table 2.2: Master-mix components for SMARTScribe reverse transcription (per sample)

Volume (µl)	Component	Initial concentration	Reference
2	SMARTScribe reverse transcriptase	100 U µl^{-1}	Appendix A.1
4	SMARTScribe first-strand buffer	5 ×	Appendix A.2
2	SmartNNNa barcoded TSA	10 µM	Appendix B.1
2	DTT	20 mM	Appendix A.2
2	dNTP mix	10 µM per nucleotide	Appendix A.2
0.5	RNasin RNase inhibitor	40 U µl^{-1}	Appendix A.1

2.2.4.3 PCR amplification and adapter addition

Following reverse transcription, UDG digestion, and cleanup, the reaction mixture underwent three successive rounds of Kapa PCR (Section 2.2.1.1, Table 2.3) each of which was followed by a further round of bead cleanups (Section 2.2.1.2, Table 2.4). The first of these PCR reactions added a second strand to the reverse-transcribed cDNA and amplified the resulting DNA molecules; the second added partial Illumina sequencing adapters and further amplified the library, and the third added complete Illumina adapters (Appendix B.3, including i5 and i7 indices [105]). Primer sequences (Appendix B.2) homologous to the template-switch adapter (M1SS and M1S) were provided by Turchaninova *et al.* [103], while those homologous to the $C_\mu 1$ constant-region exon (IGHC-B and IGHC-C) were designed using Primer3 [104].

Table 2.3: PCR protocols for *N. furzeri* immunoglobulin sequencing

PCR	Protocol details			Primers		Volumes (µl)[b]			
	# cycles	T_m (°C)	t_{ext} (s)	F	R	Template	Primers[c]	Kapa	H_2O
1	18	65	15	IGHC-B	M1SS	10.5	1 (×2)	12.5	0
2	13	65	15	M1S+P2	IGHC-C+P1	1	0.5 (×2)	12.5	10.5
3	8 to 12[d]	68	15	D50*[a]	D7**[a]	2	0.75 (×2)	12.5	9

[a] PCR3 primers selected as appropriate for each library's assigned indices; see Appendix B.3 for more information.
[b] If the number of samples to be sequenced was small, all volumes of PCR 3 were doubled for a 50 µl total PCR volume.
[c] The stated volumes apply separately to both forward and reverse primers for each reaction. All primers were diluted to and stored at an initial concentration of 10 µM.
[d] See Table 2.5 for specific cycle numbers used in each experiment.

2.2.4.4 Library pooling, size selection and sequencing

Following PCR3 and its attendant bead cleanup, the total concentration of each library was assayed with the Qubit 2.0 (Thermo Fisher, DNA HS assay kit), while the size distribution of each library was obtained using the TapeStation 4200 (Agilent, D1000 tape). To obtain the concentration of complete library molecules in each case (as opposed to primer dimers or other off-target bands), the ratio between the concentration of the desired library band (c. 620-680 bp) and the total concentration of the sample was calculated for each TapeStation lane, and the total concentration of each library as measured by the Qubit was multiplied by this ratio to obtain an estimate of the desired quantity:

$$\text{Library concentration } (L) = \text{Qubit concentration} \times \frac{\text{TapeStation concentration [main band]}}{\text{TapeStation concentration [total]}} \quad (2.1)$$

Table 2.4: Bead cleanups during *N. furzeri* immunoglobulin sequencing

Stage[a]	Sample volume	Beads volume (µl)		Elution volume (µl)[c]
		µl	×[b]	
RT	21	14.7	0.7	16.5
PCR 1	25	17.5	0.7	25
PCR 2	25	17.5	0.7	15
PCR 3	25[d]	20[d]	0.8	15[d]
Pooling	Varies[e]	Varies[e]	2.5	35

[a] Each bead cleanup takes place immediately *after* its corresponding stage.
[b] Bead volumes are usually given as multiples of the sample volume.
[c] All elutions performed in the specified volume of elution buffer (EB, Appendix A.3).
[d] If the PCR 3 reaction volume differed from 25 µl, bead and elution volumes were rescaled proportionally to sample volume as appropriate.
[e] Samples are pooled in equimolar ratio.

Table 2.5: PCR cycle numbers during *N. furzeri* immunoglobulin sequencing

Experiment	Number of PCR cycles			Reference
	PCR 1	PCR 2	PCR 3	
Pilot	18	13	9	Section 4.4
Ageing	18	13	8	Section 4.5
Gut	18	13	12	Section 4.6

All the libraries for a given experiment were then pooled, such that the estimated concentration L of each library in the final pooled sample was equal and the total mass of nucleic acid in the pooled sample was at least 240 ng. The pooled libraries underwent a final bead cleanup (Section 2.2.1.2, Table 2.4) to concentrate the samples, then underwent size selection with the BluePippin (Section 2.2.2, 1.5 % DF Marker R2, broad 400-800bp). The size-selected pooled samples underwent a final round of quality control with the Qubit and TapeStation (as above) to confirm their collective concentration (at least 1.5 nM) and size distribution (one peak at c. 620-680 bp). Finally, the pooled and size-selected libraries were sequenced on an Illumina MiSeq System (MiSeq Reagent Kit v3, 2×300 bp reads, 30 % PhiX spike-in), either at the Cologne Center for Genomics (pilot and ageing experiments) or with Admera Health (gut experiment).

2.3 Computational and analytic methods

Version details for software used in this section are given in Table E.1.

2.3.1 General data processing, pipeline structure, and data visualisation

Unless otherwise specified, processing and analysis of biological data was performed using standard Bioconductor [106] packages: Biostrings [107] and BSgenome [108] for biological sequence data, GenomicRanges [109] for sequence ranges, and genbankr [110] and rentrez [111] for GenBank files.

Smith-Waterman and Needleman-Wunsch exhaustive alignments [112, 113] were performed using the `pairwiseAlignment` function from Biostrings; percentage sequence identities were computed using the `pid` function from the same package.

Processing of tabular data was performed using the Tidyverse suite of tools, especially readr [114], dplyr [115], tidyr [116] and stringr [117]. Snakemake [118] was used to design and run data-processing pipelines.

Unless otherwise specified, standard statistical operations and comparisons were performed using built-in functions from the stats package in R [119]: `wilcox.test` for two-sample Mann-Whitney *U* tests, `kruskal.test` for multisample Kruskal-Wallis analysis-of-variance tests, `lm` for linear models, `glm` for generalised linear models, `cor` for Pearson product-moment correlation coefficients and `hclust` for hierarchical clustering. Principal co-ordinate analysis was performed using `pcoa` from the ape package [120].

Unless otherwise specified, data were visualised using ggplot2 [121]. Chromosome ideograms, locus structure visualisations, and Sashimi plots [122] were constructed using Gviz [123]. Cluster dendrograms and phylogenetic trees were drawn with ggtree [124], using utilities from ape [120] and tidytree [125]. Sequence logos were drawn with ggseqlogo [126].

2.3.2 BAC insert assembly

2.3.2.1 Identifying BAC candidates for the *Nothobranchius furzeri IGH* locus

The first group of candidate BAC clones [88] to be used in the *N. furzeri* locus assembly was identified by searching for scaffolds in a previous assembly of the *N. furzeri* genome (`NotFur1`, GenBank accession GCA_000878545.1 [77]) that contained either *IGH* gene fragments (GapFilledScaffold_8761, 8571, 16121) or genes homologous to those flanking the *IGH* locus in stickleback and medaka (GapFilledScaffold_2443 and 292). Subsequences from these scaffolds were sent to Dr Kathrin Reichwald at the FLI in Jena, who identified four BAC clones (193A03, 276N03, 209K12, 181N10) with sequenced ends close to the query sequences.

Following sequencing and assembly of these BAC inserts, a further group of BACs was identified using a second, independent genome assembly (GenBank accession GCA_001465895.2, [88]) and the database of BAC end sequences, which by then were publically available. The assembled BAC sequences were found to map within or near a large, gapped region on synteny group 3 of this

genome assembly, and BACs were selected that either intruded into this gapped region or had end sequences that mapped to another scaffold aligning to the assembled BAC inserts (scaffold01427, scaffold02214, scaffold01820). In total, 11 further BACs were sequenced and assembled in this second round (223M21, 162F04, 220O06, 248A22, 165M01, 206K13, 154G24, 208A08, 277J10, 109B21, 216D12).

2.3.2.2 Sequence trimming, filtering and correction

Demultiplexed, adapter-trimmed MiSeq sequencing data were uploaded by the sequencing provider to Illumina BaseSpace and accessed via the Illumina utility program BaseMount. Reads from each library were processed with Trimmomatic [127] to remove adapter sequences, trim low-quality sequence regions, and discard any trimmed reads below a minimum length:

```
trimmomatic PE -phred33 <forward_reads_fastq>
   ↪ <reverse_reads_fastq> <output_paths>
   ↪ ILLUMINACLIP:<adapter_directory>/TruSeq3-PE.fa:2:30:10
   ↪ LEADING:20 TRAILING:20 SLIDINGWINDOW:4:30 MINLEN:36
```

Following this, the trimmed reads were filtered to remove *E. coli* genomic DNA and other contaminants by aligning them using Bowtie 2 [128] and retaining read pairs that did not align concordantly:

```
bowtie2 --very-sensitive-local --local --reorder --un-conc
   ↪ <output_prefix> -x <ecoli_genome_index_path> -1
   ↪ <forward_reads_fastq> -2 <reverse_reads_fastq> -S
   ↪ <sam_file_prefix>
```

Before sequence assembly, the filtered reads then underwent correction, to reduce the impact of errors occurring during the library preparation and sequencing process. In order to increase the reliability of the resulting scaffolds and reduce the impact of ideosyncracies of any given correction tool, the reads were corrected in parallel using two different programs; QuorUM [129]:

```
quorum -d -q "33" -p <output_path> <interleaved_reads_files>
```

and BayesHammer (the built-in correction tool of the SPAdes genome-assembly software [130, 131]):

```
spades.py -1 <forward_reads_fastq> -2 <reverse_reads_fastq> -o
   ↪ <output_path> --disable-gzip-output
   ↪ --only-error-correction --careful --cov-cutoff auto -k
   ↪ 21,33,55,77,99,127 --phred-offset 33
```

2.3.2.3 Sequence assembly and scaffolding

Following read correction, each pair of independently-corrected reads files was passed to SPAdes [130] for *de novo* genome assembly:

```
spades.py -1 <forward_reads_fastq> -2 <reverse_reads_fastq> -o
   ↪ <output_path> --disable-gzip-output --only-assembler
   ↪ --careful --cov-cutoff auto -k 21,33,55,77,99,127
   ↪ --phred-offset 33
```

Following assembly, any *E. coli* scaffolds resulting from residual contaminating reads were identified by aligning scaffolds to the *E. coli* genome using BLASTN [132, 133], and scaffolds containing significant matches were discarded. The remaining scaffolds were then scaffolded using SSPACE [134], using jumping libraries from the two killifish genome assemblies mentioned in Section 2.3.2.1 [77, 88]:

```
SSPACE_Standard_v3.0.pl -x 0 -k 5 -a 0.7 -n 15 -z 200 -g 1 -p 0
   ↪ -l <jumping_library_config_file> -s
   ↪ <spades_scaffolds_file>
```

In order to guarantee the reliability of the assembled scaffolds, the assemblies produced with BayesHammer- and QuorUM-corrected reads were compared, and scaffolds were broken into segments whose contiguity was agreed on between both assemblies. To integrate these fragments into a contiguous insert assembly, points of agreement between BAC assemblies from the same genomic region (e.g. two scaffolds from one assembly aligning concordantly to one scaffold from another) and between BAC assemblies and genome scaffolds, were used to combine scaffolds where possible. Any still-unconnected scaffolds were assembled together through pairwise end-to-end PCR (Section 2.2.1.1, with one primer on the end of each scaffold designed using Primer3 [104]) and Sanger sequencing (Eurofins) [135].

2.3.3 Locus characterisation and assembly

2.3.3.1 Collating reference sequences

Most publications presenting characterisations of *IGH* loci do not provide easy-to-use databases of trimmed and curated gene segments, and the data that is available is often partial and heterogeneous between publications. In order to obtain standardised reference databases for locus characterisation, further analysis was performed on publically-available data from three reference species with previously-characterised *IGH* loci: medaka (*Oryzias latipes*) [25], zebrafish (*Danio rerio*) [136] and three-spined stickleback (*Gasterosteus aculeatus*) [27, 28], as described below. Following automatic

sequence extraction, the reference sequences were checked manually for any severely pathological (e.g. out-of-frame) sequences and edited before being used for inference in novel loci.

Medaka

GenBank files of the annotated medaka *IGH* locus were downloaded from the supplementary information of the medaka locus paper ([25], additional file 6) and corrected to make them parsable by genbankr. Locus sequence and annotation ranges were extracted from these GenBank files into FASTA and tab-separated tabular formats, respectively, and segment annotations were renamed to match the naming conventions used in other species. VH, DH, JH and constant-region-exon nucleotide sequences were extracted from the locus sequence using these annotations. Amino-acid sequences for VH, JH and constant-region sequences were obtained automatically by identifying the reading frames which minimised the number of stop codons in each sequence.

Stickleback

Limited sequence information on the *IGH* locus in stickleback, including VH segments and bulk (non-exon-separated) constant regions was provided in a GenBank file in the locus characterisation paper for medaka ([25], additional file 6), while additional sequence information (including DH and JH nucleic-acid sequences and amino-acid sequences of constant-region exons) was extracted manually from one of the stickleback locus papers ([27], Figure S1 to S4) into FASTA files. As with medaka, the GenBank reference file was downloaded, corrected and parsed to yield a FASTA file of the locus sequence and tab-separated tabular files of annotation ranges. VH sequences were extracted from the locus sequence using these annotation ranges and translated as specified for medaka above; JH sequences provided by Bao *et al.* [27] were translated such that the final nucleotide formed the last position of the final codon.

To obtain nucleic-acid sequences of the constant-region exons, the amino-acid sequences from Bao *et al.* [27] were aligned to the locus sequence with TBLASTN [137], with a query coverage threshold of 40 % and a maximum of three HSPs per query sequence:

```
tblastn -query <ch_aa_fasta> -subject <gac_locus_fasta>
   ↪ -qcov_hsp_perc 40 -max_hsps 3 -outfmt '<output_format>' >
   ↪ <output_path>
```

with the following standardised tabular output format:

```
6 qseqid sseqid pident qcovhsp length mismatch gapopen gaps
   ↪ sstrand qstart qend sstart send evalue bitscore qlen slen
```

To filter out alignments across subloci, any alignment of an exon upstream of the annotated boundaries of its corresponding bulk constant region (whose ranges were specified in the GenBank file) was discarded; the alignment with the highest score for each exon was then used to extract the corresponding nucleic-acid sequence from the locus. In order to control for any errors, either during manual copying

of locus sequences from the source paper or in the paper itself, these nucleic-acid sequences were then re-translated to generate new amino-acid sequences, again using the translation frame producing the fewest stop codons.

Zebrafish

GenBank files corresponding to the zebrafish *IGH* locus were provided (without segment annotations) on GenBank by Danilova *et al.* [136]; this publication also provided detailed co-ordinates for the VH, DH and JH segments (but not constant exons) on these sequences. Aligned amino-acid sequences were provided for the exons of *IGHM* and *IGHZ*, but no detailed information about *IGHD* exons could be found; as a result, reference information about *IGHD* was not used from this species.

As with stickleback, the amino-acid sequences provided were aligned to the locus sequences with TBLASTN to identify and extract exon nucleic-acid sequences, which were then translated using the frame yielding the fewest stop codons for each sequence. VH sequences were obtained using the ranges provided in Danilova *et al.* [136] and translated in the same manner. DH and JH nucleotide sequences were obtained directly from Danilova *et al.* [136]; as with stickleback, JH amino-acid sequences were obtained by translating the nucleotide sequences in the frame such that the final nucleotide formed the last position of the final codon.

2.3.3.2 Identifying putative locus sequences

In order to identify sequences in a genome assembly potentially containing part of an *IGH* locus, reference VH, JH and constant-region nucleotide and amino-acid sequences were mapped to the assembly using BLAST [132, 133]. Nucleotide sequences were aligned to the locus using the relatively permissive `blastn` algorithm:

```
blastn -task blastn -query <reference_exon_fasta> -subject
   ↪ <locus_fasta> -outfmt '<output_format>'
```

Protein sequences, meanwhile, were aligned using the standard `blastp` algorithm:

```
blastp -query <reference_exon_fasta> -subject <locus_fasta>
   ↪ -outfmt '<output_format>'
```

In both cases, the tabular output format specified in Section 2.3.3.1 was used, to provide a predictable format for downstream processing of BLAST alignment tables.

Following alignment of reference sequences, overlapping alignments to reference segments of the same segment type, isotype (if applicable) and exon number (if applicable) were collapsed together, keeping track of the number of collapsed alignments and the best E-values and bitscores obtained for each alignment group. Alignment groups with a very poor maximum E-value (> 0.001) were discarded, as were groups consisting of fewer than two alignments and groups overlapping with

much better alignments to a different sequence type, where "much better" was defined as a bitscore difference of at least 33. Following resolution of conflicts, VH and CH alignments underwent a second filtering step of increased stringency, requiring a minimum E-value of 10^{-10} to be retained.

Following alignment filtering, scaffolds containing surviving alignments to at least two distinct segment types (where VH, JH, and each type of constant-region exon each counted as one segment type), or alignments to one segment type covering at least 1 % of the scaffold's total length were retained as potential locus scaffolds. To reduce computational runtime spent processing irrelevant sequence on long scaffolds, each candidate scaffold so identified was trimmed to 100kb before the first putative gene segment and 100kb after the last one; in the case of *Nothobranchius furzeri* and *Xiphophorus maculatus*, these ranges were further reduced following more thorough segment characterisation (Section 2.3.3.4).

The exact set of reference sequences used for this extraction process differed depending on the genome being analysed. For *Nothobranchius furzeri* (Section 3.2), the reference sequences extracted from medaka, stickleback and zebrafish (Section 2.3.3.1) were used; for *Xiphophorus maculatus* (Section 3.3), gene segments inferred for *N. furzeri* were also included; and for other species (Section 3.4), the reference sequences plus those inferred for both *N. furzeri* and *X. maculatus* were used.

2.3.3.3 Locus sequence finalisation

In the case of both *Nothobranchius furzeri* and *Xiphophorus maculatus*, a single chromosome (chromosome 6 in *N. furzeri*, chromosome 16 in *X. maculatus*) was identified as bearing the *IGH* locus in that species. In the case of *X. maculatus*, this was the only segment-bearing scaffold identified in the genome, and the completed locus sequence was obtained by simply trimming the chromosomal sequence at either end of the segment-bearing region. In contrast, multiple scaffolds from the *N. furzeri* genome were also identified as bearing at least one potential *IGH* segment (Table 3.1). In order to identify which of these were in fact part of the locus and integrate them into a contiguous sequence, BAC candidates identified and assembled as described in Section 2.3.2 were incorporated into the assembly.

To do this, all assembled BAC inserts were screened for *IGH* locus segments in the same manner described for genome scaffolds (Section 2.3.3.2). Passing BACs (Table 3.2) were aligned to the candidate genome scaffolds with BLASTN and integrated manually together, giving priority in the event of a sequence conflict to (i) any sequence containing a gene segment missing from the other, and (ii) the genome scaffold sequence if neither sequence contained such a segment. BACs and scaffolds which could not be integrated into the locus sequence in this way were discarded as orphons.

2.3.3.4 Locus segment characterisation

Detailed characterisation of *IGH* gene segments was performed on finished *IGH* locus sequences for *Xiphophorus maculatus* and *Nothobranchius furzeri*, and on isolated candidate scaffolds for other species, using the same reference segment databases used to identify candidate scaffolds for that species in Section 2.3.3.2. The specific methods used depended on segment type.

VH

To identify VH segments on newly characterised loci, reference VH segments were used to construct a multiple-sequence alignment with PRANK [138]:

```
prank -d=<reference_vh_db> -o=<output_path> -gaprate=0.00001
   ↪ -gapext=0.00001 -F -termgap
```

The resulting alignment was used as an input to NHMMER [139–142], which constructs a Hidden Markov Model from a multiple-sequence alignment and uses it to identify matching sequences in a reference sequence:

```
nhmmer --dna --notextw --tblout <output_path> -T 80
   ↪ <vh_alignment> <locus_sequence_path>
```

where `-T 80` specified the minimum alignment score required to report a match. The resulting match table was used to identify candidate ranges in the locus sequence corresponding to VH segments; these ranges were extended by 9bp at either end to account for boundary errors, and the corresponding nucleotide sequences were extracted to a FASTA file. Each sequence was then checked and refined manually: 3' ends were identified by the start of the RSS heptamer sequence (consensus `CACAGTG` [36]), if present, while 5' ends and FR/CDR boundaries were identified using IMGT/DomainGapAlign [143] with the default settings. Where necessary, IMGT/DomainGapAlign was also used to IMGT-gap the VH segments in accordance with the IMGT unique numbering [33].

An initial amino-acid sequence for each VH segment was produced automatically from the extracted nucleotide sequence by identifying the reading frame which minimised the number of stop codons in the sequence; this worked well for most segments. VH amino-acid sequences were then refined (and in a few cases re-translated) using the manually-refined nucleotide sequences, including end-refinement and FR/CDR boundary identification.

Following extraction and manual curation, VH segments were grouped into families based on their pairwise sequence identity. In order to assign segments to families, the nucleotide sequence of each VH segment in a locus was aligned to every other segment using Needleman-Wunsch global alignment, and the resulting matrix of pairwise sequence identities was used to perform single-linkage hierarchical clustering on the VH segments. The resulting dendrogram was cut at 80 % sequence identity to obtain VH families. These families were then numbered based on the order of the first-

occurring VH segment from that family in the first *IGH* sublocus in which the family is represented, and each VH segment was named based on its parent sublocus, its family, and its order among elements of that family in that sublocus (Table E.4 and Tables E.11 to E.15).

JH

As with VH segments, JH segments were identified by building a multiple-sequence alignment with PRANK and using it to construct an HMM with NHMMER; the parameters used were the same as for VH segments, except that there was no minimum score for NHMMER to report a sequence match (`-T 0` instead of `-T 80`). The resulting sequence ranges were extended by 20 bp on either end and extracted into FASTA format. These sequences were then trimmed automatically by identifying the RSS heptamer sequence at the 5' end and the splice junction motif (`GTA`) at the 3' end, then checked and refined manually.

IgBLAST [144] identifies CDR3 boundaries for recombined *IGH* VDJ sequences using an auxiliary file specifying the reading frame of each JH segment, along with the co-ordinate of the conserved `TGG` codon (corresponding to the conserved W118 residue in the recombined sequence [145]) marking the CDR3/FR4 boundary. An auxiliary file for the inferred JH segments was generated automatically by searching for the conserved sequence using a series of regular-expression patterns of decreasing stringency (Table 2.6), taking the first match in each sequence as the desired residue; this determined both the reading frame and the W118 sequence co-ordinate. Once generated, the auxiliary file was then used to determine the reading frame for automatically translating the JH sequences; both the auxiliary file and amino-acid FASTA file were then edited to incorporate any manual refinements made to the JH nucleotide sequences. Curated JH sequences were named based on their order within their parent sublocus and, where applicable, on whether they were upstream of *IGHZ* or *IGHM* constant regions (Tables E.8 and E.19).

Table 2.6: Regular expressions used to search for conserved W118 residues in JH sequences

#	Pattern
1	`TGGGBNNNNGBN`
2	`TGGGBNNNGBN`
3	`TGGGBNNNNNGBN`
4	`TGGGBNNNNNNGBN`
5	`TGGGBN`

DH

Unlike VH and JH gene segments, DH segments are too short and unstructured to be found effectively using an HMM-based search strategy. Instead, DH segments in assembled loci were located using their distinctive pattern of flanking recombination signal sequences (Section 1.2.2): an antisense RSS in 5', then a short D-segment, then a sense RSS in 3'. Potential matches to this pattern were searched for using FUZZNUC from the EMBOSS collection of bioinformatics tools [146], with a high error tolerance to account for deviations from the conserved sequence in either or both of the RSSs:

```
fuzznuc -pattern
   ↪ 'GGTTTTTGTN(10,14)CACTGTGN(1,25)CACAGTGN(10,14)ACAAAAACC'
```

```
    ↪ -pmismatch 8 -rformat gff -outfile <output_path>
    ↪ <locus_sequence_path>
```

This generated a GFF file of permissive matches, representing potential DH segments; these were then arranged by sequence co-ordinate, and higher-mismatch candidates overlapping with a lower-mismatch alternative were discarded.

Automatic orientation of DH segments based on their own sequence is challenging, as the segments themselves have no clear conserved structure and the flanking RSSs are rotationally symmetric. To overcome this problem and orientate the DH segments on the locus, the table of DH candidate ranges was combined with previously-identified VH and JH ranges. Each DH candidate was then orientated based on the orientations of its flanking segments: segments with an oriented segment immediately upstream or downstream adopted the orientation of that segment, while segments with contradictory orientation information were discarded. This process was repeated until all DH candidates had either been orientated or discarded.

After orientation, the DH ranges were used to extract DH sequences in FASTA format from the locus sequence; these sequences then underwent a second, more stringent filtering step, in which sequences lacking the most conserved positions in each RSS [36] were discarded:

```
grep -B 1
    ↪ '[ACTG]\{{25,27\}}TG[ACTG]\{{1,25\}}CA[ACTG]\{{25,27\}}'
    ↪ <dh_fasta> | sed '/^--$'/d > <output_fasta>
```

Finally, the identified DH candidates were checked manually, candidates without good RSS sequences were discarded, and flanking RSS sequences were trimmed to obtain the DH segment sequences themselves. As with the JH segments, these were numbered based on their order within their parent sublocus and, when applicable, on whether they were upstream of *IGHZ* or *IGHM* constant regions (Tables E.5 and E.17).

CH

To detect and identify constant-region exons in the characterised loci, constant-region nucleotide and protein sequences from reference species were mapped to the locus sequence using BLAST [132, 133], in the same manner described for putative locus scaffolds in Section 2.3.3.2. Following alignment of reference sequences, overlapping alignments to reference segments of the same isotype and exon number were collapsed together, keeping track of the number of collapsed alignments and the best E-values and bitscores obtained for each alignment group. Alignment groups with a very poor maximum E-value (> 0.001) were discarded, as were groups overlapping with a much better alignment to a different isotype or exon type, where "much better" was here defined as a bitscore difference of at least 16.5. Where conflicting alignments to different isotypes or exon types

co-occurred without a sufficiently large difference in bitscore, both alignment groups were retained for manual resolution of exon identity.

Following resolution of conflicts, alignment groups underwent a second filtering step of increased stringency, requiring a minimum E-value of 10^{-8} and at least two aligned reference exons over all reference species to be retained. Each surviving alignment group was then converted to a sequence range, extended by 10 bp at each end to account for truncated alignments failing to cover the ends of the exon, and used to extract the corresponding exon sequence into FASTA format. These sequences then underwent manual curation to resolve conflicting exon identities, assign exon names and perform initial end refinement based on putative splice junctions.

In order to validate intron/exon boundaries and investigate splicing behaviour among *IGH* constant-region exons in *N. furzeri* and *X. maculatus*, published RNA-sequencing data (Table E.2) were aligned to the annotated locus using STAR [147]. In both cases, reads files from multiple individuals were concatenated and aligned together, in order to make the intron/exon boundary changes in mapping behaviour as clear as possible.

Before aligning the RNA-seq reads, each locus underwent basic repeat masking, using the built-in zebrafish repeat parameters from RepeatMasker [148]:

```
RepeatMasker -species danio -dir <masked_locus_dir> -s
   ↪ <unmasked_locus_path>
```

After masking, a STAR genome index was generated from each locus:

```
STAR --runMode genomeGenerate --genomeDir
   ↪ <star_index_directory_path> --genomeFastaFiles
   ↪ <masked_locus_path> --genomeSAindexNbases <sa_index>
```

where the `--genomeSAindexNbases` option determined the size of the suffix-array index and was dependent on the length of the reference sequence being indexed:

$$\text{SA index size (bits)} = \left\lfloor \frac{\log_2(\text{length of reference sequence})}{2} - 1 \right\rfloor \tag{2.2}$$

Following index generation, the RNA-seq reads were mapped to the generated index as follows:

```
STAR --genomeDir <star_index_directory_path> --readFilesIn
   ↪ <input_reads> --outFilterMultimapNmax 5 --alignIntronMax
   ↪ 10000 --alignMatesGapMax 10000
```

where the `--outFilterMultimapNmax` option excludes read pairs mapping to more than five distinct co-ordinates in the reference sequence, the `--alignIntronMax` option excludes reads spanning predicted introns of more than 10 kb, and the `--alignMatesGapMax` option excludes read pairs

mapping more than 10 kb apart. Following alignment, the resulting SAM files were processed into sorted, indexed BAM files using SAMtools [149] and visualised with Integrated Genomics Viewer (IGV [150, 151]) to determine intron/exon boundaries of predicted exons, as well as the major splice isoforms present in each dataset.

In order to reduce time and memory requirements for generating alignment figures (Figures 3.6 and 3.13), secondary alignments were performed on truncated loci consisting only of the *IGHM/D* or (where present) *IGHZ* constant regions, plus a few flanking kilobases on each side. In these cases, the additional parameters constraining multimapping, intron length and mate distance were not necessary due to the much shorter and less-repetitive reference sequence.

For species other than *N. furzeri* or *X. maculatus*, intron/exon boundaries were predicted manually based on BLASTN and BLASTP alignments to closely-related species and the presence of conserved splice-site motifs (`AG` at the 5' end of the intron, `GT` at the 3' end [152]). In cases where no 3' splice site was expected to be present (e.g. for CM4 or TM2 exons), the nucleotide exon sequence was terminated at the first canonical polyadenylation site (`AATAAA` if present, otherwise one of `ATTAAA`, `AGTAAA` or `TATAAA` [153]), while the amino-acid sequence was terminated at the first stop codon. In many cases, it was not possible to locate a TM2 exon due to its very short conserved coding sequence (typically only 2 to 4 amino-acid residues [27, 136]).

2.3.3.5 Cross-sublocus sequence comparison

Synteny between subloci in the *N. furzeri* locus was analysed using the standard synteny pipeline from the R package DECIPHER [154], which searches for chains of exact *k*-mer matches within two sequences:

```
DBPath <- tempfile()
DBConn <- dbConnect(SQLite(), DBPath)

Seqs2DB(seqs = <sublocus_1_sequence>, type = "XStringSet",
   ↪ dbFile = DBConn, identifier = "IGH1", verbose = FALSE)
Seqs2DB(seqs = <sublocus_2_sequence>, type = "XStringSet",
   ↪ dbFile = DBConn, identifier = "IGH2", verbose = FALSE)

dbDisconnect(DBConn)

SyntenyObject <- FindSynteny(dbFile = DBPath, verbose = FALSE)
```

Cross-locus sequence comparisons between gene segments were performed analogously to the comparisons involved in VH family assignment, with `pairwiseAlignment` and `pid` from Biostrings.

2.3.4 Phylogenetic trees

2.3.4.1 Species tree construction and annotation

Information about the interrelationships of most of the teleost taxa discussed in this thesis was obtained from the comprehensive teleost phylogeny of Hughes *et al.* [90], while additional, higher-resolution information on the interrelationships of African killifishes missing from that tree was provided by Cui *et al.* [155]. As no single published tree covered all the species of interest, a simple cladogram of relationships was constructed manually from the information provided by these two sources. Annotations (e.g. of clade membership or isotype status) were added using tidytree [125].

2.3.4.2 Phylogenetic inference on *IGH* locus sequences

Three phylogenetic trees were inferred from molecular data of *IGHZ* gene segments: one on the VH segments on *Nothobranchius furzeri* and *Xiphophorus maculatus*, one on CH exons from all species, and one on *IGHZ* constant-regions of *IGHZ* bearing species. In all cases, a sequence FASTA database was assembled from the relevant species. As identical sequences can cause problems during phylogenetic analysis, entries with completely identical sequences were then collapsed together into a single FASTA sequence, which was relabelled with the names of all its parent sequences.

A multiple-sequence alignment of the remaining sequences was then constructed with PRANK:

```
prank -d=<sequence_fasta> -o=<output_prefix> -DNA -termgap
```

The resulting alignment was passed to the maximum-likelihood phylogenetic inference program RAxML [156–158], using the SSE3-enabled parallelised version of the software, the standard GTR-Gamma nucleotide substitution model, and built-in rapid bootstrapping:

```
raxmlHPC-PTHREADS-SSE3 -f a -m GTRGAMMA -s <ch_prank_alignment>
    ↪ -w <output_dir> -N <n_bootstrap_replicates> -x
    ↪ <bootstrap_seed> -p <parsimony_seed> -n <output_suffix>
```

Finally, the bootstrap-annotated `RAxML_bipartitions` file was inspected and rooted manually in Figtree [159], before being annotated and visualised in R with tidytree and ggtree, respectively.

VH-segment tree

In order to build a phylogenetic tree of VH segments from the *N. furzeri* and *X. maculatus IGH* loci, all VH sequences from those loci were labelled with their origin species and combined together into a single FASTA file. Sequences with more than 25 % missing characters were discarded prior to PRANK alignment. During tree-inference with RAxML, 100 bootstrap replicates were used.

CH exon tree

To build a phylogenetic tree of CH exons, nucleotide sequences of constant exons from all species involved in this study were labelled with their origin species and combined into a single FASTA file, which was then filtered to discard transmembrane exons, secretory tails, and sequences with more than 25 % missing characters. In addition, CM4 nucleotide sequences were trimmed to the coding region, removing the 3'-UTR. As with the VH-segment tree, 100 bootstrap replicates were used during tree-inference with RAxML. As the outgroup among CH exon groups is unknown, the tree was visualised in unrooted format (Figure D.4).

IGHZ tree

To investigate the evolution of *IGHZ*, the C_ζ1-4 exons from each *IGHZ* constant region found in any of the analysed genomes from Section 2.3.3.4 were concatenated together into a single sequence and labelled with the source species and constant region. In the event of partial constant regions missing one or more C_ζ exons, the remaining exons were concatenated together in the usual order. Following database processing and alignment, RAxML tree-inference was run using 1000 bootstrap replicates, in order to increase the reliability and precision of the support values obtained. During tree visualisation, nodes with bootstrap support of less than 65 % were collapsed into polytomies.

2.3.5 IgSeq data pre-processing

Unless otherwise specified, pre-processing utilities used in the following sections were provided by the pRESTO [160] and Change-O [161] suites of IgSeq processing tools.

2.3.5.1 Sequence uploading and annotation

Demultiplexed, adapter-trimmed MiSeq sequencing data were uploaded by the sequencing provider to Illumina BaseSpace and accessed via the Illumina utility program BaseMount. Library annotation information (fish ID, sex, strain, age at death, death weight, etc.) was added to the read headers of each library FASTQ file:

```
ParseHeaders add -f <field_keys> -u <field_values> -s
   ↪ <input_reads_file>
```

The replicate and individual identity of each library were then added as additional annotations, by concatenating sets of other annotations together as specified by the user:

```
ParseHeaders.py merge -f <field_keys> -k INDIVIDUAL --act cat
   ↪ -s <annotated_reads_file>
ParseHeaders.py merge -f <field_keys> -k REPLICATE --act cat -s
   ↪ <individual_annotated_reads>
```

Following annotations, reads from different libraries were pooled together, then split by replicate identity:

```
SplitSeq.py group -f REPLICATE -s <pooled_reads_file>
```

This pooling and re-splitting process enables all reads considered to be a single replicate to be processed together even if sequenced separately, maximising the effectiveness of UMI-based pre-processing, while also allowing all replicates to be processed in parallel.

2.3.5.2 Quality control, primer masking and UMI extraction

After pooling and re-splitting, the raw read set underwent quality control, discarding any read with an average Phred score [162] of less than 20:

```
FilterSeq quality -q 20 -s <split_reads_file>
```

Following quality filtering, the reads underwent processing to identify and remove invariant primer sequences. To do this, known primer sequences were aligned to each read from a fixed starting position, and the best match on each read was identified and trimmed. To begin with, the primer sequences from the third PCR step of the library prep protocol were used, trimming off primer sequences corresponding to part of the constant $C_\mu 1$ exon and the 5' invariant part of the template-switch adapter:

```
MaskPrimers score --mode cut --start 0 -s <split_reads_3prime>
   ↪ -p <CM1_primer_file>;
MaskPrimers score --mode cut --start 0 -s <split_reads_5prime>
   ↪ -p <TSA_primer_file>
```

Following this, the forward reads underwent a second round of masking using the 3' invariant part of the TSA sequence (`CTTGGGG`), and the intervening 16 bases were extracted and recorded in each read header as that read's unique molecular identifier (UMI):

```
MaskPrimers score --mode cut --barcode --start 16 --maxerror
   ↪ 0.5 -s <masked_reads_5prime> -p <TSA_3prime_sequence_file>
```

As the match sequence for this second round of masking is shorter and more error-prone than the primer sequences used in the first round, an increased mismatch tolerance (`--maxerror 0.5`) was permitted, increasing the number of reads with successfully-extracted UMIs.

2.3.5.3 Barcode error handling

In order to reduce the level of barcode errors in each dataset, primer-masked IgSeq reads underwent barcode clustering, in which reads with the same replicate identity and highly similar UMI sequences

were grouped together into the same molecular identifier group (MIG). To do this, 5'-reads were clustered by UMI sequence using CD-HIT-EST [163, 164] with a 90 % sequence identity cutoff, with cluster identities being recorded in a new CLUSTER field in each read header [165]:

```
ClusterSets barcode -f BARCODE -k CLUSTER --cluster cd-hit-est
   ↪ --prefix B --ident 0.9 -s <remasked_reads_5prime>
```

In order to split any genuinely distinct MIGs accidentally united by this process, as well as to reduce the level of barcode collisions, the reads then underwent a second round of clustering, this time separately on the read sequences within each barcode cluster using VSEARCH (an open-source alternative to USEARCH [166, 167]). This time, the cluster dendrogram was cut at 75 % total sequence identity, and each subcluster was separated into its own distinct MIG [165]:

```
ClusterSets set -f CLUSTER -k CLUSTER --cluster vsearch
   ↪ --prefix S --ident 0.75 -s <clustered_reads_5prime>
```

These clustering thresholds (90 % for barcode clustering, 75 % for barcode splitting) were identified empirically as the values that maximise the number of reads passing downstream quality checks in turquoise-killifish data.

The cluster annotations from these two clustering steps were combined into a single annotation, uniquely identifying each MIG in each replicate. These annotations were further modified to designate the replicate identity of each read, giving each MIG a unique annotation across the entire dataset. These annotations were copied to the reverse reads, such that each read pair had a matching MIG annotation, and reads without a mate (due to differential processing of the two reads files) were discarded:

```
ParseHeaders collapse -s <reclustered_reads_5prime> -f CLUSTER
   ↪ --act cat;
ParseHeaders merge -f REPLICATE CLUSTER -k RCLUSTER --act set
   ↪ -s <annotated_reads_5prime>;
PairSeq -1 <reannotated_reads_5prime> -2 <masked_reads_3prime>
   ↪ --1f BARCODE CLUSTER RCLUSTER --coord illumina
```

2.3.5.4 Consensus-read generation and pair merging

Following barcode clustering, the IgSeq forward reads were grouped based on cluster identity, and the reads in each cluster grouping were aligned and collapsed into a consensus read sequence:

```
BuildConsensus --bf RCLUSTER --cf <header_fields> --act
   ↪ <copy_actions> --maxerror 0.1 --maxgap 0.5 -s
   ↪ <paired_reads_file>
```

Positions at which at least half the aligned reads in the MIG had a gap character were deleted from the consensus (`--maxgap 0.5`), while MIGs with a mismatch rate from the consensus of more than 10 % were discarded from the dataset (`--maxerror 0.1`). The resulting FASTQ file contained a single consensus sequence for each cluster annotation, labelled with its CONSCOUNT (the total number of reads contributing to that consensus sequence), the number of reads allocated to each barcode in the cluster, and various header fields (`<header_fields>`) propagated from the contributing reads by summing or concatenating the values from each contributing read (`<copy_actions>`). An identical consensus-read-generation step was performed on the reverse reads.

After consensus-read generation had been performed for both 5' and 3' reads, the annotations attached to each read were again unified across read pairs with matching cluster identities, and consensus reads without a mate of the same cluster identity were dropped:

```
PairSeq -1 <consensus_reads_5prime> -2 <consensus_reads_3prime>
   ↪ -coord presto
```

Following consensus-read generation and annotation unification, consensus-read pairs with matching cluster annotations were aligned and merged into a single contiguous sequence. Where possible, this was done by simply aligning the two mate sequences against each other; where this was not possible (e.g. due to the lack of a significant sequence overlap) the consensus reads were instead aligned with BLASTN [132, 133] to a reference database of VH sequences to generate a merged sequence, with `N` characters used to separate pairs that aligned in a non-overlapping manner on the same VH segment [160]:

```
AssemblePairs sequential --coord presto --scanrev --aligner
   ↪ blastn --rc tail --1f <header_fields> -1
   ↪ <paired_consensus_5prime> -2 <paired_consensus_3prime> -r
   ↪ <vh_fasta_file>
```

In either case, annotation fields were copied to the new merged sequence from the forward consensus read, with the fields to be copied specified by `--1f <header_fields>`. Sequence pairs for which both alignment approaches failed were discarded.

2.3.5.5 Collapsing identical sequences and singleton removal

To convert a dataset of MIGs (representing distinct RNA molecules) into one of unique sequences, merged consensus sequences from Section 2.3.5.4 with identical insert sequences but distinct cluster identities were collapsed together into a single FASTQ entry, recording the number, size and UMI makeup of contributing MIGs in each case in the sequence header alongside any existing annotation information:

```
CollapseSeq --inner --cf <header_fields> --act <copy_actions>
   ↪ -n 20 -s <merged_consensus_seqs>
```

The collapsed sequences from each replicate identity in the dataset were then concatenated into a single file for easier downstream processing.

As sequences represented by only a single read across all MIGs in the dataset (so-called *singleton* sequences, with a total CONSCOUNT of no more than 1) could not be corrected by UMI clustering or consensus building, they are not considered reliable for downstream processing and analysis [165]; as such, they were identified and separated from the other collapsed sequences:

```
SplitSeq group -f CONSCOUNT --num 2 -s
   ↪ <collapsed_consensus_seqs>
```

Finally, the non-singleton sequences so identified were converted into FASTA format with seqtk [168] for downstream processing:

```
seqtk seq -a <non_singleton_consensus_seqs> >
   ↪ <presto_fasta_output>
```

2.3.5.6 Assigning VDJ identities with IgBLAST

To assign VH, DH and JH identities to the sequences output by the pRESTO pipeline, the format-FASTA output from Section 2.3.5.5 was aligned to databases of reference segments with IgBLAST [144]. To do this, each reference file was converted into a BLAST database with `makeblastdb`, and the output FASTA file was aligned to these databases with `igblastn`:

```
makeblastdb -parse_seqids -dbtype nucl -in <vh_reference_fasta>
   ↪ -out <vh_db_prefix>;
makeblastdb -parse_seqids -dbtype nucl -in <dh_reference_fasta>
   ↪ -out <dh_db_prefix>;
makeblastdb -parse_seqids -dbtype nucl -in <jh_reference_fasta>
   ↪ -out <jh_db_prefix>;
igblastn -ig_seqtype Ig -domain_system imgt -query
   ↪ <presto_fasta_output> -out <igblast_output>
   ↪ -germline_db_V <vh_db_prefix> -germline_db_D
   ↪ <dh_db_prefix> -germline_db_J <jh_db_prefix>
   ↪ -auxiliary_data <jh_aux_file> -outfmt '7 std qseq sseq
   ↪ btop'
```

A JH auxiliary file (Section 2.3.3.4) was used to indicate the reading frame and CDR3 boundary co-ordinate of each JH sequence in the reference database (`-auxiliary_data <jh_aux_file>`).

2.3.5.7 Clonotype inference with Change-O

Following V/D/J identity assignment, the output FASTA file from Section 2.3.5.5, raw reference segment databases from Section 2.3.3.4 and segment assignments from Section 2.3.5.6 were used to construct a tab-delimited Change-O sequence database:

```
MakeDb igblast --regions --scores --failed --partial --cdr3 -i
   ↪ <igblast_output> -s <presto_fasta_output> -r
   ↪ <vh_reference_fasta> <dh_reference_fasta>
   ↪ <jh_reference_fasta>
```

where `--failed` indicates that invalid sequences should be included in a separate database rather than discarded outright, `--regions` and `--cdr3` indicate that the database should include comprehensive FR and CDR annotations, `--scores` indicates that the database should include alignment score metrics, and `--partial` indicates that sequences with incomplete V/D/J alignments (e.g. those without an unambiguous V- or J-assignment) should not automatically qualify as failed. Following database construction, each entry was given a unique name on the basis of its replicate identity and ordering, and the database was filtered to exclude sequences with a V-alignment score of less than 100 (Section 4.4.1).

In order to compute the appropriate distance threshold for clonotype assignment, each sequence was assigned a nearest-neighbour Hamming distance within the repertoire, using the related R packages SHazaM and Alakazam [161]:

```
tab <- readChangeoDb(<named_db_path>) %>% mutate(ROW = seq(n()))
dist_pass <- distToNearest(tab, model = "ham", normalize =
   ↪ "len", fields = "INDIVIDUAL", first = FALSE)
dist_fail <- tab %>% filter(! ROW %in% dist_pass$ROW) %>%
   ↪ mutate(DIST_NEAREST = NA)
writeChangeoDb(bind_rows(dist_pass, dist_fail),
   ↪ <db_output_path>)
```

where `model = "ham"` indicates that a single-nucleotide Hamming distance metric is to be used, `normalize = "len"` that distances should be normalised by total sequence length, `first = FALSE` determines how to handle ambiguous V/J calls, and `fields = "INDIVIDUAL"` specifies that only distances between sequences from the same source individual should be considered.

Following assignment of nearest-neighbour distances, a distance threshold for clonotyping was computed by fitting a pair of unimodal distributions to the nearest-neighbour distribution over all sequences and selecting the threshold that maximises the average of sensitivity and specificity when assigning an observation to one of these two distributions (Section 4.4.2) [169]. As with nearest-neighbour distance assignment, this was done in R using SHazaM and Alakazam; all four possible

models (fitting either a normal or gamma distribution to each of the two peaks) were tried, and the one with the highest maximum likelihood was used to compute the threshold value:

```
tab <- readChangeoDb(<changeo_db_path_with_distances>)
models <- c("gamma-gamma", "gamma-norm", "norm-gamma",
   ↪ "norm-norm")
thresholds <- numeric(length(models))
likelihoods <- numeric(length(models))
for(n in 1:length(models)){
  obj <- tryCatch(findThreshold(as.numeric(tab$DIST_NEAREST),
   ↪ method = "gmm", model = "hmm", cutoff = "opt"), error =
   ↪ function(e) return(e$message), warning = function(w)
   ↪ return(w$message))
  thresholds[n] <- ifelse(isS4(obj), obj@threshold, NA)
  likelihoods[n] <- ifelse(isS4(obj), obj@loglk, NA)
}
if (!all(is.na(thresholds)))
   ↪ write(thresholds[last(which(likelihoods ==
   ↪ max(likelihoods, na.rm = TRUE))], <threshold_output_path>)
```

Following inference of the correct distance threshold, clonotype inference was performed on the sequence database by grouping sequences by V- and J-assignment and CDR3 length, computing pairwise Hamming distances between each pair of CDR3 sequences in each group, and performing single-linkage clustering on the resulting distance matrix [170], with a maximum of one ambiguous `N` character permitted in the CDR3 junctional sequence:

```
DefineClones --act set --model ham --sym min --norm len
   ↪ --failed -d <changeo_db> --dist
   ↪ <cluster_distance_threshold> --gf INDIVIDUAL --link
   ↪ single --maxmiss 1
```

where `--maxmiss 1` specifies a maximum of 1 junctional `N` character per sequence; `--act set` tells the program how to handle ambiguous V/D/J assignments; `--model ham` specifies the clustering metric as the pairwise Hamming distance; `--norm len` indicates that the Hamming distances should be normalised by sequence length; `--sym min` specifies that, in the event of asymmetric `A -> B` and `B -> A` distances (e.g. arising from length normalisation) the minimum distance should be used; `--dist` specifies the distance threshold at which to cut the clustering dendrogram; and `--gf INDIVIDUAL` states that only sequences from the same individual can belong to the same clone.

Finally, the clonotype numbers assigned to each individual were combined with the ID of the corresponding source individual, giving a unique ID for each clonotype in the dataset.

2.3.5.8 Germline inference and segment annotation

After threshold determination and clonotyping, a so-called "full-length germline sequence" is constructed for each sequence. To do this for a given sequence, germline V/J sequences are trimmed of deleted positions and concatenated together, separated by a masked region with length corresponding to the inserted nucleotides and intervening DH sequence:

```
CreateGermlines -g dmask --cloned -d <clonotyped_changeo_db> -r
   ↪ <vh_reference_fasta> <dh_reference_fasta>
   ↪ <jh_reference_fasta> --failed
```

where `-g dmask` indicates that DH nucleotides should be masked in the germline sequence, and `--failed` indicates that sequences that fail germline assignment should be retained in a separate database. Importantly, `--cloned` indicates that sequences from the same clone should receive the same germline assignment, based on a simple majority rule among sequences in the clone; this process also enables assignment of unambiguous segment identities to ambiguously-assigned sequences within larger clones, improving segment calls in the dataset.

Following germline inference, each sequence in the dataset was annotated according to whether or not it possessed V/D/J assignments and whether these assignments were ambiguous (i.e. whether multiple possible assignments were given rather than just one), and combined VJ and VDJ assignments were obtained by concatenating the individual segment assignments as appropriate. The processed sequence databases were then passed on to downstream analysis pipelines as outlined in Section 2.3.6.

2.3.5.9 Tracking read survival

Read survival during the pre-processing pipeline was tracked for each replicate at each stage from importation of raw reads to clonotyping (no sequences were lost during germline inference). During the pRESTO pipeline, read counting was performed by extracting sequence headers into a tab-delimited table

```
ParseHeaders.py table -s <input_seqs> -f INDIVIDUAL REPLICATE
   ↪ <other_fields>
```

with other fields (e.g. CONSCOUNT) as appropriate to the stage in the pipeline. Change-O databases, meanwhile, were already in tab-delimited tabular format. Following conversion to tabular format where necessary, read survival for each replicate was assessed by aggregating the CONSCOUNT fields of all sequences assigned to each replicate; prior to consensus-building, each sequence was assigned a CONSCOUNT of 1, making this equivalent to simply counting the number of sequences.

2.3.6 Downstream analysis of antibody repertoires

2.3.6.1 Clonal counts and inter-replicate correlations

Following the completion of the pipeline from Section 2.3.5, the number and proportion of sequences successfully assigned clonotypes was evaluated by counting the number and proportion of unique sequences in the final Change-O database with missing (`NA`) clonal identities. The size of each clone in the dataset was found by counting the number of unique sequences assigned to that clonal identity, either in total or for each replicate separately; a clone was designated to be present in a replicate if at least one sequence in that replicate was assigned to that clone. Following inference of clone sizes in each replicate, the correlation between replicates was computed using a simple Pearson's product-moment correlation coefficient (implemented in `cor` in R) comparing their respective clone-size vectors.

2.3.6.2 Zipf approximation of rank/frequency distributions

To obtain the rank/frequency distributions of clones in a Change-O database, the size of each clone in each repertoire (Section 2.3.6.1) was divided by the total number of unique sequences in that repertoire to obtain a relative frequency for each clone; these frequencies were then ranked in descending order within each repertoire. The resulting distributions were then plotted on a log-log plot to visualise the underlying distribution. Best-fit Zipf distributions could then be obtained for each repertoire using maximum-likelihood estimation, either on all clones or after excluding some number of the largest clones in each repertoire:

```
lzipf <- function(f, s){
  # Compute the negative log-likelihood of a set of frequencies
  # under a Zipf distribution with parameter s
  s * sum(f * log(1:length(f))) + sum(f) * log(H(length(f), s))
}

dzipf <- function(r, s, N){
  # Return the predicted frequency of a rank r under a Zipf
  # distribution for a population with total size N
  (1/(r^s))/H(N, s)
}

compute_zipf_slope <- function(frequencies, n_exclude){
  # Estimate a Zipf exponent for a frequency vector
```

```
  m <- mle(function(s) lzipf(frequencies[-(1:n_exclude)], s),
   ↪ start = list(s=1))
  return(m@coef[["s"]])
}

# Add Zipf slope and predicted clonal frequencies to a
# pre-computed clone table
clone_table <- clntab %>% group_by(INDIVIDUAL) %>%
   ↪ arrange(CLONE_RANK) %>% mutate(S =
   ↪ compute_zipf_slope(CLONE_SIZE, n_exclude), EXP_FREQUENCY
   ↪ = dzipf(CLONE_RANK, S, n()), EXP_SIZE = sum(CLONE_SIZE) *
   ↪ EXP_FREQUENCY))
```

The resulting expected frequency from the fitted Zipf distribution could then be overlaid on the actual observed frequency to compare the fit of the inferred distribution to the actual clonal repertoire (Section 4.4.2). Meanwhile, the ranks and frequencies computed as part of the estimation process could be used to compute the observed P20 of the repertoire (i.e. the sum of frequencies of all clones with rank 20 or less), as well as the expected P20 (i.e. the sum of frequencies for those top-20 clones predicted by the fitted Zipf distribution).

2.3.6.3 Rarefaction analysis for inter-experiment comparison

Inter-individual comparisons of metrics such as clonal counts, P20, or the proportion of large clones in the repertoire make the implicit assumption that the sampling process (in particular, the sampling effort) used to obtain each repertoire was similar between the individuals being compared. When this is not the case – when the number of cycles used to amplify each library or the number of sequencing reads per library differs between individuals, for example – sample composition is liable to differ systematically between repertoires, making comparisons using such metrics unreliable. In the case of immune-repertoire sequencing, the very large number of small clones in most immune repertoires [30] means that increased sampling depth is likely to lead to larger clonal counts, lower P20 values, and smaller proportions of large clones.

In order to compare the clonal richness and P20 values of antibody repertoires from different IgSeq experiments in the turquoise killifish in a more reliable manner, rarefaction analysis [171] was performed on these repertoires at the level of unique sequences. For each of a range of sample sizes (from 10^2 to 10^4 unique sequences), sequence entries in the pre-processed Change-O database were repeatedly subsampled without replacement from the repertoire of each individual in each experiment, for a total of 20 iterations per sample size. For each experiment, individual, and sample size, the number of small (fewer than 5 unique sequences), large (5 or more unique sequences) and total clones

for each individual, as well as the P20, was computed for each subsample, and the mean and standard deviation of each measurement was computed across subsamples. The resulting rarefaction curves (of clonal count or P20 vs sample size) could then be plotted and used to compare repertoires of different experiments at any given sample size (Section 4.6).

2.3.6.4 Hill diversity spectra

The procedure for computing Hill diversity spectra (Appendix C) from a Change-O sequence database was adapted (and substantially expanded) from the code for the same purpose provided in the R package Alakazam [161, 172]. To begin with, columns in the database were specified designating how to divide the entries into an outer group (e.g. an age group), a finer inner group (e.g. an individual), and an even-finer "clone" group (by default the clonal identity of each sequence entry, but could also be a V(D)J identity or any other set of sequence categories). Counts of unique sequences were then computed for each "clone" within each inner group (`clone_tab`), and aggregated to find total sequence counts for each inner group (`group_tab`):

```
clone_tab <- data %>% group_by_(.dots = c(outer_group,
   ↪ inner_group, clone_field)) %>% dplyr::summarize(COUNT =
   ↪ n())
group_tab <- clone_tab %>% group_by_(.dots = c(outer_group,
   ↪ inner_group)) %>% dplyr::summarize_(SEQUENCES =
   ↪ interp(~sum(x, na.rm = TRUE), x = as.name("COUNT")))
```

The size of the bootstrap replicates (`NSAM`) for each sample was then computed as the minimum total sequence count across all inner groups, and a table of `nboot` bootstrap replicates of "clonal" frequencies, each of size `NSAM`, was computed for each outer-group/inner-group combination through multinomial resampling:

```
bootstrap_abundance <- function(nboot, clone_tab, group_tab,
   ↪ outer_group_name, outer_group_value, inner_group_name,
   ↪ inner_group_value, clone_field){
        # Generate independent bootstrap samples for a given
        # group combination in a clone table
        gtab <- group_tab[group_tab[[outer_group_name]] ==
                outer_group_value &
                group_tab[[inner_group_name]] ==
                inner_group_value,]
        ctab <- clone_tab[clone_tab[[inner_group_name]] ==
                inner_group_value &
                clone_tab[[outer_group_name]] ==
```

```
                outer_group_value,]
        # Get sequence number for resampling
        n <- gtab$NSAM
        # Get abundances of each clone in group
        abund_obs <- ctab$COUNT
        # Infer abundances of observed and unseen clones using
        # Chao's method (functions provided in alakazam)
        p1 <- adjustObservedAbundance(abund_obs)
        p2 <- inferUnseenAbundance(abund_obs)
        # Adjusted vector of clonal frequencies
        abund_inf <- c(p1, p2)
        # Specify clone names for known and unknown clones
        n1 <- ctab[[clone_field]]
        n2 <- paste("UNKNOWN", inner_group_value,
                seq_along(p2), sep = "_")
        names_inf <- c(n1, n2)
        # Use inferred clone distribution and resampling size to
        # generate independent bootstrap samples through
        # multinomial sampling
        sample_mat <- rmultinom(nboot, n, abund_inf)
        sample_tab <- melt(sample_mat, varnames =
                c(clone_field, "ITER"), value.name = "N")
        sample_tab[[clone_field]] <-
                names_inf[sample_tab[[clone_field]]]
        sample_tab[[outer_group_name]] <- outer_group_value
        sample_tab[[inner_group_name]] <- inner_group_value
        return(sample_tab)
}
```

Following bootstrap generation and concatenation of bootstrap tables from different inner-group/outer-group combinations, individual Hill diversity numbers (Appendix C.1.3) could be computed for each inner group for each bootstrap replicate (using alakazam's built-in `calcDiversity` function [161, 172]) for each diversity order in a vector of orders `Q`:

```
bs_tab_solo <- bootstraps %>% group_by_(.dots = c(outer_group,
   ↪ inner_group, "ITER"))
div_tab_bs_solo <- tibble()
for (q in Q){
        dt <- summarise(bs_tab_solo, Q = q,
```

```
                D = calcDiversity(N, q = q), N_GROUP = 1)
        div_tab_bs_solo <- bind_rows(div_tab_bs_solo, dt)
}
```

Similarly, different sorts of aggregate diversity spectrum (Appendix C.2) could be computed for each outer group:

```
# Gamma diversity (simple diversity over each outer group)
bs_tab_gamma <- bootstraps %>% group_by_(.dots =
   ↪ c(group_within, clone_field, "ITER")) %>% summarise(N =
   ↪ sum(N)) %>% group_by_(.dots = c(outer_group, "ITER"))
div_tab_bs_gamma <- tibble()
for (q in Q){
        dt <- summarise(bs_tab_gamma, Q = q,
                D = calcDiversity(N, q = q))
        div_tab_bs_gamma <- bind_rows(div_tab_bs_gamma, dt)
}

# Alpha diversity (average diversity across inner groups
# within each outer group)
bs_tab_alpha <- bootstraps %>% group_by_(.dots = c(outer_group,
   ↪ inner_group, "ITER"))
div_tab_split <- tibble()
for (q in Q){
        dt <- summarise(bs_tab_alpha, Q = q,
                D = calcDiversity(N, q = q))
        div_tab_split <- bind_rows(div_tab_split, dt)
}
div_tab_bs_alpha <- div_tab_bs_alpha %>% group_by_(.dots =
   ↪ c(outer_group, "ITER", "Q")) %>% summarise(D =
   ↪ ifelse(dplyr::first(Q) != 1,
   ↪ mean(D^(1-dplyr::first(Q)))^(1/(1-dplyr::first(Q))),
   ↪ exp(mean(log(D)))), N_GROUP = n())

# Beta diversity (gamma divided by alpha)
div_tab_bs_beta <- full_join(div_tab_bs_gamma,
   ↪ div_tab_bs_alpha, by = c(outer_group, "ITER", "Q"),
   ↪ suffix = c("_GAMMA", "_ALPHA")) %>% mutate(D=
   ↪ D_GAMMA/D_ALPHA) %>% select(-D_GAMMA, -D_ALPHA)
```

Finally, each diversity spectrum was grouped by diversity order and summarised across bootstrap replicates, to obtain means and standard deviations (and therefore 95 % confidence intervals) at each diversity order. Solo, alpha and gamma spectra could then be plotted (and compared between groups or against other diversity metrics) directly from these summarised tables; however, as the beta diversity of an outer group depends on the number of inner groups it contains (`N_GROUP`), beta spectra of outer groups of different sizes needed to be rescaled to a common range (0 for minimum possible beta-diversity, 1 for maximum) prior to plotting in order to be comparable (Appendix C.2.3):

```
beta_rescaled <- <summarised_beta_diversity> %>% mutate(D =
    ↪ (D-1)/(N_GROUP-1), D_UPPER = (D_UPPER-1)/(N_GROUP-1),
    ↪ D_LOWER = (D_LOWER-1)/(N_GROUP-1), D_SD =
    ↪ D_SD/(N_GROUP-1)) %>% select(-N_GROUP)
```

2.3.6.5 Repertoire Dissimilarity Index (RDI)

The Repertoire Dissimilarity Index (RDI) [173] is a pairwise distance metric between immune repertoires, based on the relative prevalence of different gene segments (or combinations of segments) in each repertoire. To compute a set of pairwise RDIs between a group of repertoires, the number of sequences assigned to each segment-choice identity is first counted for each repertoire. These counts vectors are then downsampled without replacement to the size of the repertoire with the fewest unique sequences, then normalised to sum to some constant factor. The normalised counts are then transformed with the inverse hyperbolic sine function, which is roughly linear for values close to zero and logarithmic for values greater than one [173]; this transformation puts greater weight on rare segment-choice categories to prevent them being dominated by changes in the most prevalent categories [173]. Finally, the distance between each pair of repertoires is calculated as the Euclidean distances between their respective transformed counts vectors. This process is then repeated multiple times with independent downsamplings, and the final RDI between each pair of repertoires is given as the arithmetic mean across all iterations.

In this thesis, RDIs were computed based on the VJ-assignment of each sequence in each repertoire. To compute VJ-RDIs between repertoires in a collated Change-O database, the database was first filtered to remove sequences with missing or ambiguous V- or J-calls. The database columns denoting VJ-identity and repertoire membership (by replicate for the pilot dataset, by individual for the ageing and gut dataset) were then extracted, and used to construct a VJ-counts table in the format specified by the rdi package [173] in R:

```
# Extract VJ and ID columns
genes <- pull(<filtered_changeo_db>, <vj_call_field>)
annots <- as.character(pull(<filtered_changeo_db>,
    ↪ <repertoire_id_field>))
```

```
# Create counts table
counts <- calcVDJcounts(genes = genes, seqAnnot = annots)
```

The RDI between each pair of repertoires could then be computed as described above using the provided `calcRDI` function from the rdi package [173]:

```
rdi <- calcRDI(counts, nIter = 100)
```

where `nIter = 100` specifies that the RDI measurements should be averaged over 100 iterations.

Following RDI computation, hierarchical (average-linkage) clustering was performed on the resulting distance matrix, and the dendrogram of the resulting clustering structure was visualised using ggtree. Principal co-ordinate analysis (PCoA) was performed on the RDI distance matrix (using the `pcoa` function from the R package ape [120]), then processed into standard tabular format for visualisation.

2.3.6.6 Analysing generative repertoire diversity with IGoR

The processes generating the primary antibody repertoire include VDJ recombination, P-insertion or deletion at the ends of the selected gene segments, and N-insertion between them (Section 1.2.2). The enormous potential diversity of sequences generated by these processes makes it impossible to accurately model the generative processes by explicitly assigning probabilities to each possible sequence in the primary repertoire, as in any given sample of rearranged sequences the observed frequency of the vast majority of possible sequences will be zero [48]. The program IGoR bypasses this problem by instead explicitly estimating the probability distributions of each contributing stochastic process separately, producing a generative model that can be used to estimate the entropy of the generative process, or to generate new, simulated sequences drawn from the same probability distribution [48]. In this thesis, I used IGoR to estimate these processes in turquoise killifish, in order to investigate the diversity and composition of the killifish generative repertoire and the changes that take place in the generative process with age.

Preparing sequence databases for IGoR

Before fitting a generative model with IGoR, the pre-processed Change-O database must be further processed to exclude functional sequences and (to the greatest extent possible) sequences produced by affinity maturation rather than primary sequence generation.

To begin with, the pre-processed Change-O database of all sequences in all repertoires for a given experiment is further processed with SHazaM [161] in R to collapse each clone down to a single strict-majority-rule consensus sequence:

```
changeo_db <- changeo_db[!is.na(changeo_db[[clone_field]]),]
```

```
changeo_db_consensus <- collapseClones(changeo_db, method =
   ↪ "mostCommon")
```

In order to assign functional statuses to these new consensus sequences, they were then extracted into FASTA format, assigned V/J identities and CDR3 boundaries with IgBLAST, then re-imported into Change-O database format, after which functional sequences could be identified and discarded:

```
# Extract consensus DB into FASTA Format
ConvertDb fasta -d <consensus_db_path> -o
   ↪ <consensus_fasta_path> --if SEQUENCE_ID --sf SEQUENCE_OUT
   ↪ --mf INDIVIDUAL REPLICATE <other fields>
# Assign VDJ identities etc. with IgBLAST
makeblastdb -parse_seqids -dbtype nucl -in <vh_reference_fasta>
   ↪ -out <vh_db_prefix>;
makeblastdb -parse_seqids -dbtype nucl -in <dh_reference_fasta>
   ↪ -out <dh_db_prefix>;
makeblastdb -parse_seqids -dbtype nucl -in <jh_reference_fasta>
   ↪ -out <jh_db_prefix>;
igblastn -ig_seqtype Ig -domain_system imgt -query
   ↪ <consensus_fasta_path> -out <igblast_output_path>
   ↪ -germline_db_V <vh_db_prefix> -germline_db_D
   ↪ <dh_db_prefix> -germline_db_J <jh_db_prefix>
   ↪ -auxiliary_data <jh_aux_file> -outfmt '7 std qseq sseq
   ↪ btop'
# Re-import into Change-O and split off functional sequences
MakeDb igblast --regions --scores --partial --asis-calls --cdr3
   ↪ -i <igblast_output> -s <consensus_fasta_path> -r
   ↪ <vh_reference_fasta> <dh_reference_fasta>
   ↪ <jh_reference_fasta>
ParseDb split -f FUNCTIONAL -d <new_consensus_db>
```

These operations produced a pooled database of nonfunctional, semi-naïve sequences from all individuals in all sample groups. Before running IGoR on these databases, they then needed to be split by INDIVIDUAL (for individual models) or sample group annotation (for pooled models):

```
ParseDb split -d <nonfunc_consensus_db> -f <split_field>
```

Finally, each split database was then re-extracted into FASTA format for importation by IGoR:

```
ConvertDb fasta -d <split_nonfunc_db> -o <split_nonfunc_fasta>
   ↪ --if SEQUENCE_ID --sf SEQUENCE_INPUT
```

Model inference with IGoR

Following sequence preparation, the model-inference process by IGoR took place in several steps. First, the sequences from the previous section were converted into the correct format:

```
igor -set_wd <working_dir> -read_seqs <split_nonfunc_fasta>
```

These sequences were then aligned to reference V/D/J databases, using V/J "anchor" files to specify the CDR3 boundaries for each V/J identity:

```
igor -set_wd <working_dir> -set_genomic --V
   ↪ <vh_reference_fasta> --D <vh_reference_fasta> --J
   ↪ <vh_reference_fasta> -set_CDR3_anchors --V
   ↪ <v_anchor_file> --J <j_anchor_file> -align --all
```

These alignments could then be used to infer a generative model for the repertoire, by fitting probability distributions for V/D/J-choice, N-insertions, P-insertions, and deletions:

```
igor -set_wd <working_dir> -set_custom_model <default_model>
   ↪ -set_genomic --V <vh_reference_fasta> --D
   ↪ <vh_reference_fasta> --J <vh_reference_fasta>
   ↪ -set_CDR3_anchors --V <v_anchor_file> --J <j_anchor_file>
   ↪ -infer
```

The inferred model was then *evaluated*, to generate parsable parameter and probability-distribution files:

```
igor -set_wd <working_dir> -load_last_inferred -evaluate
   ↪ -output --scenarios 5 --Pgen --coverage VJ_gene
```

Finally, having inferred and evaluated a generative model for each individual and pooled repertoire, the inferred gene-choice, insertion and deletion probability distributions, along with the inferred entropy of each component of the model, were extracted from IGoR's output files with pygor, a Python package supplied as part of the IGoR program [48].

Chapter 3

IGH locus structure and evolution in the Cyprinodontiformes

3.1 Introduction

The native structure of the immunoglobulin heavy chain (*IGH*) locus determines the state space of antibody heavy-chain diversity in a species, including the range of VH, DH and JH segment choices available in VDJ recombination [31], the relationship between VDJ recombination and isotype choice [22], and the ability of processes such as gene conversion [174] and class-switch recombination [45, 175] to affect the diversity and functionality of the antibody repertoire. The diversity produced by VDJ recombination, junctional diversity and secondary diversification processes (Sections 1.2.2 and 1.2.4) in this locus are responsible for the majority of variation in antigen-specificity within a B-cell population, while the choice of isotype among the available *IGH* constant regions determines the antibody's effector function and relationship with the rest of the immune system [6]. Understanding the native structure of the *IGH* locus is therefore essential for understanding how the adaptive immune system functions in a given vertebrate species, while comparing loci between species enables the evolutionary history of adaptive immunity across lineages to be analysed, providing crucial insight into the complex history of this essential biological system. Last but not least, by providing thorough documentation of the *IGH* gene segments present in a species, characterising the *IGH* locus in a species is an essential forerunner to quantitative analysis of adaptive immunity using immunoglobulin sequencing.

Previous work has characterised *IGH* locus structure in a number of teleost species, including zebrafish [136], medaka [25], stickleback [27, 28], rainbow trout [40], fugu [29], and Atlantic salmon [41] (Section 1.2.3). These characterisations have revealed remarkable diversity in the size, structure and functionality of teleost *IGH* loci. However, the number of loci characterised is very small compared to the total evolutionary diversity of teleost fish, and is mainly confined to major aquaculture species (trout, catfish, salmon) or research models (zebrafish, stickleback, medaka) [7, 22], with characterised species often quite distantly related to one another within the teleost clade. This relatively sparse sampling of teleost *IGH* loci has left wide swathes of teleost diversity without any characterised *IGH* loci, and has prevented higher-resolution analysis of locus structural evolution across closely related species.

In this chapter, therefore, I present complete characterisations of the *IGH* loci of two important model organisms from the Cyprinodontiformes, a diverse clade of primarily freshwater teleost fishes with no previously-characterised *IGH* loci. *Nothobranchius furzeri*, the turquoise killifish (Section 1.4), has recently emerged as an important model system for ageing research [77, 176], and is also of evolutionary and ecological interest due to its short lifespan, extreme natural environment, and unusual life history [177]. The southern platyfish *Xiphophorus maculatus*, meanwhile, is an important model organism in evolutionary ecology and population genetics [178]. Comparing the *IGH* loci of these two closely-related species (Sections 3.2 and 3.3) reveals dramatic and unexpected differences in immune structure, which when combined with information from previously-published loci from related species suggest unexpected patterns of locus evolution within this group of teleost fishes. Most

strikingly, the specialised mucosal antibody isotype *IGHZ* appears to have been convergently lost in multiple closely-related lineages.

To further investigate the history of *IGHZ* and other surprising features of *IGH* locus evolution in the Cyprinodontiformes, I performed a partial reconstruction and analysis of the *IGH* loci from ten further cyprinodontiform species (Figure 3.1), as well as from a new and improved genome assembly of medaka (*Oryzias latipes*), with a focus on the constant-region exons present in each species (Section 3.4). Phylogenetic analysis confirms the repeated independent loss of *IGHZ* in this lineage and provides evidence for multiple, independent *IGHZ* subclasses present ancestrally in the clade. Taken together, this analysis significantly extends our knowledge of constant-region diversity in teleost fish, and establishes the cyprinodontiforms, and especially the African killifishes, as a highly promising group of model systems for comparative evolutionary immunology.

3.2 The *IGH* locus of *Nothobranchius furzeri*

3.2.1 Assembling the *N. furzeri IGH* locus

In order to locate and characterise the *Nothobranchius furzeri IGH* locus, I collated databases of VH, JH, and CH exon sequences from the published locus sequences of three reference species (zebrafish [136], three-spined stickleback [27, 28], and medaka [25], Section 2.3.3.1) and aligned them to the *Nothobranchius furzeri* genome (NFZ V2.0 [89]) with BLAST [132, 133]. Genome scaffolds with high-confidence alignments to at least two distinct segment types or covering at least 1 % of the scaffold's total length were retained for downstream analysis as potential locus candidates (Section 2.3.3.2); in total, one chromosome (chr6) and 6 unincorporated scaffolds were identified in this way (Table 3.1), with chromosome 6 bearing the majority of identified gene segments.

Table 3.1: *N. furzeri* genome scaffolds containing putative *IGH* locus fragments

Scaffold	Total length (kb)	V	J	C_μ	C_δ	C_ζ	Included in locus?
chr6	6195.6	15	7	5	11	0	Yes
scf10901	1.4	0	0	0	3	0	Yes
scf21863	13.5	1	0	0	0	0	No
scf35954	16.3	3	0	0	0	0	No
scf36277	18.9	2	1	0	0	0	No
scf37083	17.7	1	0	0	0	0	No
scf9157	7.2	0	7	4	0	0	Yes

In order to determine which of the putative locus scaffolds were in fact part of the *IGH* locus, integrate these into a contiguous locus sequence, and provide additional information on any missing gene segments, I supplemented the locus assembly with insert sequences from bacterial artificial

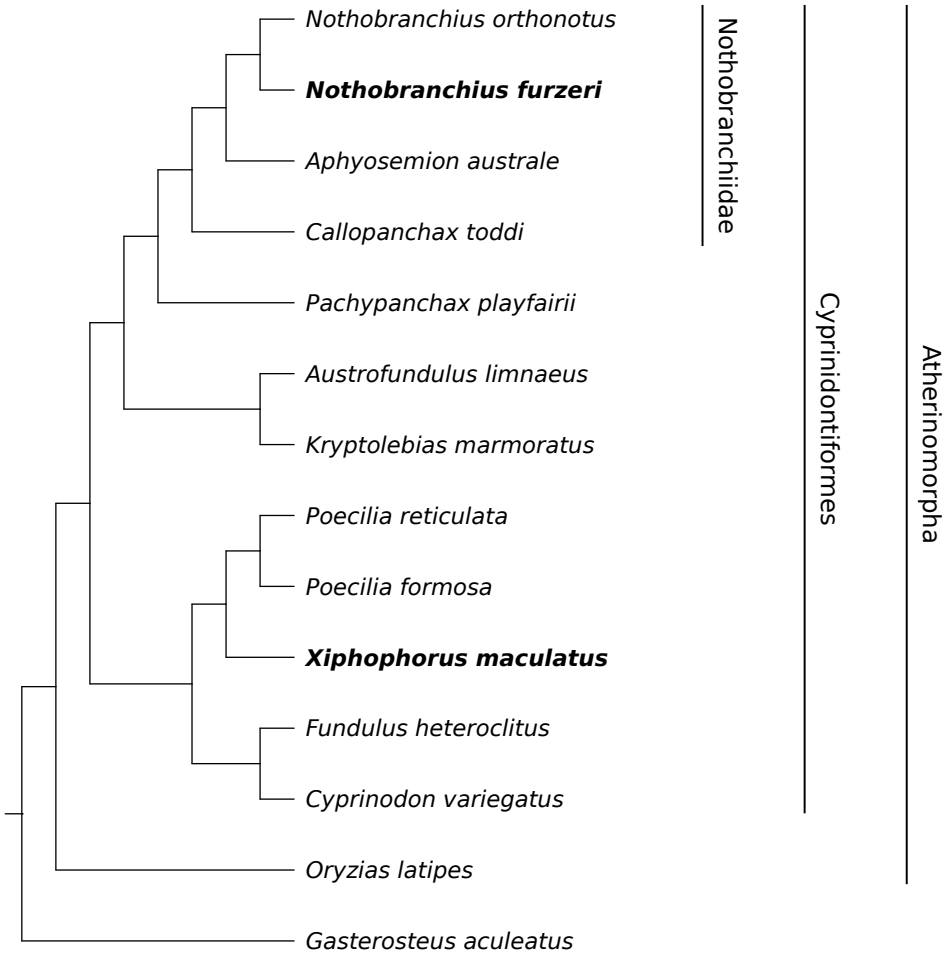


Figure 3.1: Cladogram of species included in the *IGH* locus analysis: Boldface type indicates species for which new, complete *IGH* locus assemblies were generated for this study; other species were either previously-characterised reference species (*G. aculeatus*, *O. latipes*) or underwent constant-region characterisation only (all other species). Labelled vertical bars designate higher taxa of interest.

chromosomes in the turquoise-killifish genomic BAC library [88]. BAC candidates (whose ends had already been sequenced as part of the genome project) were identified as potentially containing part of the locus sequence (Section 2.3.2.1) on the basis of their ends aligning to promising candidate scaffolds from a previous genome assembly (first round) or to the insert sequences of previously sequenced BAC inserts (second round). Once identified, BAC candidates were isolated from culture by alkaline lysis (Section 2.2.3), sequenced on an Illumina MiSeq sequencing machine, and assembled and scaffolded with SPAdes and SSPACE, respectively (Sections 2.3.2.2 and 2.3.2.3). Complete BAC insert assemblies were generated from these scaffolds by manual alignment to overlapping genome scaffolds and other BAC inserts, combined with PCR and Sanger sequencing of intervening sequences.

Having obtained complete insert sequences for promising BAC candidates, I screened them for *IGH* locus segments in the same manner described for genome scaffolds. Passing insert sequences (Table 3.2) were aligned to and integrated with the identified candidate scaffolds to produce a contiguous locus assembly (Section 2.3.3.3). To minimise the probability of losing relevant gene segments to assembly errors, I gave priority in the event of a sequence conflict between BACs and scaffolds first to any sequence containing a segment missing from the other; if neither the BAC assembly nor the genome scaffolds met this condition, I prioritised the genome scaffold over the BAC assembly. In total, 3 candidate scaffolds (including chromosome 6) and 5 BAC inserts were included in the final locus assembly, while 4 scaffolds and 6 BACs were excluded as likely representing isolated *IGH* orphon segments elsewhere in the genome. The regions of the final locus sequence contributed by different BACs and scaffolds are shown in Figure 3.2.

3.2.2 Overall locus structure

The turquoise-killifish genome contains a single *IGH* locus approximately 306 kilobases in length, located on chromosome 6 of the *N. furzeri* genome (Figure 3.3A). This locus comprises two complete subloci, *IGH1* (155 kb) and *IGH2* (118 kb), present in tandem and each occupying a classic VH-DH-JH-CH translocon configuration. This modified translocon structure, with multiple translocon subloci present in tandem, has been observed in a number of teleost *IGH* loci including catfish, medaka and stickleback (Section 1.2.3) [22]. Unusually, however, the smaller *IGH2* sublocus in *Nothobranchius furzeri IGH* is present in antisense relative to the larger *IGH1*, with the two subloci beginning at opposite ends of the locus and extending in opposite sense towards the middle (Figure 3.3B). Such a multi-orientation structure has been seen only rarely in previously-characterised teleost *IGH* loci; to my knowledge, it has only been previously observed in medaka [25] and Atlantic salmon (Figure 1.3B). As medaka is the closest relative of the turquoise killifish to have its locus characterised prior to the

Table 3.2: *N. furzeri* BAC-library inserts containing putative *IGH* locus fragments

BAC ID	Insert length (kb)	V	J	C_μ	C_δ	C_ζ	Included in locus?
154G24	106.6	17	1	0	0	0	No
162F04	119.4	5	1	0	0	0	No
165M01	110.7	15	1	0	0	0	Yes
206K13	106.7	17	1	0	0	0	No
208A08	103.2	17	1	0	0	0	Yes
209K12	133.0	1	8	4	20	0	Yes
220O06	104.8	4	1	0	0	0	No
223M21	99.3	17	1	0	0	0	No
248A22	47.3	7	0	0	0	0	No
276N03	127.9	7	0	0	0	0	Yes
277J10	120.8	17	1	0	0	0	Yes

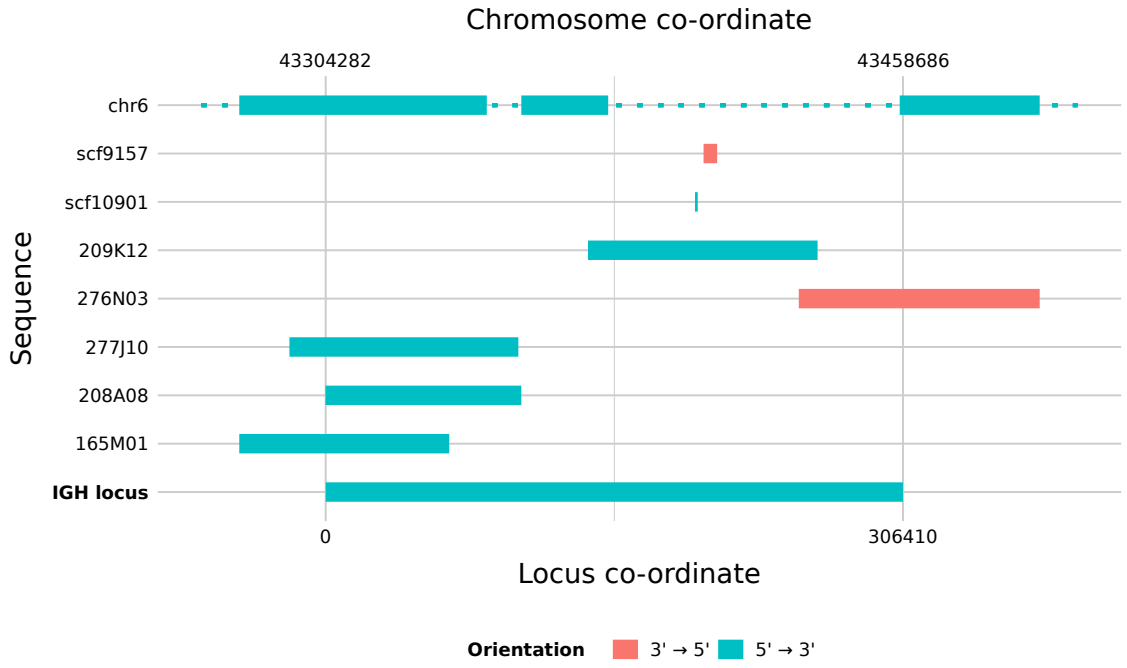


Figure 3.2: Assembling the *Nothobranchius furzeri IGH* locus: Schematic of genome scaffolds and BAC inserts contributing to the *N. furzeri IGH* locus sequence, with their corresponding place within the locus sequence (bottom axis). Internal gaps with dotted lines indicate regions on chromosome 16 with no corresponding locus sequence, as a result of intercalation of BAC or scaffold sequences.

present study, it is interesting to see this unusual feature reproduced here, raising the question of whether this ideosyncracy is homologous between the two loci.

Compared to other closely-related loci, the turquoise-killifish locus is relatively sparse, with comparatively low functional complexity given its overall size. For example, whereas the stickleback locus fits four subloci, 49 VH segments and 10 constant regions into c. 200 kb [27, 28], the turquoise-killifish locus, despite being roughly 50 % longer, contains only two subloci, four constant regions and 24 VH segments (including pseudogenised VHs). This difference is a consequence of the unusually large amount of nonfunctional sequence padding the turquoise-killifish locus, resulting in large gaps between variable segments and in some cases between constant-region exons (Figure 3.3B); this high prevalence of repetitive DNA is consistent with the rest of the turquoise-killifish genome, which comprises more than 60 % repetitive sequence [89], compared to just over 15 % in stickleback [179].

The two subloci in the turquoise-killifish locus are generally highly similar in their functional sequence, with a high degree of synteny between their functional regions (Figure 3.4). The greatest degree of divergence occurs in the VH and DH regions, with what appear to be repeated deletion events in *IGH2* resulting in a substantially lower number of VH and DH segments compared to *IGH1*;

conversely, the JH and constant regions are almost identical between two subloci. These patterns are discussed in more detail in Sections 3.2.3 and 3.2.4.

3.2.3 Constant regions

The isotype of an antibody determines its functional role within the immune system, including its possible effector functions and whether it can be secreted (Section 1.2.1) [6]. Three antibody isotypes have been identified to date in teleost fishes: *IGHM*, *IGHD* and *IGHZ* (a.k.a. *IGHT*, *IGHT/Z* or *IGHZ/T*) [7, 22, 180]. Of these, *IGHM* and *IGHD* are highly primitive within the jawed vertebrates and found in most or all other vertebrate groups; within the teleosts, both appear to be universal [7]. Conversely, *IGHZ* is a teleost-specific isotype which is absent in other vertebrate taxa; within the teleosts, most characterised *IGH* loci possess *IGHZ*, but at least two (medaka and channel catfish) have been found to lack it [7, 22]. In rainbow trout, *IGHZ* has been found to play a specialised mucosal role in the immune system analagous to that of *IGHA* in mammals [20, 26], and it is widely assumed to play this specialised role throughout the teleosts; it is as yet unclear how mucosal immunity is effected in species lacking *IGHZ* (Section 1.2.1).

In order to investigate constant regions in the *Nothobranchius furzeri IGH* locus, I identified putative exon sequences using BLAST alignments to the reference sequence databases described in Section 3.2.1. Intron/exon bounderies were refined through alignment of published RNA-sequencing data from turquoise-killifish gut ([78], BioProject accession PRJNA379208, young and old untreated groups, Table E.2) using STAR (Section 2.3.3.4 and Figure 3.6). Strikingly, the *Nothobranchius furzeri IGH* locus appears to completely lack any *IGHZ* constant region, with no C_ζ exons or *IGHZ* transmembrane exons being found on either *IGH1* or *IGH2*. Given the widespread prevalence and specialised mucosal role of *IGHZ* in teleosts, its surprising absence in turquoise killifish (Figure 3.3B) immediately raises questions about the nature, kinetics and efficacy of mucosal adaptive immunity in this species. The similar absence of *IGHZ* in medaka, which again is the closest relative of *N. furzeri* with a characterised locus, raises further questions about the evolutionary history of *IGHZ* in these species: does the shared absence of *IGHZ* in these species indicate a single ancestral deletion event, or parallel loss of this important isoform within both the Cyprinodontiformes (including the turquoise killifish) and the lineage leading to medaka? This latter question requires higher phylogenetic resolution to address effectively, and is investigated further in Section 3.3 and Section 3.4.

While *IGHZ* is completely missing from the *N. furzeri IGH* locus, *IGHM*, the most primitive and widely-found isotype in jawed vertebrates, is present in its expected location immediately downstream of the main JH-region in both subloci. This constant region occupies the standard six-exon configuration, with four C_μ exons and two transmembrane exons present in series on the chromosome (Figures 3.3B, 3.3C and 3.5A, Table E.3). As with other species, both secreted and transmembrane isoforms of *IGHM* are present in the transcriptome, with secreted *IGHM* (*IGHM-S*) consisting of C_μ1-4 (Figures 3.3C, 3.5B and 3.6A); however, the exon configuration of transmembrane

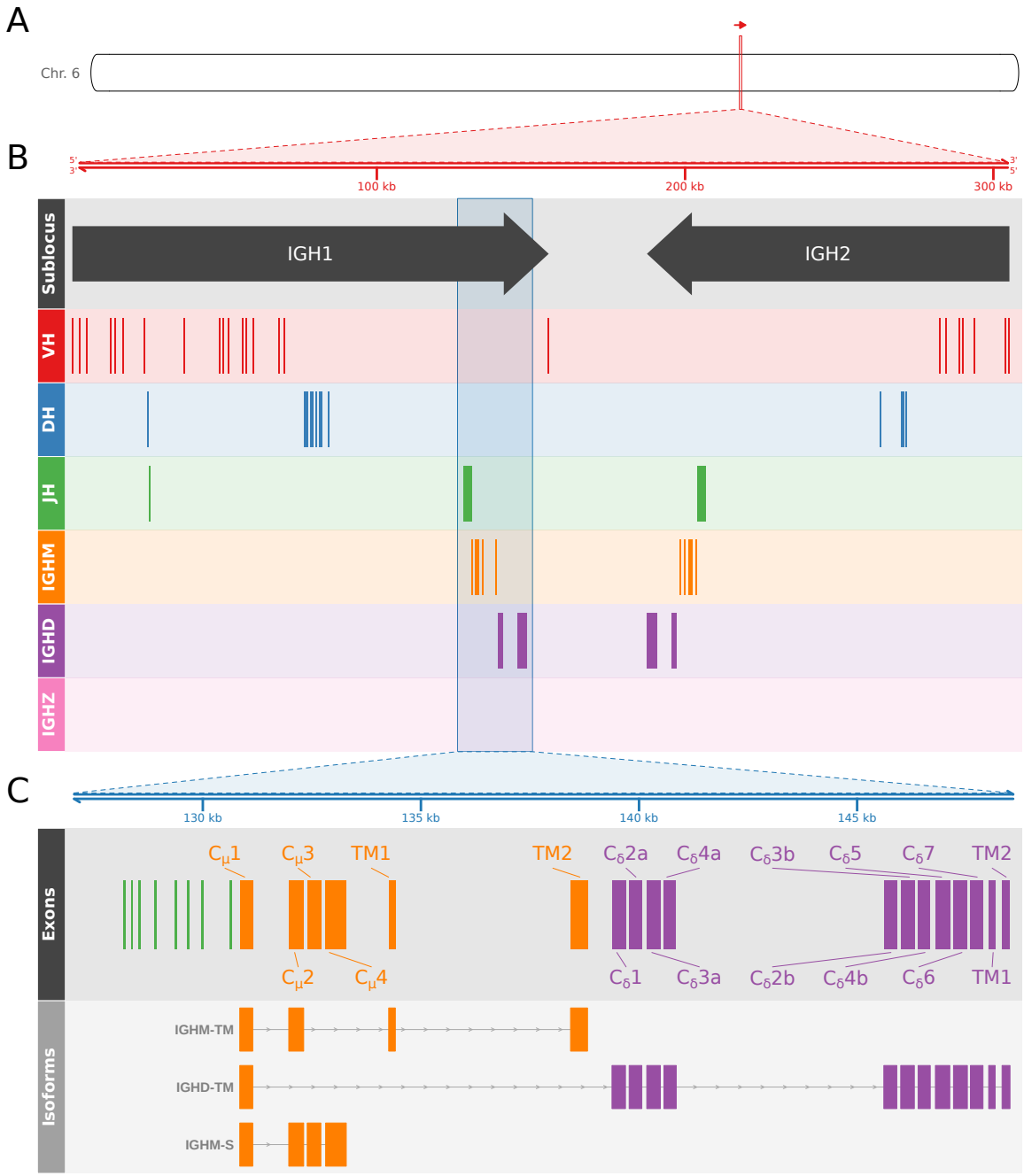


Figure 3.3: The immunoglobulin heavy chain (*IGH*) locus in *Nothobranchius furzeri*: (A) Position of the *IGH* locus on chromosome 6 of the *N. furzeri* genome. (B) Arrangement of VH, DH, JH and constant-region gene segments on the *N. furzeri IGH* locus. All segments follow the orientation of their parent sublocus, indicated in the uppermost track. (C) Detailed map of the constant regions of the *IGH1* sublocus, indicating the position and identity of the constant-region exons and the exon composition of expressed *IGH* isoforms in the turquoise killifish.

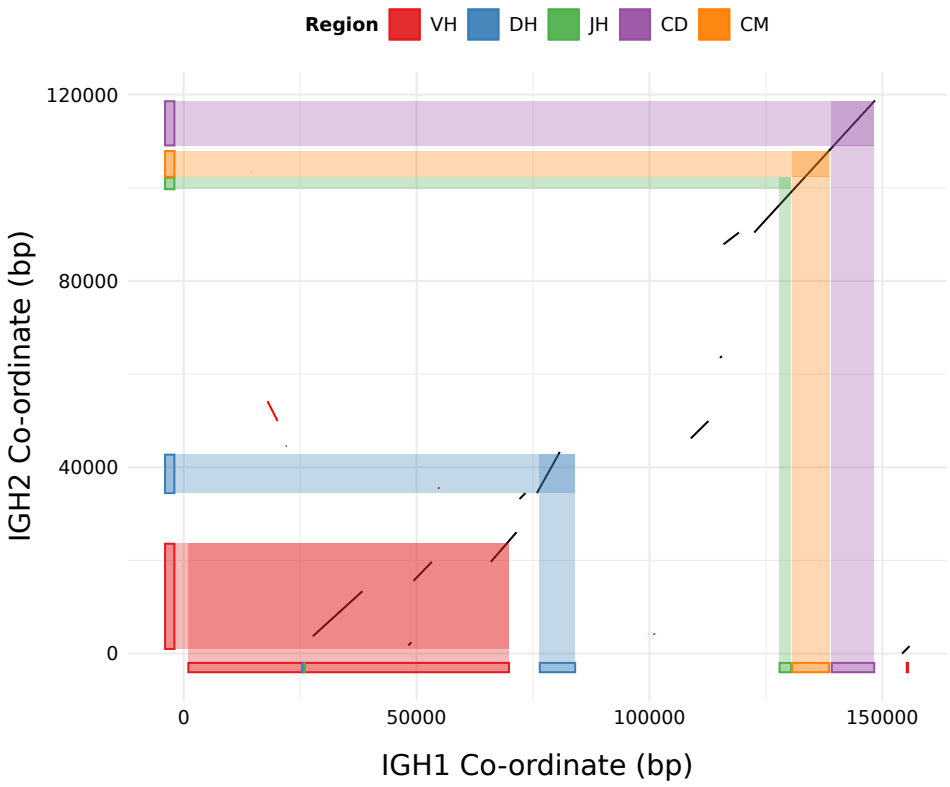


Figure 3.4: Sequence homology between subloci in *N. furzeri IGH*: Synteny dot plot of sequential best matches between *IGH1* and *IGH2* subloci, with gene segment regions indicated by coloured rectangles along each axis.

IGHM (*IGHM-TM*) deviates from both that seen in mammals (in which exon TM1 is spliced to a cryptic splice site within $C_\mu 4$) and most teleosts (in which the canonical splice site following $C_\mu 3$ is used and $C_\mu 4$ is excised) [22]. Rather, turquoise-killifish *IGHM-TM* resembles that of medaka, in which both $C_\mu 3$ and $C_\mu 4$ are excluded and the canonical splice site at the end of $C_\mu 2$ is spliced directly to TM1 (Figures 3.5C to 3.5E). This similarity to medaka again raises the possibility that this unusual feature may be a conserved feature of both lineages; however, the underlying mechanism giving rise to this difference in splicing behaviour is unknown.

Unlike *IGHM*, the exon structure of *IGHD* is highly variable across the teleosts, ranging from roughly 7-17 C_δ exons in addition to the transmembrane domains [22]. The core structure of *IGHD* comprises seven C_δ exons ($C_\delta 1$-7), but some subset of these exons may be missing or duplicated in any given species – in medaka, for example, $C_\delta 5$ is missing in all subloci [25], while in many species (e.g. zebrafish, salmon, and channel catfish) $C_\delta 2$-4 are duplicated in two or more tandem blocks [22]. This latter configuration is also observed in turquoise killifish, in which the *IGHD* constant region immediately follows *IGHM* in both subloci and has a

$C_\delta 1$-$(C_\delta 2$-$C_\delta 3$-$C_\delta 4)_2$-$C_\delta 5$-$C_\delta 6$-$C_\delta 7$-TM1-TM2

configuration, for a total of 12 exons per *IGHD* constant region (Figures 3.3B and 3.3C, Table E.3). All of these exons appear to be expressed in tandem, resulting in a much longer transcript than is observed for any isoform of *IGHM* (Figures 3.3C and 3.6B). As in other teleost species, *IGHD* in the turquoise killifish includes a chimeric $C_{\mu}1$ exon at the 5' end of the constant-region transcript, for a total of 13 exons per *IGHD-TM* mRNA (Figure 3.6B).

While the best-known form of *IGHD* in teleosts is transmembrane, secreted *IGHD* has been observed in at least two teleost species, with different mechanisms used in each case: in channel catfish, one dedicated sublocus has a dedicated *IGHD* secretory exon in place of the transmembrane exons [23], while in rainbow trout (and possibly some other species like Atlantic salmon and cod [24]) a run-on event at the end of $C_{\delta}7$ results in the production of a secretory tail in a manner similar to secretory *IGHZ* [24]. However, neither a specialised secretory exon nor a $C_{\delta}7$ secretory tail could be detected in turquoise killifish, suggesting that *IGHD* may only be expressed in transmembrane form in this species.

In the case of both *IGHM* and *IGHD*, the constant regions are present in their expected configuration in each sublocus and are highly similar in sequence between the subloci, with an average of 98.4 % nucleotide sequence identity for corresponding *IGHM* exons and 99.3 % for corresponding *IGHD* exons (Figure 3.7 and Table 3.3) in pairwise Needleman-Wunsch alignments [112]. This high level of similarity indicates either a very recent duplication event to produce the second sublocus or a high level of sequence conservation in both subloci, with the latter explanation suggesting that both subloci continue to be functional and active in the immune system.

3.2.4 Variable regions

Variable-region gene segments in the turquoise-killifish *IGH* locus were identified with a variety of methods, depending on the type of gene segment being analysed (Section 2.3.3.4). VH candidates were identified probabilistically using Hidden Markov Models constructed by nhmmer [139] from PRANK [138] multiple-sequence alignments of reference sequences, with the 3'-ends of each V-exon identified by the presence of a recombination-signal sequence (RSS) [6] and the 5'-ends refined using IMGT-DomainGapAlign [143]. JH candidates were also identified using nhmmer, with segment ends identified by the presence of an RSS (5') and a `GTA` splice-site motif (3') [25]. Finally, DH-segments, being too short and variable in sequence for HMM-based approaches to be effective, were identified by searching for pairs of flanking RSS sequences in opposite orientation, using fuzzy pattern-matching (with EMBOSS FUZZNUC [146]) to conserved RSS sequence motifs.

In total, I identified 24 VH-segments, 14 DH-segments and 17 JH-segments in the *N. furzeri* locus (Tables E.4, E.5 and E.8), of which the majority (17 VH, 10 DH and 8 JH) were present in *IGH1*. Of the VH segments identified, three contain premature stop codons, though none is out-of-frame; conversely, all the DH and JH segments identified appear to be in-frame and functional, with no

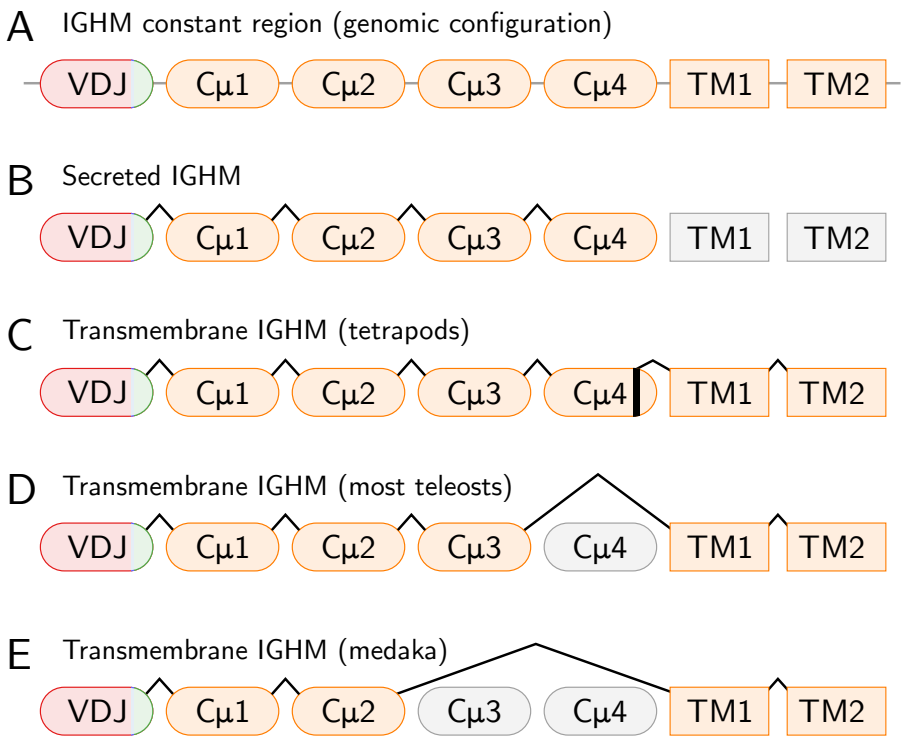


Figure 3.5: *IGHM* exon usage in bony vertebrates: Schematic of *IGHM* splice patterns in different isoforms and taxonomic groups; (A) standard genomic (pre-splicing) configuration of *IGHM*, following VDJ recombination; (B) exon configuration of secreted *IGHM* (*IGHM-S*) in tetrapods and teleosts; (C) exon configuration of transmembrane *IGHM* (*IGHM-TM*) in tetrapods, demonstrating the use of a cryptic splice site in $C_\mu 4$; (D) standard *IGHM-TM* exon configuration in teleosts, demonstrating the direct splicing of $C_\mu 3$ to TM1 and exclusion of $C_\mu 4$; (E) unusual *IGHM-TM* exon configuration observed in medaka, in which both $C_\mu 3$ and $C_\mu 4$ are excluded. Adapted from Figure 3 of [22].

premature stop codons. However, in all cases a minority of segments contain RSS sequences that deviate significantly from the expected consensus sequence (Tables E.4, E.6, E.7 and E.9); it is unclear whether these sequences can recombine to successfully produce mature VDJ sequences *in vivo*. In the case of the VH segments, of the six sequences without clearly functional RSS sequences, three also contain premature stop codons, suggesting the changes to the RSS in these cases may arise from relaxed purifying selection on already-pseudogenised sequences.

Apart from these few exceptions, however, the recombination signal sequences (RSS) marking the ends of the VH, DH and JH gene segments in the *N. furzeri* locus otherwise strongly resemble those of other characterised teleosts, which in turn resemble those of non-teleost loci (Figures 3.10 and D.1). The overall heptamer and nonamer consensus sequences (`CACAGTG` for heptamers and `ACAAAAACC` for nonamers) closely matched those expected from the literature [6], while in 88 % of cases the spacer region was within 1bp of the expected length (12bp for DH-RSSs, 23bp for VH- and JH-RSSs); interestingly, the greatest number of VH-RSSs had a 22bp (rather than 23bp) spacer, but

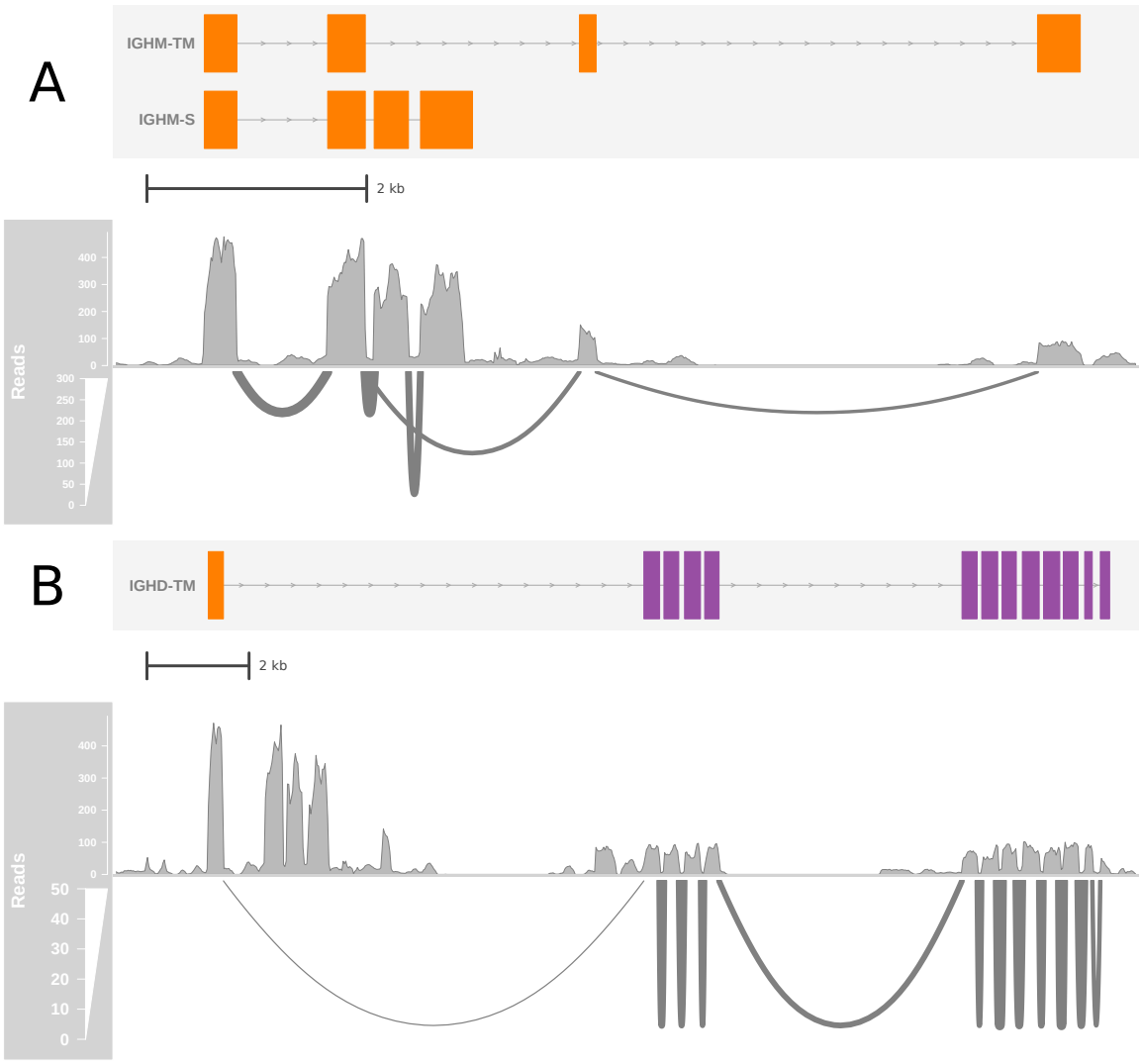


Figure 3.6: Constant-region isoforms in *N. furzeri*: Coverage and Sashimi plots [122] of STAR-aligned RNA-seq reads from *N. furzeri* gut samples [78], demonstrating the splicing behaviour of *IGH1* constant-region isoforms and showing the read coverage of each exon and splice junction. (A) *IGHM* exon splicing, showing alternative splicing patterns of *IGHM-TM* and *IGHM-S*; (B) *IGHD* exon splicing, showing chimeric splicing of $C_{\mu}1$ to $C_{\delta}1$.

this is unlikely to interfere with RSS functionality. Overall, the RSSs in the turquoise killifish appear to be supporting the normal operation of VDJ-recombination in this species.

Of the VH, DH and JH segments identified, all but one of each type of segment is located within contiguous V-, D-, and J-regions within each sublocus, supporting a modified translocon configuration for turquoise-killifish *IGH*. The exceptions to this are *IGH1D01* and *IGH1J01*, which are embedded within the *IGH1* V-region, and a single VH segment located in between the *IGHD* constant regions of the two subloci (Figure 3.3B). The unusual location of *IGH1D01* and *IGH1J01* may represent the result of a transposition event within the *IGH* locus; however, their close colocalisation and 5' position

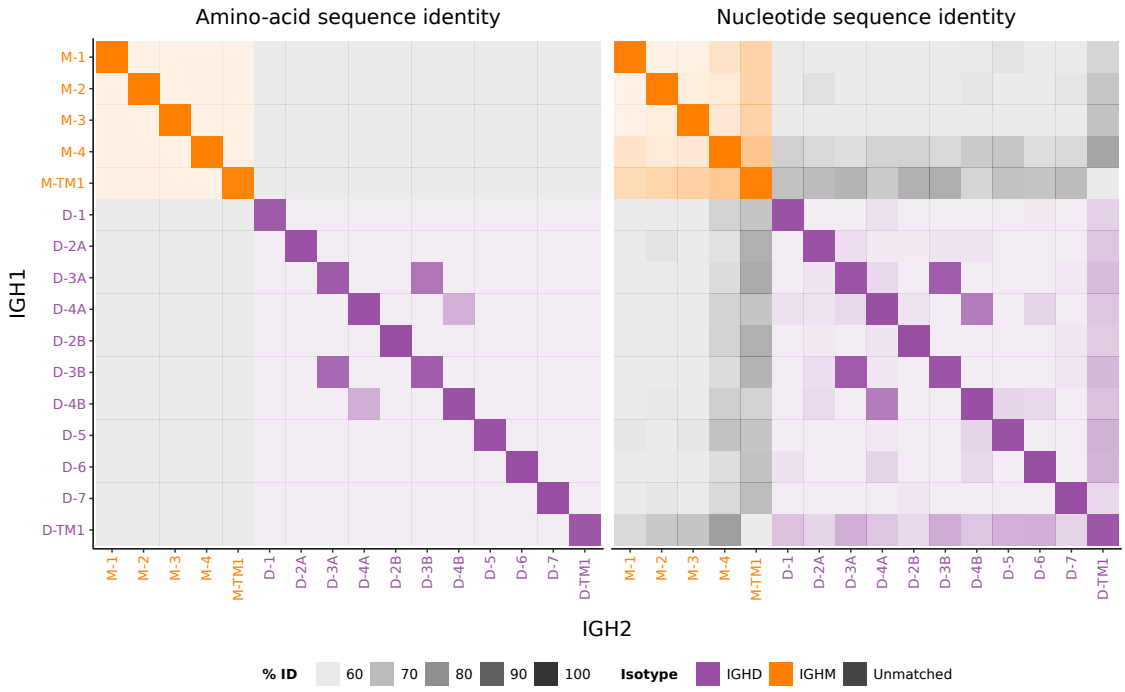


Figure 3.7: Cross-sublocus sequence similarity of constant-region exons in *N. furzeri IGH*: Heatmap of percentage sequence identity between amino-acid (left) and nucleotide (right) sequences of constant-region exons (excluding *IGHM-TM2* and *IGHD-TM2*) from the two subloci of *N. furzeri IGH*, calculated using pairwise Needleman-Wunsch global alignments.

within the *IGH1* sublocus, as well as the fact that neither has a close paralogue in *IGH2* (Figure 3.9B), suggest that they may instead represent the remnant of a formerly present *IGHZ* constant region, as these typically have dedicated D/J segments independent of those serving *IGHM* (Section 1.2.3). Given its forward orientation, meanwhile, I assigned the orphaned VH-segment to the *IGH1* sublocus as *IGH1V1-07*; however, if annotated correctly, it is unlikely to successfully recombine with segments in either sublocus due to its unusual location.

VH sequences within an *IGH* locus are conventionally grouped into families on the basis of nucleotide sequence identity, with a typical identity cutoff of 80 % [180]. In order to group the *N. furzeri* VH genes into families, I performed pairwise Needleman-Wunsch global alignments on each pair of VH sequences to obtain pairwise identity scores, followed by single-linkage clustering on the resulting identity matrix (Section 2.3.3.4). Cutting the dendrogram at 80 % sequence identity revealed a total of six VH families, of which four contained more than one VH segment (Figure 3.8); this number of VH families in the *N. furzeri* locus is roughly in line with those found in related species (Table 3.4). Of these, V1 and V2 make up the bulk (42 % and 29 % respectively) of the VH segments in the locus. V2 and V4 are highly similar, and all the members of V4 are pseudogenised by premature

Table 3.3: Cross-sublocus sequence similarity of constant-region exons in *N. furzeri*: Percentage sequence identities of pairwise Needleman-Wunsch global alignments between nucleotide (NT) or amino-acid (AA) sequences of corresponding exons from the two subloci of *N. furzeri IGH*.

Isotype	Exon	NT	AA
M	1	99.66	100.00
M	2	100.00	100.00
M	3	100.00	100.00
M	4	100.00	100.00
M	TM1	99.34	98.00
M	TM2	91.67	100.00
D	1	99.03	97.06
D	2A	98.97	98.96
D	3A	98.72	97.09
D	4A	99.65	98.92
D	2B	100.00	100.00
D	3B	98.72	96.12
D	4B	99.64	98.91
D	5	99.09	99.08
D	6	100.00	100.00
D	7	100.00	100.00
D	TM1	97.99	97.96
D	TM2	99.44	100.00

stop codons; it may therefore be more appropriate to regard V4 as a pseudogenised subfamily of V2 than as a VH family in its own right.

The total number of functional VH segments in the turquoise-killifish locus is unusually small in comparison to the total numbers observed in many other teleost species (Table 3.4); however, the number of segments per sublocus is in line with the numbers seen in closely-related species (2 to 12 in medaka [25], 6 to 18 in stickleback [27, 28]), with the overall difference mainly arising from a difference in the number of subloci per locus. A similar pattern is observed with JH segments, with similar numbers of segments per sublocus in turquoise killifish and closely-related species, especially medaka. It therefore appears that the per-sublocus segment diversity available to the turquoise killifish is similar to that of previously characterised species, with any difference in total available diversity at this level arising from differences in the number of functional subloci rather than the size of the V/D/J-regions *per se*.

As can be seen from Figure 3.4, much of the V-, D- and especially J-region sequence in the *N. furzeri* locus is syntenic between the two *IGH* subloci, with downstream portions of the *IGH2* V-region corresponding to downstream parts of the *IGH1* region. Of the seven VH segments in *IGH2*, six have a corresponding segment on *IGH1* with which they share at least 97 % sequence identity (Figure 3.8), and these partner segments are largely (though not entirely) colinear in their ordering

between the two subloci. A similar pattern can be observed for the D- and (especially) the J-regions: of the four DH segments detectable in *IGH2*, three (*IGH2D02* to *IGH2D04*) are identical with another block of adjacent DH segments in *IGH1* (*IGH1D05* to *IGH1D07*), while the JH-regions exhibit almost complete sequence identity between the eight JH segments of the main JH region in *IGH1* and the eight JH segments in *IGH2* (Figure 3.9).

Nevertheless, as is clear from Figure 3.4, there are large portions the *IGH1* variable region, including the first 25 kilobases of the V-region, for which no corresponding sequence exists in *IGH2*, and there are many VH and DH segments in *IGH1* (and a much smaller number in *IGH2*) for which no close homologue exists in the other sublocus. Taken together, these data are consistent with a model in which *IGH2* was produced via duplication and inversion of all or part of *IGH1*, followed by subsequent deletion events in the redundant, and structurally volatile, *IGH2* VH and DH regions. However, it is not clear at present how to distinguish between this model and an alternative one of expansion in *IGH1*, or how to explain why the JH region is so much more conserved between subloci than either the VH or DH regions.

Table 3.4: Number of functional VH-segments and VH-families in other teleost species

Common Name	Species	# Functional VH Segments	# VH Families	Source
Zebrafish	*Danio rerio*	39	13 [1]	[180]
Grass carp	*Ctenopharyngodon idella*	8	5 [2]	[42]
Fugu	*Takifugu rubripes*	34	3	[180]
Medaka	*Oryzias latipes*	35	6	[22, 25]
Stickleback	*Gasterosteus aculeatus*	49	4	[180]
Turquoise killifish	*Nothobranchius furzeri*	21 [3]	6	–

[1] VH families in zebrafish were identified based on 70 % (rather than 80 %) sequence identity.

[2] It is not clear what clustering method or threshold was used to identify VH families in grass carp.

[3] Excluding VH segments with nonsense or frameshift mutations, but not those with uncertain or missing RSS sequences.

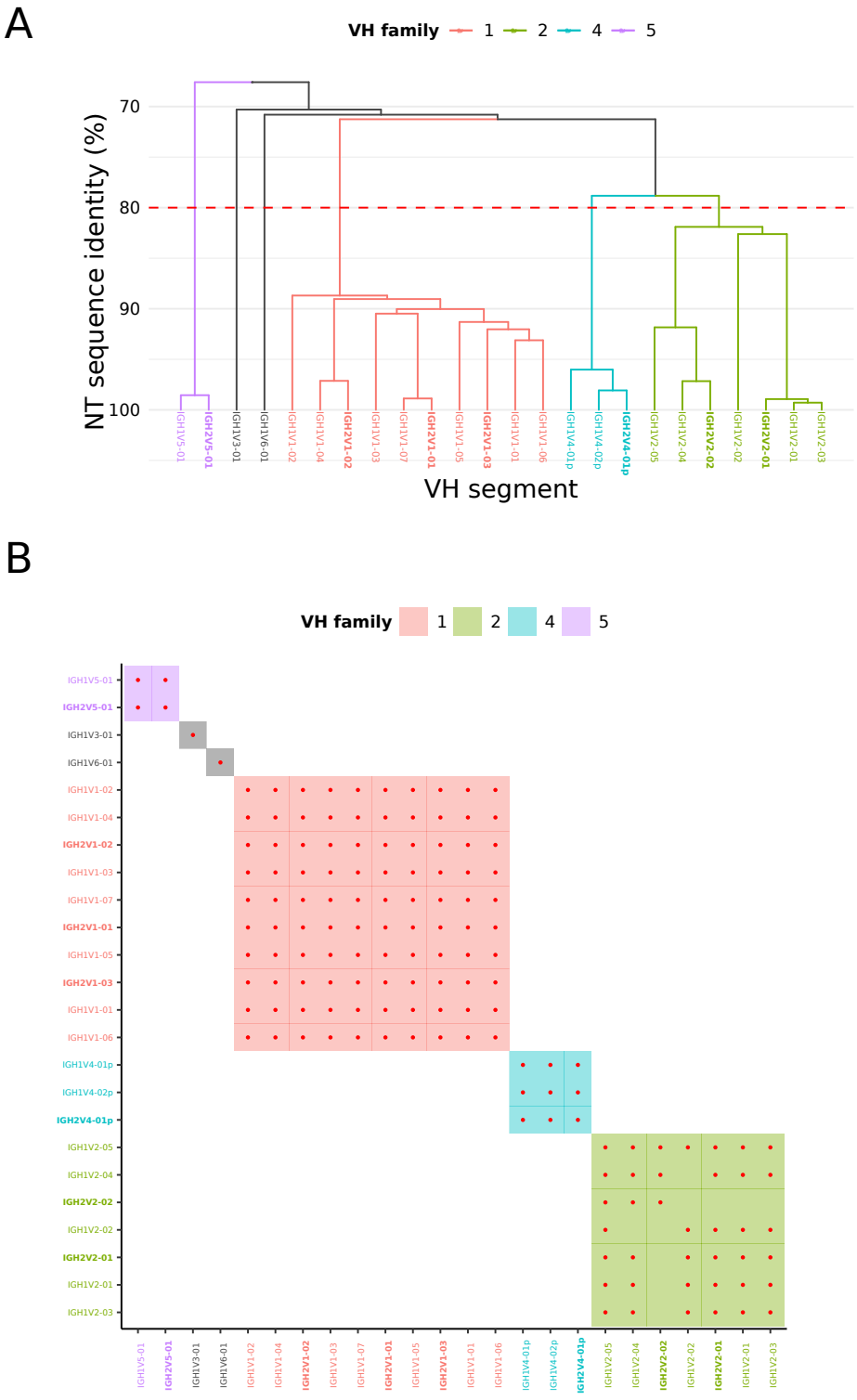


Figure 3.8: VH families in the *N. furzeri IGH* locus: (A) Dendrogram of sequence similarity of VH segments in the *N. furzeri IGH* locus, arranged by single-linkage clustering on nucleotide sequence identity. The red line indicates the 80 % cutoff point for family assignment. (B) Heatmap of family relationships among *N. furzeri* VH segments, with shaded squares indicating families and red dots indicating pairwise nucleotide sequence identity of at least 80 %. In both subfigures, VH families containing multiple segments are uniquely coloured, single-segment families are in grey, and segments from the *IGH2* sublocus are displayed in boldface.

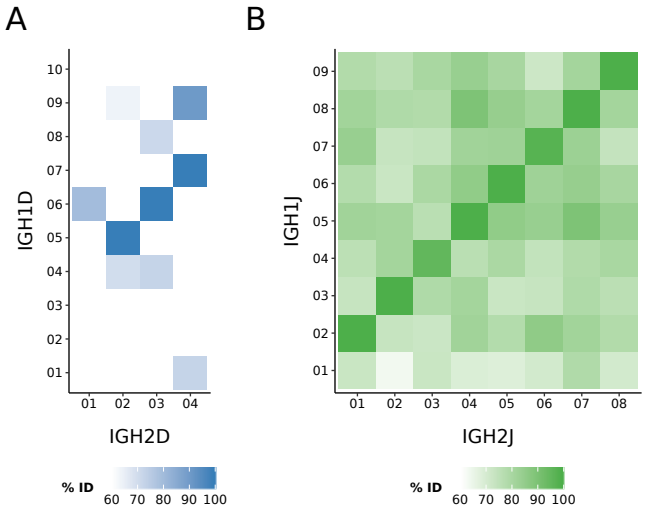


Figure 3.9: Cross-sublocus sequence similarity of DH and JH gene segments in *N. furzeri IGH*: Heatmap of percentage nucleotide sequence identities of Needleman-Wunsch global alignments between (A) DH and (B) JH gene segments in *IGH1* vs *IGH2*, revealing syntenic runs of highly similar sequences across both subloci.

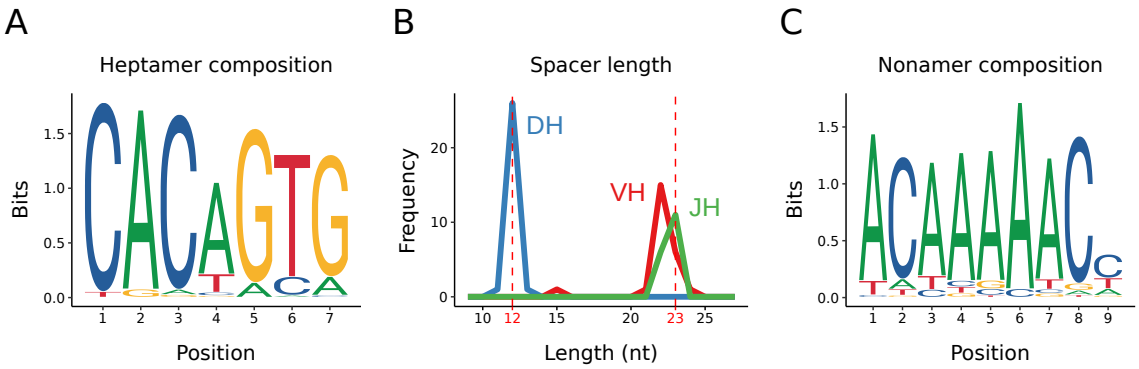


Figure 3.10: Recombination signal sequences in *N. furzeri IGH*: (A) Sequence composition of conserved heptamer sequences across all *N. furzeri* heavy-chain RSSs; (B) length distribution of unconserved spacer sequences in *N. furzeri* heavy-chain RSSs; (C) sequence composition of conserved heptamer sequences across all *N. furzeri* heavy-chain RSSs.

3.3 The *IGH* locus in *Xiphophorus maculatus*

The turquoise-killifish *IGH* locus shares many features with other characterised teleost loci, including a modified tandem-translocon configuration with intact VH, DH, JH and constant regions (Figure 3.3B), a four-exon secreted configuration of *IGHM* (Figure 3.6A), an expanded *IGHD* constant region with tandem C_δ-exon block repeats (Figures 3.3B and 3.6B,), a conserved RSS structure (Figure 3.10), and a chimeric $C_\mu 1$ in *IGHD* (Figure 3.6B). However, it also exhibits many ideosyncratic features that differ from those observed in most characterised teleost loci, including an unusually small number of VH segments (Figure 3.8 and Table 3.4), a four-exon $C_\mu 1$-$C_\mu 2$-TM1-TM2 configuration of transmembrane *IGHM* (Figure 3.6A), an inverted sublocus present in antisense (Figure 3.3B), and a complete absence of *IGHZ*.

Many of these peculiarities, including the unusual *IGHM-TM* splicing pattern, inverted sublocus, and lack of *IGHZ*, are shared with the *IGH* locus of medaka (*Oryzias latipes*), which is the closest relative of *Nothobranchius furzeri* to have its immunoglobulin heavy chain locus characterised prior to this study [25]. Given the close relationship between the two species, the shared unusual features of their *IGH* loci suggested a common origin of these traits in the common ancestor of both species. If this hypothesis were correct, one would expect *IGHZ* to also be absent in any other descendents of this common ancestor, including other cyprinodontiform species.

To investigate this hypothesis further, I performed a complete characterisation of the *IGH* locus in the platyfish *Xiphophorus maculatus*, another cyprinodontiform species that has seen widespread use as a model organism [178]. Surprisingly, the *X. maculatus* locus possessed none of the unusual features shared between the turquoise-killifish and medaka loci, strongly suggesting independent loss of *IGHZ* in *N. furzeri* and medaka and implying a high level of volatility in *IGH* locus structure in this group of teleost fishes.

3.3.1 Overall structure

As was the case with the *N. furzeri IGH* locus, I identified candidate genome scaffolds from the most recent *Xiphophorus maculatus* genome assembly (Genbank accession GCA_002775205.2) by aligning them to *IGH* gene segments from zebrafish, stickleback and medaka, supplemented in this case with segments from the newly-characterised *N. furzeri* locus itself. In contrast to the more fragmented results in *N. furzeri*, this process identified a single sequence region on one chromosome of the *X. maculatus* locus, which I extracted and characterised as described for the assembled *N. furzeri* locus (Section 3.2) without the need for further sequencing or assembly.

The *X. maculatus IGH* locus so identified occupies roughly 293 kb on chromosome 16 (scaffold NC_036458.1; Figure 3.11A). Unlike in turquoise killifish and medaka, all identified gene segments share a common orientation; no evidence of a second sublocus in antisense could be identified. In stark

Table 3.5: Sequence similarity between *IGHZ* constant-regions in *X. maculatus*: Percentage sequence identities of pairwise Needleman-Wunsch global alignments between nucleotide (NT) or amino-acid (AA) sequences of corresponding C_ζ exons from the two *IGHZ* constant regions of *X. maculatus IGH*.

Isotype	Exon	NT	AA
Z	1	59.14	44.57
Z	2	63.93	53.41
Z	3	66.19	43.48
Z	4	65.15	50.49

contrast with both turquoise killifish and medaka, the single "sublocus" comprising *X. maculatus IGH* contains not one but two *IGHZ* constant regions, along with a hugely extended V-region extending over almost 250 kb and containing more than 120 VH-segments (Figure 3.11B). This enormous VH-diversity exceeds that of almost all previously-characterised teleost *IGH* loci (Section 1.2.3), while the presence of multiple *IGHZ* constant regions without intervening *IGHM* or *IGHD* is also highly unusual [22].

Even cursory examination of the *X. maculatus IGH* locus is therefore sufficient to reveal a unique and highly interesting structure with many unexpected differences from both turquoise killifish and medaka (Figure 3.12, columns 1-5). In particular, since *X. maculatus* is more closely related to *N. furzeri* than either is to medaka, the presence of *IGHZ* in the former strongly suggests at least two independent loss events in the lineage containing turquoise killifish and medaka, indicating an unexpected level of volatility in the evolution of this important isotype.

3.3.2 Constant regions

As discussed briefly in Section 3.3.1, the *X. maculatus IGH* locus contains two distinct *IGHZ* constant regions: one in the usual position immediately preceding the *IGHM*-associated D- and J-regions, the other, unexpectedly, at the far 5'-extremity of the locus (Figure 3.11B). Both *IGHZ* constant regions occupy the expected configuration, with four C_ζ exons, two transmembrane exons, and a secretory tail (Figure 3.11C, Table E.10). However, in contrast to the duplicate constant regions in *N. furzeri*, the two *IGHZ* constant regions in *X. maculatus* are quite distinct from each other in sequence, with an average of only 64 % nucleotide and 48 % amino-acid sequence identity between corresponding C_ζ exons (Figure D.5, Table 3.5). This unexpectedly high level of sequence divergence suggests a relatively ancient duplication event, and raises the possibility that the lineage giving rise to *N. furzeri* may have lost not one, but two distinct *IGHZ* constant regions.

While the state of *IGHZ* constant regions differs markedly between *X. maculatus* and *N. furzeri*, the configurations of the *IGHM* and *IGHD* constant regions of the two species are quite similar, with a C_μ1-C_μ2-C_μ3-C_μ4-TM1-TM2 configuration for *IGHM* and a C_δ1-(C_δ2-C_δ3-C_δ4)$_2$-C_δ5-C_δ6-C_δ7-TM1-TM2 configuration for *IGHD* (Figures 3.11B and 3.11C, Table E.10). In the *X. maculatus*

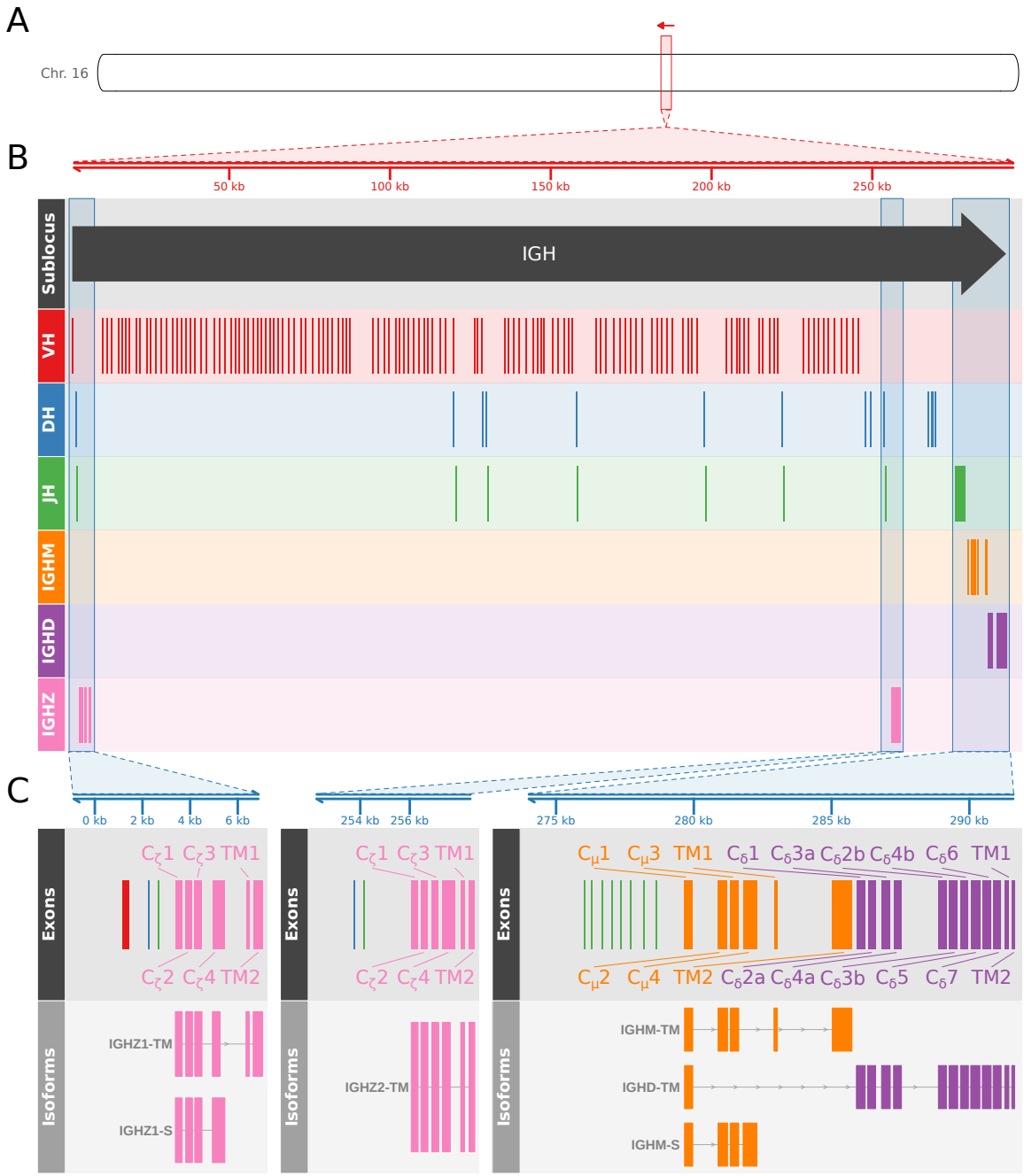
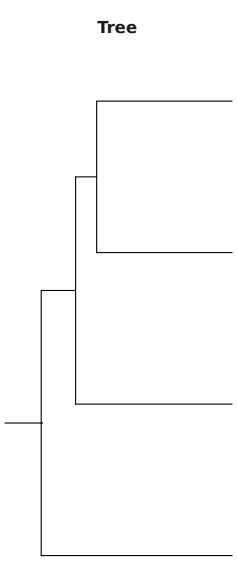


Figure 3.11: The immunoglobulin heavy chain (*IGH*) locus in *Xiphophorus maculatus*: (A) Position of the *IGH* locus on chromosome (group) 16 of the *X. maculatus* genome. (B) Arrangement of VH, DH, JH and constant-region gene segments on the *X. maculatus IGH* locus. (C) Detailed map of the *IGHZ1*, *IGHZ2* and *IGHM/D* constant regions, indicating the position and identity of the constant-region exons and the exon composition of expressed *IGH* isoforms in *X. maculatus*. Note change of orientation between subfigures (A) and (B-C).



Species	Common name	IGHZ?	Inverted sublocus?	# IGHM-TM exons	# CD(2,3,4) duplications
Xiphophorus maculatus	Southern platyfish	Yes	No	5	**2**
Nothobranchius furzeri	Turquoise killifish	**No**	**Yes**	**4**	**2**
Oryzias latipes	Medaka	**No**	**Yes**	**4**	1
Gasterosteus aculeatus	Three-spined stickleback	Yes	No	5	1

Figure 3.12: Important *IGH* phenotypes in turquoise killifish, southern platyfish, and medaka: Cladogram of the evolutionary relationship between southern platyfish (*Xiphophorus maculatus*), turquoise killifish (*Nothobranchius furzeri*) and medaka (*Oryzias latipes*), with three-spined stickleback (*Gasterosteus aculeatus*) as an outgroup. The state of various *IGH* phenotypes of interest are annotated to the right of the tree; states deviating from the expected teleost configuration are in bold.

locus, these constant regions and *IGHZ2* adopt the standard configuration seen in comparatively simple teleost *IGH* loci like those of zebrafish and fugu, with a VH-DH-JH-**CZ**-DH-JH-**CM-CD** arrangement that allows the choice between *IGHZ* and *IGHM/D* usage to be made via the choice of DH segment during VDJ-recombination. However, whether such a mechanism is also responsible for the choice between these constant regions and *IGHZ1*, which lies more than 200 kb away and upstream of the great majority of VH segments in the locus (Section 3.3.3) is questionable.

In order to investigate the expressed isoforms present in *X. maculatus*, I aligned published RNA-sequencing reads from various platyfish tissues (BioProject accession PRJNA420092, all libraries, Table E.2) to the *IGHZ* and *IGHM/D* constant regions with STAR. The results indicate the expected six-exon transmembrane configuration in both *IGHZ1* and *IGHZ2*, as well as a secretory form of *IGHZ1* comprising $C_\zeta 1$ to $C_\zeta 4$ plus a 23 bp secretory tail formed by a transcriptional run-on event from $C_\zeta 4$ (Figures 3.13A and 3.13B). However, while an in-frame secretory tail of similar length (20 bp) can be found in *IGHZ2*, it does not appear to be expressed in the read sets analysed here, suggesting that *IGHZ2* may only be expressed in transmembrane form in the individuals sampled (Figure 3.13B).

Meanwhile, the results for *IGHD* (Figure 3.13D) indicate a similar configuration to that observed in turquoise killifish, with a chimeric $C_\mu 1$ followed by 10 C_δ exons and two transmembrane exons; as in *N. furzeri*, neither a dedicated *IGHD* secretory exon nor a post-$C_\delta 7$ secretory tail was identified,

suggesting that *IGHD* may be produced solely in transmembrane form in *X. maculatus*. Secretory *IGHM* (*IGHM-S*) was also found to occupy the same four-exon configuration seen in turquoise killifish and elsewhere. However, the configuration observed for transmembrane *IGHM* (*IGHM-TM*) did not correspond to the four-exon structure shared between turquoise killifish and medaka (Figures 3.6A and 3.13C); rather, *IGHM-TM* in *X. maculatus* occupies the five-exon configuration seen in most characterised teleosts (Figure 3.5D). This surprising difference indicates that two different splice configurations of *IGHM-TM* persist in the cyprinodontiform lineage, and raises the question of what, if any, functional difference arises from the presence or absence of $C_\mu 3$ in transmembrane *IGHM* in different species. However, it remains unclear whether this pattern of exon usage (Figure 3.12) is the result of independent changes in medaka and turquoise killifish or of a reversion in *X. maculatus* to the primitive teleost configuration.

3.3.3 Variable regions

In total, I identified 125 VH segments, 14 DH segments and 15 JH segments in the *X. maculatus IGH* locus (Figure 3.11B). Of these, exactly one VH (*IGHV01-01*), DH (*IGHDZ01*) and JH (*IGHJZ01*) lie upstream of the *IGHZ1* constant region, indicating that the variable-region sequence diversity available to this isotype is limited to a single VDJ combination. In contrast, the variable region between the end of *IGHZ1* and the start of *IGHZ2* is highly expanded, with 124 tightly-clustered VH segments – more than five times the total number seen in *N. furzeri*, and more than seven times the number in the largest *N. furzeri* sublocus. Of these 124 VH segments, 106 (86 %) are apparently functional, with the remainder pseudogenised by a variety of frameshift mutations, nonsense mutations, or truncation events (Tables E.11 to E.15); it remains to be seen whether *IGHV01-01* is also capable of recombining with DH segments downstream of the *IGHZ1* constant region, and so constitutes part of the range of VDJ combinations available to the other constant regions. The VH sequences in the *X. maculatus* locus are much more tightly packed than in the *N. furzeri* locus, consistent with a lower overall prevalence of repetitive regions (21 %) in the *X. maculatus* genome [179].

In total, the VH regions in *X. maculatus IGH* fall into 23 families, of which eight contain multiple segments (Figures 3.15 and D.2); strikingly, the single VH segment serving *IGHZ* (*IGHV01-01*) represents a separate family which is distinct from any other segment in the locus. To further investigate the evolutionary history of these families, I aligned the VH segments from both the *X. maculatus* and *N. furzeri IGH* loci together with PRANK, and used the resulting alignment to construct a phylogenetic tree with RAxML [156–158]; the resulting tree (Figure 3.16) revealed a clear interrelationship between the largest families in both loci (*X. maculatus* V02 and *N. furzeri* V1), with a similar relationship observed for the second-largest families (*X. maculatus* V03 and *N. furzeri* V2). In accordance with the close sequence relationship noted in Section 3.2.4, *N. furzeri* V4 falls comfortably within the V03/V2 subtree, supporting its status as a pseudogenised subfamily of *N. furzeri* V2.

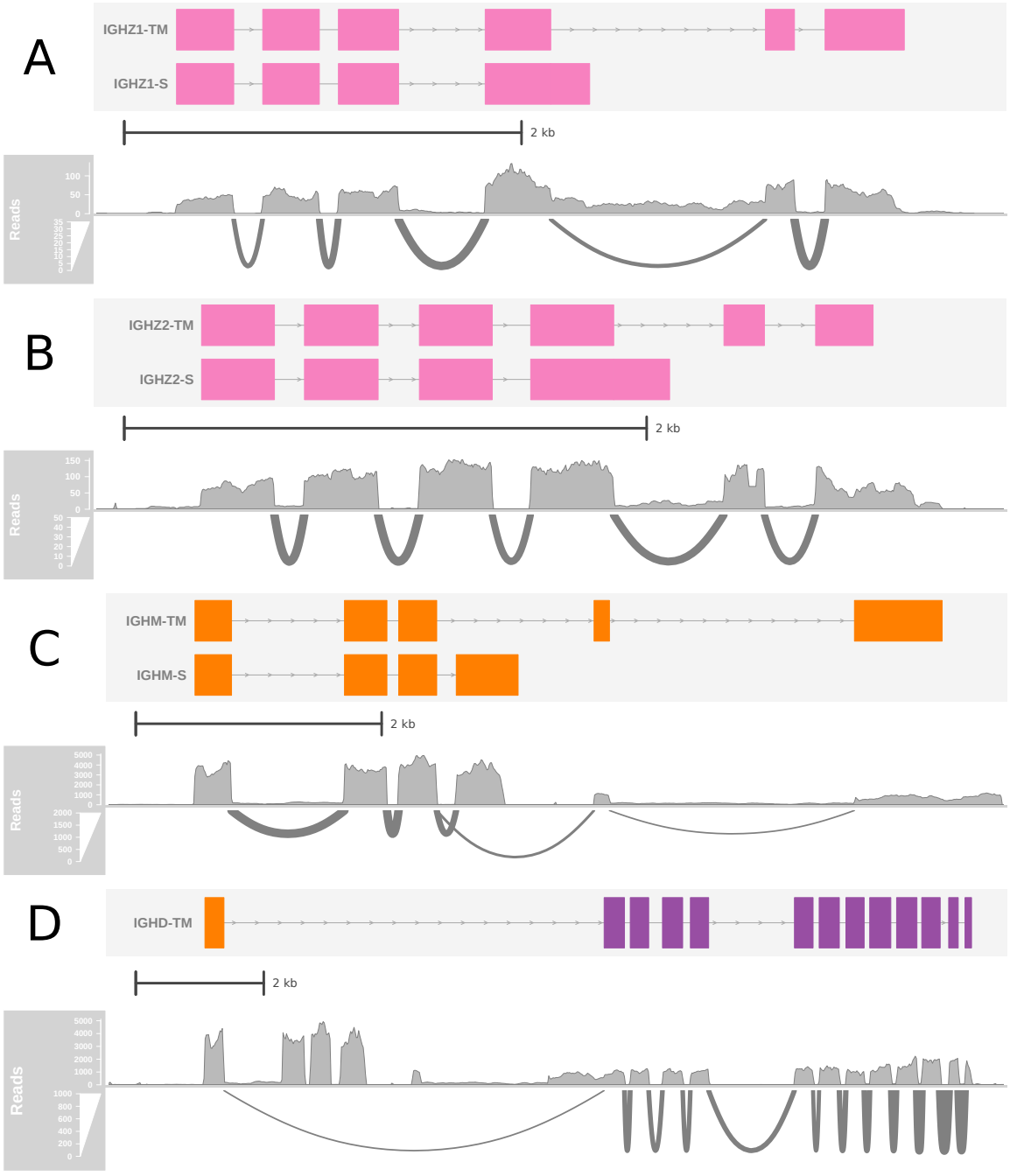


Figure 3.13: Constant-region isoforms in *X. maculatus*: Coverage and Sashimi plots [122] of STAR-aligned RNA-seq reads from *X. maculatus* samples, demonstrating the splicing behaviour of *IGH* constant-region isoforms and showing the read coverage of each exon and splice junction. (A) *IGHZ1* exon splicing, showing alternative use of the $C_{\zeta}4$/TM1 splice junction and the post-$C_{\zeta}4$ secretory tail; (B) *IGHZ2* exon splicing, showing the apparent lack of expression of the *IGHZ2* secretory isoform; (C) *IGHM* exon splicing, showing alternative splicing patterns of *IGHM-TM* and *IGHM-S*; (D) *IGHD* exon splicing, showing chimeric splicing of $C_{\mu}1$ to $C_{\delta}1$.

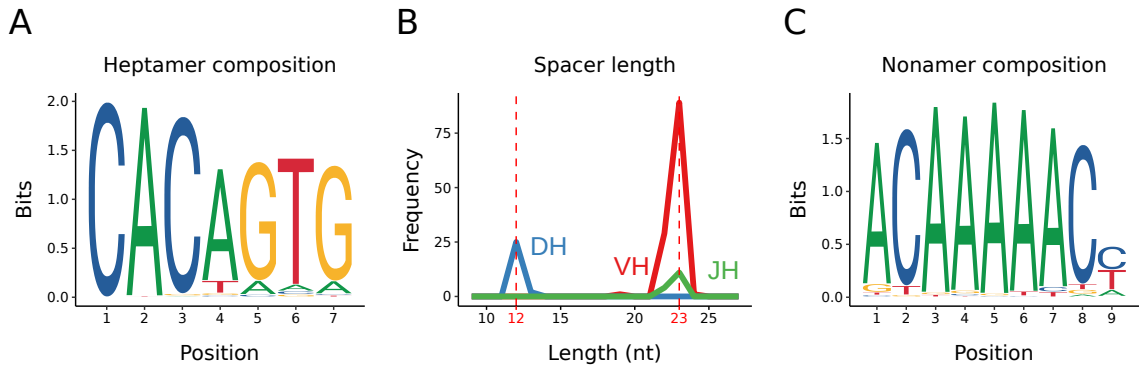


Figure 3.14: Recombination signal sequences in the *X. maculatus IGH* locus: (A) Sequence composition of conserved heptamer sequences across all *X. maculatus* heavy-chain RSSs; (B) length distribution of unconserved spacer sequences in *X. maculatus* heavy-chain RSSs; (C) sequence composition of conserved heptamer sequences across all *X. maculatus* heavy-chain RSSs.

In addition to its highly expanded VH region, the variable region of the *X. maculatus* locus is unusual in the arrangement of its DH and JH segments (Tables E.17 and E.19): in addition to the relatively densely-packed blocks of four DH and eight JH segments between *IGHZ2* and *IGHM*, and the smaller groups of three DH segments and one JH segment between the last VH segment and *IGHZ2*, small numbers of DH and JH segments are interspersed between blocks of VH segments in the extended V-region between *IGHZ1* and *IGHZ2* (Figure 3.11B). Many of these segments are arranged such that groups of one or two DH segments are closely associated with a single JH segment, raising the possibility of a more cluster-like behaviour in which each VDJ block acts as a distinct recombination unit. However, the presence of larger D- and J-regions more closely upstream of the constant regions suggests a more conventional translocon behaviour; it remains to be seen which of these traditional models of antigen-receptor structure more closely matches the *in vivo* recombination behaviour of this locus.

Finally, as is the case with *N. furzeri*, the recombination signal sequences (RSSs) in *X. maculatus IGH* correspond closely to the standard expectations across the vertebrates, with the expected heptamer and nonamer consensus sequences and spacer length distributions (Figures 3.14 and D.3, 97.6 % of RSS spacers within 1 bp of the expected conserved length).

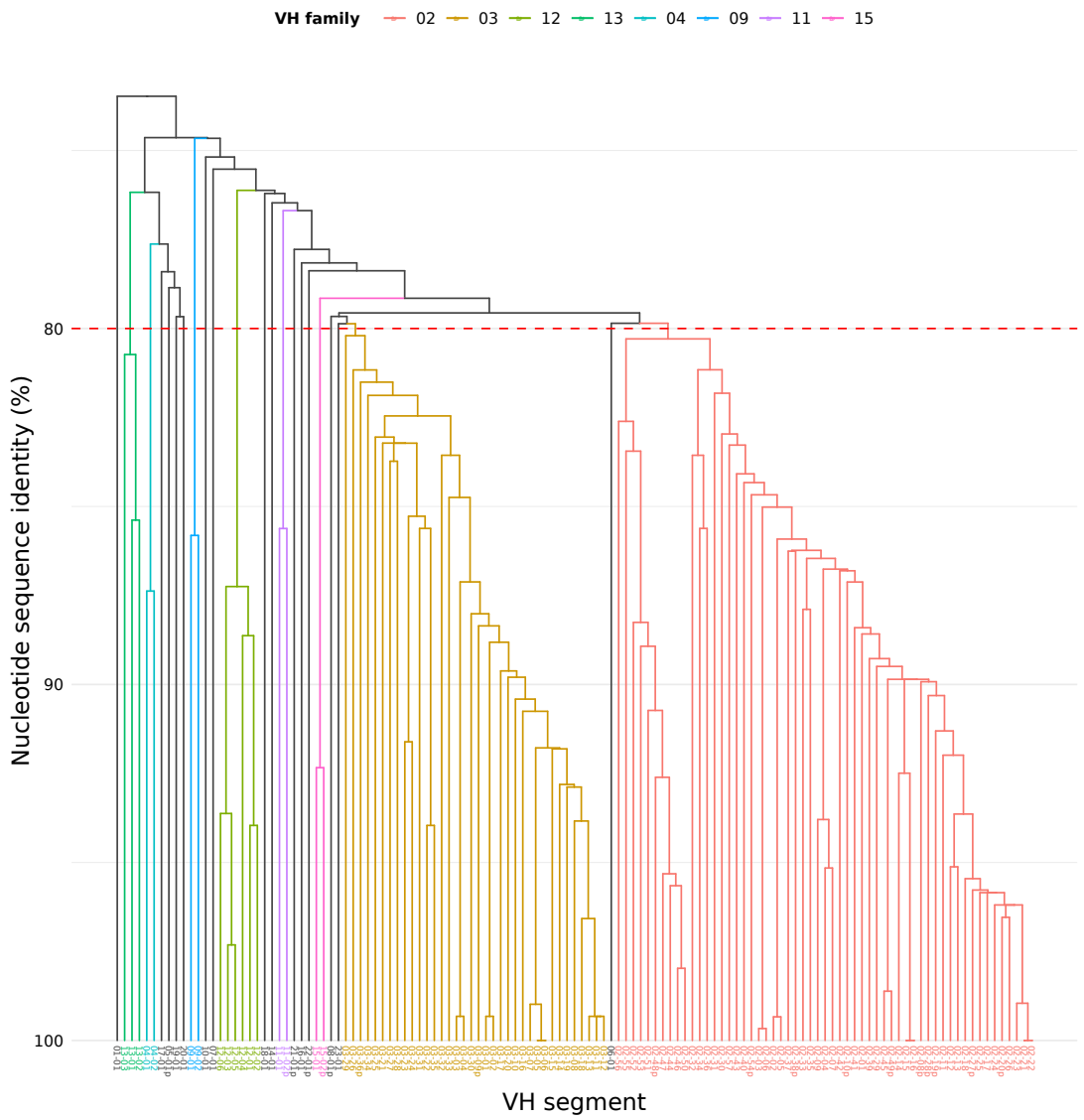


Figure 3.15: Dendrogram of VH families in the *X. maculatus IGH* locus: Dendrogram of sequence similarity of VH segments in the *X. maculatus* locus, arranged by single-linkage clustering on nucleotide sequence identity. The red line indicates the 80 % cutoff point for family assignment, while branch colour indicates family membership: VH families containing multiple segments are uniquely coloured, while single-segment families are in grey.

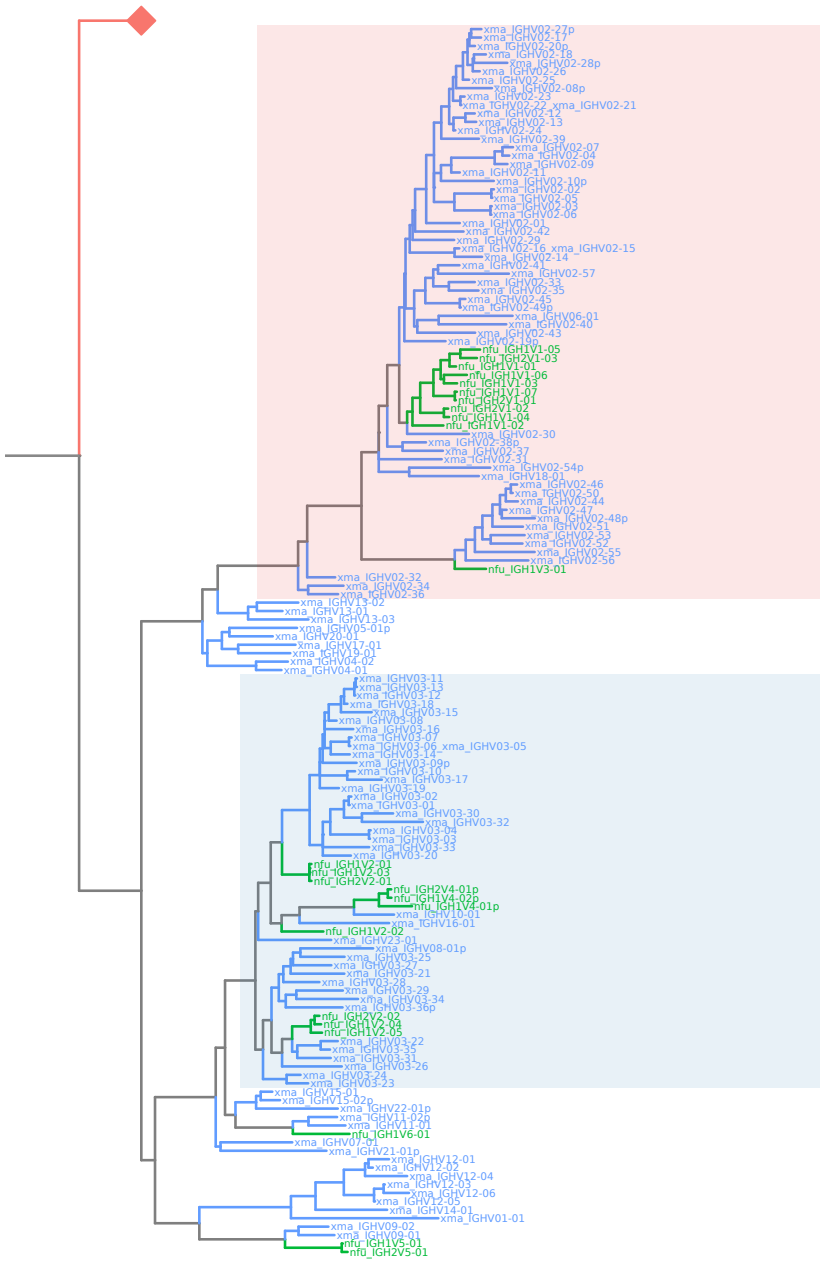


Figure 3.16: Evolutionary relationships between VH families in *X. maculatus* and *N. furzeri*: Phylogenetic tree of evolutionary relationships between *IGH* VH segments in *N. furzeri* and *X. maculatus*, as inferred from the nucleotide sequences of VH segments from both loci. Note the close interrelationship between the largest VH families in each species (red zone), and similarly between the second-largest families in each species (blue zone). The red diamond indicates the location of the outgroup, which is composed of zebrafish *TRB* V-segments.

3.4 *IGH* constant-region evolution in the Cyprinodontiformes

The characterised *IGH* loci of *Nothobranchius furzeri* and *Xiphophorus maculatus* together reveal a high degree of variability in structure and function across the Atherinomorpha, the parent clade of the Cyprinidontiformes (including *N. furzeri* and *X. maculatus*) and medaka (Figure 3.1). Several unusual features (including the loss of *IGHZ*, an inverted sublocus, and an unusual splicing pattern of *IGHM-TM*) are shared between medaka and turquoise killifish, but absent in *X. maculatus* (Figure 3.12), indicating either an independent origin in the former two species or a reversion to the primitive state in the latter. In addition, the copy number, exon usage and orientation of constant regions of other isotypes differs among the three species, raising the further question of how, when and why these changes occurred.

In order to investigate a subset of these questions with a greater degree of phylogenetic resolution, I identified and analysed *IGH* constant regions in the genomes of ten additional cyprinodontiform species (Figure 3.1, Table 3.6), as well as in a new and improved genome assembly of medaka (Genbank accession GCA_002234675.1), using the same methods described for *N. furzeri* and *X. maculatus* (Sections 3.2.3 and 3.2.4). I then grouped the constant-region exons so identified (Figure D.4) by order, exon type and spatial proximity to identify contiguous constant-regions, enabling the presence/absence and number of constant regions of each isotype to be estimated for each species. The results of this analysis (Figure 3.17, Tables E.21 to E.23) demonstrate that every species investigated possesses at least one complete *IGHM* and *IGHD* constant region in tandem, with several species exhibiting multiple such regions. Apart from *N. furzeri* and medaka, only *Nothobranchius orthonotus* was identified as clearly possessing adjacent constant regions in opposite orientation, indicating the presence of at least one sublocus in antisense; however, the fragmented nature of the *IGH* locus assembly in many analysed species prevented a confident exclusion of such configurations in other cases.

In addition to at least one *IGHM* and *IGHD* constant region, the majority of species analysed (8 out of 13) were also found to possess at least one complete *IGHZ* constant region; of the exceptions, *A. limnaeus* exhibits an orphaned, pseudogenised *IGHZ-TM1* exon but no C_ζ exons in the current genome assembly (Figure 3.17, Table E.22), while *O. latipes*, *A. australe*, *N. furzeri* and *N. orthonotus* possess no *IGHZ* exons at all. Annotating the tree from Figure 3.1 with the *IGHZ* status of each species (Figure 3.18) confirms that the loss of *IGHZ* in turquoise killifish (and related species) and medaka represent two distinct deletion events, with *Austrofundulus limnaeus* potentially representing a second independent loss of *IGHZ* within the Cyprinodontiformes and a third within the Atherinomorpha.

Apart from its repeated loss within the lineage, a second striking feature of *IGHZ* within the Cyprinodontiformes is its frequent presence in multiple copies per *IGH* locus, with a geometric mean of 1.929 regions per *IGHZ*-bearing locus. On average (geometric mean), the species analysed have approximately 1.62 *IGHZ* constant regions per *IGHM* constant region, and the same ratio vs *IGHD*,

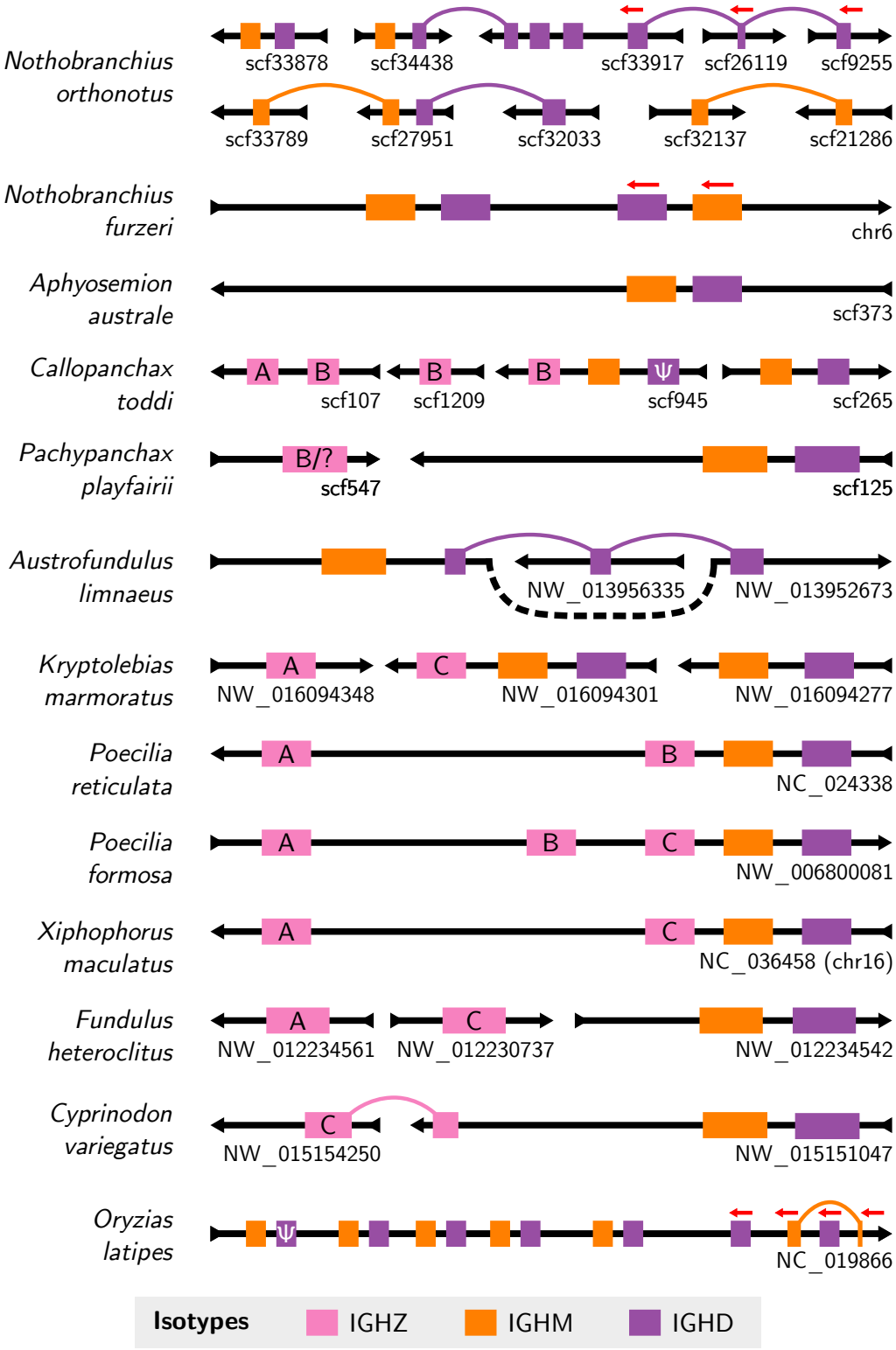


Figure 3.17: Constant-region organisation in the Atherinomorpha: Schematic of *IGH* constant regions in the genomes of thirteen species from the Atherinomorpha. Scaffold orientation is given by the black arrows; constant regions are oriented left-to-right unless otherwise specified (red arrows). Links between regions on different scaffolds indicate that exons from what appears to be the same constant region are distributed across multiple scaffolds in the order indicated; the order of unlinked scaffolds is arbitrary. The isotype of each region is given by its colour; *IGHZ* regions are further annotated with their subclass (Figure 3.18). Clearly pseudogenised constant regions are indicated by Ψ. Isotype length, scaffold length, and scaffold position are not to scale. Variable regions and lone, isolated constant-region exons are not shown.

suggesting a more complex evolutionary history than can be captured by a simple presence/absence metric. Concordantly, phylogenetic analysis (Figure 3.19, tree built using PRANK and RAxML on concatenated $C_\zeta 1$–$C_\zeta 4$ exon sequences) reveals three distinct lineages (or subclasses) of *IGHZ* constant regions in the Cyprinidontiformes, *IGHZA* to *C*, each of which is present in multiple different species and appears to have been present in the common ancestor of the eight *IGHZ*-bearing species analysed (Figure 3.20).

Only one *IGHZ* constant region from the analysed species could not be confidently assigned to one of these three subclasses, namely the single *IGHZ* of *Pachypanchax playfairii* (Figure 3.19). In order to more closely investigate the relationships of *IGHZ* in this species, I aligned the exon sequences of *P. playfairii* $C_\zeta 1$–$C_\zeta 4$ separately to the C_ζ exons of all other *IGHZ*-bearing species and plotted the distribution of alignment scores in each case (Figure 3.21). The results show a striking difference in alignment behaviour between the exons, with $C_\zeta 1$ and $C_\zeta 2$ aligning significantly more strongly to exons from the B subclass and $C_\zeta 3$ and $C_\zeta 4$ showing more ambiguous affinity for either A- or C-subclass sequences. This unexpected behaviour indicates that the *P. playfairii IGHZ* sequence is the result of a deletion or fusion event combining the first two exons of a B-subclass *IGHZ* constant region with the latter exons of a constant region from another subclass, resulting in a chimeric gene with ambiguous ancestry.

In summary, in addition to the still-universal primitive antibody classes *IGHM* and *IGHD*, the cyprinodontiforms ancestrally possessed at least three variants of *IGHZ*, giving rise to multiple subclasses of *IGHZ* constant regions evolving in parallel across the clade. Each of these subclasses appears to have been lost in multiple cyprinodontiform species, with different species showing distinct

Genus	**Species**	**Common Name**	**Bioproject/ GenBank Accession**
Nothobranchius	*furzeri*	Turquoise killifish	PRJNA599375
Xiphophorus	*maculatus*	Southern platyfish	GCA_002775205.2
Austrofundulus	*limnaeus*	–	GCA_001266775.1
Fundulus	*heteroclitus*	Mummichog	GCA_000826765.1
Poecilia	*formosa*	Amazon molly	GCA_000485575.1
Poecilia	*reticulata*	Guppy	GCA_000633615.1
Cyprinodon	*variegatus*	Sheepshead minnow	GCA_000732505.1
Kryptolebias	*marmoratus*	Mangrove rivulus	GCA_001649575.1
Aphyosemion	*australe*	Lyretail panchax	GCA_006937985.1
Callopanchax	*toddi*	–	GCA_006937965.1
Pachypanchax	*playfairii*	Golden panchax	GCA_006937955.1
Nothobranchius	*orthonotus*	Spotted killifish	GCA_006942095.1
Oryzias	*latipes*	Medaka	GCA_002234675.1

Table 3.6: Genome assemblies used to identify *IGH* locus sequences in cyprinodontiform fishes

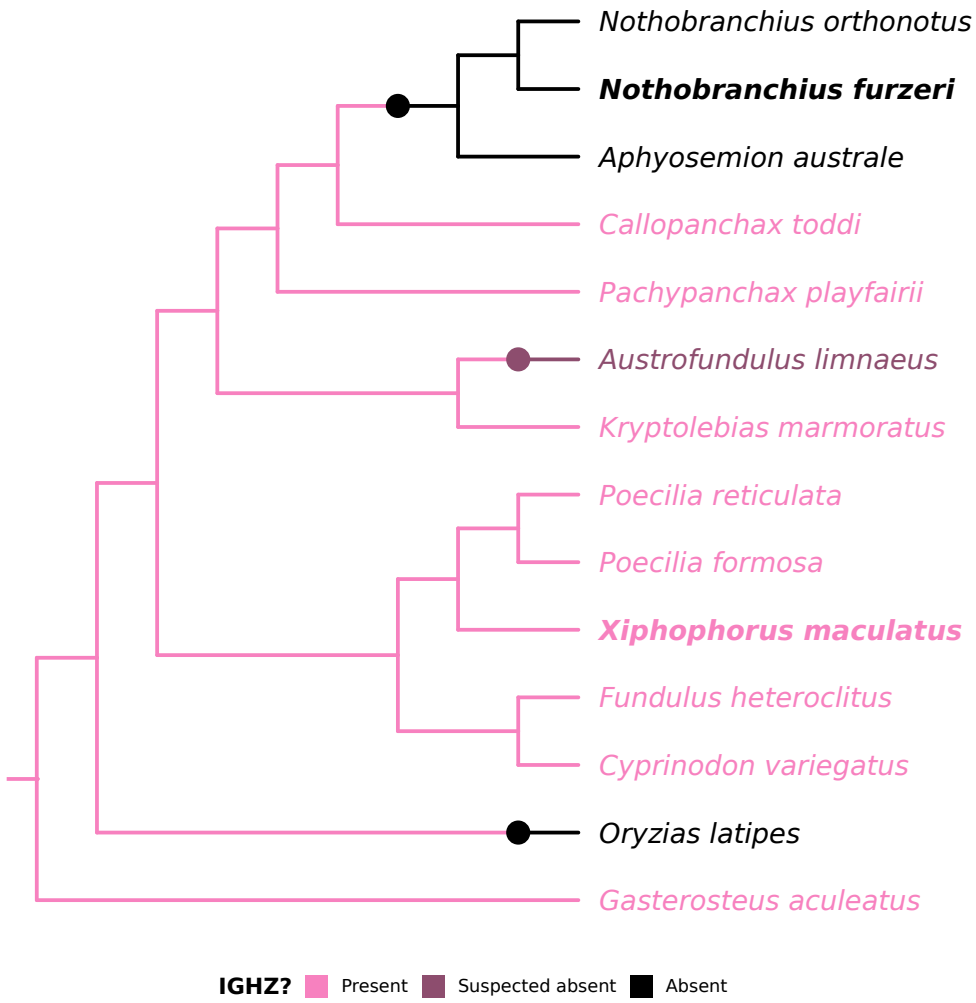


Figure 3.18: *IGHZ* has been lost multiple times independently in the Atherinomorpha: Cladogram of species reproduced from Figure 3.1, annotated according to the known (tip nodes) or inferred (internal nodes) presence or absence of intact *IGHZ* constant regions in each species. Large coloured points on the cladogram denote sites of hypothesised state changes; *IGHZ* is assumed to be primitively present in the clade and losses to be irreversible. The currently-available genome assembly of *A. limnaeus* (dark pink) contains one pseudogenised *IGHZ-TM1* exon and no C_ζ exons.

patterns of retention and loss, and in at least one lineage – that of *Pachypanchax playfairii* – two different *IGHZ* lineages have fused to produce a chimeric isotype. All three subclasses are missing from a subset of species in the Nothobranchiidae (including *Nothobranchius furzeri*), and also appear to have been independently lost in *Austrofundulus limnaeus*. Taken together, these data suggest a high degree of complexity and volatility in the evolution of mucosal adaptive immunity in the Cyprinodontiformes.

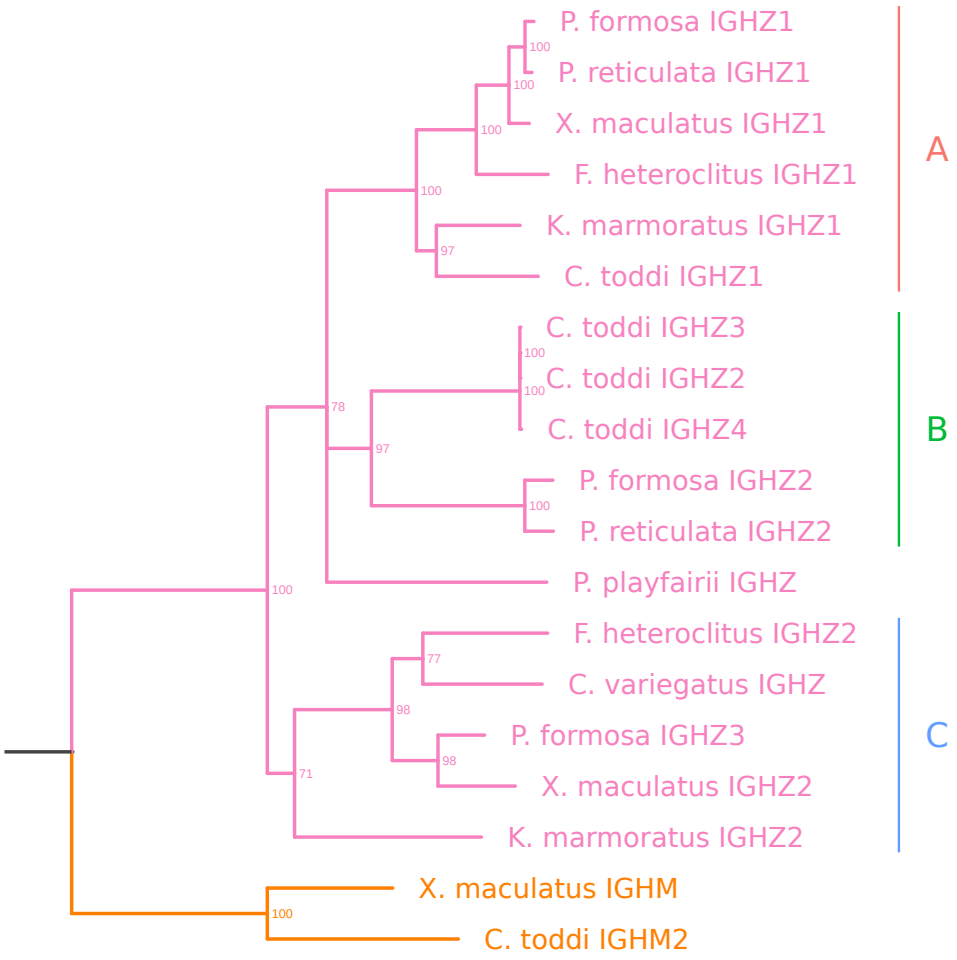
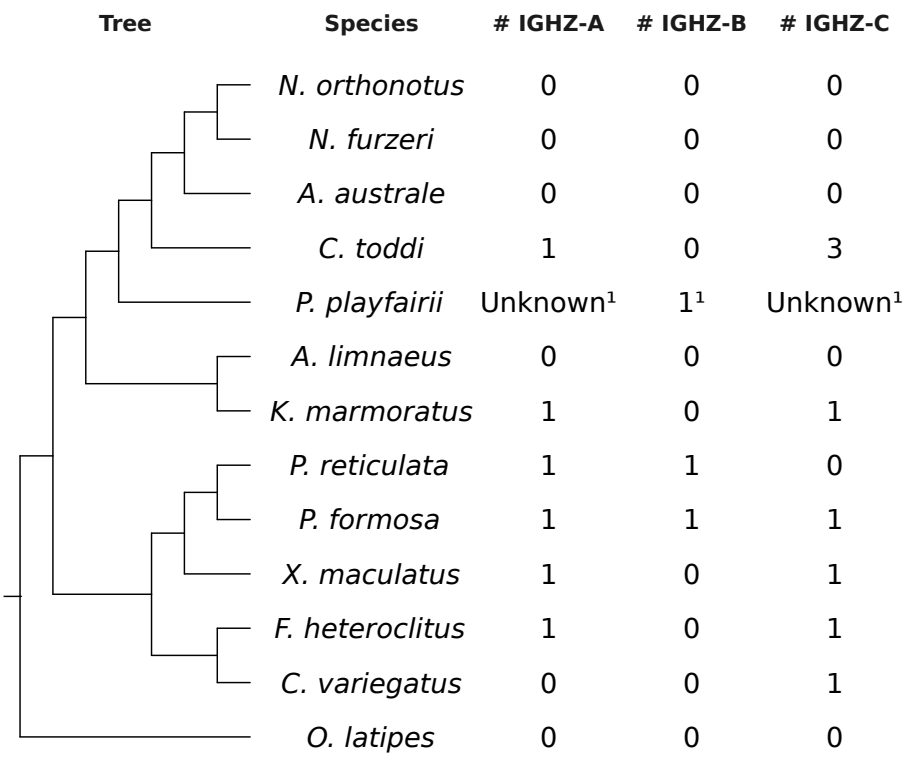


Figure 3.19: *IGHZ* constant regions in the Cyprinodontiformes constitute three distinct subclasses: Phylogram of concatenated $C_\zeta 1$–$C_\zeta 4$ nucleotide sequences from *IGHZ*-bearing species in Table 3.6, with $C_\mu 1$–$C_\mu 4$ sequences from two species as an outgroup (in orange). Nodes with bootstrap support of less than 65 % are collapsed into polytomies. Major clades (A–C) are annotated on the right. Support values indicate the result of rapid bootstrapping by RAxML across 1000 replicates.

3.5 Discussion

The teleost fishes are the largest and most diverse group of vertebrates, with nearly 30,000 species comprising almost half of extant vertebrate diversity [181]. Of this vast variety of teleosts, only a few species, mostly those used extensively in aquaculture or scientific research, have undergone extensive study with regard to their immunoglobulin gene loci (Section 1.2.3). Studies in these model species have revealed a high level of structural diversity among teleost *IGH* loci, with huge variation in both size and organisation, as well as the existence of three distinct antibody isotypes in fish, of which two (*IGHM* and *IGHD*) are shared with tetrapods and one (*IGHZ*) appears to be teleost-specific. However,



Tree	Species	# IGHZ-A	# IGHZ-B	# IGHZ-C
	N. orthonotus	0	0	0
	N. furzeri	0	0	0
	A. australe	0	0	0
	C. toddi	1	0	3
	P. playfairii	Unknown[1]	1[1]	Unknown[1]
	A. limnaeus	0	0	0
	K. marmoratus	1	0	1
	P. reticulata	1	1	0
	P. formosa	1	1	1
	X. maculatus	1	0	1
	F. heteroclitus	1	0	1
	C. variegatus	0	0	1
	O. latipes	0	0	0

[1] *P. playfairii IGHZ* $C_\zeta 1$–$C_\zeta 2$ appear to be derived from *IGHZ-B*, while the other exons are of uncertain subclass origin (Figure 3.21).

Figure 3.20: Distribution of *IGHZ* subclasses in the Atherinomorpha: Cladogram of atherinomorph species with characterised *IGH* constant regions, annotated with the number of regions belonging to each *IGHZ* isotype in each species. All three subclasses are present in at least one species in both major branches of the cyprinodontiform clade, suggesting that they were all present in the common ancestor of this grouping.

in many large and important teleost lineages, the genetic basis of B-cell-mediated humoral immunity remains unknown.

In this chapter, I presented two complete and ten partial assemblies of *IGH* loci from the Cyprinodontiformes, a diverse clade of primarily freshwater fishes for which no such loci have previously been characterised. The two complete assemblies were of the *IGH* loci of the turquoise killifish *Nothobranchius furzeri* and the southern platyfish *Xiphophorus maculatus*, two important model species with an estimated divergence time of less than 80 Mya [90]. Despite their close relationship, these species show radically different locus organisations, with huge differences in VDJ number (24 VH segments in *N. furzeri* versus 125 in *X. maculatus*), locus organisation (two small subloci in opposite sense in *N. furzeri*, one large unitary locus in *X. maculatus*) and isotype availability (no *IGHZ* in *N. furzeri*, two distinct *IGHZ* regions in *X. maculatus*), as well as more subtle but still-important distinctions like differences in constant-region splicing behaviour (four exons in *N. furzeri IGHM-TM*, five in *X. maculatus*). These results are consistent with previous findings of highly-diverse teleost loci and support a process of rapid evolution in the *IGH* locus. Characterisation of the constant regions of

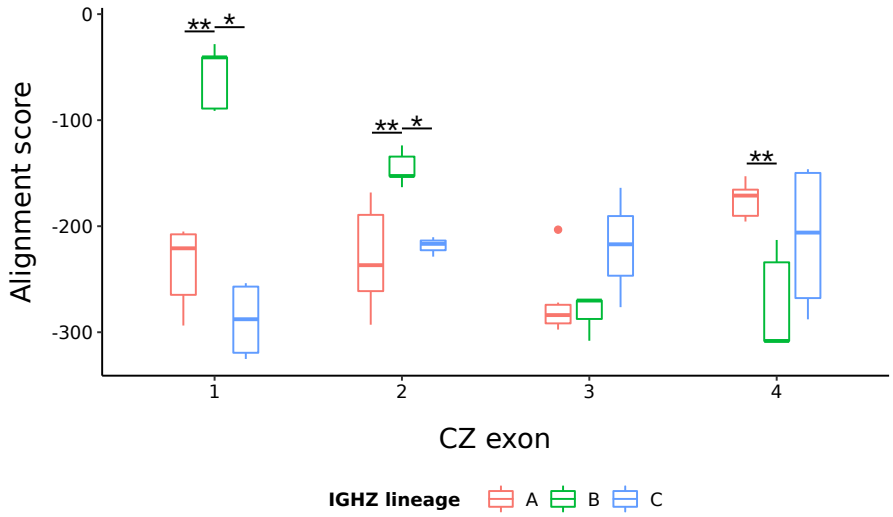


Figure 3.21: Subclass affinity of *IGHZ* in *Pachypanchax playfairii*: Boxplots of Needleman-Wunsch alignment scores between the amino-acid sequences of *P. playfairii* C_ζ exons and those of seven other *IGHZ*-bearing cyprinodontiform species, demonstrating the differing affinity of different *P. playfairii* exons for each of the three *IGHZ* subclasses. Pairwise p-values were computed using nonparametric Mann-Whitney U tests ($* : 0.01 < p \leq 0.05$; $** : 0.001 < p \leq 0.01$; $*** : p \leq 0.001$).

a further ten cyprinodontiform species confirmed this finding, with several groups of closely-related species (e.g. *Nothobranchius furzeri*, *Nothobranchius orthonotus* and *Callopanchax toddi*) showing highly divergent locus structures and constant-region availability.

It is interesting to speculate on the origins of this extremely rapid diversification in gene structure. Very little is known about the relationship between environmental context and immune locus structure; it is possible that part of the variety in *IGH* gene locus structure in the Cyprinodontiformes represents divergent adaptations to different immune environments. Alternatively, this diversification may be primarily the result of unusually high rates of stochastic, non-adaptive changes in gene structure in germline *IGH*. Finally, at least some of the difference between locus structures in different species is likely to be attributable to differences in assembly quality; for example, the characterisation of medaka constant regions presented here contains many fewer unusual or incomplete constant regions than that presented in the published medaka *IGH* locus [25], primarily due to the increased quality of the more recent medaka genome assemblies.

Before the publication of this work, only two teleost species (medaka and channel catfish) were known or thought to lack the *IGHZ* antibody isotype from their *IGH* loci, out of more than ten species with published locus characterisations. This relative rarity of observed absence, combined with the apparent importance of *IGHZ* in teleost mucosal immunity, suggested that the loss of *IGHZ* was likely to be a rare and unusual event. However, in addition to confirming the absence of *IGHZ* in medaka,

this study identified four new teleost species (*Nothobranchius furzeri*, *Nothobranchius orthonotus*, *Aphyosemion australe* and *Austrofundulus limnaeus*) that appear to lack *IGHZ* constant regions in their *IGH* loci, representing two distinct and previously unknown loss events independent from that affecting the closely-related medaka. This finding, which triples the number of known teleost species without *IGHZ* and doubles the number of known loss events, is even more striking when combined with the discovery that the cyprinidontiform common ancestor likely had no fewer than three distinct *IGHZ* constant regions, all of which would have had to be lost on the way to any *IGHZ*-free lineage. The high level of observed variability in *IGHZ* prevalence among the cyprinidontiforms suggests that the presence/absence of *IGHZ* in the wider teleost clade may be much more volatile than suggested by previously available locus data, and raises the possibility that, given sufficiently high-density analysis of other teleost lineages, a similar frequency of *IGHZ*-lacking species may also be found elsewhere. However, it may also be the case that the apparently high frequency of *IGHZ* loss events in the Atherinomorpha is a special case, arising from chance, an unusual selective environment, or limitations in the available genome assemblies.

The absence of *IGHZ* from so many species in the Atherinomorpha naturally raises the important question of how the mucosal adaptive immune system in these species differs from that of their *IGHZ*-bearing relatives. Data from rainbow trout suggest that *IGHT* (an alternative name for *IGHZ* in some species) plays a specialised role in antibody immune responses at multiple mucosal surfaces, with increased prevalence of $IGHT^+$ B-cells and secreted IGHT antibodies relative to serum, a primarily *IGHT*-dependent response to mucosal infections, and a much higher rate of bacterial coating by IGHT in skin and gut flora relative to IGHM [20, 26]. If these findings hold for other teleost species, it is not clear how *IGHZ*-lacking teleost species carry out specialised immune functions at mucosal barriers: how, and to what extent, can *IGHM* compensate for the loss of a specialised mucosal isotype? This question is especially interesting in the case of *IGHZ*-lacking species with close *IGHZ*-bearing relatives (e.g. *Nothobranchius furzeri* and *Callopanchax toddi*, or perhaps *Austrofundulus limnaeus* and *Kryptolebias marmoratus*); if it is the case that mucosal immune responses differ systematically between these species, such that *IGHM* takes up some or all of the roles normally played by *IGHZ*, then uncovering the mechanisms by which this shift is regulated could reveal important new insights into decision-making and control of humoral adaptive immunity.

One important difference between the *X. maculatus* and *N. furzeri* loci whose evolution is more difficult to investigate using genomic data is the exon usage behaviour of the different splice isoforms present in the transcriptome of each species. In *X. maculatus*, transmembrane *IGHM* adopts the same configuration as that seen in most teleosts for which this has been investigated: a five-exon isoform in which the end of $C_\mu 3$ is spliced to the start of TM1 and $C_\mu 4$ is excised. Conversely, in *N. furzeri IGHM-TM* adopts the same four-exon configuration observed in medaka, in which $C_\mu 3$ is also excluded. Given that *X. maculatus* adopts the primitive configuration, the recurrence of the same unusual configuration in both medaka and turquoise killifish is surprising, and indicates that both configurations are present in the Cyprinodontiformes; however, without more information about the

mechanisms and genomic sequence correlates underlying this difference, it is impossible to distinguish an independent origin of the derived phenotype in medaka and *N. furzeri* from a reversion to the primitive phenotype in *X. maculatus*. It is also not clear at present what functional differences, if any, arise from this difference in exon usage, although it seems unlikely that the shorter four-exon form of *IGHM-TM* would persist in multiple species if it prevented effective antibody development, selection, or antigen response.

As a result of the research findings presented in this chapter, a number of previously-uncharacterised teleost species now have databases of constant regions available. As a result, primer design for targeted RNA-sequencing of expressed antibody sequences is now possible for these taxa, enabling quantitative immune-repertoire sequencing approaches in a large number of closely-related cyprinidontiform species. In addition to the special interest of immune-repertoire data from any one of these new species (e.g. *N. furzeri* for immune-repertoire ageing or *X. maculatus* for ecological and evolutionary research), the possibility of sequencing the repertoires of several related species adds an exciting comparative dimension. This comparative element would be especially interesting in the context of investigating the repertoire responses of closely related species with different *IGHZ* genotypes. In combination with the genomic and functional comparisons discussed above, the novel possibility of large-scale comparative repertoire studies arising as a result of this research establishes the Cyprinidontiformes, and especially the African killifishes, as a highly-promising group of model species for comparative evolutionary immunology.

Chapter 4

Immunoglobulin sequencing in *Nothobranchius furzeri*

4.1 Introduction

The antibody repertoire of an individual is the complete collection of immunoglobulin sequences (either DNA, RNA or protein) expressed or secreted by the complete population of B-lymphocytes it contains [182]. This enormous population of sequences contains both a high degree of sequence diversity and a great deal of functionally-relevant internal structure: naïve sequences in the primary repertoire (Section 1.2.2) are produced by the same underlying generative and selective process and so share common statistical properties, while many sequences in the secondary repertoire (Section 1.2.4) are related to one another as members of clones descended from a common naïve B-cell ancestor. The diversity and structure of the antibody repertoire are both essential for the functionality of the humoral adaptive immune system in vertebrates, affecting both the ability of an organism to respond to a wide range of novel pathogenic threats and the extent and effectiveness of its long-term immune memory. However, the extensive changes that occur in B-cell immunity with age are believed to have a substantial and damaging effect on the antibody repertoire, disrupting its composition, reducing its diversity and impairing its ability to respond effectively to infections and other antigenic challenges (Section 1.3).

By sampling the repertoire of antibody sequences present in an individual and reconstructing its underlying recombinatorial and clonal structure, much can be learned about the history and current state of humoral adaptive immununity in that organism. The advent of modern high-throughput-sequencing technologies has enabled the development of specialised sequencing techniques designed to interrogate the antibody repertoire, typically via targeted sequencing of heavy-chain transcripts [103, 183]. From the beginning, these immunoglobulin-sequencing (or IgSeq) studies included teleost model organisms, with early IgSeq studies investigating the antibody repertoire of developing and adult zebrafish [63, 67]. Studies of antibody-repertoire ageing, however, have largely been confined to human peripheral blood, where they have revealed a range of age-related changes including a loss in within-repertoire sequence diversity, an increase in between-individual variability in repertoire composition, and an accumulation of memory-cell clonal expansions (Section 1.3). More comprehensive research into antibody-repertoire ageing, however, has been restricted by the limitations of peripheral blood as a sample of the whole-body repertoire, the difficulty in obtaining enough high-quality samples of immune organs from humans of diverse ages, and the relatively long lifespans of most common vertebrate model organisms (Section 1.3).

As a remarkably short-lived, experimentally-tractable vertebrate species with a complex adaptive immune system, the turquoise killifish (*Nothobranchius furzeri*, Sections 1.4 and 3.2) represents a unique opportunity to investigate questions of adaptive immunosenescence that have previously been difficult to address with a sufficient degree of resolution, reproducibility and depth. Using the *IGH* locus structure and sequence unveiled in Chapter 3, it is now possible to design primers and pipelines to perform IgSeq in the turquoise killifish for the first time, and thus to investigate the

antibody repertoire of this emerging model system for a variety of sample types, ages and experimental conditions.

In this chapter, therefore, I present the first analysis of antibody repertoires in turquoise killifish, revealing diverse, reproducible and individualised repertoires even within a highly-inbred laboratory strain (Section 4.4). These repertoires exhibit significant changes in diversity structure with age, including a loss of alpha (within-individual) diversity and increase in beta (between-individual) diversity, consistent with the changes observed in human peripheral blood (Section 4.5). Extending the analysis from whole-body killifish samples to total RNA isolated from killifish guts, I perform the first specific analysis of antibody repertoire ageing in a mucosal immune organ, demonstrating that the ageing phenotypes observed in the whole body are even stronger in the gut mucosal repertoire, a difference perhaps explained by the much greater prevalence of large, expanded clones in mucosal antibody repertoires compared to the body as a whole (Section 4.6). Together, these results firmly establish the turquoise killifish as a model for quantitative immunology and vertebrate adaptive immunosenescence, opening up an enormous range of future experiments with the potential to greatly expand our knowledge of the kinetics, underlying processes, and potential amelioration of antibody repertoire ageing.

4.2 Collecting killifish samples for immunoglobulin sequencing

To obtain samples for establishing and validating an immunoglobulin-sequencing protocol in turquoise killifish, as well as to investigate changes in killifish repertoires with age, male GRZ-AD turquoise killifish from a single hatching cohort were raised under standard husbandry conditions (Section 2.1) and sacrificed by anaesthesia, followed by flash-freezing in liquid nitrogen and preservation at $-80\,^{\circ}\mathrm{C}$. In total, thirty-two fish were sacrificed at four time points (Tables 4.1 and E.24): regular groups of ten fish each were sacrificed at roughly 5.5 weeks (shortly after reproductive maturation), 8 weeks (middle adulthood) and 10.5 weeks (close to median lifespan) post-hatching, while two surviving fish were sacrificed eight weeks later at roughly 18 weeks post-hatching.

In order to obtain a representative sample of the whole-body antibody repertoires from these individuals (as opposed to the specific repertoire of a particular organ), the frozen fish were homogenised

Table 4.1: Summary of killifish used in IgSeq pilot and ageing experiments. All fish are GRZ-AD strain and male.

Group	# Fish	Hatch date	Sacrifice date	Age (days)	Age (weeks)	Mean weight (g)
1	10	2016-05-09	2016-06-17	39	5.57	1.3
2	10	2016-05-09	2016-07-04	56	8	1.37
3	10	2016-05-09	2016-07-21	73	10.4	1.76
4	2	2016-05-09	2016-09-14	128	18.3	2.3

in liquid nitrogen with a mortar and pestle, and total RNA was isolated from the resulting powder with guanidinium thiocyanate-phenol-chloroform extraction (Section 2.2.1.4).

4.3 Establishing IgSeq in the turquoise killifish: principles & protocols

4.3.1 Preparing the sequencing library

Immunoglobulin sequencing is a precise and quantitative method which is highly sensitive to errors and biases arising during both library preparation (especially PCR) and sequencing. In order to minimise the impact of these errors and biases on the results, unique molecular identifiers (UMIs) are widely used in IgSeq library-preparation protocols [103, 183] to uniquely label each input molecule, enabling sequences in the dataset arising from the same input molecule to be identified, grouped and collapsed into a single consensus sequence. This process effectively corrects for biases in apparent abundance arising from differential amplification during library preparation or differential clustering during sequencing, and allows identical sequences descended from the same input molecule to be distinguished from those descended from different input molecules with the same sequence [183]. In addition, as long as enough reads arising from a given input sequence are present, the generation of a consensus sequence from each UMI group effectively corrects for sequence errors arising during library preparation and sequencing, enabling the original input sequence to be reconstructed much more accurately (Figure 4.1) [103, 183].

In the IgSeq protocol established here for the turquoise killifish, the addition of UMIs is achieved through the use of a procedure adapted from Turchaninova *et al.* [103]. This procedure takes advantage of the intrinsic terminal-transferase activity exhibited by reverse-transcriptase enzymes derived from Moloney Murine Leukemia Virus (MMLV). Upon reaching the end of an RNA template, such viruses add a variable number of untemplated deoxyribonucleotides to the 3'-end of the new cDNA molecule, with a strong bias for cytidine residues [184]. If a template-switch adapter (TSA) oligo ending in riboguanosines is added to the reaction mixture, it will pair with these untemplated terminal cytidines to form a new priming site for the reverse transcriptase. The enzyme will then re-attach to this new priming site and process to the end of the paired oligo, adding the sequence of the TSA to the 3'-end of the cDNA [184]. This procedure, known as *template switching* (Figure 4.2), enables semi-arbitrary sequences to be prepended to the reverse-transcribed mRNA sequence, including a primer sequence and (in this case) a UMI [103].

In the library-preparation protocol used for immunoglobulin sequencing in the turquoise killifish, therefore, reverse transcription is performed on total RNA using a gene-specific primer (GSP) homologous to the constant-region sequence of the isotype of interest and an MMLV-derived reverse-transcriptase enzyme optimised for its terminal-transferase activity (Section 2.2.4.2). A template-switch adapter containing an invariant 5' primer sequence and random 3' UMI sequence (Figure 4.3)

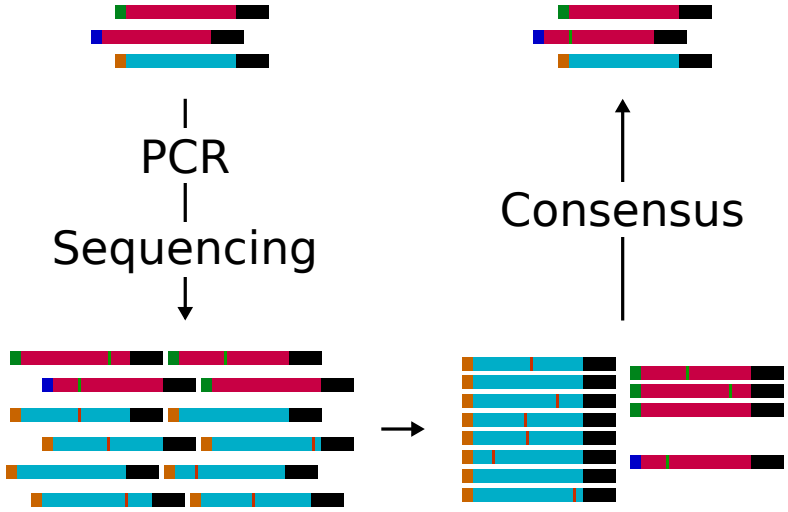


Figure 4.1: Correcting errors and biases with unique molecular identifiers: In this simplified schematic, three template RNA molecules representing two distinct sequences (red and blue rectangles) are tagged with unique molecular identifiers (UMIs, coloured left-hand ends) prior to PCR-based library preparation and sequencing. As a result of differential amplification bias between the sequences, the less-abundant input sequence gives rise to a larger number of sequencing reads; however, by aligning and collapsing matching UMI groups into consensus reads, the original proportions of the sample can be reconstructed. Consensus-read generation also enables the correction of various PCR and sequencing errors (thin red/green bars), the majority of which are present in only a minority of sequence copies within a UMI group; in this example, only the dark-blue UMI group, for which only a single sequencing read was obtained, cannot be effectively corrected in this way.

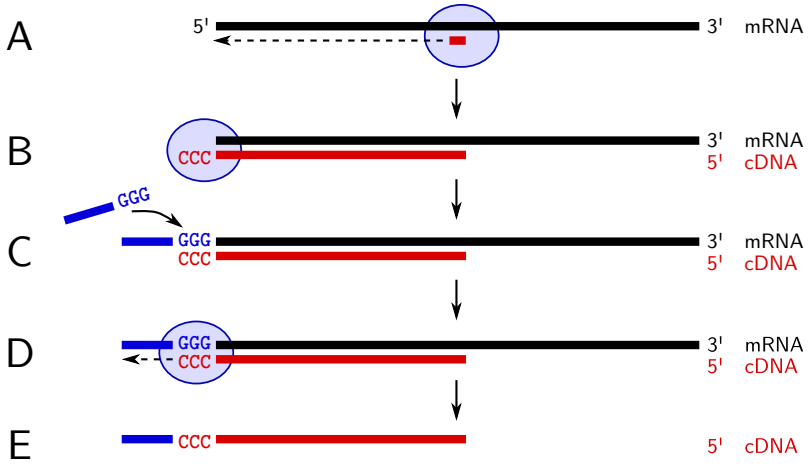


Figure 4.2: Addition of a known 5'-sequence to cDNA with template switching: In this simplified schematic, an MMLV-derived reverse-transcriptase enzyme (blue oval) binds a primed RNA template and processes to the 5-prime end (A), where it deposits additional untemplated cytidine residues at the 3'-end of the cDNA (red bar) using its terminal-transferase activity (B). A template-switch adapter (TSA) with complementary terminal guanosine residues (blue bar) pairs with these terminal cytidines (C), creating a new double-stranded priming site to which the reverse-transcriptase enzyme can bind (D). Processing of the enzyme to the end of the TSA sequence produces a cDNA molecule with an additional, known 3' sequence (E).

AAGCAGUGGTAUCAACGCAGAGUNNNN–
–UNNNNUNNNNUCTTrGrGrGrG

Figure 4.3: The SmartNNNa template-switch adapter: Annotated sequence of the SmartNNNa barcoded template-switch adapter (TSA) used in template-switch reverse transcription for IgSeq library preparation [103]. The 5'-terminal purple characters represent an invariant sequence used for primer-binding in downstream PCR steps, while the green N characters represent the random nucleotides constituting the unique molecular identifier (UMI), each of which could take any value from A, C, G or T. The U residues represent deoxyuridine, which is specifically digested after reverse transcription to remove residual TSA oligos from the reaction mixture (Section 2.2.4.2). The orange, 3'-terminal rG characters indicate riboguanosine residues, which pair with terminal-transferase-added cytidine residues to enable template switching.

is included in the reaction mixture and added to the cDNA sequence via template switching. In addition to enabling UMI-based clustering and correction as described above, this approach has the advantage of bypassing the variable region of the *IGH* transcript and thus avoiding the use of multiplexed V-segment primers, which may introduce additional biases through differential binding affinity between V-segments [185].

Following reverse transcription, the reaction mixture is treated with uracil-DNA glycosylase (UDG) to specifically remove residual TSAs through digestion of deoxyuridine residues (which are absent in the template sequence). The library then undergoes three successive rounds of PCR amplification (Figure 4.4, Section 2.2.4.3), which respectively serve to amplify the reverse-transcribed cDNA sequence; amplify further while adding partial Illumina TruSeq adapter sequences; and add complete adapter sequences including library-specific P1 and P2 index sequences [183]. In the first two cases, the PCR primer sequences are homologous to the invariant part of the TSA sequence and the $C_\mu 1$ exon, respectively; as all known forms of *IGH* in turquoise killifish (*IGHM-TM*, *IGHM-S* and *IGHD-TM*) share this exon, the isotype- and isoform-specificity of the library prep can be altered simply by changing the position of the reverse-transcription GSP, with all other primer sequences left unchanged. In all experiments included in this chapter, a GSP on the $C_\mu 2$ exon was used, resulting in a library including *IGHM-TM* and *IGHM-S* sequences but excluding *IGHD*.

Libraries to be sequenced together are then pooled in equimolar ratio and the pooled sample undergoes size-selection to remove residual primer-dimers and other unwanted sequences (Section 2.2.4.4). The complete library-preparation protocol reliably produces a single peak in the range of 650-680 bp, consistent with the known lengths of VH, DH and JH gene segments in the *N. furzeri* locus. Following size-selection and quality control, pooled libraries are sequenced on an Illumina MiSeq sequencing machine with 2×300 bp reads (Section 2.2.4.4); this longer read length is necessary to completely cover the variable region. To avoid problems associated with the low sequence complexity of single-amplicon sequencing libraries [186], a large proportion of PhiX spike-in (30 %) was used to increase the sequence complexity of the libraries.

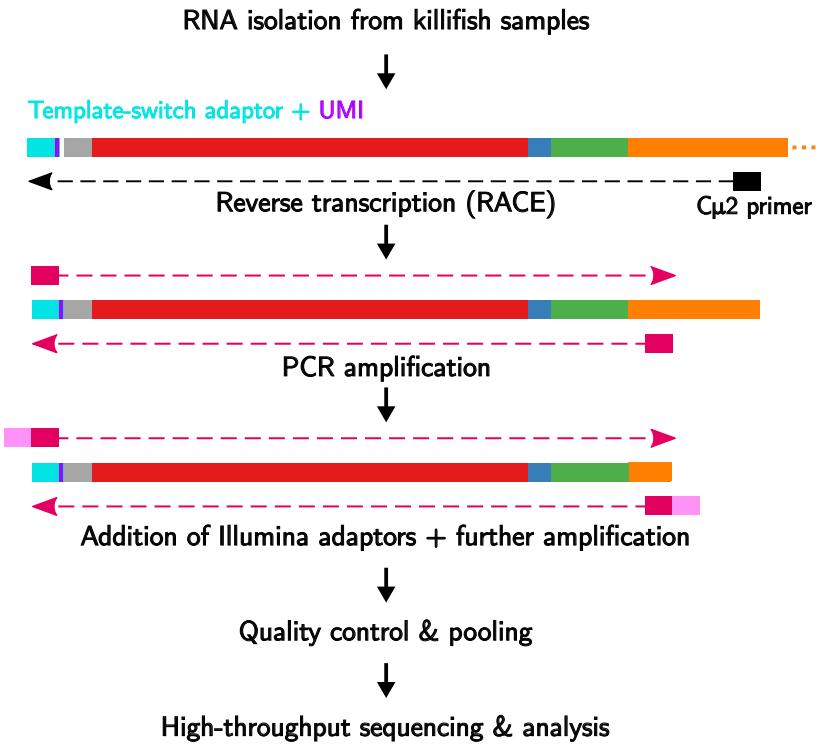


Figure 4.4: Summary of the IgSeq library-preparation protocol for turquoise-killifish samples: Schematic summary of library-preparation protocol used to prepare immunoglobulin-sequencing libraries from turquoise-killifish samples. In order to capture all *IGHM* transcripts (both secreted and transmembrane), the reverse-transcription primer was homologous to the $C_{\mu}2$ exon, while the PCR primers were homologous to $C_{\mu}1$. The third step ("Addition of Illumina adapters + further amplification") represents two distinct rounds of PCR amplification and bead purification, for stepwise addition of Illumina TruSeq adapter sequences.

4.3.2 Sequence pre-processing with pRESTO

Raw IgSeq data from the protocol described in Section 4.3.1 takes the form of a large number of paired-end sequencing reads, each of which represents a single biased and error-prone observation from the set of input sequences in the original sample. To get from this fragmented and unreliable dataset to a set of complete, error-corrected and bias-adjusted *IGH* variable-region sequences, extensive pre-processing (Section 2.3.5) must be performed on the raw data. In this case, this pre-processing was largely carried out with pRESTO [160], part of the Immcantation suite of repertoire-sequencing analysis tools (Figure 4.5).

To begin with, each read pair in the dataset is annotated with various information about the source individual (ID, strain, sex, age and weight at death, etc.) as well as information about its place in the replicate structure of the experiment (Section 2.3.5.1). The sequences are then filtered to remove low-quality sequences (with a mean Phred score of less than 20 [162]). Invariant primer sequences are trimmed from the ends of the reads, and the UMI sequence of each forward read (containing

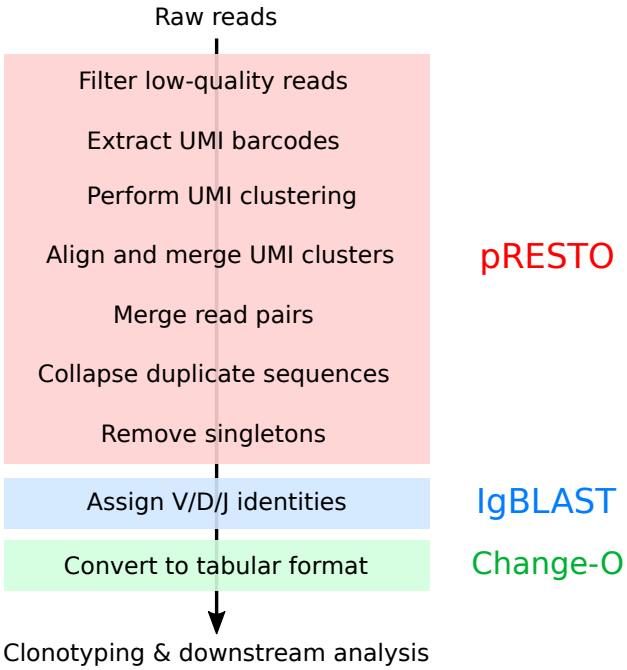


Figure 4.5: Summary of the IgSeq pre-processing pipeline for turquoise-killifish data: Schematic summary of post-sequencing analysis pipeline of killifish IgSeq data, up to and including conversion from FASTA to Change-O tabular format. Clonotyping and additional downstream analysis methods are described in Sections 2.3.5.7 and 4.4.

the TSA sequence) is extracted into a sequence annotation and trimmed from the read sequence (Section 2.3.5.2).

As discussed in Section 4.3.1, the use of UMI sequences enables biases and errors in library insert sequences to be corrected by taking the consensus sequence of all reads sharing a given UMI. However, PCR and sequencing errors can also affect the sequence of the UMI itself, in which case reads that in fact belong to a single group will be spuriously separated during pre-processing; this can result in spuriously low UMI group sizes, spuriously high numbers of unique sequences, and avoidable loss of sequencing data due to reads with erroneous barcodes being discarded (as low-quality, low-read-count unique sequences) at various points in the pre-processing pipeline [187]. In addition to these barcode *errors*, barcode *collisions* can occur, in which multiple distinct sequences are labelled with the same UMI sequence and spuriously grouped together during UMI grouping. This can lead to spuriously large UMI groups and spuriously low numbers of unique sequences, and in extreme cases can lead to the rejection and loss of entire UMI groups due to an insufficiently high level of internal sequence identity during consensus-read generation [187].

In order to reduce the effect of such barcode errors and collisions on the pre-processing pipeline, primer-trimmed forward reads in this pipeline undergo clustering following extraction of UMI sequences (Section 2.3.5.3). Firstly, reads are clustered by UMI sequence, and those with sufficiently

similar UMIs are grouped together into a single cluster even if their UMIs differ slightly. Following this, the insert sequences (as opposed to UMI sequences) of the reads in each cluster are themselves clustered, and those with sufficiently different insert sequences are split apart into separate clusters. Following these clustering steps, each cluster consists of reads with highly similar barcode sequences as well as similar insert sequences; it is then these clusters, rather than the raw UMI sequences, on which consensus-read generation is performed.

Following cluster inference as described above, annotations (including barcode and cluster annotations) are copied from each TSA-bearing forward-read to its mate among the reverse reads. The forward and reverse reads were then separately grouped by cluster identity and collapsed into a consensus sequence (Section 2.3.5.4), based on the quality score of each aligned base call at each position [160]. As most PCR and sequencing errors should be present in only a minority of the sequences descended from a given input sequence, this process effectively corrects these errors, while also removing the effect of biased amplification on observed sequence abundance. The more sequences present in a given cluster, the more effective is consensus-read generation at correcting these errors; as such, when sequencing IgSeq libraries there is a trade-off inherent in the amount of oversequencing of the molecules in each library, with more reads per molecule improving error correction but reducing the amount of useful information gained per sequencing read [103].

Following consensus-read generation, pairs of forward and reverse consensus reads with matching cluster annotations are assembled into a single contiguous sequence, ideally covering the entire variable-region sequence of the template molecule (Section 2.3.5.5); forward or reverse consensus reads lacking a mate in the other read set are discarded. At this point in the pipeline, each entry is assumed to represent a distinct RNA template molecule in the original sample. Sequences with different cluster annotations but matching insert sequences are then collapsed together into a single sequence, which is annotated with the number of contributing consensus reads (Section 2.3.5.5); each sequence entry now represents a unique sequence in the dataset. Finally, unique sequences represented by only a single read pair (which could not be corrected by consensus-read generation and are therefore highly unreliable) are discarded.

At the end of the pRESTO pre-processing pipeline, the raw data has been processed into a set of complete variable-region sequences, each of which is annotated with the number of contributing reads and the number of distinct instances of that sequence found in the dataset. These sequences are then assigned V/D/J-identities through alignment to reference databases with IgBLAST (Section 2.3.5.6) [144]. Finally, the sequences and their metadata, including annotations and V/D/J-identities, are converted by Change-O (another program from the Immcantation suite) [161] into tabular format for efficient downstream processing and analysis (Figure 4.5).

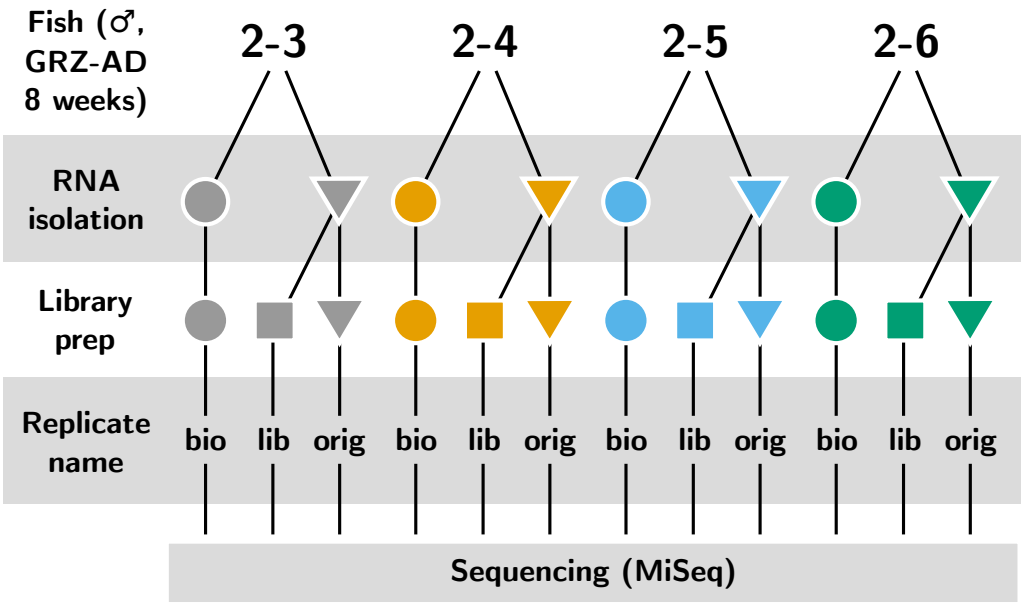


Figure 4.6: Experimental design of killifish IgSeq pilot study, showing relationship between replicates for each individual. Colour denotes individual of origin, while shape denotes replicate type.

4.4 Establishing IgSeq in the turquoise killifish: pilot study

In order to refine and functionally validate the library-preparation protocol and processing pipeline described in Section 4.3, as well as assess the state of the turquoise-killifish antibody repertoire in mature adults, I selected a group of four killifish from the second, eight-week-old sample group (specifically, fish 2-03, 2-04, 2-05 and 2-06 from Table E.24) for a pilot study. In this experiment, total RNA was isolated twice independently from each fish, and independent library preps were performed once on the first RNA isolate and twice on the second, for a total of three replicates per individual (Figure 4.6). These twelve replicates were sequenced together in a single MiSeq run, yielding a total of 13.2 million read pairs (0.8 to 1.3 million pairs per replicate, 3 to 3.5 million pairs per individual). These data were then used to develop the downstream analysis pipeline for killifish immunoglobulin-sequencing data, validate the performance and replicability of the protocol, and investigate the state of the antibody repertoire in mature adult killifish.

4.4.1 Read survival and composition

Figure 4.7 shows the absolute and relative read survival for each of the twelve replicates throughout the pre-processing pipeline, up to and including VDJ assignment and Change-O table construction. The twelve replicates show relatively consistent behaviour, with 69.8 % to 78.4 % of reads surviving the entire process. Of those that do not, the biggest losses typically occur during quality filtering and

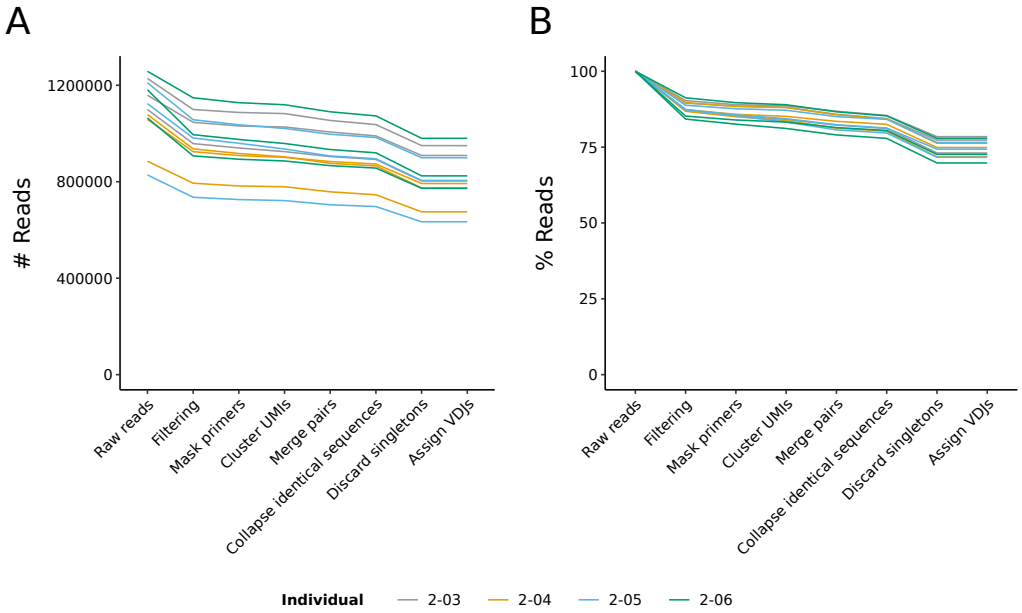


Figure 4.7: Read survival during initial pre-processing of the IgSeq pilot dataset: Line graphs of absolute (A) and relative (B) read survival during pre-processing of the IgSeq pilot dataset, up to VDJ assignment and Change-O table construction.

removal of singleton sequences, with most other steps giving rise to relatively little sequence loss (Figure 4.7B).

In total, the pre-processed sequence repertoires of the pilot replicates contain between 6000 and 12000 unique sequences per replicate, corresponding to between 21000 and 26000 unique sequences per individual killifish and 90957 unique sequences in total. Of these, 50.1 % of sequences (corresponding to 75 % of surviving sequencing reads) are annotated by Change-O as functional, meaning they have successfully been assigned V- and J-identities, their V- and J-sequences are in-frame, and they do not contain any stop codons (Figure 4.8A). A further 34.4 % of sequences (corresponding to 12.2 % of surviving reads) failed to be assigned even an uncertain J-identity, meaning that no JH sequence in the reference database aligned to the insert sequence; the remaining 15.4 % of sequences (corresponding to 12.8 % of surviving reads) have a J-assignment but are rendered nonfunctional by an internal stop codon and/or a frameshift between their V- and J-sequences (Figure 4.8A), arising during primary diversification (Section 1.2.2) or possibly SHM.

As genuinely rearranged but nonfunctional sequences would be expected to have undergone V(D)J recombination and so have a complete J-sequence, the lack of assigned J-identities for a significant minority of sequences suggests that this subset of sequences may be artefactual, erroneous or otherwise malformed. Supporting this assumption, sequences without J-assignments overwhelmingly have very low V-alignment scores reported by IgBLAST, with an average score of 24.1 $\pm$ 35.5 (mean $\pm$ standard deviation), compared to 287 $\pm$ 199 for other nonfunctional sequences and 433 $\pm$ 58.4 for

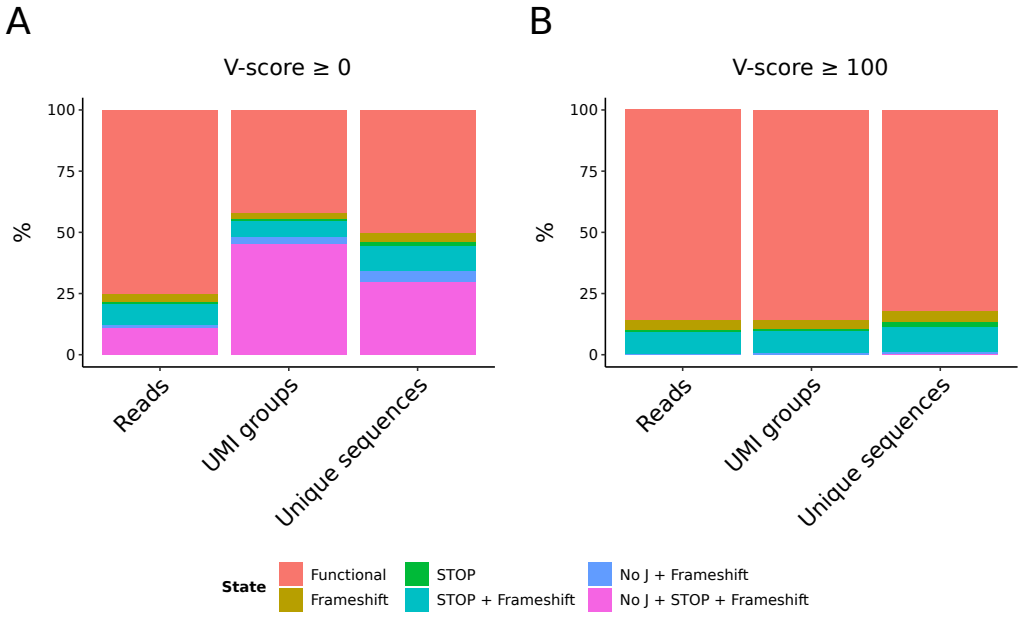


Figure 4.8: Functional composition and V-score filtering in the IgSeq pilot dataset: Proportion of input reads, UMI groups and unique sequences in the IgSeq pilot dataset belonging to different (non)functional categories, before (A) and after (B) filtering on V-alignment score.

functional sequences (Figure 4.9). A simple V-score cut-off of 100, therefore, effectively removes the vast majority of these low-quality sequences, while leaving the population of functional sequences intact (Figure 4.8B).

In total, 35703 unique sequences, corresponding to 19.4 % of input reads, were removed in this way. As a result, including this final filtering step, between 57.4 % and 71.3 % of reads per input library survived up to this point in the pipeline (Figure D.6), a good range suggesting both that the library-preparation protocol successfully captured antibody-repertoire sequencing data from turquoise killifish for the first time, and that the subsequent computational pipeline was able to completely and successfully process the majority of the resulting data. Of the remaining 55254 unique sequences present in the dataset, 45362 (82.1 %) are functional, 9334 (16.9 %) are rendered nonfunctional by a stop codon or frameshift, and only 558 (1 %) could not be assigned a J-identity (Figure 4.8B). These results confirm that the data from this first IgSeq experiment in the turquoise killifish are of sufficient quality to proceed to clonotyping and analysis of repertoire diversity.

4.4.2 Clonotyping and clonal repertoire diversity

Following assignment of VDJ identities and quality filtering, sequences in an antibody repertoire can be assigned to clones: groups of B-cells descended from a single naïve B-cell ancestor, and therefore sharing a single VDJ-recombination event. Sequences in the same clone are said to share a *clonotype*. In Change-O, clonotyping is performed by dividing sequences into groups sharing a

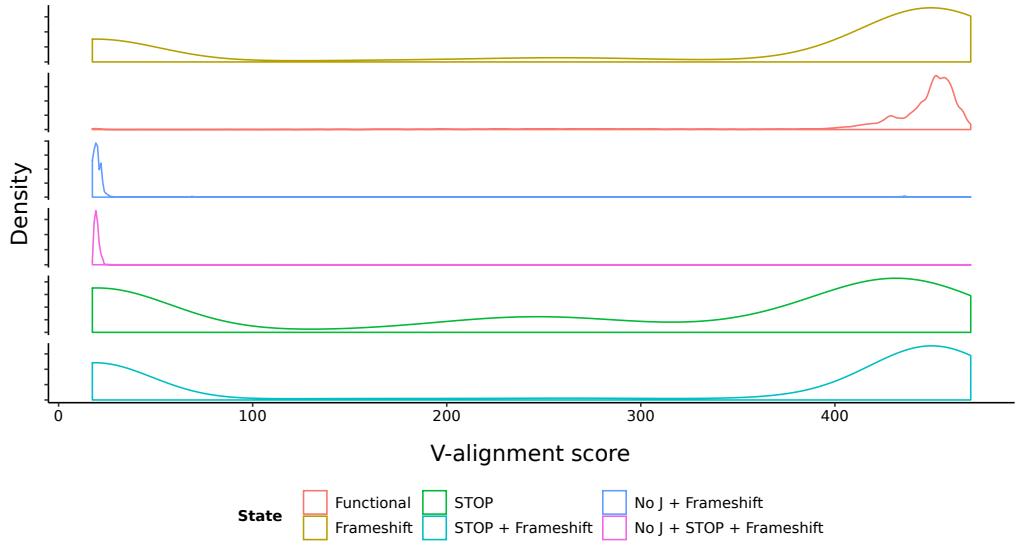


Figure 4.9: V-score distributions of functional and nonfunctional sequences: Kernel density plots of distributions of V-alignment scores among unique sequences in the IgSeq pilot dataset. Vertical axes are not to scale between sequence categories.

consistent V-assignment, J-assignment and CDR3 length, then performing single-linkage clustering on each group of sequences based on Hamming distances between CDR3 sequences [170]. To identify a distance threshold for cutting the cluster dendrogram into clones, each unique sequence in the repertoire is assigned a nearest-neighbour distance based on the length-normalised Hamming distance to the most similar sequence in the repertoire. The resulting nearest-neighbour distribution is typically bimodal, with the lower peak (representing more-similar sequences) indicating members of the same clonotype and the higher peak (representing less-similar sequences) indicating members of different clones; by fitting a pair of gamma or normal distributions to these two peaks, a distance threshold for clonotype membership can be determined according to the desired levels of sensitivity and specificity [169] (Section 2.3.5.7). By cutting the cluster dendrogram at this threshold, each group of repertoire sequences can be separated into some number of distinct clones, each of which is assumed to share a unique naïve B-cell ancestor.

One disadvantage of single-linkage clustering in this context is that non-informative `N` positions can result in artifactual links between unrelated sequences. As such, sequences with a large number of junctional `N` positions can significantly disrupt the clonotyping process. On the other hand, as over 9 % of unique sequences in the pilot dataset contain at least one such junctional `N` position, excluding them all would represent a significant loss of data. In order to minimise the number of discarded sequences while also minimising the disrupting effects of sequences with junctional `N`s, sequences with exactly one junctional `N` position (comprising 6.13 % of total sequences and 67.9 % of sequences with at least one junctional `N`; Table 4.2) were included in the clonotyping process for the pilot dataset, while

Table 4.2: Distribution of junctional `N` positions in the V-score-filtered pilot dataset

# junctional Ns	# unique sequences	% of all sequences	% of sequences with >0 junctional Ns
0	49134	88.924	0.00
1	3388	6.132	67.94
2	961	1.739	19.27
3	324	0.586	6.50
4	125	0.226	2.51
5	93	0.168	1.86
>5	96	0.174	1.93

those with two or more junctional `N` positions were excluded. This procedure successfully assigned clonal identities to 95.2 % of unique sequences in the V-score-filtered dataset, with fewer than 3000 sequences (corresponding to 0.6 % of input reads) lost during the clonotyping phase of the pipeline.

In total, 4800 to 6400 clones were identified per individual fish in the pilot dataset. As expected, the clone-size distribution is overwhelmingly dominated by small clones (Figure 4.10A): across all individuals, 57.5 % of clones are observed as just a single unique sequence across all replicates, while 93.2 % contain fewer than five unique sequences. As a result, the great majority of clones (90.34 %) are observed in only a single replicate per individual, with 7.22 % present in two replicates and only 90.34 % shared across all three. Unsurprisingly, however, larger clones are much more likely to be shared across multiple replicates (Figure 4.10B), consistent with a model in which a very large number of small clones are sampled only rarely while a much smaller number of large clones is sampled much more often [30]. Overall, the level of agreement between the replicates is high (Figure 4.11), with an average inter-replicate correlation in clone size of $r = 0.89$, indicating that, despite the problem of undersampling many very small clones, the clonal composition of the killifish can be captured reproducibly by immunoglobulin sequencing.

One of the most strikingly reproducible findings of immune-repertoire-sequencing studies has been the approximately power-law distribution of the clonal repertoire, a phenomenon observed in both antibody and TCR repertoires across multiple different species [30, 189]. More precisely, the frequency of the kth largest clone in a repertoire dataset containing N total clones can often be roughly predicted by a Zipf distribution [190] of the form:

$$f(k|N,s) = \frac{k^{-s}}{H_{N,s}} \tag{4.1}$$

for some exponent parameter $s > 0$, where $H_{N,s}$ is the Nth generalised harmonic number of order s (Section 2.3.6.2). This power-law distribution of clone sizes is inconsistent with clonal expansion under neutral selection and instead suggests that clone sizes evolve within a complex and fluctuating fitness landscape [189]. Before going on with more detailed investigation of the diversity structure

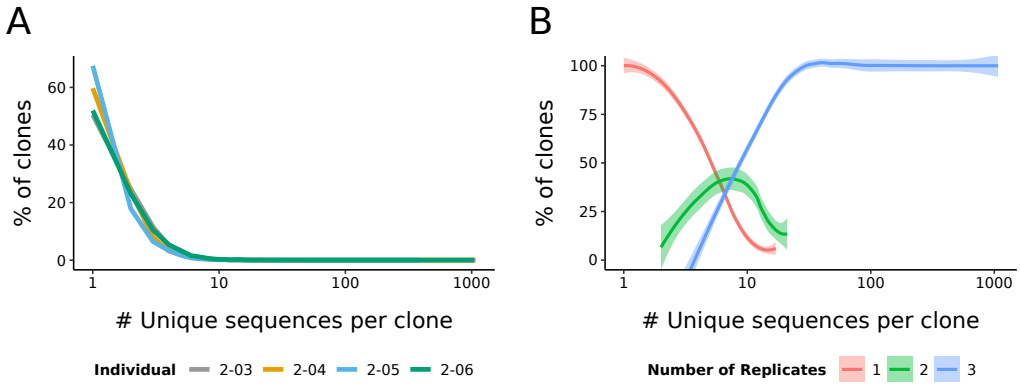


Figure 4.10: Clone size and cross-replicate reproducibility in the IgSeq pilot dataset: (A) Proportion of clones of different sizes for each individual in the pilot dataset, measured in unique sequences per clone. (B) LOESS-smoothed curves [188] showing proportion of clones of each size found across one, two or all three replicates of the appropriate individual.

of turquoise killifish antibody repertoires, it would be interesting to test whether the phenomenon of roughly Zipf-distributed clonal frequencies persists in this species as well.

Figure 4.12 shows the rank/frequency distributions of the clonal repertoires of the four individuals from the pilot dataset. From roughly the tenth-largest clone onwards, these distributions appear approximately linear on a log-log plot and can be reasonably approximated by a power-law distribution; however, the largest clones in each repertoire clearly deviate from this pattern, and are much larger than would be predicted by a power-law approximation. In order to fit power-law curves to these distributions, I performed simple maximum-likelihood estimation of the Zipf exponent for each individual repertoire in R (Section 2.3.6.2). When the five largest clones in each repertoire were excluded from the calculation, the resulting Zipf distributions (Figure 4.13A) provided a good approximation of the remaining points and were highly consistent across individuals, with exponents ranging from 0.525 to 0.542. Conversely, when the largest clones are included, the resulting Zipf approximations follow the line of the remaining clones much less accurately in several individuals, and their exponents range more widely from 0.566 to 0.694. Hence, while the bulk of the clonal repertoire in turquoise killifish corresponds to the expected power-law behaviour seen in other species, the largest clones, for unknown reasons, consistently deviate strongly from this pattern.

In addition to the exponent of the best-fit Zipf distribution, another quick and simple way to measure the extent to which a clonal repertoire is dominated by the largest clones is to measure the P20, the proportion of all unique sequences in the repertoire belonging to the 20 largest clones; in humans, for example, the P20 is typically a few percent in healthy individuals, but can reach up to 90 % in patients with B-cell malignancies [191]. In the killifish pilot dataset, the P20 of the clonal repertoires ranges from 9.9 % to 21.6 % (Figure 4.14), much higher than that observed in healthy humans. These high P20 values are primarily due to the highly expanded state of the largest few

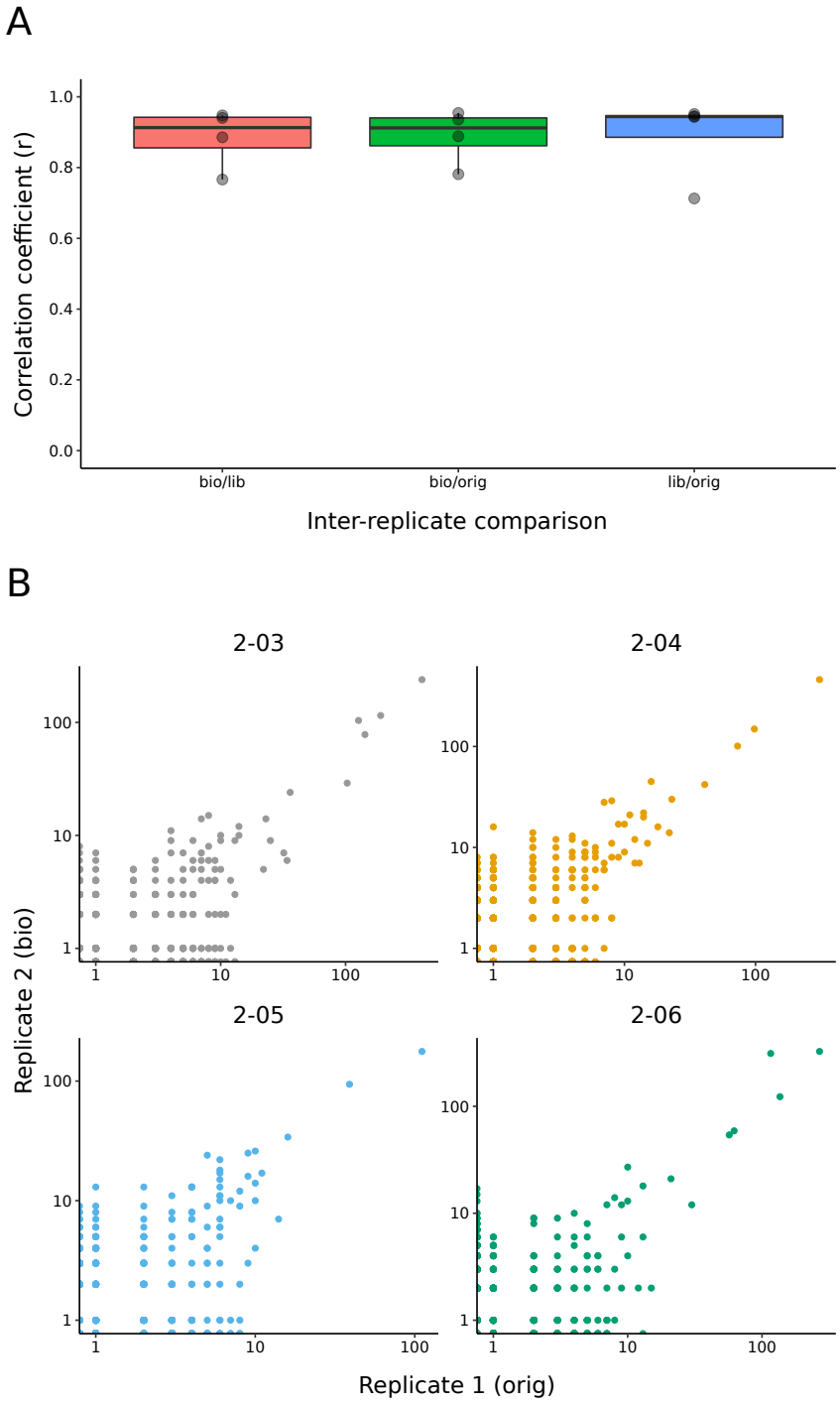


Figure 4.11: Clone-size correlation between pilot replicates: (A) Boxplots showing distribution of Pearson product-moment correlation coefficients of clone sizes for each pair of replicates, as measured by the number of unique sequences per clone in each replicate. (B) Scatter plots comparing clone sizes across biological replicates for each individual in the pilot dataset. In both subfigures, clones absent in a given replicate were given a size of zero.

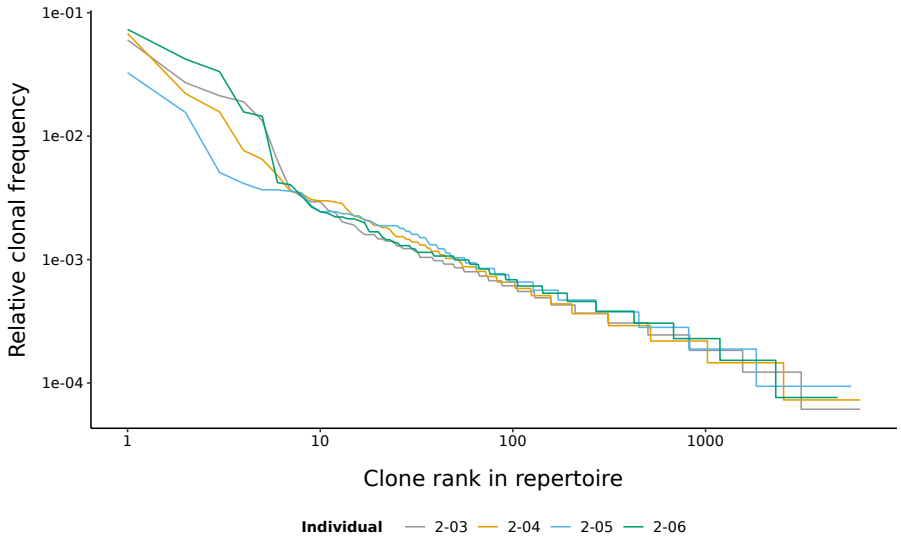


Figure 4.12: Rank/frequency distributions of pilot clonal repertoires: Log-log line plots of rank/frequency distributions of the individual clonal repertoires in the IgSeq pilot dataset, showing the roughly linear (i.e. power-law-distributed) distribution of smaller clones (roughly rank 10 onwards) and the clear deviation from this distribution in the largest clones.

clones in each repertoire; when the expected clonal frequencies from the Zipf distributions fitted in Figure 4.13A are used instead of the actual values, the P20 values fall to between 5.8 % and 6.7 % (Figure 4.14).

In addition to investigating the P20, Rosenfeld *et al.* [191] recommend evaluating clonal expansions in antibody repertoires on the basis of both the relative clonal frequency of the largest clones in the repertoire and the extent to which each such clone is larger than the next-largest clone, with recommended cutoffs for human data of 5 % for the former and threefold for the latter. In the pilot killifish dataset, clones exceeding 5 % of unique sequences in the repertoire occur in three out of four individuals (2-03, 2-04 and 2-06), while clones exceeding the size of the next-largest clone by at least threefold occur in a different three individuals (2-03, 2-05 and 2-06). However, only one individual, 2-04, exhibits a potentially medically-relevant clonal expansion by the standards of Rosenfeld *et al.*; this clone (2-04_8143) occupies 6.8 % of the repertoire and is roughly 3.1-fold larger than the next-largest clone, and is identified reproducibly as a clonal expansion in both the pooled dataset (Figure 4.15) and each of the separate replicates (Figure D.7) from individual 3-04. These results could be taken as demonstrating that potentially medically-relevant clonal expansions are present, albeit not common, in middle-aged adult male turquoise killifish; however, without functional validation it is not clear to what extent the recommended guidelines derived from human data apply to this species, especially when the body size (and hence the potential clonal richness) of turquoise killifish is so much smaller.

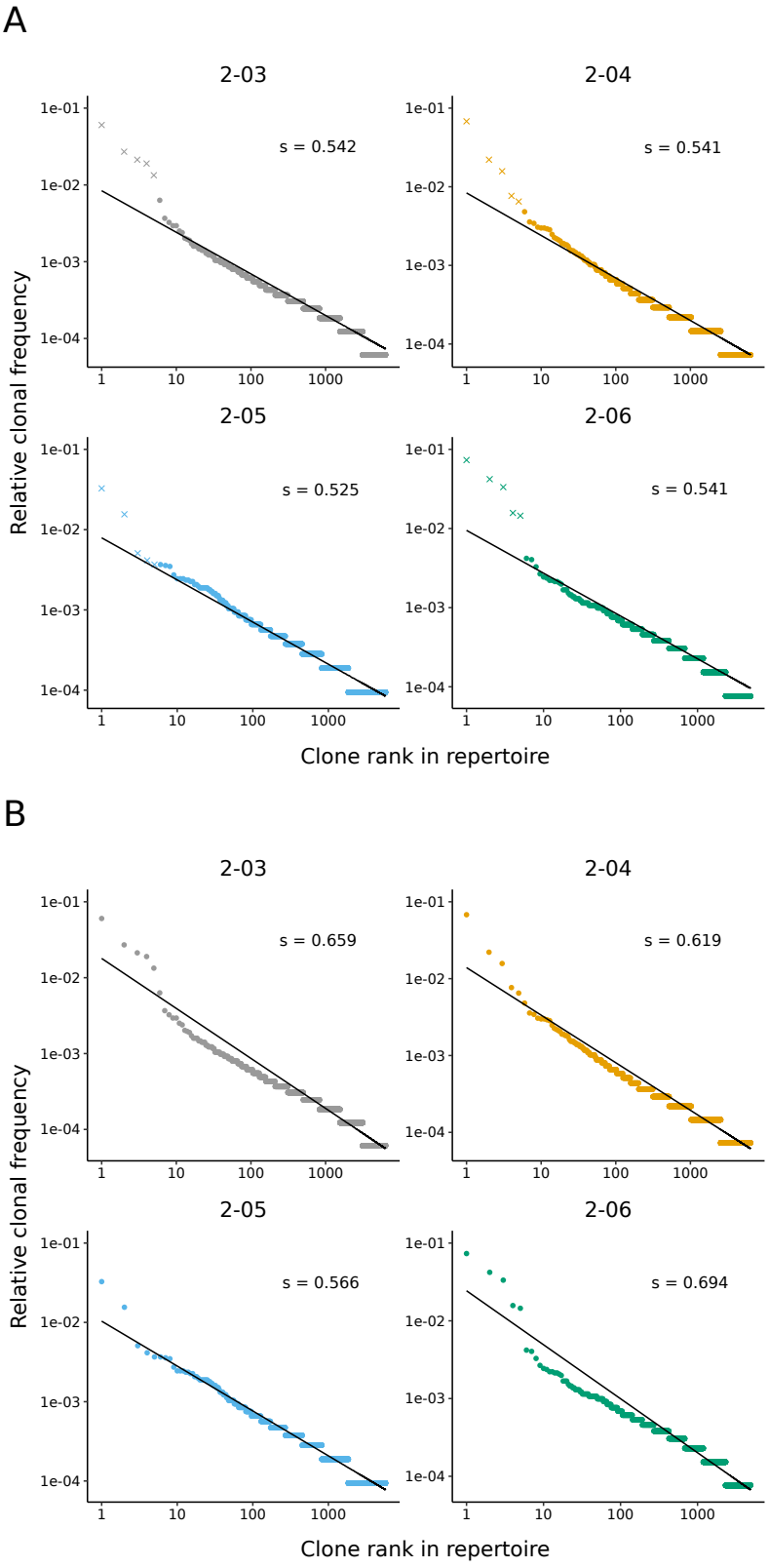


Figure 4.13: Best-fit Zipf distributions of pilot clonal repertoires: Log-log scatter plots of the rank/frequency distributions for each clonal repertoire in the pilot dataset, overlaid with a maximum-likelihood estimate of the underlying Zipf distribution computed (A) with the five largest clones excluded or (B) with all clones included. In (A), the clones excluded from the Zipf estimation are plotted as crosses.

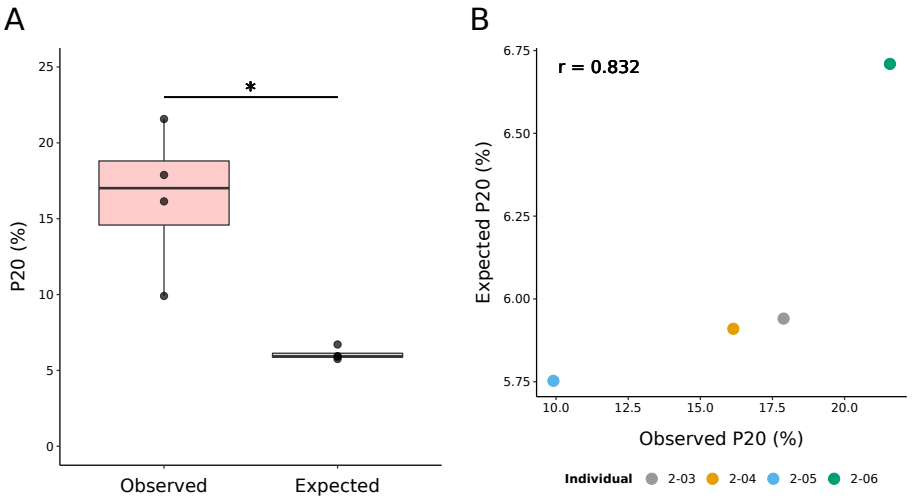


Figure 4.14: Clonal P20 values in *N. furzeri* pilot repertoires: (A) Boxplots of observed and expected (from the Zipf distributions fitted in Figure 4.13A) P20 distributions of the clonal repertoires from the IgSeq pilot dataset. Pairwise p-value computed using nonparametric Mann-Whitney U test ($* : 0.01 < p \leq 0.05$; $** : 0.001 < p \leq 0.01$; $*** : p \leq 0.001$) . (B) Scatter plot comparing observed vs expected P20 for each individual, annotated with the correlation (r) between the two sets of P20 values.

Having investigated the clonal abundance distribution and various measures of clonal expansion, we now come to the question of the *diversity* of the killifish clonal repertoire. As discussed in detail in Appendix C, the diversity of a population can be interpreted – and measured – in multiple ways, with different measures laying different amounts of emphasis on the richness (number of species) and evenness (relative distribution of species) of that population and giving different degrees of consideration to common vs rare species (Appendix C.1.2). In order to visualise many different aspects of repertoire diversity simultaneously and on a common scale, Hill diversity spectra [192] can be used. For any population X of elements (or *individuals*) divisible into some disjoint set of categories (or *species*) S, the *Hill numbers* or "true diversities" of that population (Appendix C.1.3) [192, 193] are given by

$$ {}^{q}D(X) = \begin{cases} \left(\sum_{s \in S} p_s^q \right)^{\frac{1}{1-q}} & q \neq 1 \\ \exp\left(-\sum_{s \in S} p_s \cdot \ln p_s \right) & q = 1 \end{cases} \tag{4.2} $$

where p_s is the proportion of individuals in X belonging to a species s. The parameter q in this equation denotes the *order* of diversity: the greater the order, the more rare species are downweighted compared to common ones when computing the diversity of the population. By plotting ${}^{q}D$ as a function of q across a wide range of diversity orders, we can therefore observe how the measured diversity of the population changes as we put progressively more or less emphasis on species of

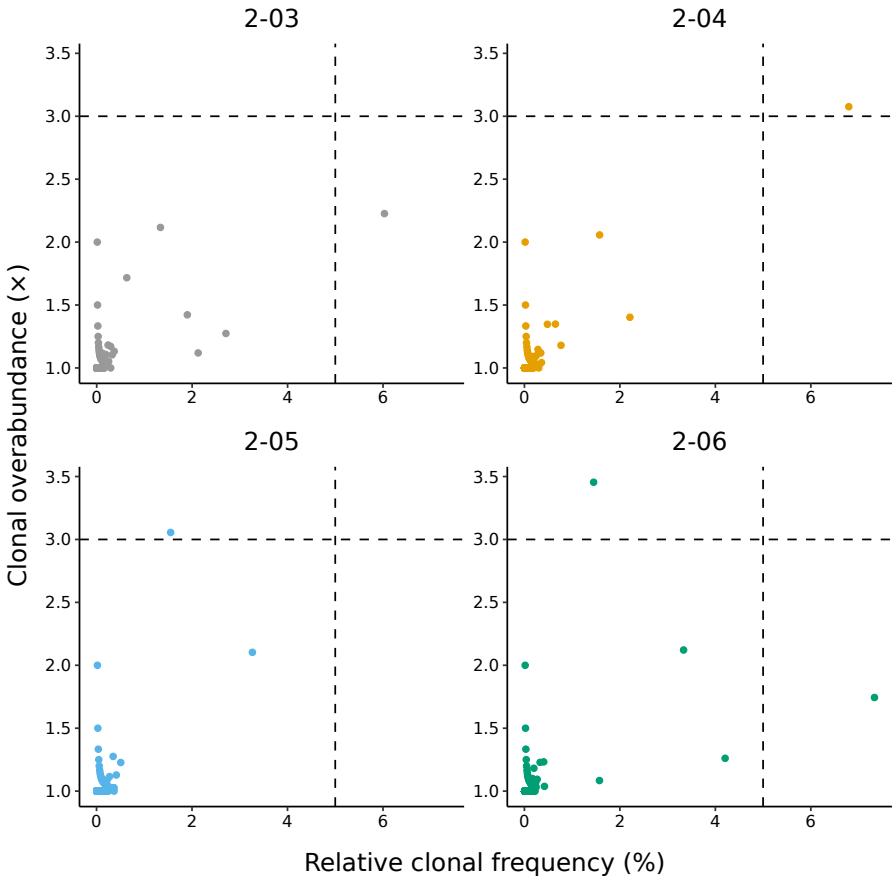


Figure 4.15: Clonal expansions in *N. furzeri* pilot repertoires: Scatter plots of clonal abundance for each individual in the IgSeq pilot dataset, measured in terms of the proportion of unique sequences in the repertoire (*x*-axis) and the abundance relative to the next-largest clone (*y*-axis). Thresholds for identifying clonal expansions (5 % and 3-fold for the *x*- and *y*-axis, respectively) suggested by Rosenfeld *et al.* [191].

different abundances, and so simultaneously visualise many different aspects of population diversity (Appendices C.1.3 and C.2.2).

Figure 4.16 shows the diversity of clone sizes in each individual in the pilot dataset, as measured using Hill diversity spectra across each individual's three replicates (Figure 4.6). Figure 4.16A gives the alpha diversity, or average *within-replicate* diversity, while Figure 4.16B shows the beta diversity, or variation in clone-size distribution *between replicates* for each individual (Appendices C.2.1 and C.2.2). In both cases, each curve represents a single individual and gives the corresponding diversity measure at a range of different diversity orders. As beta diversity (unlike alpha diversity) is sensitive to the number of sub-populations being compared, it has been rescaled here such that 0 represents the minimum possible beta diversity at each order and 1 represents the maximum (Appendix C.2.3).

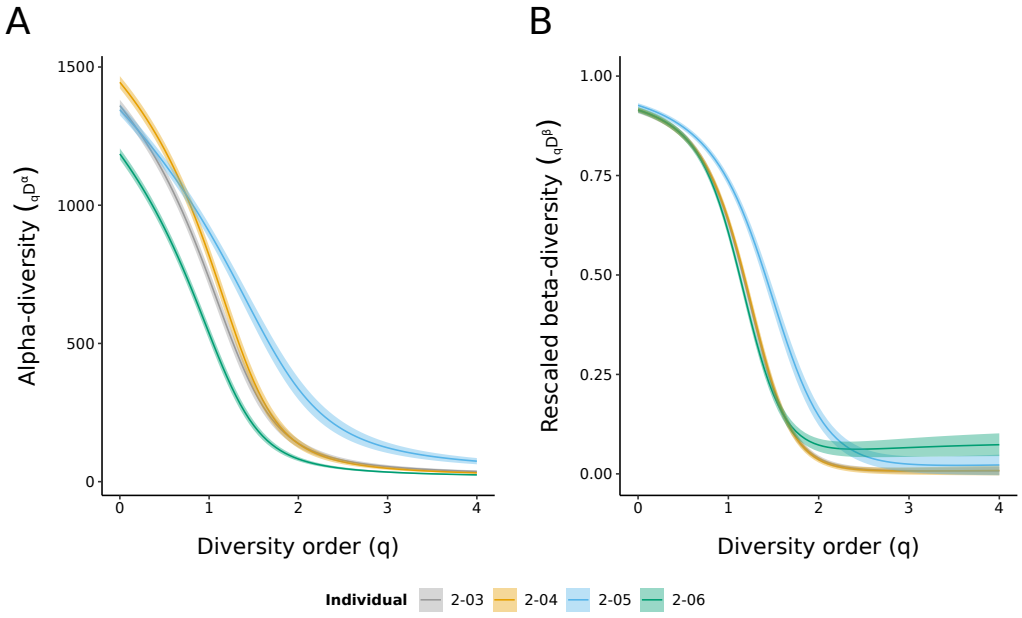


Figure 4.16: Clonal-diversity spectra for the IgSeq pilot dataset: Bootstrapped Hill diversity spectra of clone sizes (as measured by number of unique sequences per clone) over replicates for each individual in the pilot dataset. (A) alpha diversity across replicates; (B) beta diversity across replicates, rescaled to between 0 (minimum) and 1 (maximum) for each individual. Shaded regions in both subfigures represent 95 % confidence intervals, estimated using bootstrapping.

The results from the alpha diversity spectrum (Figure 4.16A) suggest there may be significant differences in diversity between individuals in middle-aged killifish: in particular, individual 2-06 appears to be less diverse than the other individuals across the whole range of diversity orders (suggesting that its clonal repertoire both contains fewer clones overall and is more uneven in its clone sizes) while individual 2-05 appears to be more diverse at higher diversity orders (suggesting that its clonal repertoire is of similar richness to the other individuals, but more even). This would accord with previous measures of P20 and Zipf exponent in these repertoires (Figures 4.13B and 4.14), both of which suggest that the clonal repertoire of 2-06 is substantially more dominated by expanded clones than the other individuals in the dataset, and that of 2-05 less so.

By-eye comparisons of apparent diversities, however, are not sufficient for concluding that a significant difference in diversity exists between sample groups. Ideally, the entire shape of the diversity spectrum would be compared between groups to test for a significant difference in distribution; however, the unusual and nonparametric nature of the Hill spectrum makes such a holistic comparison difficult, and established tests for such a difference do not, to my knowledge, yet exist. However, it is possible to use standard statistical methods to test for a difference in Hill diversity at one or more specific diversity orders, and so to obtain a partial overview of significant differences for particular aspects of each group's diversity profile.

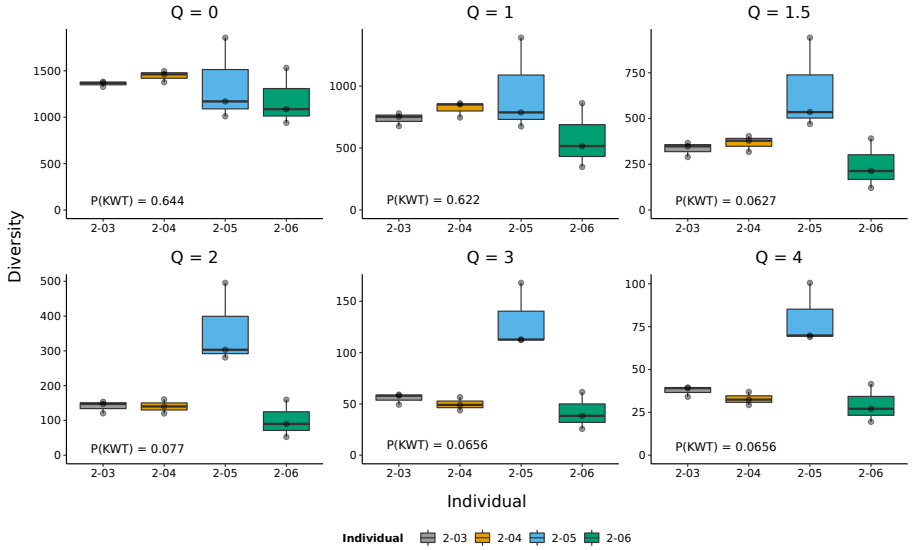


Figure 4.17: Comparing clonal alpha diversities between individuals in the IgSeq pilot dataset: Boxplots of Hill diversity values for the clonal repertoires of replicates from each individual in the IgSeq pilot dataset at a sample of diversity orders. Pairwise p-values were computed using nonparametric Mann-Whitney U tests ($*: 0.01 < p \leq 0.05$; $**: 0.001 < p \leq 0.01$; $***: p \leq 0.001$). Annotated p-values ($P(KWT)$) indicate the statistical significance of the estimated individual effect on diversity at each order under a Kruskal-Wallis test.

To that end, I computed the clonal diversity spectra separately for each replicate in the pilot dataset (Figure D.8), and extracted the distribution of diversity values obtained for each individual at each of six diversity orders (0, 1, 1.5, 2, 3 and 4). I compared these distributions pairwise using Mann-Whitney U nonparametric tests, as well as performing a Kruskal-Wallis nonparametric analysis-of-variance test for an effect of source individual on repertoire diversity. Neither of these tests returned a significant result for any of the diversity orders tested (Figure 4.17). From the data available, therefore, it is not possible to conclude that clonal repertoires differ significantly between individuals in middle-aged adult male turquoise killifish, though given the relatively low Kruskal-Wallis p-values obtained at higher diversity orders (Figure 4.17) it is possible such a difference would be observed if more replicates per individual were available.

Unlike alpha diversity, the beta diversity spectrum of a population does not necessarily decline monotonically with increasing diversity order; nevertheless, in Figure 4.16B the between-replicate beta diversity is much higher at low diversity orders (where it approaches the maximum) than at higher ones (where it is close to the minimum). This indicates that replicates from the same individual have very different clonal content when all clones are included, but become increasingly similar as more and more weight is put on the largest clones in each replicate. This result is consistent with the findings in Figure 4.10 that each repertoire contains a small number of large clones (which are shared reproducibly between replicates) and a much larger number of much smaller clones (which are not);

the difference in patterns of cross-replicate reproducibility between small and large clones observed in Figure 4.10B gives rise to the patterns of beta diversity observed in Figure 4.16B.

Finally, given the apparent correspondence between a high P20 or fitted Zipf exponent and low effective diversity as measured using Hill spectra, I investigated to what extent Hill numbers can substitute for these metrics as a measure of repertoire diversity. To that end, I fitted Zipf distributions separately to the clonal repertoires of each replicate in the dataset (both including all clones and excluding the five largest in each repertoire, as above) as well as computing the P20 values for these repertoires. I then computed Pearson product-moment correlation coefficients between these metrics and Hill diversity at each diversity order, enabling correlation spectra to be plotted for each of the three potential proxy metrics. I did this both for the correlation between per-replicate proxy metrics and per-replicate repertoire diversity (Figure 4.18A), and for the correlation between per-individual averaged metrics and per-individual alpha diversity (Figure 4.18B).

At the per-replicate level, the correspondence between Hill diversity and filtered Zipf exponent was fairly poor at all diversity orders, while the all-clones Zipf exponent and P20 both correlated relatively poorly with Hill diversity at very low orders but correlated well with higher-order diversity measures ($q \geq 1$), with P20 consistently outperforming either Zipf exponent as a predictor of Hill diversity; the best correlation of all was found for P20 at $q = 1.45$, with a correlation coefficient of -0.86. When comparing averaged metrics with average (alpha) per-individual diversity, the pattern for the all-clones Zipf exponent and P20 is broadly similar, albeit reaching much lower correlation values: at $q = 1.3$, the average P20 approaches a near-perfect correlation of -1. Conversely, the filtered Zipf exponent behaves very differently in the average case compared to the per-replicate case, with even worse correlation at very low diversity orders but improving greatly at higher orders, to the point where, from $q = 1.75$ upwards, it actually correlates better with Hill diversity than either P20 or the all-clones Zipf exponent, with an optimum correlation of -0.98 at $q = 2.35$.

It therefore appears that P20 and Zipf metrics of clonal alpha diversity can be well-predicted by higher-order Hill-diversity measures, with the closest correspondence found at diversity orders between 1 and 1.5 (for P20 and all-clones Zipf exponent) or 2 and 2.5 (for the filtered Zipf exponent). This finding aptly demonstrates the earlier point that, while most measures of repertoire diversity capture only one aspect of a population's species composition, the Hill spectra can simultaneously depict a whole range of different aspects of population diversity and so substitute for many different diversity measures at once. In addition to high-order metrics of clonal diversity like P20 or the filtered Zipf exponent, the Hill spectrum also captures lower-order metrics like species richness, providing a valuable overview of the entire diversity structure of the clonal repertoire.

In conclusion, therefore, adult turquoise killifish express diverse clonal B-cell repertoires, with each fish containing several thousand detected clones. As in other species, this repertoire comprises a few very large expanded clones and a much larger number of very small clones (Figure 4.10), the latter of which are highly vulnerable to undersampling; it is therefore likely that, as in most

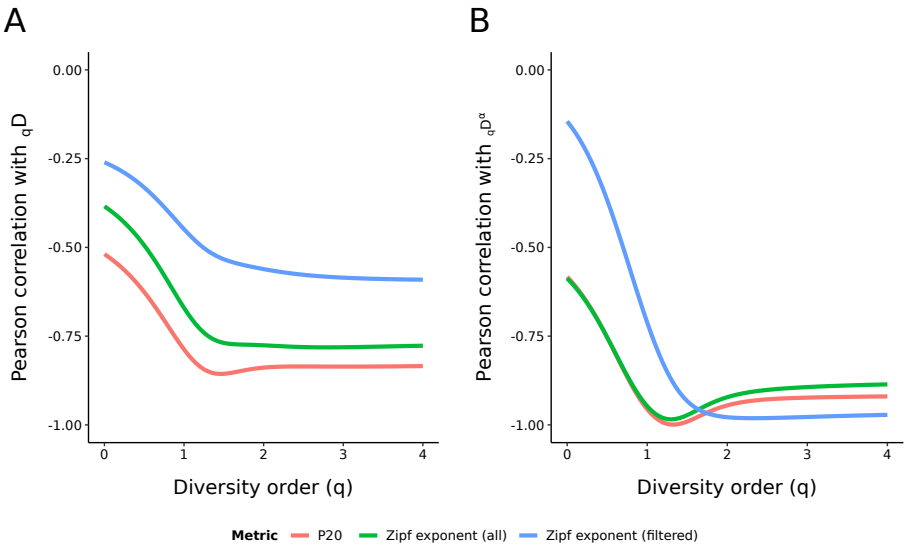


Figure 4.18: Correlation between Hill diversity and proxy metrics of clonal expansion: Pearson product-moment correlation coefficients between clonal Hill diversity and various proxy metrics (P20, fitted Zipf exponent over all clones, fitted Zipf exponent excluding largest clones) over a range of diversity orders for the IgSeq pilot dataset. (A) Per-replicate correlation between proxy metrics and per-replicate diversity. (B) Per-individual correlation between mean proxy metrics (averaged over replicates for each individual) and alpha diversity.

repertoire-sequencing experiments [30], the total clonal richness of the repertoire is substantially higher than the richness measures used here suggest. Whether or not the number, size composition or diversity of clones in the killifish antibody repertoire change as a function of age, however, cannot be inferred from this initial pilot dataset.

4.4.3 V(D)J segment usage and segment-repertoire diversity

The clonal repertoire of an organism reflects the history of naïve B-cell production and subsequent clonal expansion in that individual, and as such is unique to each individual repertoire: by definition, clones cannot be shared between individuals. As such, while an essential part of any repertoire analysis, clonal measures are limited in their ability to meaningfully compare repertoire composition between different individuals. In contrast, the range of V(D)J-combinations available within an antibody repertoire is defined by the corresponding gene locus (Section 3.2), and is therefore largely shared across individuals of a given species. This is particularly true for inbred lines of laboratory model organisms, for which the high level of polymorphism observed in the V-regions of wild populations [194, 195] is not an issue. As such, the V(D)J usage distributions of antibody repertoires represent an alternative metric for measuring repertoire composition and diversity, which is more amenable to comparison between individuals (and groups of individuals) of the same species.

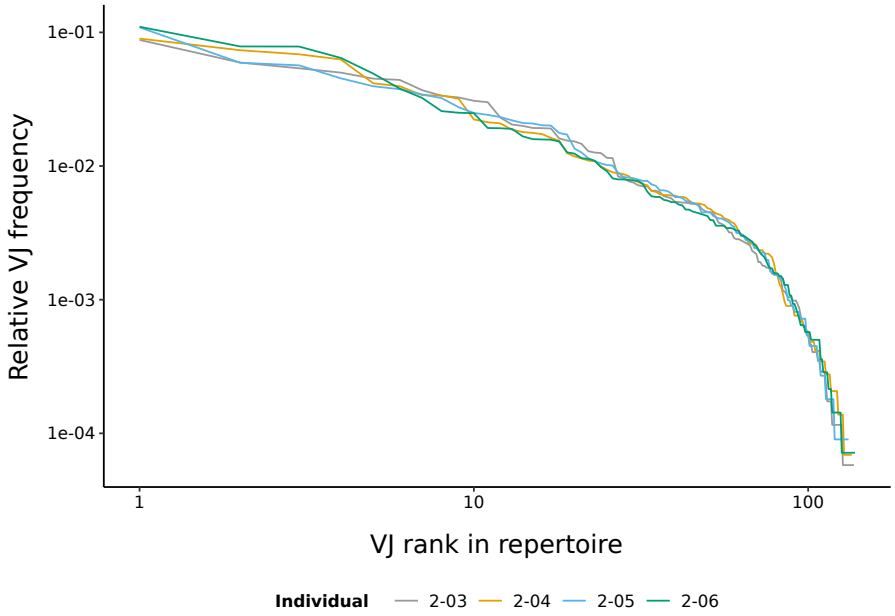


Figure 4.19: Rank/frequency distributions of pilot VJ-repertoires: Log-log line plots of rank/frequency distributions of the individual VJ-repertoires in the IgSeq pilot dataset, showing the rapid decline in frequency of the smallest V/J-assignment categories and the resulting clear deviation from a power-law distribution.

In order to analyse the composition of the V(D)J segment repertoire in an individual, I needed unambiguous segment calls for as many unique sequences in that repertoire as possible. In the processed pilot dataset, 98.1 % of unique sequences (corresponding to 99.6 % of surviving input reads) were assigned an unambiguous VJ-identity, while only 68.7 % of unique sequences (corresponding to 70.3 % of surviving input reads) were assigned an unambiguous VDJ-identity. This difference arises from the fact that 31 % of unique sequences (corresponding to 29.7 % of surviving input reads) were assigned either no D-identity at all, or an ambiguous one with two or more possible D-assignments. To avoid any distortions or loss of resolution caused by the loss of almost a third of the dataset, I therefore used the VJ-repertoire to investigate segment-choice diversity in killifish antibodies.

Figure 4.19 shows the rank/frequency distribution of the VJ-repertoire for each individual in the pilot dataset. Unlike with the clonal repertoire, these distributions are clearly not well-modelled by a power law even when the largest V/J combinations are excluded; in particular, the frequency of smaller V/J combinations declines much more quickly than would be predicted by Zipf's law. These VJ-repertoires are very strongly dominated by the largest combinations in each individual, with P20 values ranging from 66.6 % to 68.8 %. In total, of the 288 possible theoretically distinguishable V/J combinations in the turquoise-killifish repertoire, 154 (53.5 %) are observed in at least one individual, with the number observed in any single individual ranging from 132 (45.8 %) to 138 (47.9 %).

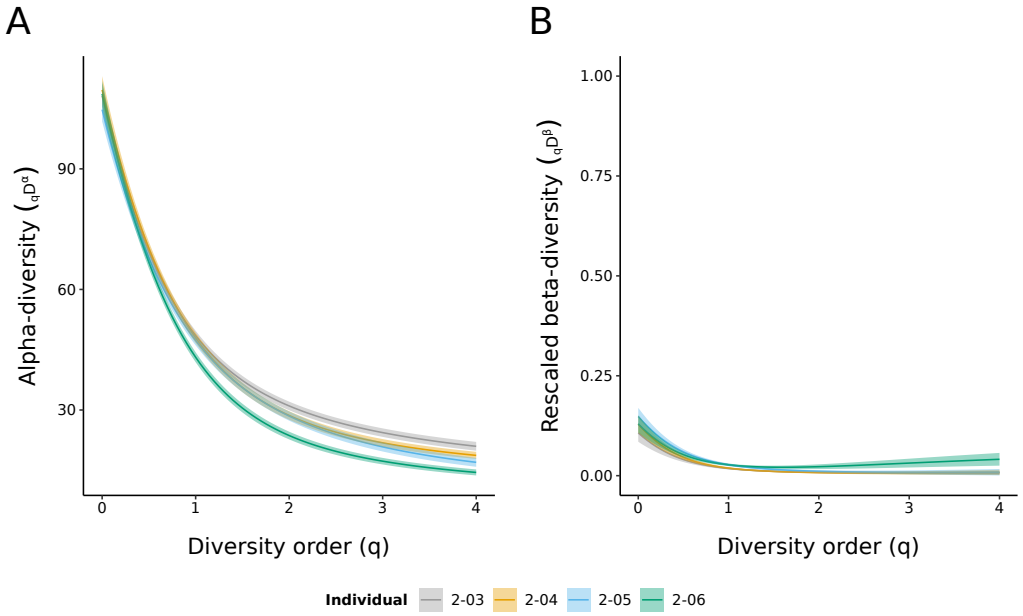


Figure 4.20: VJ-diversity spectra for the IgSeq pilot dataset: Bootstrapped Hill diversity spectra of VJ usage (as measured by number of unique sequences per unambiguous V/J combination) over replicates for each individual in the IgSeq pilot dataset. (A) alpha diversity across replicates; (B) beta diversity across replicates, rescaled to between 0 (minimum) and 1 (maximum) for each individual. Shaded regions in both subfigures represent 95 % confidence intervals, estimated using bootstrapping.

As with the clonal repertoire, the diversity of the VJ-repertoire can be assessed in a wide variety of ways, among the most flexible and informative of which is the Hill diversity spectrum (Appendix C). Figure 4.20 shows the alpha- (Figure 4.20A) and beta- (Figure 4.20B) diversity spectra of the VJ-repertoire for each individual in the pilot dataset. Unsurprisingly, the alpha-diversity measures of the VJ-repertoire were much lower than for the clonal repertoire, reflecting the relatively restricted range of possible V/J choices compared to the huge number of total naïve B-cells. As with the clonal repertoire, there were no significant pairwise differences in alpha diversity between individuals, although in this case a Kruskal-Wallis test was able to find a marginally significant overall effect of individual identity on VJ diversity at high diversity orders (Figures 4.17, D.8 and D.9).

The beta-diversity spectra, meanwhile, are close to the minimum value across all diversity orders, indicating that the V/J usage distribution in each individual's secondary repertoire is highly consistent across replicates for both large and small V/J combinations. To verify that this low beta diversity is the result of similarity between replicates from the same individual, rather than homogeneity in V/J usage between individuals, I made use of the Repertoire Dissimilarity Index (RDI) [173], a method for computing a distance between any two repertoires based on the Euclidean distance between their respective V(D)J-abundance vectors (Section 2.3.6.5). Average-linkage clustering on the basis of the VJ-RDI distances between replicates in the pilot dataset correctly re-groups all replicates from all

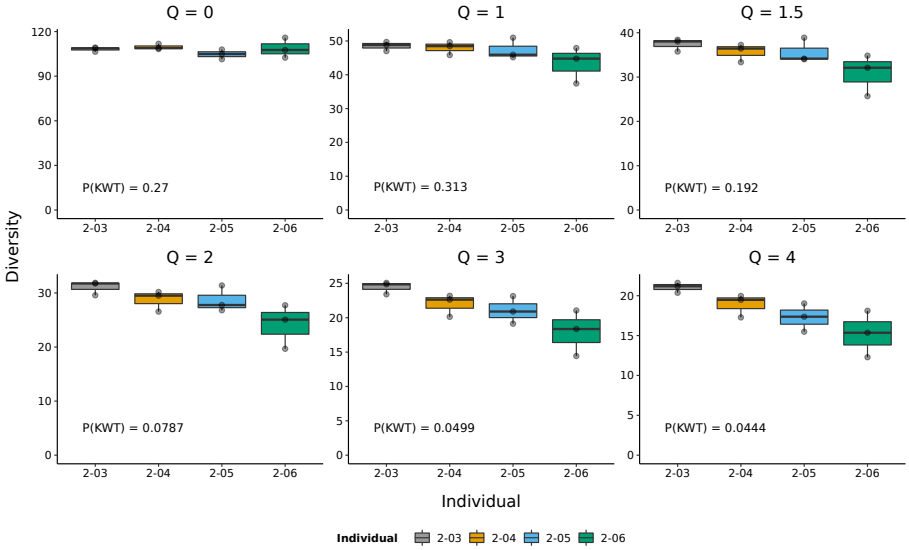


Figure 4.21: Comparing VJ alpha diversities between individuals in the IgSeq pilot dataset: Boxplots of Hill diversity values for the VJ-repertoires of replicates from each individual in the IgSeq pilot dataset at a sample of diversity orders. Pairwise p-values were computed using nonparametric Mann-Whitney U tests ($* : 0.01 < p \leq 0.05$; $** : 0.001 < p \leq 0.01$; $*** : p \leq 0.001$). Annotated p-values ($P(KWT)$) indicate the statistical significance of the estimated individual effect on diversity at each order under a Kruskal-Wallis test.

four individuals (Figure 4.22A), demonstrating that the information in the VJ-repertoire is sufficient to distinguish repertoires from different individuals.

4.4.4 Generative models and potential repertoire entropy

Sections 4.4.2 and 4.4.3 analysed the clonal and VJ-repertoires of adult killifish as actually observed in the pilot dataset. For various reasons, including primary selection, clonal expansion, and selection during affinity maturation, this observed secondary repertoire will differ in its statistical properties (V/J prevalence, CDR3 length, etc.) from the original generative process giving rise to *IGH* sequences in the primary lymphoid organs [38]. In order to investigate the statistical makeup of this generative repertoire, it is necessary to bypass these distortions and infer a model for the primary generative process itself.

To investigate these generative repertoires in turquoise killifish, I made use of IGoR, a tool for inferring generative models of the VDJ-recombination and junctional-diversity process from immune-repertoire-sequencing data [48]. This program exploits the fact that B-cells which fail their first attempt at VDJ-recombination but suceed in their second retain their first, nonfunctional, sequence rearrangement, which can then be detected at low levels in the repertoire. As these nonfunctional sequences are not used to produce BCR or antibody proteins, they do not take part in primary

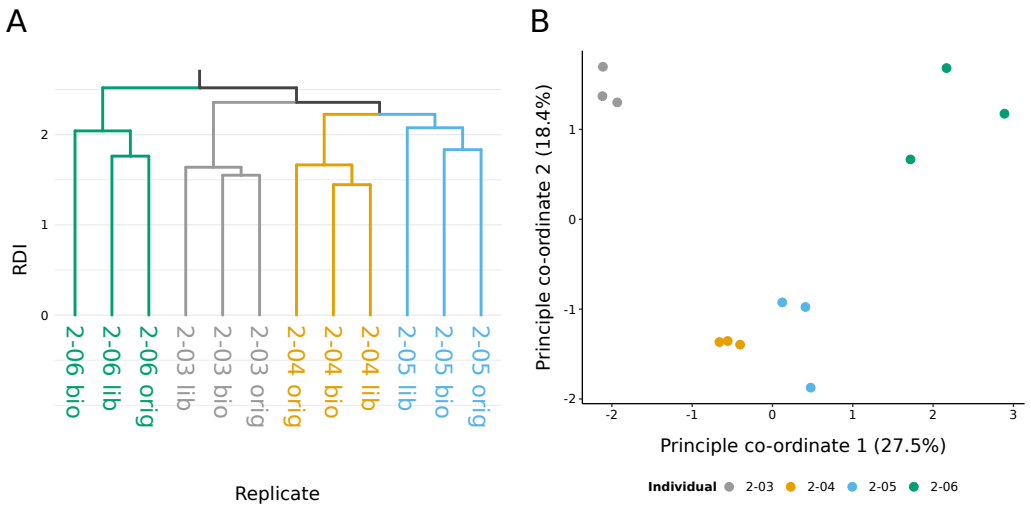


Figure 4.22: Repertoire dissimilarity index (RDI) analysis of IgSeq pilot repertoires: (A) Dendrogram of pilot replicates produced through average-linkage clustering on pairwise VJ-RDI distances. (B) Principal co-ordinate analysis (PCoA) of pairwise VJ-RDI distances between pilot replicates, coloured by source individual.

selection; as a result, unlike the functional repertoire, the statistical properties of the nonfunctional sequence population are not systematically altered by this selective process [38]. By using only nonfunctional repertoire sequences, therefore, IGoR can reconstruct the original parameters of the sequence-generation process [48].

In order to obtain something as close as possible to naïve, nonfunctional sequences from the pilot repertoire dataset, I first took the consensus sequence of each clone in each individual repertoire, then filtered these by functional status to retain only sequences marked by Change-O as nonfunctional (Section 2.3.6.6). This yielded a total of 4893 sequences, or 1067 to 1484 per individual repertoire. Unfortunately, the numbers of sequences obtained here per individual are not sufficient to infer a reliable generative model with IGoR [48, 196]; however, as the parameters of the generative process tend to be highly similar between individuals (and would be expected to be even more so in a genetically-homogeneous inbred line), pooled sequences across multiple individuals can be used to infer a more reliable model [196]. I therefore used IGoR to infer a generative model of VDJ recombination using the entire pooled collection of nonfunctional consensus sequences from the pilot dataset, as well as separate models for each individual.

Figure 4.23 shows the probability distributions inferred by IGoR for V/D/J segment choice during VDJ recombination, while Figure 4.24 shows the distributions of N-insertions and P-insertions/deletions arising from junctional diversification; as P-insertions and deletions are mutually exclusive (Section 1.2.2), the former can be modelled as negative deletions and displayed as part of the same probability distribution [38]. These plots confirm a high-level of statistical similarity in the generative process between different individuals, and between individuals and the pooled dataset. The segment-choice distributions demonstrate that, regardless of which segment combinations are favoured by

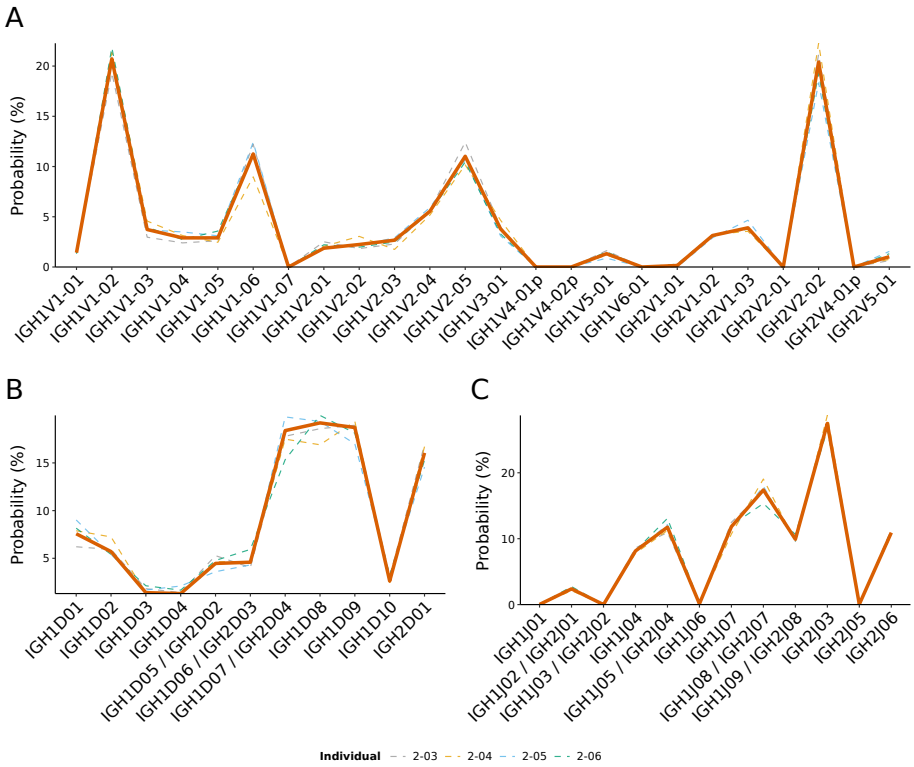


Figure 4.23: Generative segment-choice distributions in the IgSeq pilot dataset: Probability distributions of segment choice for (A) VH-, (B) DH- and (C) JH-segments during VDJ recombination in adult male turquoise killifish, inferred from the IgSeq pilot dataset using IGoR. Thin dashed lines represent the distributions inferred for individual killifish, while the thick solid line represents those inferred from pooled data from all individuals. DH and JH segments with identical sequences (which cannot be distinguished in the repertoire data even in principle) are collapsed together.

positive selection or affinity maturation, certain segments are strongly favoured over others even during the initial VDJ-recombination process. The junctional-deletion profiles, meanwhile, roughly resemble those inferred for human blood repertoires, albeit with more of a skew towards deletion and away from P-insertion (Figure 4.24C to Figure 4.24F) [38]. Conversely, however, the inferred distributions of N-insertions differ sharply from those found in human repertoires: whereas in humans the distributions peak at about 5 N-insertions and often yield 10-20 insertions per junction [38], in killifish there is a strong peak at 0 insertions, and sequences with more than 5 inserted nucleotides in either junction are very rare (Figures 4.24A and 4.24B). These results suggest that, apart perhaps from differences in the number of gene segments available to VDJ-recombination, the most dramatic difference between human and killifish generative repertoires lies in the much smaller number of N-insertions in most killifish sequences.

After inferring probability distributions for gene choice, insertions and deletions, and some other components of the generative process, IGoR can also infer the Shannon entropy of that process (Appendix C.1.2), giving an estimate of the theoretical sequence diversity that could be produced

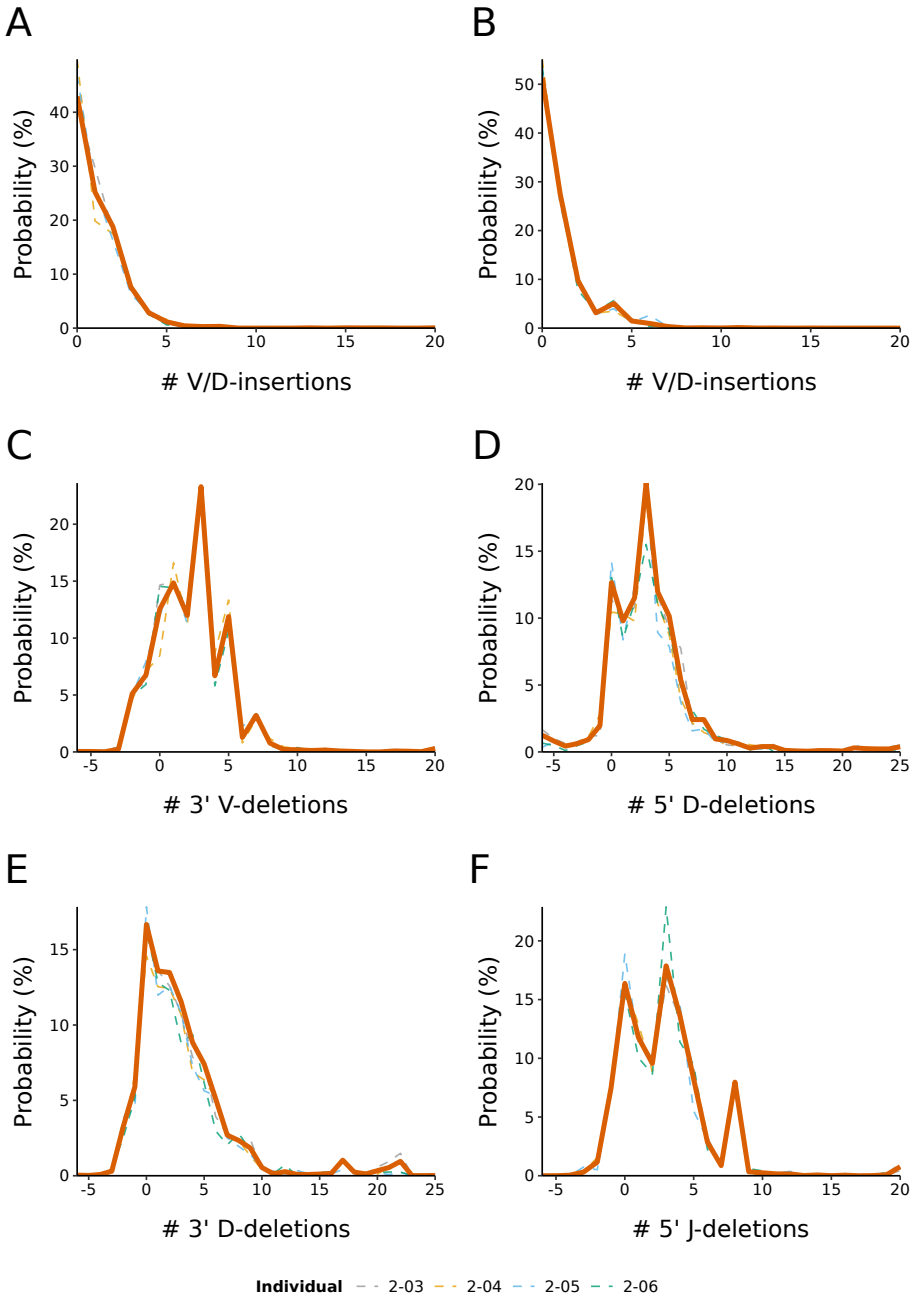


Figure 4.24: Generative insertion/deletion distributions in the IgSeq pilot dataset: Probability distributions of the number of N-insertions (A-B) or P-insertions/deletions (C-F) following VDJ recombination in adult male turquoise killifish, inferred from the IgSeq pilot dataset using IGoR. P-insertions are modelled as negative deletions. Thin dashed lines represent the distributions inferred for individual killifish, while the thick solid line represents those inferred from pooled data from all individuals.

by heavy-chain sequence generation in that species [48]. Importantly, this is not the same as (and is typically much higher than) the actual observed diversity of the naïve antibody repertoire in an individual, as the latter is limited by the number of naïve B-cells produced in an organism during its lifetime, as well as their proliferation history and the sampling depth of a given IgSeq experiment [30]. The generative repertoire discussed here is also distinct from the theoretical diversity of the functional primary repertoire, as it does not take into account reductions in diversity arising due to convergent recombination events, nonfunctional sequences, or primary selection [38]. Nevertheless, the magnitude and composition of the entropy of the generative repertoire is a highly valuable measure of the complexity of the primary diversification process in a given organism, which can be used to compare this process across species, ages, or treatment groups.

Figure 4.25 shows the generative entropy of the antibody repertoire in turquoise killifish from the pilot dataset, broken down into its constituent diversification processes. In total, the killifish generative repertoire has an entropy of roughly 32 bits, of which roughly 8 bits arise from differences in gene choice, 12 bits from variability in the number and sequence composition of N-insertions, and the rest from P-insertions and deletions. The killifish generative process is therefore capable of generating in excess of 10^9 unique recombination events (the first-order Hill diversity corresponding to the total entropy of the model) – as expected, a vastly higher number than the observed clonal alpha diversity of the pilot samples (Figure 4.16).

Despite its impressive diversity, this generative entropy of the killifish repertoire is nevertheless far lower than that of humans, whose peripheral repertoires exhibit generative entropies in the range of 70 to 80 bits (Figure 4.26) [38], corresponding to a first-order Hill diversity of 10^{21} to 10^{24}. Suprisingly, however, the entropies of gene choice and P-insertions/deletions are only slightly lower in killifish than in humans, suggesting that these processes are of roughly comparable diversity in the two species (Figures 4.25 and 4.26). The major difference in the diversity of the human and killifish generative repertoires arises from the vastly greater entropy arising from N-insertions in humans (53 bits, compared to 12 bits in killifish), a result consistent with the strong differences in insertion distributions observed in Figures 4.24A and 4.24B. The reasons for such a striking difference in insertion diversity are unclear at present, but it may result from differences in the activity of the TdT enzyme (Section 1.2.2) between humans and killifish.

In conclusion, both the generative and realised antibody repertoires of adult male turquoise killifish exhibit substantial sequence diversity, with thousands of detectable clones per individual (Section 4.4.3) and billions of unique potential recombination events. While the generative process appears to be highly similar between individuals, the unique life history of the humoral adaptive immune system in each individual provides sufficient information for the secondary repertoires of different individuals to be distinguished based on their V/J-usage profiles (Section 4.4.3). Despite its short lifespan, therefore, the turquoise killifish expresses a diverse population of antibody sequences well

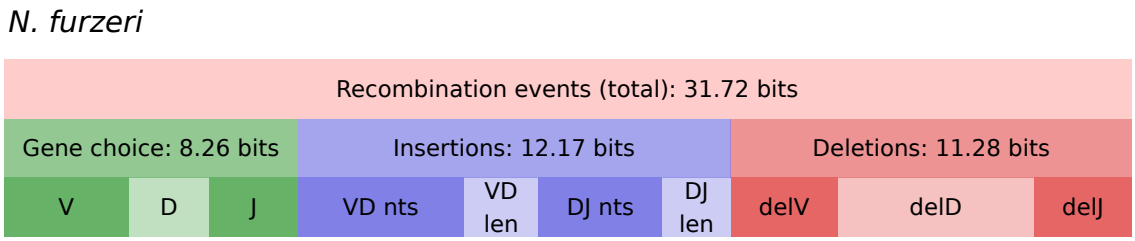


Figure 4.25: Entropy composition of the killifish generative repertoire: Entropy composition of different diversification mechanisms in the generative process of the turquoise-killifish antibody repertoire, inferred from the IgSeq pilot dataset using IGoR. "VD len" and "DJ len" denote sequence variability arising from the number of N-insertions, while "VD nts" and "DJ nts" refer to variability in the nucleotide sequences so inserted; "delV", "delD" and "delJ" refer to variability arising from deleted (or P-inserted) nucleotides at the boundaries of the selected gene segments.

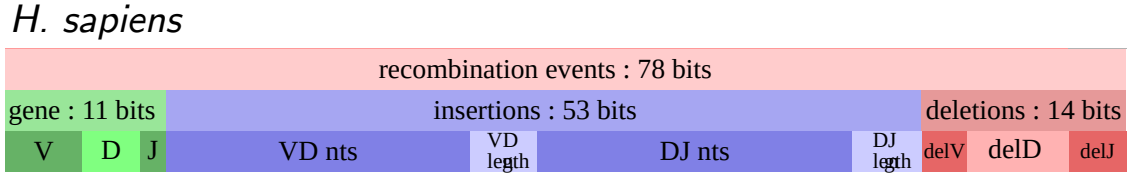


Figure 4.26: Entropy composition of a human generative repertoire: Example breakdown of the generative entropy of a human peripheral repertoire, adapted from Figure 8 of Elhanati *et al.* [38].

into middle adulthood, demonstrating its utility as a model organism to investigate the development and ageing of the antibody repertoire.

4.5 The effect of ageing on killifish antibody repertoires

The pilot study into the antibody repertoires of mature adult turquoise killifish (Section 4.4) demonstrated that these repertoires are diverse and individualised, with a complex generative process giving rise to naïve sequences and an underlying clonal structure resembling that of other species. This pilot dataset, however, consisted of killifish from a single age and treatment group, and therefore gave no information as to the effect of ageing on the composition and diversity of killifish repertoires. In order to investigate the effects of ageing on the killifish antibody repertoire, a second IgSeq experiment was performed using all 32 fish described in Tables 4.1 and E.24, comprising 10 fish each sacrificed at 39, 56 and 73 days post-hatching and two fish sacrificed at 128 days post-hatching. Whole-body RNA samples from these fish underwent two independent library preparations as described in Section 2.2.4 and Section 4.3.1, performed by me and Aleksandra Placzek, for a total of 64 pooled libraries; these were then sequenced together in two MiSeq runs, yielding a total of 26.7 million read pairs (0.2 to 0.9 million pairs per replicate, 0.5 to 1.3 million pairs per individual), and the resulting reads underwent pre-processing, filtering and clonotyping as described in Sections 4.3.2, 4.4.1 and 4.4.2.

Figure 4.27 shows the absolute and relative read survival for each of the sixty-four sequencing libraries throughout this process. As with the pilot dataset, the replicates showed relatively consistent behaviour up to and including VDJ assignment and Change-O database construction, with 66.9 % to 88.6 % of reads surviving up to this stage in the pipeline. However, a somewhat larger number of sequences (19.4 % in total, up to 33.4 % per replicate) were lost during V-score filtering compared to the pilot dataset (9.5 % in total, up to 14.3 % per replicate). This inconsistency was due to a greater preponderance in the ageing dataset of malformed, J-identity-lacking sequences, which in this case actually made up an absolute majority of unique sequences in the pRESTO-processed dataset (Figure D.10A). After filtering on V-score, however, the functional composition of the ageing dataset was similar to that of the pilot data (Figures 4.8B and D.10B).

Following V-score filtering, 98.3 % of remaining unique sequences in the ageing dataset were successfully assigned clonal identities. The number of clones inferred per individual ranged from 700 to 5600, with a median of 3020; there was a non-significant (Kruskal-Wallis analysis of variance, $p = 0.45$) decline in number of clones per individual with age (Figure D.11). These clonal counts were much lower than the median of 6020 clones per individual for the pilot study, reflecting the lower number of reads per individual available in this dataset; for comparison, for the four individuals included in the pilot study, the number of identified clones in the ageing dataset ranged from 2000 to 3600. Concordantly, the distribution of clone sizes detected in the ageing dataset was even more skewed towards small clones than in the pilot dataset: 83 % of clones were observed as just a single unique sequence across all replicates, while 98.3 % contained fewer than five unique sequences (Figure D.12A). As in the pilot dataset, larger clones were much more likely to be observed across both replicates for a given individual (Figure D.12B).

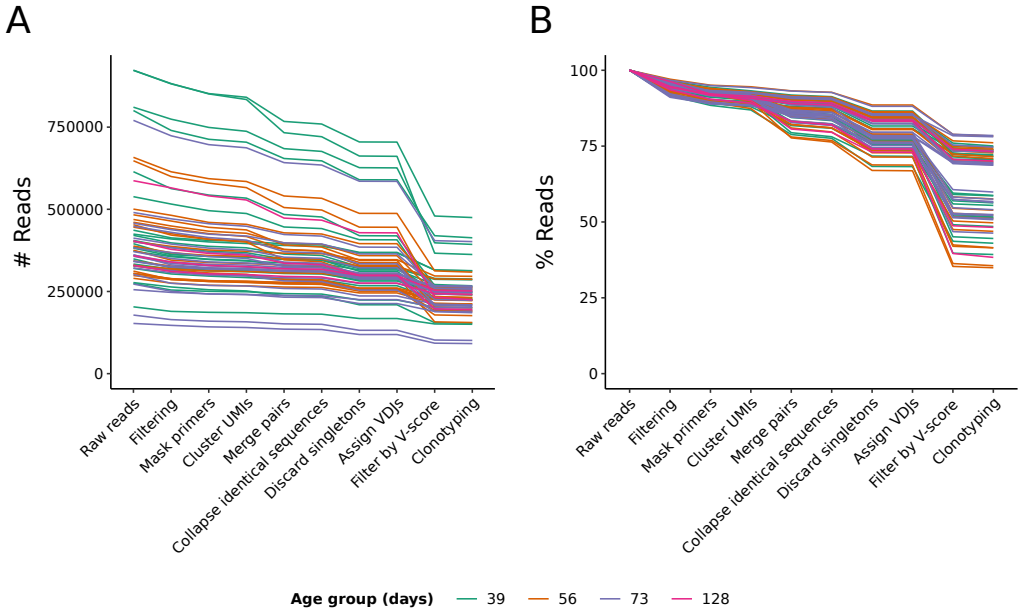


Figure 4.27: Read survival during pre-processing of the IgSeq ageing dataset: Line graphs of absolute (A) and relative (B) read survival during pre-processing of the IgSeq ageing dataset, up to and including clonotyping.

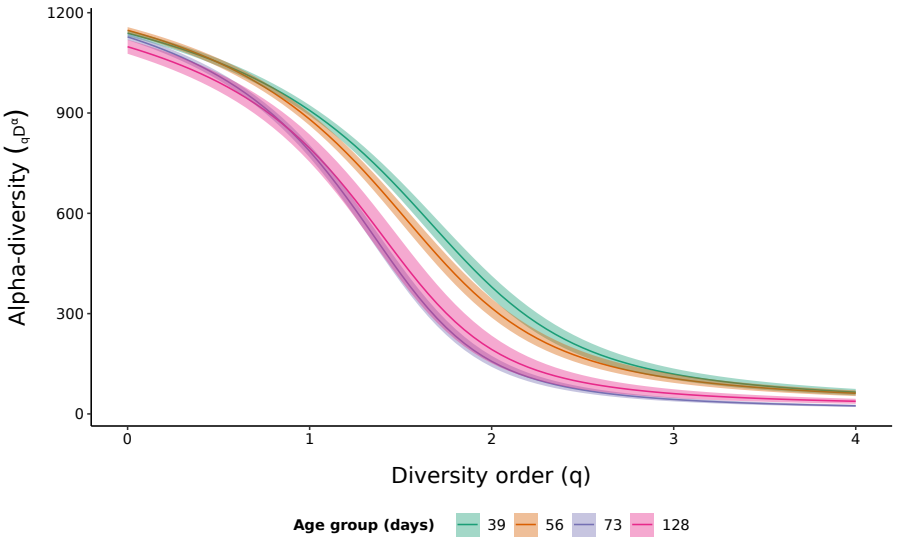


Figure 4.28: Clonal alpha-diversity spectra for the IgSeq ageing dataset: Bootstrapped alpha-diversity spectra of clone sizes for each age group in the IgSeq ageing dataset, as measured by number of unique sequences per clone. Shaded regions in both subfigures represent 95 % confidence intervals, estimated using bootstrapping.

In order to assess the effect of age on the clonal diversity of killifish antibody repertoires, the alpha diversity spectrum of each age group (Appendix C.2.2) was computed as described in Sections 2.3.6.4 and 4.4.2. The resulting spectra, shown in Figure 4.28, indicate that repertoire diversity declines monotonically between 5.5 and 10.5 weeks of age in the turquoise killifish, with the most rapid decline observed between 8 and 10.5 weeks: across the entire spectrum, the alpha diversity of the 8-week group is the same as or lower than that of the 5.5 week group, the 10.5-week group is markedly less diverse than the 8-week group, and the 18-week and 10.5-week groups exhibit similar diversity. These observations suggest a model in which repertoire diversity begins declining before or at reproductive maturation (c. 5 weeks post-hatching), declines rapidly in later adulthood, and reaches a plateau of low diversity late in life; however, the smaller size of the 18-week-old group in comparison to the others means that the patterns observed in this group cannot be taken with confidence, and it is possible that a further decline after 10.5 weeks would be observed in a larger sample.

In order to test the significance of these differences in clonal diversity between age groups, the Hill-diversity measurements from each individual spectrum (Figure D.13) were compared across age groups at each of six diversity orders (0, 1, 1.5, 2, 3 and 4) using two distinct methods. In the first, a Gamma-family generalised linear model (GLM) of diversity vs age-at-death was fitted to each diversity order under an inverse link function, and the resulting model was compared to a null (intercept-only) model to test for a significant ageing effect; in the second, a nonparametric Kruskal-Wallis analysis of variance was used, again to test for a significant effect of age-at-death on the distribution of Hill diversities at a given diversity order. Both results gave roughly consistent results (Figure 4.29), as did linear models and inverse-Gaussian GLM-based methods (Figures D.14 and D.15): in all cases, a significant effect of age on diversity was found for all tested diversity orders above 1.5, but not for order 0 or 1. This pattern, which indicates a significant age effect at higher but not lower diversity orders, suggests that such an effect is driven more by expansion of large clones (which primarily affects higher-order diversity measures) than a decrease in the number of smaller clones (which primarily affects lower-order measures like species richness), a result consistent with the results in Figure D.11; if this is the case, it would accord with earlier findings of persistent clonal expansions of activated cells in aged B-cell repertoires in other model systems (Section 1.3).

While a significant decline in clonal diversity was found at various diversity orders, no such decline was observed in the VJ-repertoires of the different age-groups, either when visually comparing the alpha-diversity spectra (Figure 4.30A) or statistically comparing the distributions at different diversity orders (Figures 4.31 and D.16). This marked difference between the clonal and VJ-repertoires is at first surprising, as a distortion in the clonal repertoire caused by an increased dominance of a few large clones might be expected to cause a similar distortion in the VJ-repertoire by expanding those dominant clones' V/J combinations; however, given that an average of 84.3 % of unique sequences in each repertoire were contained in small clones of fewer than 5 unique sequences (Figure D.19), and that this average was not found to change significantly with age (Kruskal-Wallis analysis of variance, $p = 0.4$), it may simply be the case that any age-related change in VJ-diversity arising from

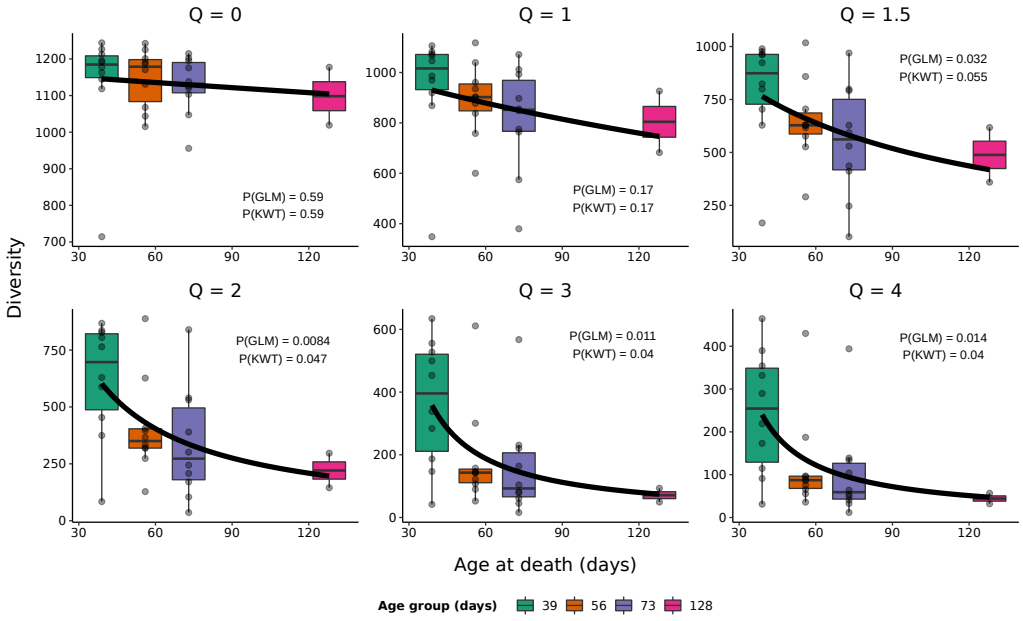


Figure 4.29: Comparing clonal alpha diversities between age groups in the IgSeq ageing dataset: Boxplots of Hill diversity values for the antibody repertoires of individuals of each age group in the IgSeq ageing dataset at a sample of diversity orders, overlaid with with the predictions at each order of the best-fit Gamma-family GLM under an inverse link function. Annotated p-values indicate the statistical significance of the estimated age effect on diversity under the GLM ($P(GLM)$) and a Kruskal-Wallis test ($P(KWT)$) for each diversity order.

the minority of sequences sampled from expanded clones was dominated by an absence of change in the V/J-usage distribution of naïve B-cells. To test this hypothesis, I repeated the estimation of VJ-diversity spectra while excluding small clones (containing fewer than five unique sequences); as expected, a clear pattern of reduced alpha diversity with age emerged (Figure 4.32A), with significant declines in diversity with age observed across all diversity orders tested (Figure 4.33).

The finding that a significant age-related change in VJ-alpha diversity is only observed when small clones are excluded suggests that, while a the secondary antibody repertoire changes significantly with age in the turquoise killifish, the primary generative process does not, at least as far as its statistical composition is concerned. To test this, I inferred generative models for each age group with IGoR in the manner described in Section 4.4.4. Following consensus-sequence inference across clones and discarding of functional sequences, the number of sequences available per individual for this

Table 4.3: Unique nonfunctional sequences from the IgSeq ageing dataset available for model inference with IGoR

Age group (days)	# Unique sequences
39	6107
56	5214
73	5443
128	702

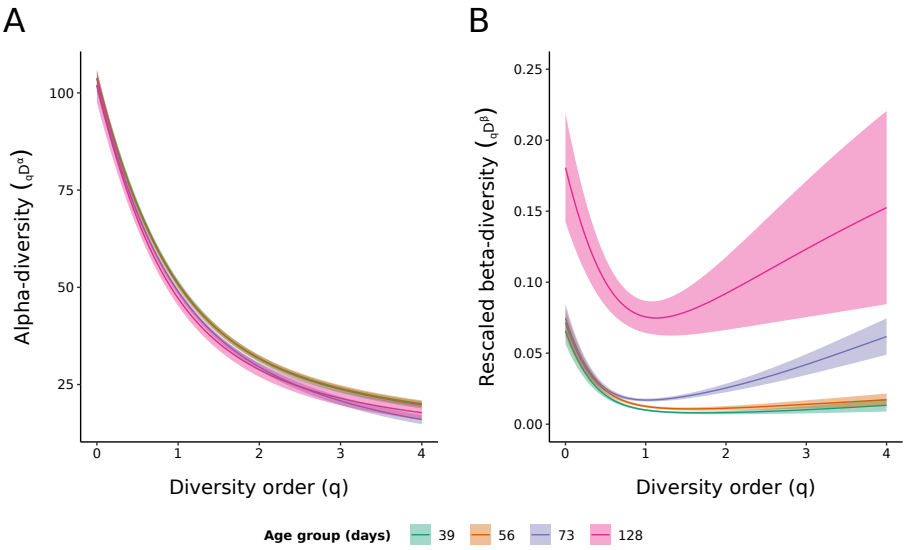


Figure 4.30: VJ-diversity spectra for the IgSeq ageing dataset: Bootstrapped Hill diversity spectra of VJ usage (as measured by number of unique sequences per unambiguous V/J combination) over individuals for each age group in the IgSeq ageing dataset. (A) alpha diversity across individuals; (B) beta diversity across individuals, rescaled to between 0 (minimum) and 1 (maximum) in accordance with the number of individuals in each age group. Shaded regions in both subfigures represent 95 % confidence intervals, estimated using bootstrapping.

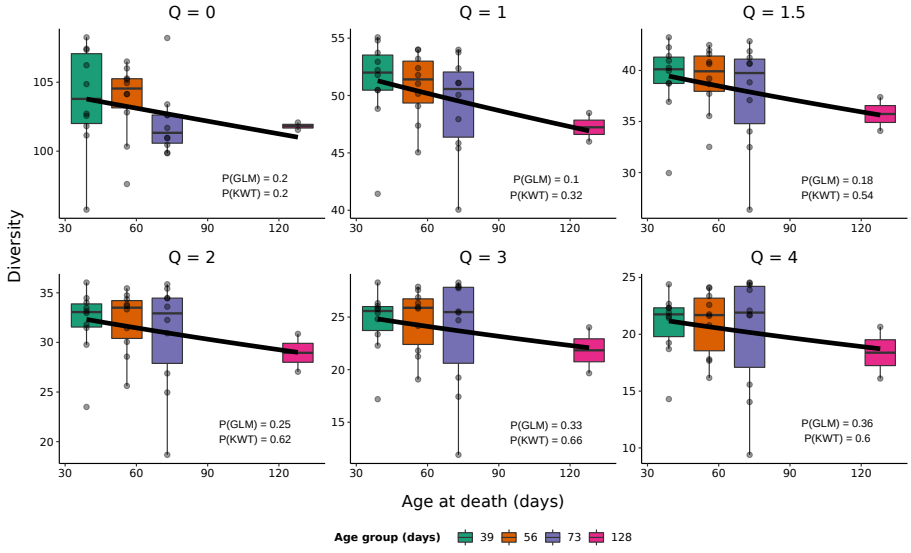


Figure 4.31: Comparing VJ alpha diversities between age groups in the IgSeq ageing dataset: Boxplots of Hill diversity values for the VJ-repertoires of individuals of each age group in the IgSeq ageing dataset at a sample of diversity orders, overlaid with the predictions at each order of the best-fit Gamma-family GLM under an inverse link function. Annotated p-values indicate the statistical significance of the estimated age effect on diversity under the GLM ($P(GLM)$) and a Kruskal-Wallis test ($P(KWT)$) for each diversity order.

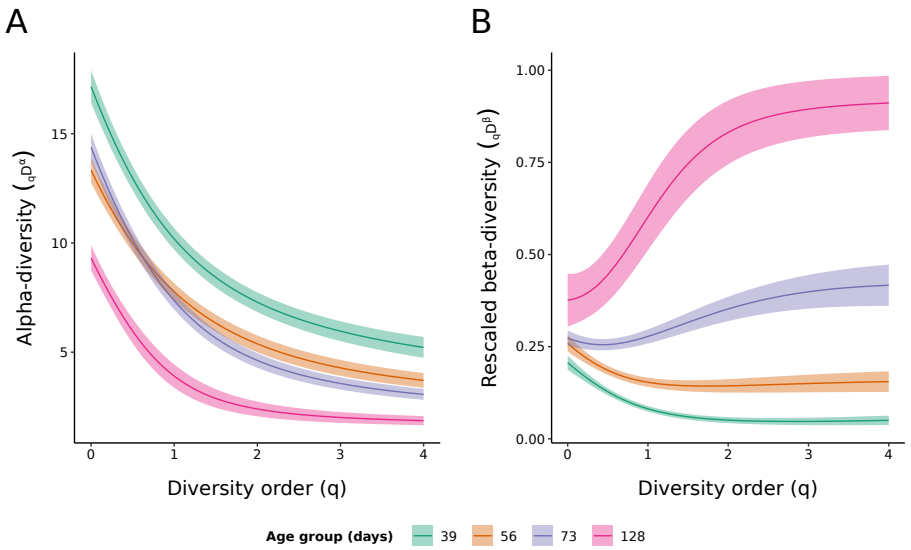


Figure 4.32: VJ-diversity spectra for expanded clones in the IgSeq ageing dataset: Bootstrapped Hill diversity spectra of VJ usage (as measured by number of unique sequences per unambiguous V/J combination) over individuals for each age group in the IgSeq ageing dataset, excluding clonal groups with fewer than five unique sequences. (A) alpha diversity across individuals; (B) beta diversity across individuals, rescaled to between 0 (minimum) and 1 (maximum) in accordance with the number of individuals in each age group. Shaded regions in both subfigures represent 95 % confidence intervals, estimated using bootstrapping.

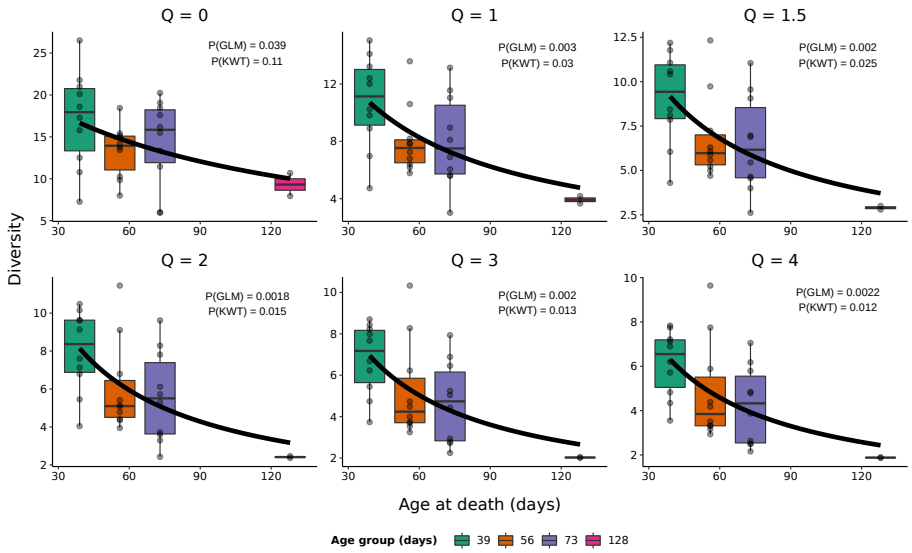


Figure 4.33: Comparing VJ alpha diversities between age groups for expanded clones in the IgSeq ageing dataset: Boxplots of Hill diversity values for the VJ-repertoires of individuals of each age group in the IgSeq ageing dataset at a sample of diversity orders, excluding clonal groups with fewer than five unique sequences, overlaid with the predictions at each order of the best-fit Gamma-family GLM under an inverse link function. Annotated p-values indicate the statistical significance of the estimated age effect on diversity under the GLM ($P(GLM)$) and a Kruskal-Wallis test ($P(KWT)$) for each diversity order.

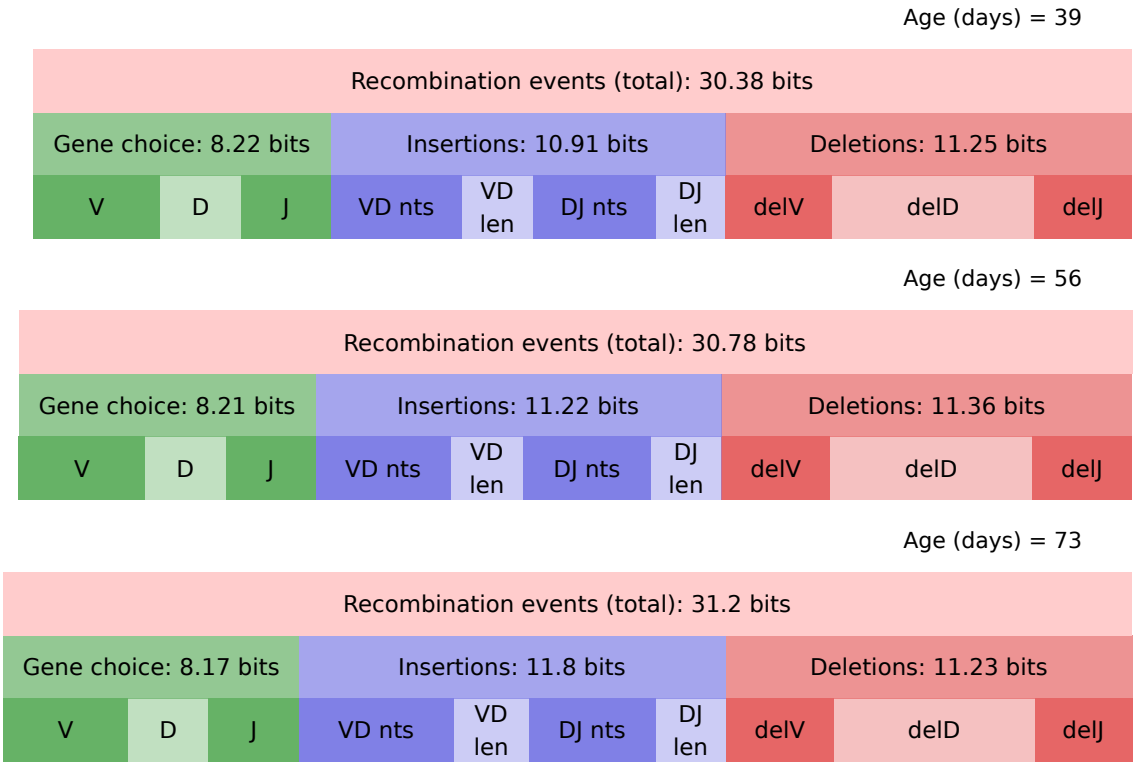


Figure 4.34: Entropy composition of the killifish generative repertoire in different age groups: Entropy composition of different diversification mechanisms in the generative process of the turquoise-killifish antibody repertoire, inferred using IGoR for the first three age groups in the IgSeq ageing dataset. "VD len" and "DJ len" denote sequence variability arising from the number of N-insertions, while "VD nts" and "DJ nts" refer to variability in the nucleotide sequences so inserted; "delV", "delD" and "delJ" refer to variability arising from deleted (or P-inserted) nucleotides at the boundaries of the selected gene segments. The fourth age group in the dataset (128 days) contained too few nonfunctional sequences to fit a reliable model (Table 4.3) and was excluded from the analysis.

purpose ranged from 190 to 1013; all far too low to infer a reliable model with IGoR. However, pooling these repertoires by age group yielded over 5000 unique sequences for the first three age groups, though the oldest, smallest group yielded only c. 700 sequences total (Table 4.3). I therefore investigated the entropy composition of the generative repertoire for the first three age groups in the dataset (Figure 4.34). The results supported the hypothesis that the generative process remains intact throughout the killifish lifespan, with very similar entropy compositions for all three age groups; neither the total entropy nor any of the major components differs by as much as 1 bit between any two of the three models. Consistent with this, the segment-choice, insertion, and deletion distributions inferred for the generative process are highly similar across all individuals and age groups, despite the very small size of many of the datasets used to infer these models (Figures D.17 and D.18).

While the alpha diversity of the entire VJ-repertoire (reflecting average diversity within each individual repertoire) was not found to change significantly with age in the turquoise killifish, the beta diversity (reflecting inter-individual variability in VJ-expression) showed dramatic differences between age groups, especially at higher diversity orders, for both the entire repertoire (Figure 4.30B), and the subset of the repertoire contained in large clones (Figure 4.32B): in both cases, while the diversity of the 5.5-week-old group was close to the theoretical minimum for that sample size, the 10.5-week-old group exhibited a substantial increase in inter-individual variability across a wide range of diversity orders, and the 18-week old group showed an even more dramatic increase. This increase in inter-individual diversity with age is corroborated by inter-individual RDI distances computed on VJ-repertoires within each age group: both the overall distribution of inter-individual distances (Figure 4.35A), and the distribution of nearest-neighbour distances within each age group (Figure 4.35B), increased significantly with age. This tendency for individuals from increasing age groups to become more distinct in their VJ-repertoires can also be observed using principal co-ordinate analysis: in Figure 4.35D, the great majority of individuals from younger age groups are visibly more tightly clustered than those from older age groups, which drift apart progressively as age-at-death increases. As in ageing humans (Section 1.3), therefore, the ageing killifish appears to become more individualised and distinct in its antibody repertoire, presumably as the result of an accumulating history of unique individual responses to antigen exposure.

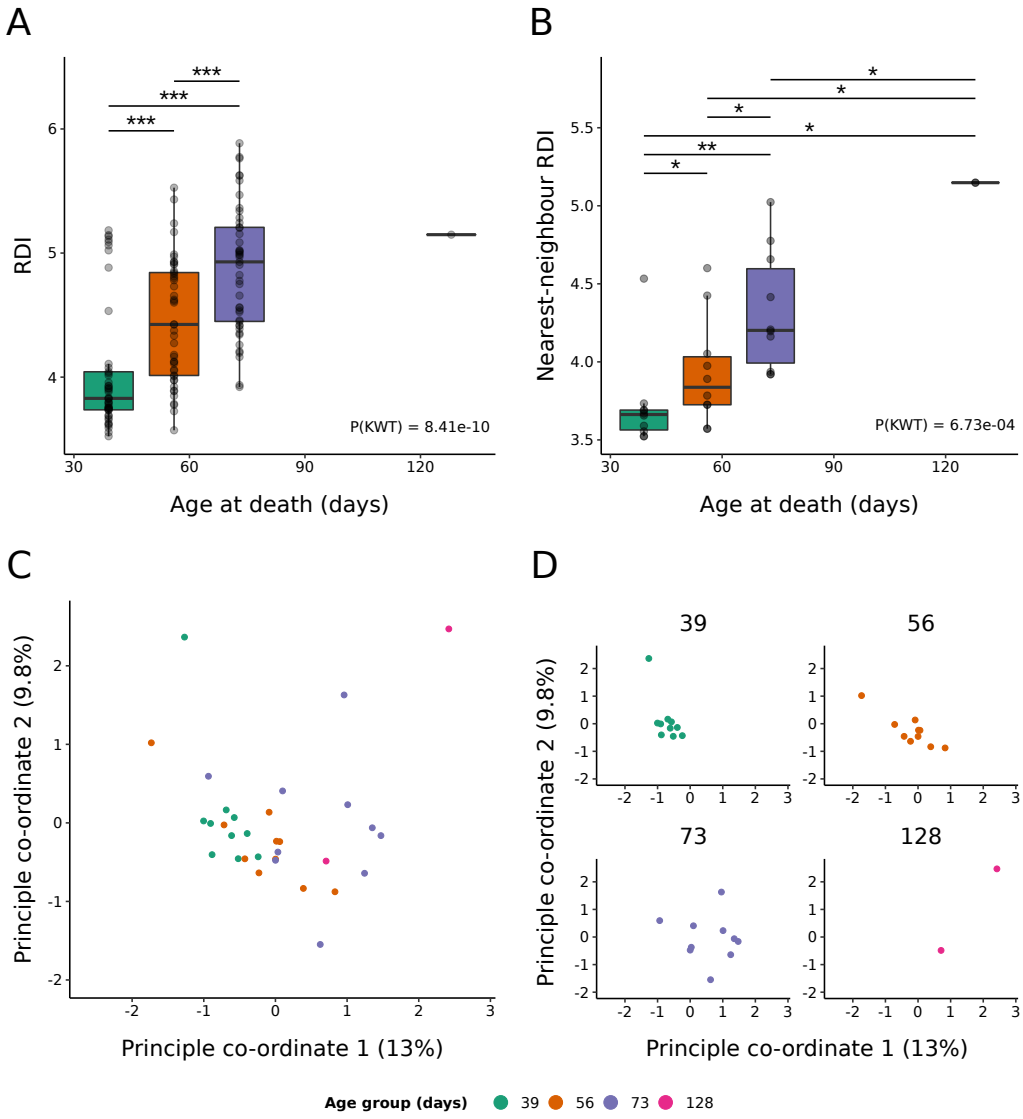


Figure 4.35: Inter-individual VJ-RDI distances in the IgSeq ageing dataset: (A-B) Boxplots of overall (A) and nearest-neighbour (B) inter-individual VJ-RDI distances for each age group in the IgSeq ageing dataset. Pairwise p-values were computed using nonparametric Mann-Whitney U tests ($*: 0.01 < p \leq 0.05$; $**: 0.001 < p \leq 0.01$; $***: p \leq 0.001$). $P(KWT)$ indicates the p-value of a Kruskal-Wallis analysis of variance test for a difference in RDI distribution with age. (C-D) Principal co-ordinate analysis (PCoA) of pairwise inter-individual VJ-RDI distances in the IgSeq ageing dataset, coloured by age group and displayed together (C) or separately by age group (D).

4.6 Gut-microbiota transfer and the killifish mucosal repertoire

The results from Section 4.5 demonstrate that the whole-body antibody repertoire of adult male turquoise killifish declines in clonal alpha diversity with age, while increasing in VJ beta diversity, with no change observed in the generative process underlying the primary repertoire. These results demonstrate that the turquoise killifish antibody repertoire undergoes rapid age-related changes, in keeping with the more general rapid ageing phenotype observed in this species (Section 1.4). However, while these results demonstrate significant age-related changes in repertoire composition at the level of the entire body, they give no specific information about the ageing process exhibited by the specialised repertoires of particular immune organs, which could differ substantially as a result of their distinct antigenic environments and B-cell-subpopulation composition (Section 1.3). In particular, the gut mucosal repertoire, which polices the interface between the host organism and the gut microbiota, represents a highly important B-cell subpopulation with an intense and distinctive experience of antigenic exposure [197]. It would therefore be interesting to investigate whether the pattern of repertoire ageing observed in the gut accords with that seen in the wider body, or exhibits its own distinct ageing phenotypes.

Smith *et al.* [78] demonstrated that gut microbiota transfer from young to middle-aged turquoise killifish significantly extends lifespan in this species, as well as significantly altering microbiotal composition and gene expression in the gut. As part of these experiments, total RNA was isolated from the intestines of a number of male turquoise killifish of the GRZ-Bellemans substrain (a closely-related substrain to the GRZ-AD substrain used in Sections 4.4 and 4.5) at different ages and following different experimental interventions (Figure 4.36 and Tables 4.4 and E.25). By using these RNA samples to perform immunoglobulin sequencing, I was able to investigate the effects of both age and microbiota transfer on the diversity of the gut mucosal repertoire; in particular, given the observed lifespan effect of gut-microbiota transfer and the intimate relationship between the gut microbiota and the mucosal adaptive immune system, I hypothesised that gut-microbiota transfer might significantly ameliorate the ageing of the adaptive-immune repertoire observed in Section 4.5, either through delaying the decline in diversity observed in older fish or by stimulating a renewal of repertoire diversity in mucosal B-cells.

Table 4.4: Summary of killifish used in IgSeq gut experiment. All fish are GRZ-Bellemans strain and male.

Group	Age (weeks)	# Fish (Sequenced/Total)	Antibiotics?	Microbiota Transfer?
YI_6	6	4 / 4	No	No
WT_16	16	3 / 4	No	No
ABX_16	16	4 / 4	Yes	No
SMT_16	16	3 / 4	Yes	Yes (9.5-week-old donor)
YMT_16	16	4 / 4	Yes	Yes (6-week-old donor)

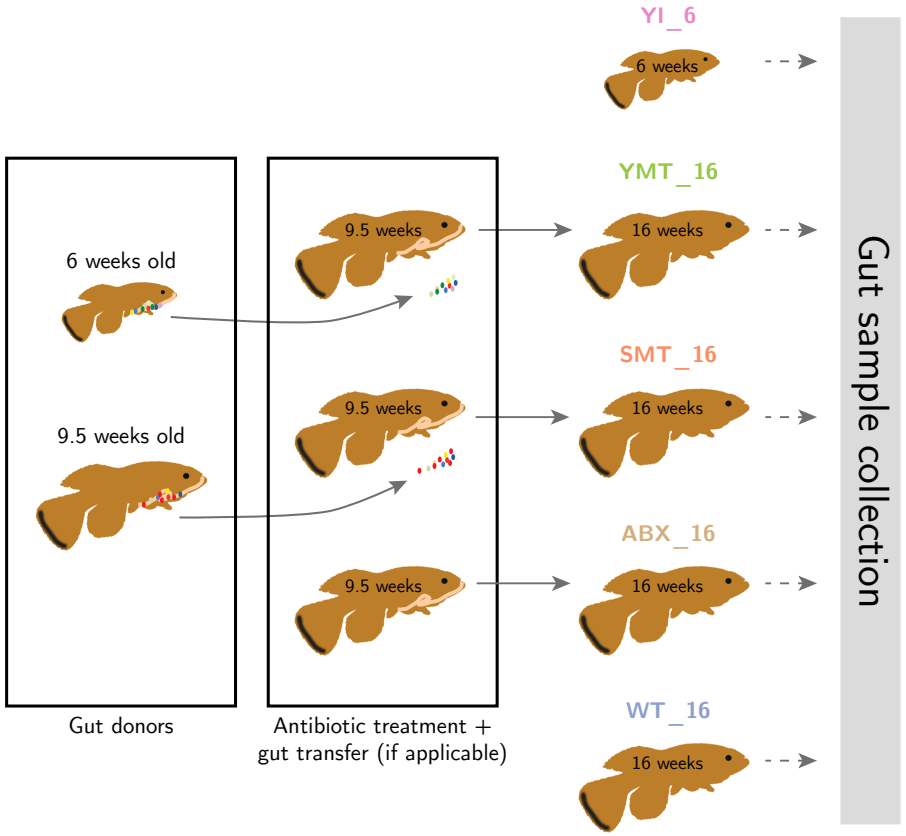


Figure 4.36: Experimental design of the killifish gut-microbiota transfer study: Schematic of the design of the gut-microbiota transfer experiment in Smith *et al.* [78]. Groups YI_6 and WT_16 were sacrificed without experimental intervention at 6 and 16 weeks, respectively, while the others recieved antibiotic treatment at 9.5 weeks followed by either no further intervention (ABX_16) or gut-microbiota transfer from a 6-week-old (YMT_16) or 9.5-week-old (SMT_16) donor. Adapted from [78], Figure 4.

Of the twenty samples outlined in Tables 4.4 and E.25, one (fish 400, from the WT_16wk group) contained too little RNA to undergo IgSeq library preparation, while another (fish 1005, from the OMT_16wk group) was too degraded for a useful library to be obtained. The remaining 18 samples underwent IgSeq library preparation, performed by Aleksandra Placzek and Michael Poeschla using the protocol I designed in Section 4.3.1. Several of the other samples were also somewhat degraded, with RNA integrity numbers between 5.5 and 7.0 (Table E.25), but succeeded in producing useable libraries; these samples were included in the sequencing pool, but the RNA integrity numbers (RIN values) from each sample were retained for comparison with downstream quality-control measures following immunoglobulin sequencing. The resulting libraries were sequenced together in two MiSeq runs, yielding a total of 23.9 million read pairs (0.8 to 2.6 million pairs per individual), and the resulting reads underwent pre-processing, filtering and clonotyping as described in Sections 4.4 and 4.5.

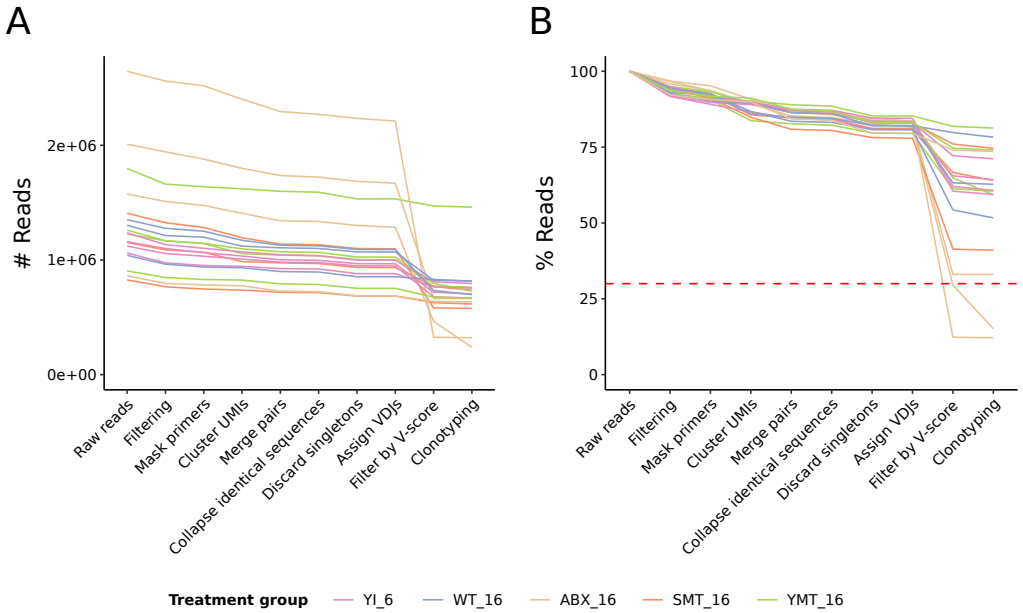


Figure 4.37: Read survival during pre-processing of the IgSeq gut dataset: Line graphs of absolute (A) and relative (B) read survival during pre-processing of the IgSeq gut dataset, up to and including clonotyping. The dotted red line in (B) indicates the 30 % read-survival cutoff, below which samples were discarded prior to downstream analysis.

Compared to the datasets in those sections, the gut dataset exhibited highly consistent behaviour up to and including VDJ assignment and Change-O database construction, with 77.9 % to 85.3 % of reads surviving up to this stage in the pipeline (Figure 4.37). However, a substantially higher proportion of reads (29.4 %) were lost during V-score filtering (Figure D.20) and clonotyping, with some individuals losing as much as 71.4 % of their input reads during these stages. The proportion of reads lost at this stage did not appear to have any relationship with the RNA integrity of the samples (Figure D.21, $r \approx -0.1$, $p \approx 0.7$), and following V-score filtering the functional composition of surviving sequences was similar to that of the other datasets (Figures 4.8B, D.10B and D.20B). Nevertheless, to avoid any problems associated with very low numbers (and possibly low quality) of surviving input reads, individuals with fewer than 30 % of input reads surviving through filtering and clonotyping were excluded from downstream analysis; two individuals (1274 and 1309, both from the ABX_16 antibiotics-treated group) were excluded in this way (Figure 4.37B).

Following V-score filtering and exclusion of high-read-loss individeals, 95.6 % of remaining unique sequences in the ageing dataset were successfully assigned clonal identities, with the number of clones inferred per individual ranging from 72 to 8994, with a median of 289; as in the ageing dataset, there was a non-significant (Kruskal-Wallis analysis of variance, $p = 0.069$) decline in number of clones per individual with age (Figure 4.38). With two exceptions (visible as outliers in Figure 4.38), these clonal counts are dramatically lower than those of either the pilot or ageing dataset, both in absolute terms (Figure 4.39A) and relative to the number of UMI groups or unique

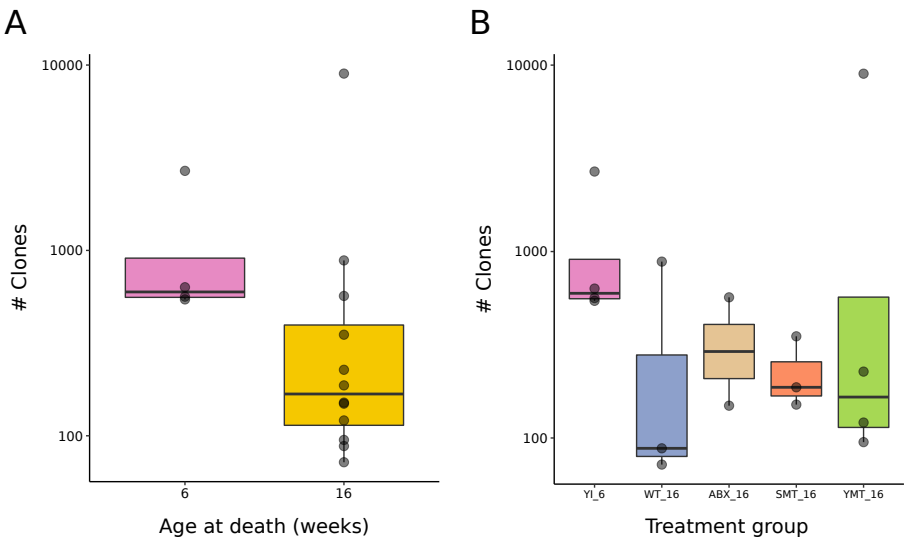


Figure 4.38: Number of clones in the IgSeq gut dataset: Boxplots of clonal counts for each individual in the IgSeq gut dataset, grouped by age at death (A) or treatment group (B). The apparent decline in clonal count with age is not significant (Kruskal-Wallis analysis of variance, $p = 0.069$), and there is also no significant effect of treatment group on clonal count (Kruskal-Wallis analysis of variance, $p = 0.43$).

sequences in each repertoire (Figure 4.39B). This suggests, not entirely surprisingly, that killifish guts contain many fewer B-cell clones than whole-body killifish samples. However, the metrics presented in Figure 4.39 are strongly dependent on the size of each dataset and cannot be taken at face value: for example, even though the pilot dataset contains the same type of sample and even a subset of the same individuals from the ageing dataset, their distributions in Figure 4.39 are very different.

In order to determine whether the apparent difference in clonal richness between the gut and other datasets is real, therefore, I performed rarefaction analysis, measuring the number of clones and P20 in repeated downsamplings of UMI groups from each dataset. The results for the rarefied clonal counts (Figure 4.40A) confirmed the results from Figure 4.39: at any given downsampling size, the great majority of repertoires from the gut dataset contained far fewer clones than the repertoires in the pilot or ageing datasets. The gut repertoires also showed a much greater degree of dominance by the largest clones in each repertoire (Figure 4.40B), with over 85 % of gut samples exhibiting greater asymptotic P20 than over 90 % of ageing or pilot individuals and over 65 % exhibiting higher asymptotic P20 than any sample in either of the other experiments. In both cases (clonal counts and P20), a small minority of individuals deviated strongly from the general trend, exhibiting clonal counts and P20 values more consistent with those seen in the pilot and ageing experiments; however, only one individual (1412, from the YI_6 6-week-old untreated group) was an outlier in both clonal count and P20 value.

The combination of low clonal counts and elevated P20 indicated by Figures 4.40A and 4.40B suggests that the intestinal antibody repertoire of the turquoise killifish contains many fewer small, naïve clones than the whole-body repertoire, and is consequently much more strongly dominated by

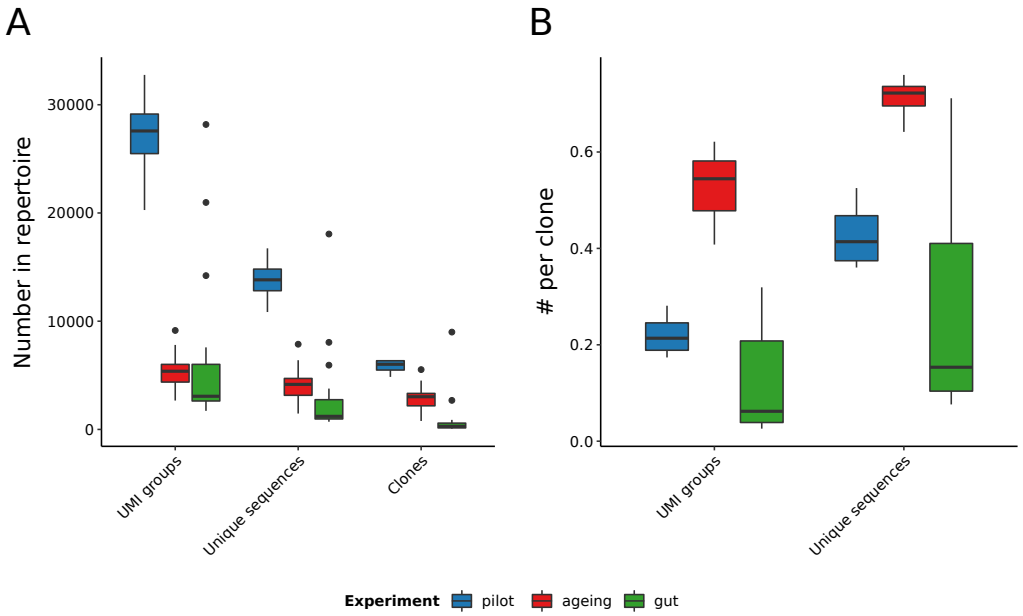


Figure 4.39: Comparison of clonal counts between IgSeq experiments: Boxplots comparing different count metrics between the IgSeq pilot, ageing and gut datasets, demonstrating the reduced absolute and relative clonal counts in the latter. (A) Absolute numbers of UMI groups, unique sequences and clones in each dataset. (B) Relative numbers of UMI groups and unique sequences per clone.

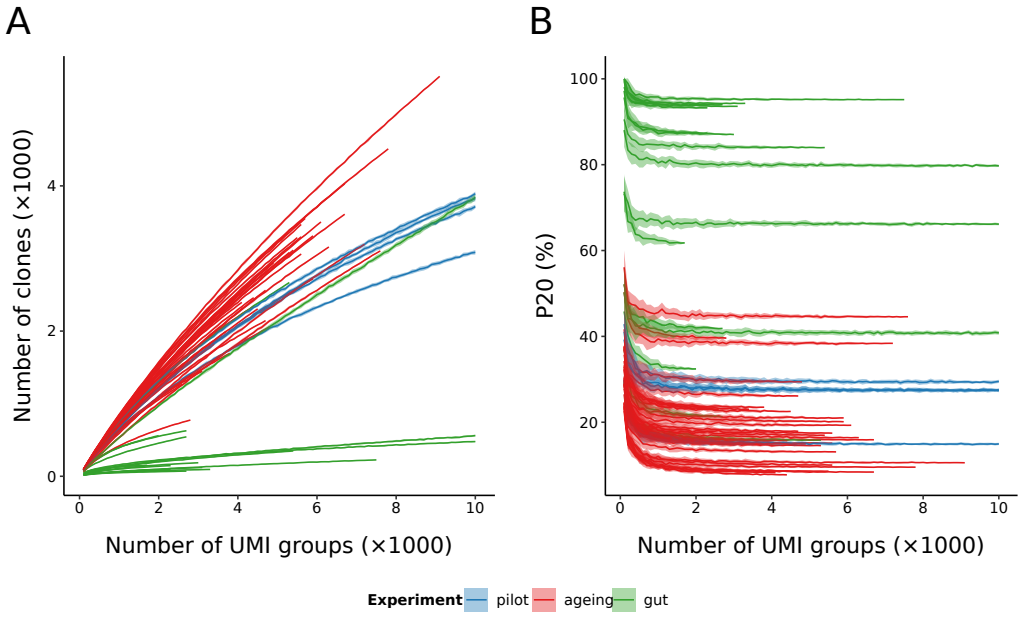


Figure 4.40: Comparative rarefaction analysis of clonal counts and P20 in IgSeq experiments: Rarefaction analysis of (A) clonal counts and (B) P20 in turquoise killifish repertoires from the IgSeq pilot, ageing and gut-microbiota-transfer experiments. Lines and shaded regions indicate the mean and standard deviation, respectively, over twenty replicates per sample size.

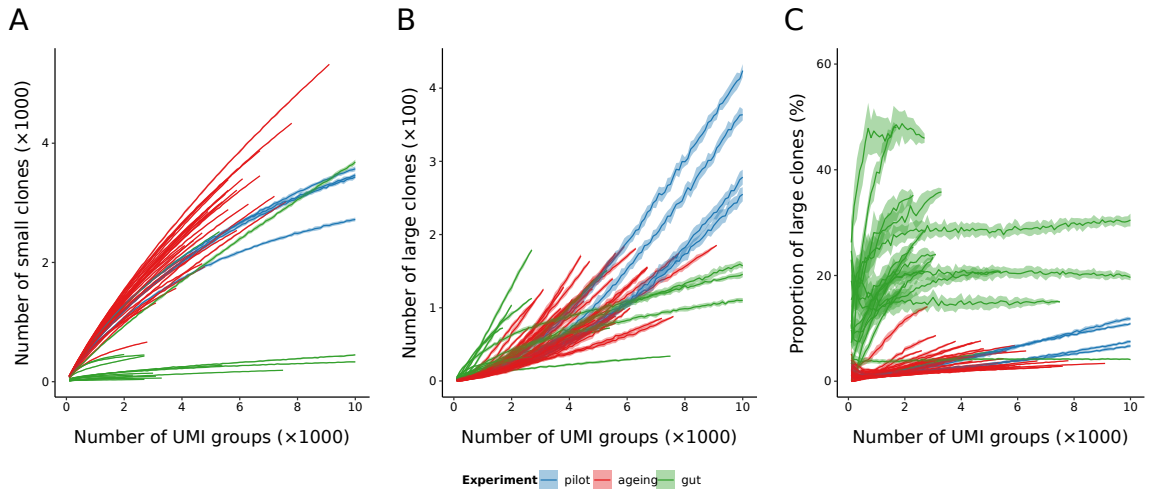


Figure 4.41: Comparative rarefaction analysis of clonal size composition in IgSeq experiments: Rarefaction curves of (A) the number of small clones (< 5 unique sequences), (B) the number of large clones (≥ 5 unique sequences) and (C) the proportion of clones which are large in each individual, coloured by source experiment. Lines and shaded regions indicate the mean and standard deviation, respectively, over twenty replicates per sample size.

a comparatively small number of expanded clones. This hypothesis is confirmed when the rarefied clonal counts are separated by clone size (Figure 4.41): while the number of small clones (containing fewer than 5 unique sequences) in the gut repertoires was again much smaller than in the other experiments (Figure 4.41A), the number of large clones (containing at least 5 unique sequences) was roughly the same (Figure 4.41B), and as a consequence the proportion of large clones was much higher in most gut repertoires than in the whole body (Figure 4.41C). This difference between the gut and whole-body repertoires makes sense: unlike the gut repertoire, the whole-body repertoire includes clones from primary lymphoid organs (in particular, the anterior kidney), and so could be expected to contain a much larger number of small, naïve clones. Furthermore, as a site of extensive interaction between the host and the gut microbiota, the gut provides many opportunities for its resident B-cells to encounter foreign antigens [198], and the consequent high rate of clonal expansion is likely to further increase the extent to which the gut repertoire is dominated by large clones.

The clonal alpha-diversity spectra for the gut dataset are shown in Figure 4.42. At all diversity orders, there is a clear and drastic difference between the young (6-week-old) and old (16-week-old) groups, while the different 16-week-old treatment groups do not appear to show a strong difference in diversity. These observations are confirmed by statistical comparison of the distributions of individual diversity measurements at different diversity orders, which indicate highly significant differences in clonal repertoire diversity between young and old guts across the diversity spectrum (Figures 4.43A and D.22A), but no difference between treatment groups (Figures 4.43B and D.22B). A similar pattern is observed for V/J-usage diversity (Figure 4.44), with a large and significant difference between

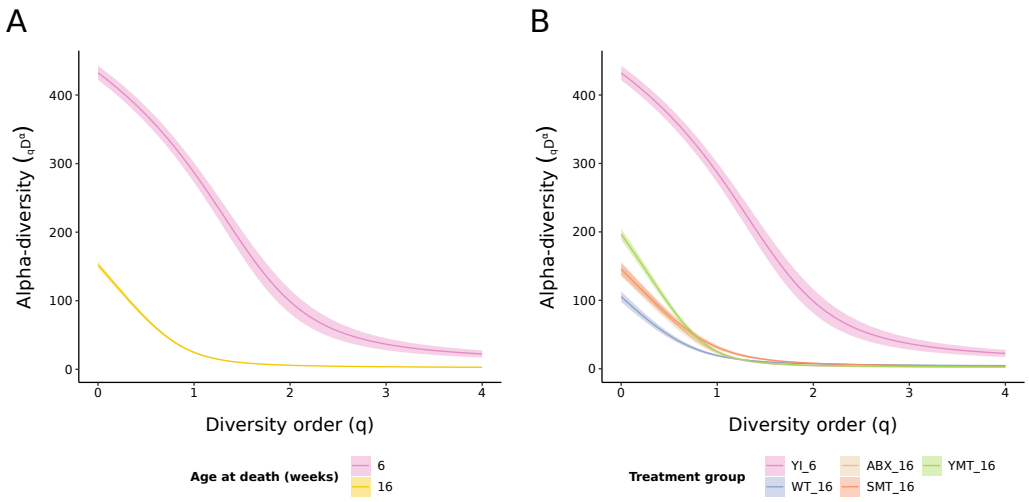


Figure 4.42: Clonal alpha-diversity spectra for the IgSeq gut dataset: Bootstrapped alpha-diversity spectra of clone sizes for each (A) age group and (B) treatment group in the IgSeq gut dataset, as measured by number of unique sequences per clone. Shaded regions in both subfigures represent 95 % confidence intervals, estimated using bootstrapping.

young and old cohorts at many different diversity orders (Figures 4.45A and D.23A) but no significant differences between 16-week-old treatment groups (Figures 4.45B and D.23B).

In terms of beta diversity, the Hill-spectrum results (Figure 4.46) are equivocal, with large differences in diversity between age groups observed at intermediate diversity orders (c. 1 to 2.5) but not at very low or very high orders (Figure 4.46A). Different treatment groups, meanwhile, exhibit very different patterns of beta diversity at higher orders (Figure 4.46B), but not in any clear pattern: the antibiotic-treated and same-age-transfer groups show near-maximal high-order beta diversity, indicating dramatic differences between individuals in their VJ-usage profiles, while the untreated and young-donor-transfer groups show much lower between-individual variablity. To more closely investigate differences in inter-individual variability between sample groups, I again computed repertoire dissimilarity index (RDI) distances between each pair of individual repertoires in the dataset, and visualised the results by age cohort (Figure 4.47) and treatment group (Figure 4.48). As in the ageing dataset, these results indicate large differences in VJ-usage variability between young and old intestinal repertoires in the turquoise killifish, with young samples clustering much more closely together (Figure 4.47B) and consequently exhibiting much lower pairwise RDI distances (Figures 4.47C and 4.47D); conversely, there is no significant difference in RDI distribution between the different 16-week-old treatment groups (Figures 4.48C and 4.48D), indicating that the different groups are similar in their inter-individual variability. These RDI results contrast markedly with the differences in beta-diversity spectra observed in Figure 4.46B, suggesting that the latter may not be reliable measures of significant differences in intra-group variablity when sample sizes are small.

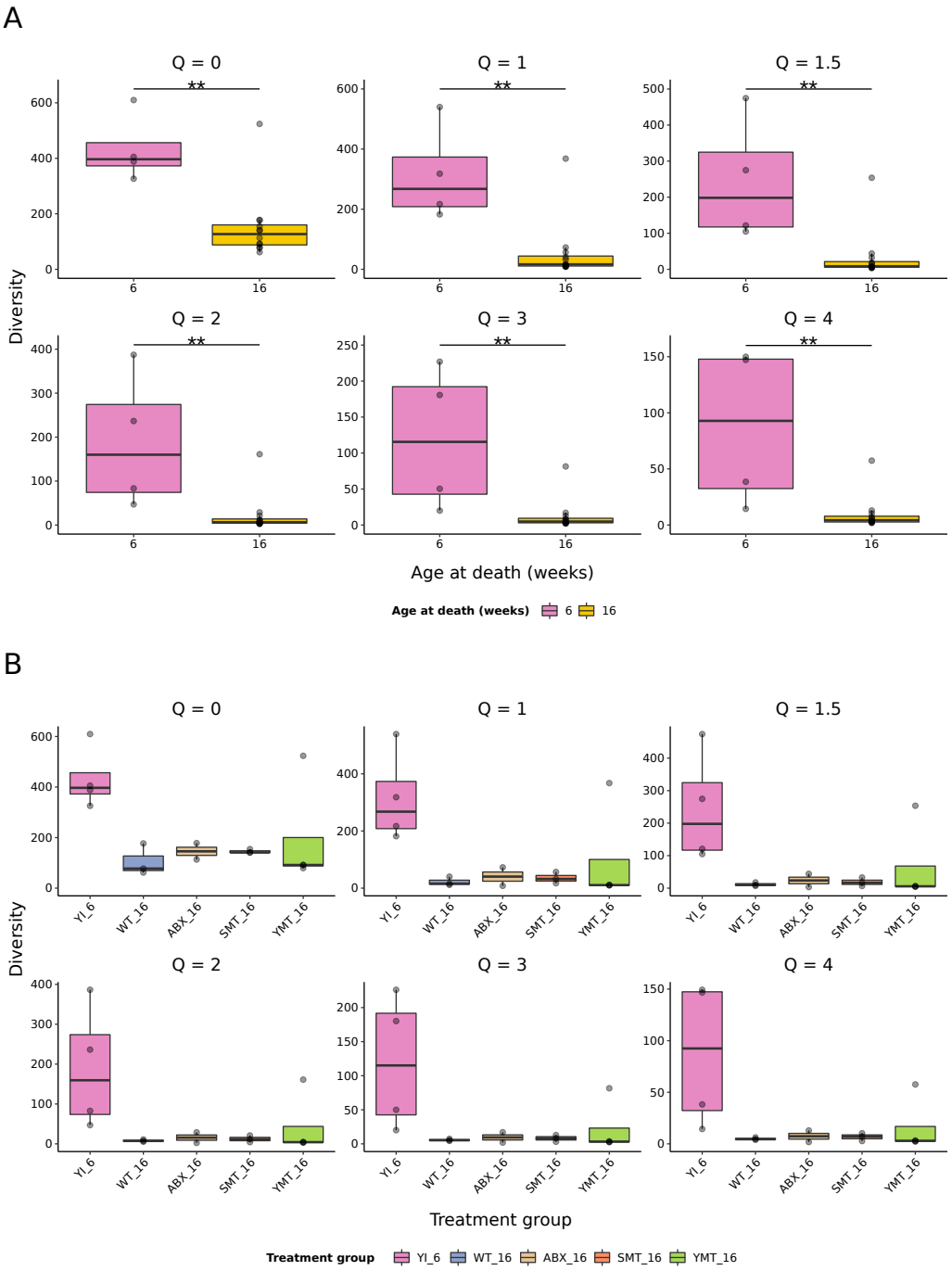


Figure 4.43: Comparing clonal alpha diversities between age and treatment groups in the IgSeq gut dataset: Boxplots of clonal Hill diversity values for the antibody repertoires of individuals of each (A) age group and (B) treatment group in the IgSeq gut dataset at a sample of diversity orders. Pairwise p-values were computed using nonparametric Mann-Whitney U tests ($*: 0.01 < p \leq 0.05$; $**: 0.001 < p \leq 0.01$; $***: p \leq 0.001$).

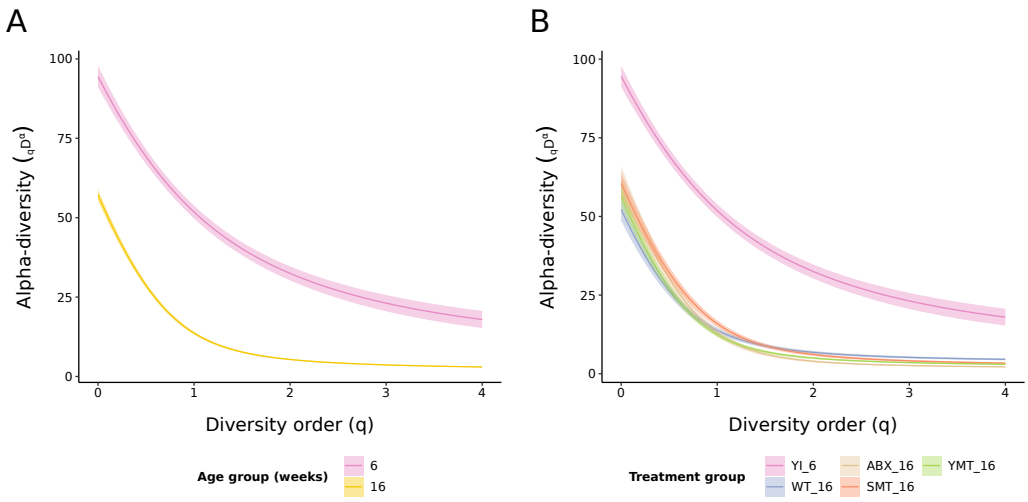


Figure 4.44: VJ alpha-diversity spectra for the IgSeq gut dataset: Bootstrapped alpha-diversity spectra of VJ usage for each (A) age group and (B) treatment group in the IgSeq gut dataset, as measured by number of unique sequences per unambiguous VJ identity. Shaded regions in both subfigures represent 95 % confidence intervals, estimated using bootstrapping.

In terms of repertoire ageing, therefore, the killifish intestinal repertoire not only reproduces the whole-body phenotype of reduced alpha and increased beta diversity with age, but in fact shows a stronger ageing phenotype than the whole-body repertoires reported in Section 4.5. In addition to showing a much stronger age-related decline in clonal alpha diversity with age than that exhibited in the ageing dataset (Figures 4.28, 4.29, 4.42A and 4.43A), the gut repertoire also shows a significant age-related decline in VJ alpha diversity, something not observed in the whole-body data (Figures 4.30A, 4.31, 4.44A and 4.45A). This latter difference is likely due to the much lower clonal richness and higher oligoclonality exhibited by the gut repertoires (Figures 4.40A, 4.40B and 4.41): when the local repertoire contains fewer small naïve clones and exhibits a greater degree of domination by the few largest clones, differences in VJ-usage between these large, activated clones may outweigh unchanged VJ-usage distributions in small, naïve clones to a greater extent than in the much-more-polyclonal whole-body repertoire. The much more dramatic loss of clonal diversity in intestinal repertoires, meanwhile, may be attributable to the especially-intense antigenic challenge at the mucosal surface, resulting in high levels of clonal expansion and progressive domination of the mucosal B-cell niche by a small number of highly expanded clones [198]; this phenomenon could in turn be exacerbated by the observed loss of taxonomic diversity (but not total abundance) in the gut microbiota [78], which might progressively reduce the variety of antigenic exposure at the mucosal surface and so further encourage a restriction in clonal diversity. If so, the strength of the ageing phenotype in these repertoires may have important consequences for the ability of the organism to regulate its mucosal microbial environment. More research is needed, however, to determine the mechanisms, kinetics and importance of mucosal-repertoire ageing in turquoise killifish.

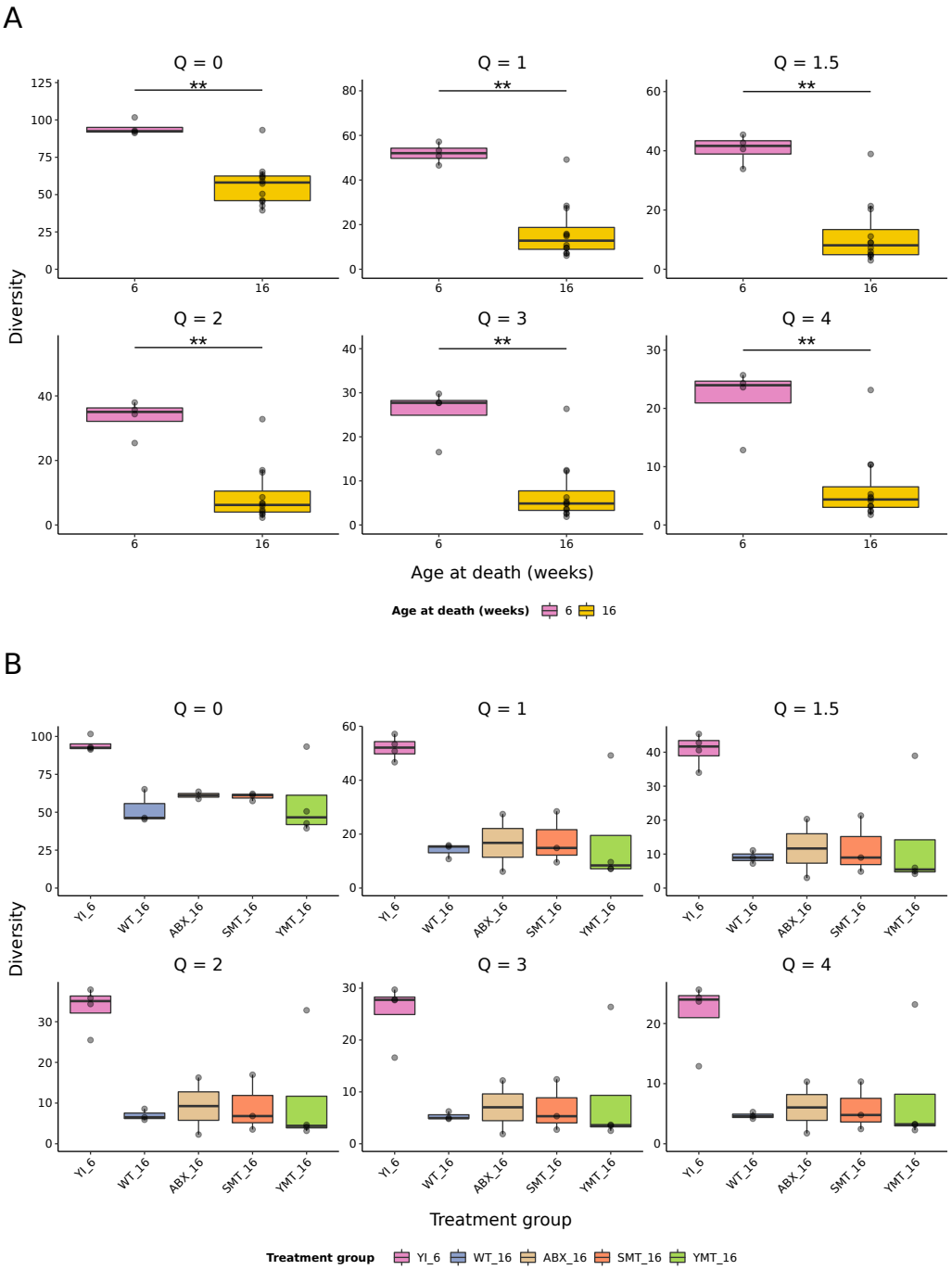


Figure 4.45: Comparing VJ alpha diversities between age and treatment groups in the IgSeq gut dataset: Boxplots of VJ Hill diversity values for the antibody repertoires of individuals of each (A) age group and (B) treatment group in the IgSeq gut dataset at a sample of diversity orders. Pairwise p-values were computed using nonparametric Mann-Whitney U tests ($*: 0.01 < p \leq 0.05$; $**: 0.001 < p \leq 0.01$; $***: p \leq 0.001$).

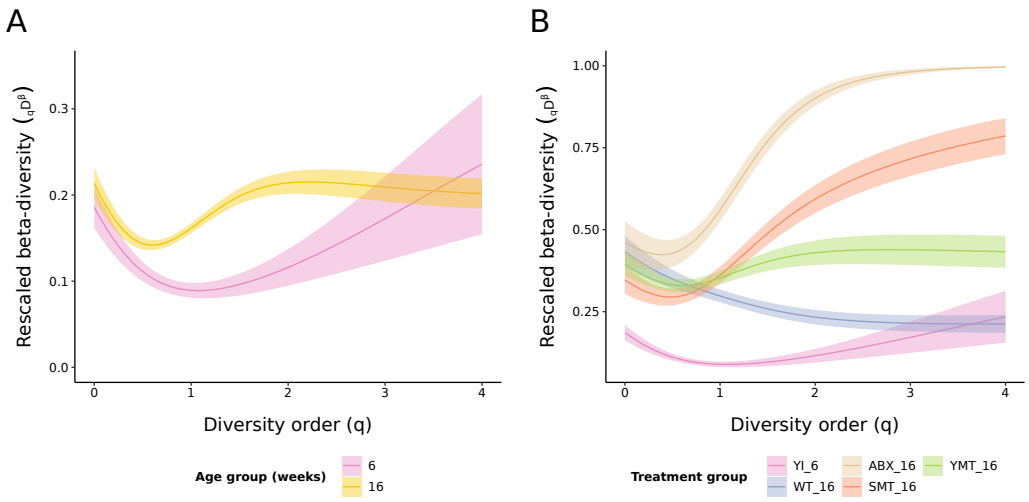


Figure 4.46: VJ beta-diversity spectra for the IgSeq gut dataset: Bootstrapped beta-diversity spectra of VJ usage for each (A) age group and (B) treatment group in the IgSeq gut dataset, as measured by number of unique sequences per unambiguous VJ identity, rescaled to between 0 (minimum) and 1 (maximum) for each individual. Shaded regions in both subfigures represent 95 % confidence intervals, estimated using bootstrapping.

In sharp contrast to the age-related changes observed in the gut antibody repertoire, there is no clear evidence from these analyses supporting the hypothesis that gut-microbiota transfer from young fish rejuvenates the antibody repertoire. While the alpha diversity spectrum of the YMT_16 group appears to be slightly higher than the SMT_16/ABX_16 groups at very low diversity orders (Figure 4.42B), these indications are not borne out by subsequent statistical analysis, and it is precisely these low-order diversity metrics that are most vulnerable to sampling bias and undersampling (Appendix C.1.3), a pervasive problem in immune-repertoire sequencing [30]. There is also no obvious difference in VJ alpha diversity between treatment groups, while the beta diversity of the YMT_16 group is non-significantly *higher* than that of the untreated control group (Figures 4.46B and 4.48C). Overall, whatever mechanism underlies the effect of gut microbiota transfer on killifish lifespan [78], it does not appear to be operating through modulation of gut antibody-repertoire diversity.

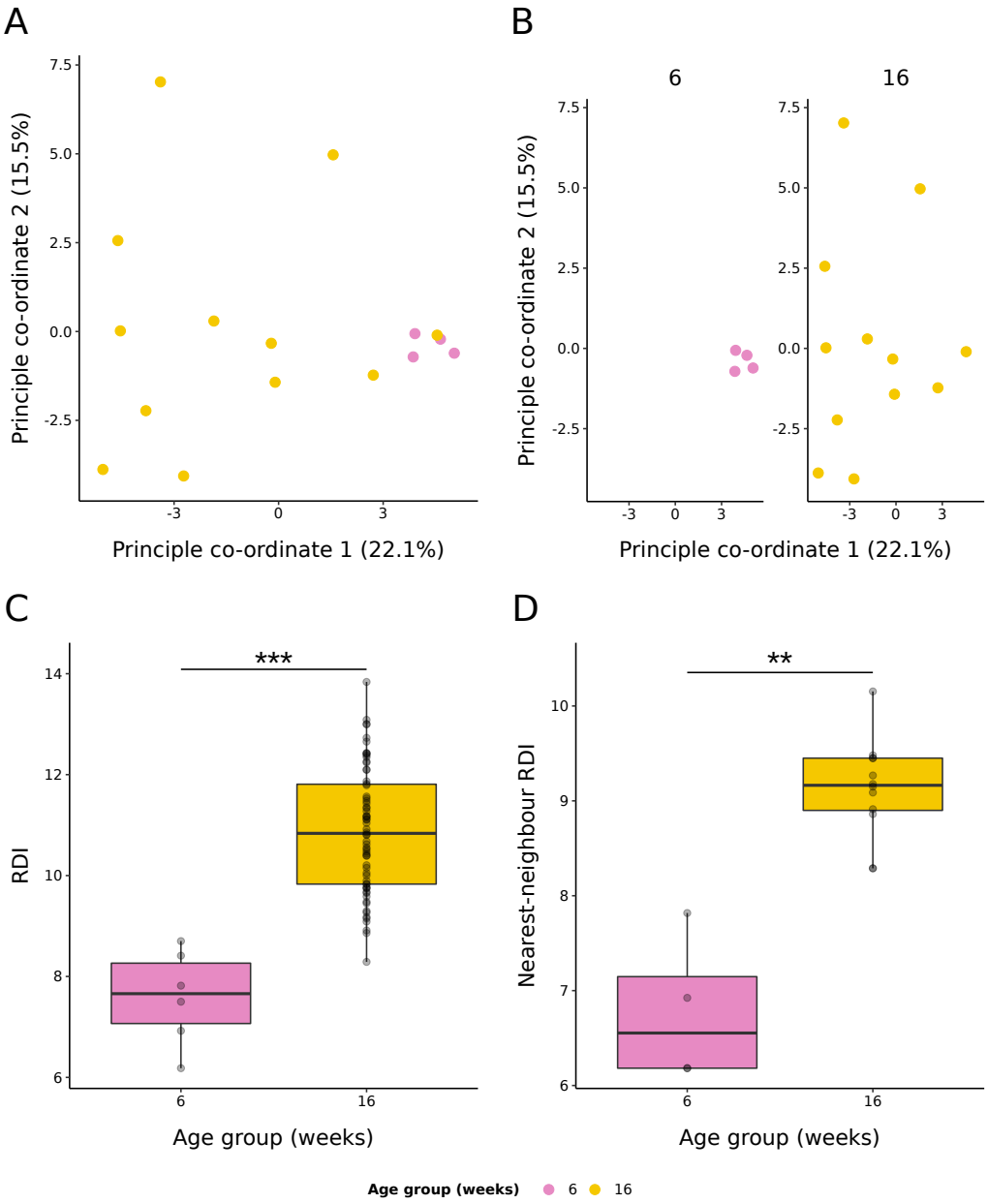


Figure 4.47: Intra-age-group variability in VJ expression in the IgSeq gut dataset: (A-B) Principal co-ordinate analysis (PCoA) of pairwise inter-individual VJ-RDI distances in the IgSeq gut dataset, coloured by age group and displayed together (A) or separately by age group (B). (C-D) Boxplots of overall (C) and nearest-neighbour (D) inter-individual VJ-RDI distances for each age group in the dataset. Pairwise p-values were computed using nonparametric Mann-Whitney U tests ($*: 0.01 < p \leq 0.05$; $**: 0.001 < p \leq 0.01$; $***: p \leq 0.001$).

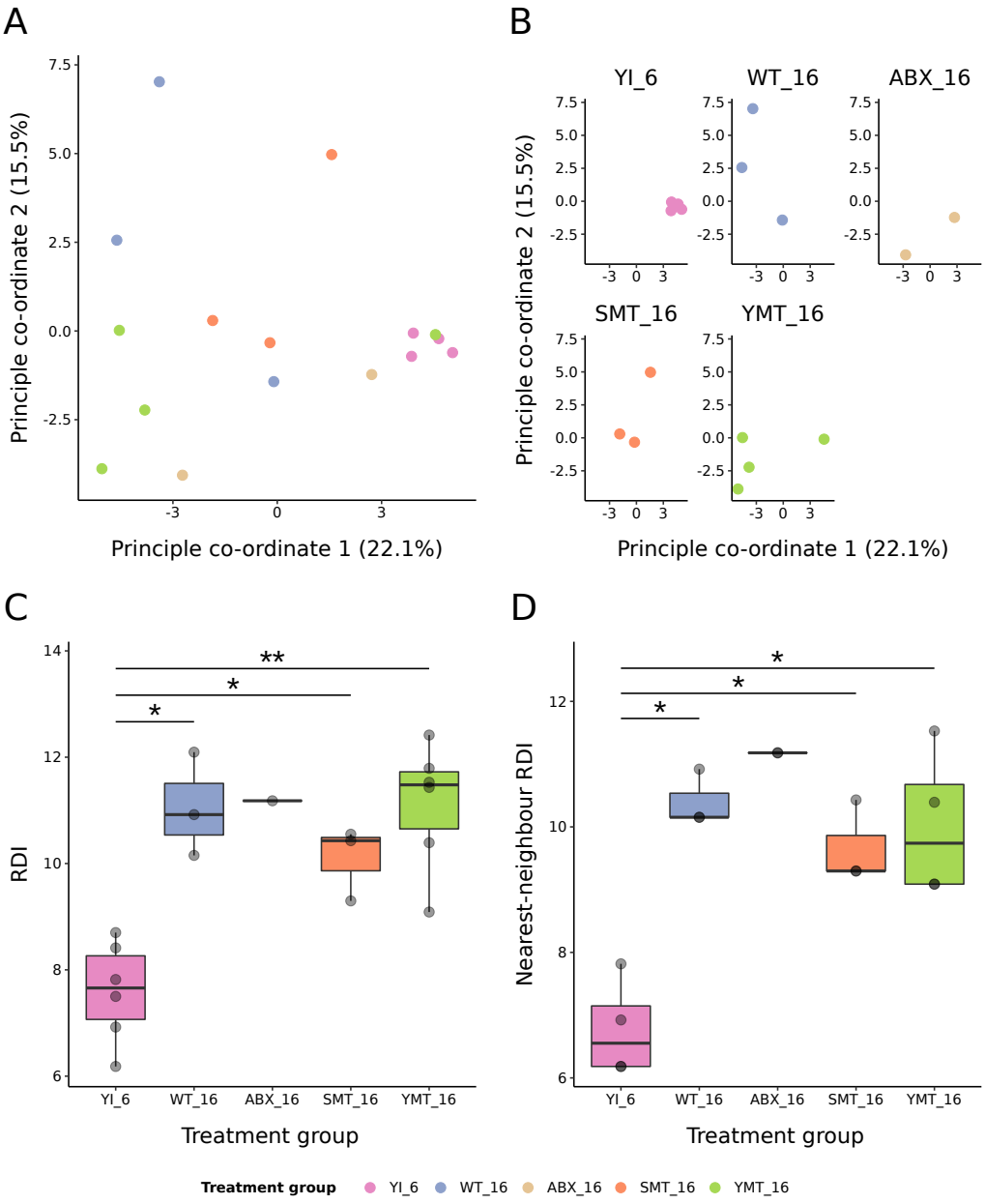


Figure 4.48: Intra-treatment-group variability in VJ expression in the IgSeq gut dataset: (A-B) Principal co-ordinate analysis (PCoA) of pairwise inter-individual VJ-RDI distances in the IgSeq gut dataset, coloured by treatment group and displayed together (A) or separately by treatment group (B). (C-D) Boxplots of overall (C) and nearest-neighbour (D) inter-individual VJ-RDI distances for each treatment group in the dataset. Pairwise p-values were computed using nonparametric Mann-Whitney U tests ($*: 0.01 < p \leq 0.05$; $**: 0.001 < p \leq 0.01$; $***: p \leq 0.001$).

4.7 Discussion

Apart from the nervous system itself, no other system in the vertebrate body exhibits such complex learning behaviour as the adaptive immune system. Through the combination of an extremely diverse generative repertoire, antigen-dependent clonal expansion, and extremely long-lasting immune memory, the adaptive immune system progressively learns to model its immune environment and predict – and defend against – future pathogenic attacks [4]. During ageing, however, this predictive and protective ability progressively declines through a combination of a loss of adaptivity and overexpansion of previously established memory clones, leaving the body unable to effectively protect itself from novel infections or establish adequate immune memory in response to vaccination or other antigenic challenges (Section 1.3). Learning more about the mechanisms, spatial distribution and temporal progression of this age-related decline in adaptive immunity constitutes an essential part of the study of vertebrate ageing.

In this chapter, I established the first working library-preparation protocol and sequence-analysis pipeline for immunoglobulin sequencing in turquoise killifish (*Nothobranchius furzeri*), a highly promising emerging model for vertebrate ageing and immunosenescence (Section 1.4), and investigated the composition, diversity and ageing of *IGHM* antibody repertoires in adult males of this species. The results demonstrated that killifish express diverse and individualised antibody repertoires, with thousands of clones and billions of potential unique sequences per individual, which are both reproducible between replicates from the same individual and distinguishable between different individuals of the same age and inbred strain (Sections 4.4.2 and 4.4.4). With the exception of the very largest clones, these repertoires conformed to the heavy-tailed Zipf-like distribution of clone sizes observed in other species, and thus exhibited a clonal structure divided between a very large number of very small clones and a much smaller number of highly expanded clones.

Over the course of adulthood, these larger clones exhibit a progressive loss of diversity due to accumulating clonal expansions, leading to reduced alpha diversity of the clonal repertoire at higher diversity orders (when small clones are strongly downweighted compared to large ones) and a progressive loss in alpha VJ-diversity among large clones (Section 4.5). At the same time, the most expanded V/J-combinations become progressively more different in older killifish, resulting in an increase in beta diversity which is particularly strong at high diversity orders. This loss of alpha diversity and increase in beta diversity with age in turquoise killifish is in line with observations made previously in human peripheral blood (Section 1.3). Small clones, by contrast, appear to exhibit much less change in repertoire composition with age in whole-body killifish samples: there is a non-significant decline in total clonal count and low-order clonal alpha-diversity in older killifish antibody repertoires, and any significant change in alpha VJ-diversity disappears when small clones are included. In addition, the complexity of the generative process giving rise to novel antibody sequences does not appear to change between age groups, suggesting that newly-produced naïve B-cells may be of similar diversity in young and old killifish.

These results do not necessarily demonstrate that the repertoire composition of small clones in turquoise killifish is unchanged with age; several metrics, including clonal richness, decline non-significantly with age, and it may be the case that significant changes would become apparent with larger sample sizes or a different set of data-generation and -analysis methods. A comprehensive assessment of naïve B-cell repertoires in ageing killifish would require more specialised laboratory techniques (such as cell sorting of B-cell subtypes) and examination of processes such as primary selection that have been neglected in this analysis. One important alternative explanation relates to body size: unlike humans and mice, turquoise killifish grow continuously with age (Tables 4.1 and E.24), and it is possible that an age-related decline in the rate of B-cell output by the primary lymphoid organs is offset by a continuous increase in the size of these organs with age. Nevertheless, these results do suggest that the bulk of the change in the *IGHM* repertoire of ageing killifish arises from changes in the size and composition of expanded antigen-experienced clones rather than changes in the naïve compartment. If this is the case, it raises the question of to what extent these changes, including the observed loss in diversity, represent functional declines as a result of ageing, rather than the expected and functional result of progressive immune adaptation to a controlled laboratory environment. Such a question is an important one, but lies outside the scope of the experiments and analyses presented in this chapter.

As discussed above, the inferred entropy of the heavy-chain generative process in turquoise killifish does not appear to change much with age, with estimates varying between 30 and 32 bits (Section 4.4.4). Antibody generation in turquoise killifish is therefore a highly complex and diverse process, with the number of potentially-generable sequences exceeding 10^9, but is nevertheless vastly less diverse than the equivalent process in humans (which can potentially produce in excess of 10^{21} different sequences). A lower potential repertoire diversity in killifish compared to humans is not necessarily surprising, given the much greater size and number of gene segments in the human *IGH* locus; unexpectedly, however, most of the observed difference in diversity does not reflect a general discrepency across all contributing processes, but rather a specific and very large (in excess of 40 bits) difference in the complexity of the N-insertion process alone. This specific difference reflects the very different N-insertion distributions in humans and killifish, with the former showing much larger mean, median and modal numbers of insertions compared to the latter [38]. The cause of this difference is not known, nor whether it is specific to turquoise killifish or a more general feature of teleost antibody repertoires; however, it seems likely to involve the structure, expression or regulation of the terminal dideoxy transferase enzyme (TdT) responsible for the N-insertion process. At least one orthologue of human TdT is annotated in the turquoise killifish genome, and more detailed investigation of the structure and function of this enzyme in *Nothobranchius furzeri* and related species may reveal unexpected insights into the mechanism and regulation of the N-insertion process during primary antibody diversification.

As early as 2009, Caruso *et al.* [198] commented that the mucosal adaptive immune system might exhibit particularly strong loss of diversity with age, as a result of especially frequent antigenic

challenge. Nevertheless, to my knowledge, no previous study has specifically investigated antibody-repertoire ageing in a mucosal immune organ such as the gut. In this chapter, I performed the first such analysis, investigating changes in antibody diversity in RNA samples from the guts of young and old killifish (Section 4.6). The results strongly confirmed Caruso *et al.*'s prediction: not only did gut mucosal antibody repertoires exhibit a much greater degree of clonal expansion compared to the whole-body repertoire (as shown by differences in both P20 and the proportion of large clones), but both the clonal and VJ-repertoires exhibited a dramatic drop in diversity with age across all diversity orders, to a much larger extent than observed in whole-body samples; as in the whole-body repertoire, these changes were also accompanied by a large increase in beta VJ-diversity between older individuals. These results demonstrate that the killifish gut exhibits dramatic changes in both clonal structure and VJ-usage with age; given the much greater predominance of expanded clones in the gut repertoire, such differences in ageing phenotypes between the gut and whole-body repertoires further support the conclusion that changes in repertoire composition with age in the killifish are primarily driven by expanded, antigen-experienced clones rather than the naïve compartment. Given these results, it would be interesting to investigate other mucosal immune organs (such as the gills and skin [26]) to see if these ageing phenotypes represent a general pattern or are specific to the intestinal mucosa.

In stark contrast to the striking drop in diversity with age in the killifish intestinal repertoire, no change in repertoire diversity was observed as a result of gut microbial transfer from young donor fish. In part, the small size of each treatment group among the old gut samples makes drawing any strong conclusions difficult, but nevertheless there was no indication that the young-donor group showed any delay or reversal in any of the ageing phenotypes observed in the mucosal antibody repertoire. This lack of change in the repertoire is to some extent surprising, as microbiota transfer has been shown to affect gut microbial diversity in older fish, which might be expected to in turn affect the diversity of the antibody repertoire; the lack of effect observed in this chapter suggests that the link between microbial and immune-repertoire diversity is slower, weaker or more indirect than this simple description would suggest, though more research is needed to determine exactly what the connection between the diversities of the two systems might in fact be.

Despite the apparent lack of effect on intestinal antibody-repertoire diversity resulting from gut microbial transfer, the effect of anti-ageing interventions on age-related changes in the antibody repertoire remains an important and underexplored topic of study. To my knowledge, nothing is known about how well-established pro-longevity interventions like dietary restriction, rapamycin treatment or reduced insulin/IGF1 signaling [199] affect the ageing of the antibody repertoire, whether in the gut, peripheral blood or elsewhere. Given its short lifespan, highly scalable husbandry conditions, and demonstrated age-dependent changes in the antibody repertoire, the turquoise killifish would be an ideal model organism in which to elaborate this intersection between the biology of ageing and vertebrate immunology, with the potential to identify new and important interrelationships between longevity pathways and the adaptive immune system.

As suggested by the foregoing paragraph, the work included in this chapter is only the tip of an iceberg of potential research into adaptive immunosenescence that could be performed with the turquoise killifish. Even using the datasets already generated for this chapter, a great deal of research remains to be done on the state and ageing of antibody repertoires in this species. Particularly important topics left unexplored in this thesis include the rate, distribution and ageing of somatic hypermutation in killifish repertoires; the efficacy of clonal selection during affinity maturation (Section 1.2.4); and whether any age-related changes affect the primary-selection process separating the generation of new antibody sequences from the naïve antibody repertoire. Further experiments could substantially improve our understanding of the origins and ageing of killifish antibody repertoires by incorporating light-chain sequence and heavy-chain/light-chain pairing information or by supplementing bulk repertoire-sequencing data with higher-resolution single-cell data from sorted naïve or antigen-experienced B-cells [200], while studies involving infection, vaccination or other immune interventions could explore the functional implications of a loss of repertoire diversity with age.

Such experiments, however, would require the development of important additional resources for the turquoise killifish model, including reliable antibodies and cell-surface markers, characterisation of the light-chain gene loci, and optimised single-cell repertoire-sequencing techniques, work that is still ongoing in several killifish labs worldwide. Simpler but still-valuable experiments could investigate a greater range of timepoints and immune organs, incorporate data from *IGHD* or separate secretory and transmembrane *IGHM* repertoires, or extend our knowledge of killifish antibody repertoires to development and early adulthood. Eventually, a whole-body, whole-lifespan atlas of adaptive-immune ageing in the killifish could reveal a great deal of information about how and why the antibody repertoires of ageing vertebrates exhibit the compositional and functional changes they do, with important consequences for biogerontological research, comparative immunology, and drug development.

Conclusion

One of the most remarkable features of the adaptive immune system is its sheer proteanism: at all levels of vertebrate biology and evolution, it is constantly changing, shifting in response to both stochastic processes and the adaptive pressures imposed by pathogens, internal selection mechanisms and the rest of the immune system. At the level of a single lymphocyte clone, the dynamic processes of VDJ recombination and junctional diversity give rise to a controlled yet hugely variable diversity of gene sequences, while affinity maturation results in repeated flowerings of even more sequence diversity from single naïve B-cell ancestors (Section 1.2). At the level of the organism, the constant production of new antigen-receptor sequences combined with the rapid evolution of lymphocyte lineages gives rise to a constantly fluctuating population of breathtaking complexity, exhibiting both a huge naïve sequence diversity and constantly-accumulating layers of long-term immune memory. Over the entirety of an individual's lifespan, the effects of ageing add additional levels of change to this already-mercurial population, resulting in immune repertoires in old age which differ systematically from those expressed in the young (Sections 1.3 and 4.5). And at the level of a species lineage, the rapid genomic evolution of antigen-receptor genes results in still more variability, with closely-related species and even conspecific individuals [195] differing in their locus organisation, their VH segments, or even the constant regions they possess (Chapter 3).

In this thesis, I have made use of an important emerging model organism, the turquoise killifish (*Nothobranchius furzeri*), and its close relatives to investigate several of these levels of variability in one of the most important and diverse antigen-receptor genes in the adaptive immune system, the immunoglobulin heavy chain (*IGH*). I have shown that this gene is highly variable in its composition and structure in the Cyprinidontiform lineage of teleost fishes, with large differences in sublocus organisation and V/D/J segment availability between species and even repeated gains and losses of an important antibody isoform, the mucosally-specialised *IGHZ* (Chapter 3). Focusing on the turquoise killifish itself, I have shown that this species exhibits an intact and highly diverse heavy-chain repertoire, with thousands of unique clones and billions of potential unique antibody sequences per individual and highly individualised V/J segment usage (Section 4.4). Consistent with the extremely short lifespan and rapid ageing phenotypes of the turquoise killifish, its secondary antibody repertoire exhibits a rapid age-associated decline in diversity over a timescale of weeks, both in the whole-body repertoire and the specialised mucosal repertoire of the gut, as well as increased divergence in

repertoire composition between older individuals (Sections 4.5 and 4.6). These changes, however, appear to be concentrated in expanded clones, with only non-significant declines in diversity observed in the small, naïve clones that make up most of the clonal richness of the repertoire. Whether this distinction represents a genuine preservation of primary-repertoire diversity with age, a confounding of falling primary diversity by increasing body size, or some other phenomenon remains to be seen; in my opinion, any of these cases would prove to be an interesting discovery.

In Chapter 1, I described the objective of this thesis as to "establish the turquoise killifish as a model for the study of comparative immunology and humoral adaptive immunosenescence". My goal has always been to start something new, not finish it. To my knowledge, no previously-published research has ever investigated *IGH* locus organisation in so many closely-related species (Section 3.4), characterised antibody-repertoire ageing in such a short-lived species (Section 4.5), or looked at repertoire ageing specifically in a mucosal immune organ (Section 4.6). In all cases, however, I've only scratched the surface of what can be learned with these models, methods and mindset. Despite everything that we've learned about the adaptive immune system and its ageing, the sheer distributed complexity of both immunity and ageing makes really *understanding* the intersection of the two a daunting task. My hope is that the work I've done here can help contribute to a better understanding of this intersection, not just in humans and mice but in all organisms possessing this unique and remarkable suite of adaptations we call adaptive immunity.

References

1. *Pathogen-Host Interactions: Antigenic Variation V. Somatic Adaptations* (eds Hsu, E. & Du Pasquier, L.) (Springer, 2015).
2. Jack, R. S. *Evolution of Immunity and Pathogens* in *Pathogen-Host Interactions: Antigenic Variation V. Somatic Adaptations* (eds Hsu, E. & Du Pasquier, L.) 1–20 (Springer, 2015).
3. Kurosaki, T., Kometani, K. & Ise, W. Memory B cells. *Nature Reviews Immunology* **15,** 149–159 (2015).
4. Mayer, A. *et al.* How a well-adapting immune system remembers. *arXiv,* 1806.05753 (2018).
5. Sompayrac, L. *How the Immune System Works* 4th ed. (Wiley-Blackwell, 2012).
6. Schroeder, H. W. & Cavacini, L. Structure and function of immunoglobulins. *Journal of Allergy and Clinical Immunology* **125,** S41–S52 (2010).
7. Bengtén, E. & Wilson, M. *Antibody Repertoires in Fish* in *Pathogen-Host Interactions: Antigenic Variation V. Somatic Adaptations* (eds Hsu, E. & Du Pasquier, L.) 193–234 (Springer, 2015).
8. Van Zelm, M. C. *et al.* An Antibody-Deficiency Syndrome Due to Mutations in the CD19 Gene. *New England Journal of Medicine* **354,** 1901–1912 (2006).
9. Sambhara, S. & McElhaney, J. E. *Immunosenescence and Influenza Vaccine Efficacy* in *Vaccines for Pandemic Influenza* (eds Compans, R. W. & Orenstein, W. A.) 413–429 (Springer, 2009).
10. Georgiou, G. *et al.* The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature Biotechnology* **32,** 158–168 (2014).
11. Gerhard, G. S. *et al.* Life spans and senescent phenotypes in two strains of Zebrafish (*Danio rerio*). *Experimental Gerontology* **37,** 1055–1068 (2002).
12. Bowler, J. K. *Longevity of reptiles and amphibians in North American collections as of 1 November, 1975* (Society for the Study of Amphibians and Reptiles, 1977).
13. Yuan, R. *et al.* Aging in inbred strains of mice: study design and interim report on median lifespans and circulating IGF1 levels. *Aging Cell* **8,** 277–287 (2009).
14. Boehm, T. Design principles of adaptive immune systems. *Nature Reviews Immunology* **11,** 307–317 (2011).
15. Kasahara, M. *Variable Lymphocyte Receptors: A Current Overview* in *Pathogen-Host Interactions: Antigenic Variation V. Somatic Adaptations* (eds Hsu, E. & Du Pasquier, L.) 175–192 (Springer, 2015).
16. Mix, E., Goertsches, R. & Zett, U. K. Immunoglobulins—Basic considerations. *Journal of Neurology* **253,** v9–v17 (2006).
17. Sunyer, J. O. Fishing for mammalian paradigms in the teleost immune system. *Nature Immunology* **14,** 320–326 (2013).

18. Shirai, H., Kidera, A. & Nakamura, H. H3-rules: identification of CDR-H3 structures in antibodies. *FEBS Letters* **455,** 188–197 (1999).
19. Mashoof, S. & Criscitiello, M. F. Fish Immunoglobulins. *Biology* **5,** 45 (2016).
20. Zhang, Y.-A. *et al.* IgT, a primitive immunoglobulin class specialized in mucosal immunity. *Nature Immunology* **11,** 827–835 (2010).
21. Senger, K. *et al. Antibody Isotype Switching in Vertebrates* in *Pathogen-Host Interactions: Antigenic Variation V. Somatic Adaptations* (eds Hsu, E. & Du Pasquier, L.) 295–324 (Springer, 2015).
22. Fillatreau, S. *et al.* The astonishing diversity of Ig classes and B cell repertoires in teleost fish. *Frontiers in Immunology* **4,** 28 (2013).
23. Bengtén, E. *et al.* Structure of the catfish IGH locus: analysis of the region including the single functional *IGHM* gene. *Immunogenetics* **58,** 831–844 (2006).
24. Ramirez-Gomez, F. *et al.* Discovery and Characterization of Secretory IgD in Rainbow Trout: Secretory IgD Is Produced through a Novel Splicing Mechanism. *The Journal of Immunology* **188,** 1341–1349 (2012).
25. Magadán-Mompó, S., Sánchez-Espinel, C. & Gambón-Deza, F. Immunoglobulin heavy chains in medaka (*Oryzias latipes*). *BMC Evolutionary Biology* **11,** 165 (2011).
26. Xu, Z. *et al.* Teleost skin, an ancient mucosal surface that elicits gut-like immune responses. *PNAS* **110,** 13097–13102 (2013).
27. Bao, Y. *et al.* The immunoglobulin gene loci in the teleost *Gasterosteus aculeatus*. *Fish & Shellfish Immunology* **28,** 40–48 (2010).
28. Gambón-Deza, F., Sánchez-Espinel, C. & Magadán-Mompó, S. Presence of an unique IgT on the IGH locus in three-spined stickleback fish (*Gasterosteus aculeatus*) and the very recent generation of a repertoire of VH genes. *Developmental & Comparative Immunology* **34,** 114–122 (2010).
29. Savan, R. *et al.* Discovery of a new class of immunoglobulin heavy chain from fugu. *European Journal of Immunology* **35,** 3320–3331 (2005).
30. Mora, T. & Walczak, A. Quantifying lymphocyte receptor diversity. *arXiv,* 1604.00487 (2016).
31. David Jung *et al.* Mechanism and Control of V(d)j Recombination at the Immunoglobulin Heavy Chain Locus. *Annual Review of Immunology* **24,** 541–570 (2006).
32. Kogut, I. *et al.* B cell maintenance and function in aging. *Seminars in Immunology* **24,** 342–349 (2012).
33. Lefranc, M.-P. *et al.* IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Developmental & Comparative Immunology* **27,** 55–77 (2003).
34. Ruiz, M. *et al.* The human immunoglobulin heavy diversity (IGHD) and joining (IGHJ) segments. *Experimental and Clinical Immunogenetics* **16,** 173–184 (1999).
35. Schatz, D. G. & Swanson, P. C. V(D)J Recombination: Mechanisms of Initiation. *Annual Review of Genetics* **45,** 167–202 (2011).
36. Hesse, J. E. *et al.* V(D)J recombination: a functional definition of the joining signals. *Genes & Development* **3,** 1053–1061 (1989).
37. Lefranc, M. P. Nomenclature of the human immunoglobulin heavy (IGH) genes. *Experimental and Clinical Immunogenetics* **18,** 100–116 (2001).
38. Elhanati, Y. *et al.* Inferring processes underlying B-cell repertoire diversity. *Phil. Trans. R. Soc. B* **370,** 20140243 (2015).

39. Flaherty, D. K. *Antibody Diversity* in *Immunology for Pharmacy* 79–86 (Mosby, 2012).
40. Hansen, J. D., Landis, E. D. & Phillips, R. B. Discovery of a unique Ig heavy-chain isotype (IgT) in rainbow trout: Implications for a distinctive B cell developmental pathway in teleost fish. *PNAS* **102,** 6919–6924 (2005).
41. Yasuike, M. *et al.* Evolution of duplicated IgH loci in Atlantic salmon, *Salmo salar*. *BMC Genomics* **11,** 486 (2010).
42. Xiao, F. S. *et al.* Ig heavy chain genes and their locus in grass carp *Ctenopharyngodon idella*. *Fish & Shellfish Immunology* **29,** 594–599 (2010).
43. Dunn-Walters, D. K. & Ademokun, A. A. B cell repertoire and ageing. *Current Opinion in Immunology* **22,** 514–520 (2010).
44. Noia, J. M. D. & Neuberger, M. S. Molecular Mechanisms of Antibody Somatic Hypermutation. *Annual Review of Biochemistry* **76,** 1–22 (2007).
45. Magor, B. G. Antibody Affinity Maturation in Fishes—Our Current Understanding. *Biology* **4,** 512–524 (2015).
46. Victora, G. D. & Nussenzweig, M. C. Germinal Centers. *Annual Review of Immunology* **30,** 429–457 (2012).
47. Howard, W. A., Gibson, K. L. & Dunn-Walters, D. K. Antibody Quality in Old Age. *Rejuvenation Research* **9,** 117–125 (2006).
48. Marcou, Q., Mora, T. & Walczak, A. M. High-throughput immune repertoire analysis with IGoR. *Nature Communications* **9,** 561 (2018).
49. Segre, D. & Segre, M. Age-related changes in B and T lymphocytes and decline of humoral immune responsiveness in aged mice. *Mechanisms of Ageing and Development* **6,** 115–129 (1977).
50. Ademokun, A., Wu, Y.-C. & Dunn-Walters, D. The ageing B cell population: Composition and function. *Biogerontology* **11,** 125–137 (2010).
51. Montecino-Rodriguez, E., Berent-Maoz, B. & Dorshkind, K. Causes, consequences, and reversal of immune system aging. *Journal of Clinical Investigation* **123,** 958–965 (2013).
52. Labrie, J. E. *et al.* Bone Marrow Microenvironmental Changes Underlie Reduced RAG-mediated Recombination and B Cell Generation in Aged Mice. *Journal of Experimental Medicine* **200,** 411–423 (2004).
53. Mehr, R. & Melamed, D. Reversing B cell aging. *Aging* **3,** 438–443 (2011).
54. Aberle, J. H. *et al.* Mechanistic insights into the impairment of memory B cells and antibody production in the elderly. *AGE* **35,** 371–381 (2013).
55. Frasca, D. *et al.* Age effects on B cells and humoral immunity in humans. *Ageing Research Reviews* **10,** 330–335 (2011).
56. Blomberg, B. B. & Frasca, D. Age effects on mouse and human B cells. *Immunologic Research* **57,** 354–360 (2013).
57. Siegrist, C.-A. & Aspinall, R. B-cell responses to vaccination at the extremes of age. *Nature Reviews Immunology* **9,** 185–194 (2009).
58. Tabibian-Keissar, H. *et al.* Aging affects B-cell antigen receptor repertoire diversity in primary and secondary lymphoid tissues. *European Journal of Immunology* **46,** 480–492 (2016).
59. Sasaki, S. *et al.* Limited efficacy of inactivated influenza vaccine in elderly individuals is associated with decreased production of vaccine-specific antibodies. *The Journal of Clinical Investigation* **121,** 3109–3119 (2011).

60. Frasca, D. & Blomberg, B. B. Effects of aging on B cell function. *Current Opinion in Immunology* **21,** 425–430 (2009).
61. Henry, C. *et al.* Influenza Virus Vaccination Elicits Poorly Adapted B Cell Responses in Elderly Individuals. *Cell Host & Microbe* **25,** 357–366 (2019).
62. Gibson, K. L. *et al.* B-cell diversity decreases in old age and is correlated with poor health status. *Aging Cell* **8,** 18–25 (2009).
63. Weinstein, J. A. *et al.* High-Throughput Sequencing of the Zebrafish Antibody Repertoire. *Science* **324,** 807–810 (2009).
64. Jiang, N. *et al.* Lineage Structure of the Human Antibody Repertoire in Response to Influenza Vaccination. *Science Translational Medicine* **5,** 171ra19 (2013).
65. Wang, C. *et al.* Effects of Aging, Cytomegalovirus Infection, and EBV Infection on Human B Cell Repertoires. *The Journal of Immunology* **192,** 603–611 (2014).
66. De Bourcy, C. F. A. *et al.* Phylogenetic analysis of the human antibody repertoire reveals quantitative signatures of immune senescence and aging. *PNAS* **114,** 1105–1110 (2017).
67. Jiang, N. *et al.* Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *PNAS* **108,** 5348–5353 (2011).
68. Krasnov, A., Jørgensen, S. M. & Afanasyev, S. Ig-seq: Deep sequencing of the variable region of Atlantic salmon IgM heavy chain transcripts. *Molecular Immunology* **88,** 99–105 (2017).
69. Lund, H. *et al.* A time-course study of gene expression and antibody repertoire at early time post vaccination of Atlantic salmon. *Molecular Immunology* **106,** 99–107 (2019).
70. Fu, X. *et al.* High-throughput sequencing of the expressed torafugu (Takifugu rubripes) antibody sequences distinguishes IgM and IgT repertoires and reveals evidence of convergent evolution. *Frontiers in Immunology* **9,** 251 (2018).
71. Belkaid, Y. & Hand, T. W. Role of the microbiota in immunity and inflammation. *Cell* **157,** 121–141 (2014).
72. Valdesalici, S. & Cellerino, A. Extremely short lifespan in the annual fish *Nothobranchius furzeri*. *Proc. R. Soc. Lond. B* **270,** S189–S191 (2003).
73. Genade, T. *et al.* Annual fishes of the genus *Nothobranchius* as a model system for aging research. *Aging Cell* **4,** 223–233 (2005).
74. Jubb, R. A new *Nothobranchius* (Pisces, Cyprinodontidae) from Southeastern Rhodesia. *Journal of the American Killifish Association* **8,** 314–321 (1971).
75. Terzibasi, E. *et al.* Large Differences in Aging Phenotype between Strains of the Short-Lived Annual Fish *Nothobranchius furzeri*. *PLOS One* **3,** e3866 (2008).
76. Kirschner, J. *et al.* Mapping of quantitative trait loci controlling lifespan in the short-lived fish *Nothobranchius furzeri* – a new vertebrate model for age research. *Aging Cell* **11,** 252–261 (2012).
77. Valenzano, D. R. *et al.* The African Turquoise Killifish Genome Provides Insights into Evolution and Genetic Architecture of Lifespan. *Cell* **163,** 1539–1554 (2015).
78. Smith, P. *et al.* Regulation of life span by the gut microbiota in the short-lived African turquoise killifish. *eLife* **6,** e27014 (2017).
79. Harel, I., Valenzano, D. R. & Brunet, A. Efficient genome engineering approaches for the short-lived African turquoise killifish. *Nature Protocols* **11,** 2010–2028 (2016).
80. Valenzano, D. R. *et al.* Resveratrol Prolongs Lifespan and Retards the Onset of Age-Related Markers in a Short-Lived Vertebrate. *Current Biology* **16,** 296–300 (2006).

81. Valenzano, D. R. & Cellerino, A. Resveratrol and the Pharmacology of Aging: A New Vertebrate Model to Validate an Old Molecule. *Cell Cycle* **5,** 1027–1032 (2006).
82. Di Cicco, E. *et al.* The short-lived annual fish *Nothobranchius furzeri* shows a typical teleost aging process reinforced by high incidence of age-dependent neoplasias. *Experimental Gerontology* **46,** 249–256 (2011).
83. Valenzano, D. R. *et al.* Temperature affects longevity and age-related locomotor and cognitive decay in the short-lived fish *Nothobranchius furzeri*. *Aging Cell* **5,** 275–278 (2006).
84. Terzibasi, E. *et al.* Effects of dietary restriction on mortality and age-related phenotypes in the short-lived fish *Nothobranchius furzeri*. *Aging Cell* **8,** 88–99 (2009).
85. Terzibasi, E., Valenzano, D. R. & Cellerino, A. The short-lived fish *Nothobranchius furzeri* as a new model system for aging studies. *Experimental Gerontology* **42,** 81–89 (2007).
86. Reichwald, K. *et al.* High tandem repeat content in the genome of the short-lived annual fish *Nothobranchius furzeri*: a new vertebrate model for aging research. *Genome Biology* **10,** R16 (2009).
87. Valenzano, D. R. *et al.* Mapping Loci Associated With Tail Color and Sex Determination in the Short-Lived Fish *Nothobranchius furzeri*. *Genetics* **183,** 1385–1395 (2009).
88. Reichwald, K. *et al.* Insights into Sex Chromosome Evolution and Aging from the Genome of a Short-Lived Fish. *Cell* **163,** 1527–1538 (2015).
89. Willemsen, D. *et al.* Intra-Species Differences in Population Size shape Life History and Genome Evolution. *bioRxiv,* 852368 (2019).
90. Hughes, L. C. *et al.* Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. *PNAS* **115,** 6249–6254 (2018).
91. Dodzian, J. *et al.* A protocol for laboratory housing of turquoise killifish (*Nothobranchius furzeri*). *J. Vis. Exp.* **134,** 57073 (2018).
92. Carter, K. M., Woodley, C. M. & Brown, R. S. A review of tricaine methanesulfonate for anesthesia of fish. *Reviews in Fish Biology and Fisheries* **21,** 51–59 (2011).
93. Paul, N., Shum, J. & Le, T. *Hot Start PCR* in *RT-PCR Protocols* (eds King, N. & O'Connell, J.) 2nd ed., 301–318 (Humana Press, 2010).
94. Hawkins, T. L. *et al.* DNA purification and isolation using a solid-phase. *Nucleic Acids Research* **22,** 4543–4544 (1994).
95. DeAngelis, M. M., Wang, D. G. & Hawkins, T. L. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Research* **23,** 4742–4743 (1995).
96. Lennon, N. J. *et al.* A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454. *Genome Biology* **11,** R15 (2010).
97. Fisher, S. *et al.* A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biology* **12,** R1 (2011).
98. Zumbo, P. *Phenol-chloroform extraction.* Weill Cornell Medical College (2012).
99. Zumbo, P. *Ethanol precipitation.* Weill Cornell Medical College (2012).
100. Chomczynski, P. & Sacchi, N. The single-step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction: twenty-something years on. *Nature Protocols* **1,** 581–585 (2006).
101. Magloire, A. *BluePippin DNA Size Selection System Operations Manual.* Sage Science Inc. (2016).

102. Birnboim, H. C. & Doly, J. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Research* **7,** 1513–1523 (1979).

103. Turchaninova, M. A. *et al.* High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nature Protocols* **11,** 1599–1616 (2016).

104. Untergasser, A. *et al.* Primer3—new capabilities and interfaces. *Nucleic Acids Research* **40,** e115 (2012).

105. Illumina. *Illumina Adapter Sequences,* version 9. URL: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/experiment-design/illumina-adapter-sequences-1000000002694-09.pdf (2018).

106. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods* **12,** 115–121 (2015).

107. Pagès, H. *et al. Biostrings: Efficient manipulation of biological strings.* R package. URL: https://bioconductor.org/packages/release/bioc/html/biostrings.html (2019).

108. Pagès, H. *BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs.* R package. URL: https://bioconductor.org/packages/release/bioc/html/bsgenome.html (2018).

109. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology* **9** (2013).

110. Becker, G. & Lawrence, M. *genbankr: Parsing GenBank files into semantically useful objects.* R package. URL: https://bioconductor.org/packages/release/bioc/html/genbankr.html (2018).

111. Winter, D. J. rentrez: an R package for the NCBI eUtils API. *The R Journal* **9,** 520–526 (2017).

112. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48,** 443–453 (1970).

113. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *Journal of Molecular Biology* **147,** 195–197 (1981).

114. Wickham, H., Hester, J. & Francois, R. *readr: Read Rectangular Text Data.* R package. URL: https://cran.r-project.org/package=readr (2019).

115. Wickham, H. *et al. dplyr: A Grammar of Data Manipulation.* R package. URL: https://cran.r-project.org/package=dplyr (2019).

116. Wickham, H. & Henry, L. *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions.* R package. URL: https://cran.r-project.org/package=tidyr (2019).

117. Wickham, H. *stringr: Simple, Consistent Wrappers for Common String Operations.* R package. URL: https://cran.r-project.org/package=stringr (2018).

118. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28,** 2520–2522 (2012).

119. R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. URL: https://www.r-project.org/ (2018).

120. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35,** 526–528 (2019).

121. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).

122. Katz, Y. *et al.* Sashimi plots: Quantitative visualization of RNA sequencing read alignments. *arXiv,* 1306.3466 (2013).

123. Hahne, F. & Ivanek, R. *Visualizing Genomic Data Using Gviz and Bioconductor* in *Statistical Genomics: Methods and Protocols* (eds Mathé, E. & Davis, S.) 335–351 (Springer, 2016).

124. Yu, G. *et al.* Two methods for mapping and visualizing associated data on phylogeny using ggtree. *Molecular Biology and Evolution* **35,** 3041–3043 (2018).

125. Yu, G. *tidytree: A Tidy Tool for Phylogenetic Tree Data Manipulation.* R package. URL: https://cran.r-project.org/package=tidytree (2019).

126. Wagih, O. *ggseqlogo: A ggplot2 Extension for Drawing Publication-Ready Sequence Logos.* R package. URL: https://cran.r-project.org/package=ggseqlogo (2019).

127. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).

128. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012).

129. Marçais, G., Yorke, J. A. & Zimin, A. QuorUM: An Error Corrector for Illumina Reads. *PLOS One* **10,** e0130821 (2015).

130. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* **19,** 455–477 (2012).

131. Nikolenko, S. I., Korobeynikov, A. I. & Alekseyev, M. A. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics* **14,** S7 (2013).

132. Altschul, S. F. *et al.* Basic local alignment search tool. *Journal of Molecular Biology* **215,** 403–410 (1990).

133. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25,** 3389–3402 (1997).

134. Boetzer, M. *et al.* Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27,** 578–579 (2011).

135. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *PNAS* **74,** 5463–5467 (1977).

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

137. Gertz, E. M. *et al.* Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. *BMC Biology* **4,** 41 (2006).

138. Löytynoja, A. *Phylogeny-aware alignment with PRANK* in *Multiple Sequence Alignment Methods* (ed Russell, D. J.) 155–170 (Humana Press, 2014).

139. Wheeler, T. J. & Eddy, S. R. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29,** 2487–2489 (2013).

140. Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Computational Biology* **7,** e1002195 (2011).

141. Eddy, S. R. *A new generation of homology search tools based on probabilistic inference* in *Genome Informatics 2009* (eds Morishita, S., Lee, S. Y. & Sakakibara, Y.) 205–211 (Imperial College Press, 2009).

142. Eddy, S. R. A Probabilistic Model of Local Sequence Alignment That Simplifies Statistical Significance Estimation. *PLOS Computational Biology* **4,** e1000069 (2008).

143. Ehrenmann, F. & Lefranc, M.-P. IMGT/DomainGapAlign: IMGT Standardized Analysis of Amino Acid Sequences of Variable, Constant, and Groove Domains (IG, TR, MH, IgSF, MhSF). *Cold Spring Harbor Protocols* **2011,** 737–749 (2011).

144. Ye, J. *et al.* IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research* **41,** W34–W40 (2013).

145. Lefranc, M.-P. Immunoglobulins: 25 Years of Immunoinformatics and IMGT-ONTOLOGY. *Biomolecules* **4,** 1102–1139 (2014).

146. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **16,** 276–277 (2000).

147. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

148. Smith, A., Hubley, R. & Green, P. *RepeatMasker Open-4.0.* URL: https://www.repeatmasker.org (2018).

149. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

150. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology* **29,** 24–26 (2011).

151. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14,** 178–192 (2013).

152. Shapiro, M. B. & Senapathy, P. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Research* **15,** 7155–7174 (1987).

153. Ulitsky, I. *et al.* Extensive alternative polyadenylation during zebrafish development. *Genome Research* **22,** 2054–2066 (2012).

154. Wright, E. S. Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. *The R Journal* **8,** 352–359 (2016).

155. Cui, R. *et al.* Relaxed selection limits lifespan by increasing mutation load. *Cell* **178,** 1–15 (2019).

156. Stamatakis, A., Ludwig, T. & Meier, H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21,** 456–463 (2005).

157. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).

158. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).

159. Rambaut, A. *FigTree* version 1.4. URL: https://tree.bio.ed.ac.uk/software/figtree/ (2018).

160. Vander Heiden, J. A. *et al.* pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* **30,** 1930–1932 (2014).

161. Gupta, N. T. *et al.* Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* **31,** 3356–3358 (2015).

162. Ewing, B. & Green, P. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Research* **8,** 186–194 (1998).

163. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).

164. Fu, L. *et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28,** 3150–3152 (2012).

165. Vander Heiden, J. A. Personal communication. 2018.

166. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461 (2010).

167. Rognes, T. *et al.* VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4,** e2584 (2016).

168. Li, H. *seqtk: A fast and lightweight tool for processing FASTA or FASTQ sequences.* URL: https://github.com/lh3/seqtk (2016).

169. Nouri, N. & Kleinstein, S. H. Optimized Threshold Inference for Partitioning of Clones From High-Throughput B Cell Repertoire Sequencing Data. *Frontiers in Immunology* **9** (2018).

170. Gupta, N. T. *et al.* Hierarchical Clustering Can Identify B Cell Clones with High Confidence in Ig Repertoire Sequencing Data. *The Journal of Immunology* **198,** 2489–2499 (2017).

171. Gotelli, Nicholas J. & Colwell, Robert K. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters* **4,** 379–391 (2001).

172. Stern, J. N. H. *et al.* B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Science Translational Medicine* **6,** 248ra107 (2014).

173. Bolen, C. R. *et al.* The Repertoire Dissimilarity Index as a method to compare lymphocyte receptor repertoires. *BMC Bioinformatics* **18,** 155 (2017).

174. Wysocki, L. J. & Gefter, M. L. Gene Conversion and the Generation of Antibody Diversity. *Annual Review of Biochemistry* **58,** 509–527 (1989).

175. Patel, B. *et al.* Diversity of Immunoglobulin (Ig) Isotypes and the Role of Activation-Induced Cytidine Deaminase (AID) in Fish. *Molecular Biotechnology* **60,** 435–453 (2018).

176. Harel, I. *et al.* A platform for rapid exploration of aging and diseases in a naturally short-lived vertebrate. *Cell* **160,** 1013–1026 (2015).

177. Cellerino, A., Valenzano, D. R. & Reichard, M. From the bush to the bench: the annual *Nothobranchius* fishes as a new model system in biology. *Biological Reviews* **91,** 511–533 (2016).

178. Schartl, M. *et al.* The genome of the platyfish, *Xiphophorus maculatus*, provides insights into evolutionary adaptation and several complex traits. *Nature Genetics* **45,** 567–572 (2013).

179. Yuan, Z. *et al.* Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments. *BMC Genomics* **19,** 141 (2018).

180. Magadan, S., Sunyer, O. J. & Boudinot, P. *Unique Features of Fish Immune Repertoires: Particularities of Adaptive Immunity Within the Largest Group of Vertebrates* in *Pathogen-Host Interactions: Antigenic Variation V. Somatic Adaptations* (eds Hsu, E. & Du Pasquier, L.) 235–264 (Springer, 2015).

181. Ravi, V. & Venkatesh, B. The divergent genomes of teleosts. *Annual Review of Animal Biosciences* **6,** 47–68 (2018).

182. Miho, E. *et al.* Computational Strategies for Dissecting the High-Dimensional Complexity of Adaptive Immune Repertoires. *Frontiers in Immunology* **9,** 224 (2018).

183. Vollmers, C. *et al.* Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *PNAS* **110,** 13463–13468 (2013).

184. Zajac, P. *et al.* Base Preferences in Non-Templated Nucleotide Incorporation by MMLV-Derived Reverse Transcriptases. *PLOS One* **8,** e85270 (2013).

185. Rosati, E. *et al.* Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnology* **17,** 61 (2017).

186. Illumina. *What is nucleotide diversity and why is it important?* URL: https://support.illumina.com/bulletins/2016/07/what-is-nucleotide-diversity-and-why-is-it-important.html (2018).

187. Shlemov, A. *et al.* Reconstructing antibody repertoires from error-prone immunosequencing reads. *The Journal of Immunology* **199,** 3369–3380 (2017).

188. Cleveland, W., Grosse, E. & Shyu, W. *Local regression models* in *Statistical models in S* 309–376 (Routledge, 1991).

189. Desponds, J., Mora, T. & Walczak, A. M. Fluctuating fitness shapes the clone-size distribution of immune repertoires. *PNAS* **113,** 274–279 (2016).

190. Mora, T. *et al.* Maximum entropy models for antibody diversity. *PNAS* **107,** 5405–5410 (2010).

191. Rosenfeld, A. M. *et al.* Computational Evaluation of B-Cell Clone Sizes in Bulk Populations. *Frontiers in Immunology* **9,** 1472 (2018).

192. Hill, M. O. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology* **54,** 427–432 (1973).

193. Lou Jost. Entropy and diversity. *Oikos* **113,** 363–375 (2006).

194. Gadala-Maria, D. *et al.* Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *PNAS* **112,** E862–E870 (2015).

195. Corcoran, M. M. *et al.* Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nature Communications* **7,** 13642 (2016).

196. Marcou, Q. Personal communication. 2019.

197. Spencer, J. & Sollid, L. The human intestinal B-cell response. *Mucosal immunology* **9,** 1113–1124 (2016).

198. Caruso, C. *et al.* Mechanisms of immunosenescence. *Immunity & Ageing* **6,** 10 (2009).

199. López-Otín, C. *et al.* The Hallmarks of Aging. *Cell* **153,** 1194–1217 (2013).

200. Friedensohn, S., Khan, T. A. & Reddy, S. T. Advanced Methodologies in High-Throughput Sequencing of Immune Repertoires. *Trends in Biotechnology* **35,** 203–214 (2017).

201. Robert K. Peet. The measurement of species diversity. *Annual Review of Ecology and Systematics* **5,** 285–307 (1974).

202. Berger, W. H. & Parker, F. L. Diversity of Planktonic Foraminifera in Deep-Sea Sediments. *Science* **168,** 1345–1347 (1970).

203. Simpson, E. H. Measurement of Diversity. *Nature* **163,** 688 (1949).

204. Caruso, T. *et al. The Berger–Parker index as an effective tool for monitoring the biodiversity of disturbed soils: a case study on Mediterranean oribatid (Acari: Oribatida) assemblages* in *Biodiversity and Conservation in Europe* (eds Hawksworth, D. L. & Bull, A. T.) 35–43 (Springer, 2008).

205. Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal* **27,** 379–423 (1948).

206. Mora, T. & Walczak, A. M. Rényi entropy, abundance distribution, and the equivalence of ensembles. *Physical Review E* **93,** 052418 (2016).

207. Lou Jost. Partitioning diversity into independent alpha and beta components. *Ecology* **88,** 2427–2439 (2007).

208. Bradshaw, W. J. & Valenzano, D. R. Extreme genomic volatility characterises the evolution of the immunoglobulin heavy chain locus in teleost fishes. *bioRxiv,* 752063 (2019).

Appendix A

Solutions and buffers

A.1 Enzymes

Enzyme	Concentration	Manufacturer	Product code
KAPA HiFi HotStart ReadyMix PCR Kit	2 ×[a]	Kapa Biosystems	KR0370
SMARTScribe Reverse Transcriptase	100 U µl^{-1}	Clontech Laboratories	639537
RNasin RNase inhibitor	40 U µl^{-1}	Promega	N2511
Uracil DNA glycosylase (UDG)	5 U µl^{-1}	NEB	M0280S
RNase A	100 mg ml^{-1}	QIAGEN	19101

[a] KAPA HiFi HotStart DNA Polymerase present at 0.04 U µl^{-1}.

A.2 Non-enzyme reagents and components

Reagent	Concentration	Manufacturer	Product code
SMARTScribe first-strand buffer	5 ×	Clontech Laboratories	639537[a]
Dithiothreitol (DTT)	20 mM	Clontech Laboratories	639537[a]
dNTP mix	10 µM each[b]	NEB	N0447L
1 µm Sera-Mag Magnetic SpeedBeads	50 mg ml^{-1}	GE Healthcare	65152105050250
QIAzol Lysis Reagent	1 ×	QIAGEN	79306
BluePippin electrophoresis buffer	1 ×	Sage Science	BDF1510[c]
BluePippin R2 loading solution / marker mix	1 ×	Sage Science	BDF1510[c]
Roti-Phenol/chloroform/isoamyl alcohol	1 ×	Roth	A156.2

[a] Supplied with SMARTScribe Reverse Transcriptase (Appendix A.1).

[b] i.e. 10 µM each of dATP, dGTP, dCTP and dTTP.

[c] Supplied with BluePippin 1.5 % agarose dye-free cassettes.

A.3 Prepared buffers

Name	Purpose	Composition	pH	Storage temperature
TET	Washing Sera-Mag beads	• 10 mM Tris base • 1 mM Na_2-EDTA • 0.05 % (v/v) Tween 20	8.0[a]	Room temperature
iSB	Preparing SeraSure bead suspension	• 4.2 M NaCl • 16.8 mM Tris base • 1.68 mM Na_2-EDTA	8.0[a]	Room temperature
EB	Buffering nucleic-acid solutions	• 10 mM Tris-HCl	8.5[a]	Room temperature
P1	Resuspending cultured *E. coli* cells	• 50 mM Tris-HCl • 10 mM Na_2-EDTA • 100 µg ml^{-1} RNase-A[b]	8[a]	4 °C[c]
P2	Cell lysis	• 200 mM Sodium hydroxide • 1 % (v/v) Sodium dodecyl sulfate	–	Room temperature
P3	Precipitation of cell lysate	• 3 M Potassium acetate	5.5[d]	Room temperature

[a] Adjust to the required pH with hydrochloric acid (HCl).

[b] Appendix A.1.

[c] Can be stored at room temperature before addition of RNase-A.

[d] Adjust to the required pH with glacial acetic acid.

Appendix B

Primers and oligonucleotides

B.1 Template-switch-adapter oligos for reverse transcription

Name	Sequence	Source
SmartNNNa	AAGCAGUGGTAUCAACGCAGAGUNNNNUNNNNUNNNNUCTTrGrGrGrG	[103]

B.2 PCR and reverse-transcription primers

Name	Sequence	Purpose	Source[a]
RT1	TGGTCTTGCCAGCTGGTGATTTCCGCC	IgSeq $C_\mu 2$ reverse-transcription primer	–
M1SS	AAGCAGTGGTATCAACGCA	IgSeq PCR1 forward primer	[103]
IGH-B	CCACATGGCACCAGAGGAAAC	IgSeq PCR1 reverse primer	–
M1S+P2	GTGACTGGAGTTCAGACGTGTGCTCTTC–CGATCTCAGTGGTATCAACGCAGAG	IgSeq PCR2 forward primer	[103]
IGH-C+P1	ACACTCTTTCCCTACACGACGCTCTTC–CGATCTATGGCACCAGAGGAAACACAAC	IgSeq PCR2 reverse primer	–

[a] Items without a specified source were designed by the author using Primer3 [104].

B.3 Illumina TruSeq adaptor sequences

P1/i5 adaptor sequences

Base sequence: AATGATACGGCGACCACCGAGATCTACAC**NNNNNNNN**–
ACACTCTTTCCCTACACGACGC

Index sequences:

Name	Index sequence[a]
D501	AGGCTATA
D502	GCCTCTAT
D503	AGGATAGG
D504	TCAGAGCC
D505	CTTCGCCT
D506	TAAGATTA
D507	ACGTCCTG
D508	GTCAGTAC

[a] [105]

P2/i7 adaptor sequences

Base sequence: ACAAGCAGAAGACGGCATACGAGAT**NNNNNNNN**–
GTGACTGGAGTTCAGACGTGTGC

Index sequences:

Name	Index sequence[a]
D701	CGAGTAAT
D702	TCTCCGGA
D703	AATGAGCG
D704	GGAATCTC
D705	TTCTGAAT
D706	ACGAATTC
D707	AGCTTCAG
D708	GCGCATTA
D709	CATAGCCG
D710	TTCGCGGA
D711	GCGCGAGA
D712	CTATCGCT

[a] [105]

Appendix C

Hill numbers and antibody repertoire diversity

The question of how to measure the diversity of clones or sequences within an antibody repertoire, as well as the degree of divergence in composition between pairs or groups of repertoires, closely parallels related questions in ecology, information theory, and other fields. Over time, a great many different conceptions of diversity have been developed in these fields [201], each with its own cohort of measurement and approximation methods. Many of these diversity indices, however, can be unified into a common framework of so-called "true diversities" based on Hill numbers [192, 193], providing a more intuitive and comprehensive insight into the diversity structure of a population. In this appendix, I describe the concepts and motivations underlying this conception of population of diversity, both for individual (unitary) populations (Appendix C.1) and for more complex populations with internal subpopulation structure (Appendix C.2).

C.1 Diversity in unitary populations

C.1.1 Terminology

Though in this appendix I use terminology derived from the ecological literature, the concepts and measures discussed here can be applied to any situation in which a set (or *population*) of elements (or *individuals*) is partitioned among some number of mutually-exclusive categories (or *species*). Considered abstractly, these terms could be used to refer to coloured balls in an urn, species in a rainforest, or sequences in a repertoire.

Let X be a unitary population of individuals, each of which is assigned to some species s from a set of possible species S. The term "unitary" here denotes that X is considered to have no internal

structure except the species identity of its constitutent individuals. Let n_s denote the number of individuals in X belonging to s, and N denote the total number of individuals in X. Then

$$p_s = \frac{n_s}{N} \tag{C.1}$$

denotes the proportion (or *relative frequency*) of individuals in X belonging to s, or equivalently the probability that a randomly-selected individual from X belongs to s. The *species richness* of P is the total number of species $|S|$, while the *evenness* of P is the degree to which different species are similar in their p_s values: a population containing one very abundant species and $n-1$ very rare species, for example, is much less even than a population containing n equally-abundant species.

C.1.2 Simple diversity indices

The diversity of a unitary population X of some total size N is generally considered to increase with both the richness and evenness of X. Different measures of within-population diversity, however, place different amounts of weight on the richness of X compared to its evenness when evaluating its diversity. At one extreme, the species richness itself is used as a measure of diversity, albeit one that entirely ignores evenness; other commonly-used diversity measures, such as Simpson's index, Shannon entropy, and the Berger-Parker index are affected to varying degrees by both richness and evenness [201, 202]. In addition to their relative emphasis on richness vs evenness, different indices will also place different amounts of weight on common vs rare species when computing diversity: at the extremes, species richness gives the same amount of weight to all species regardless of frequency, while the Berger-Parker index gives zero weight to all species except the most common. As a result, any given diversity index captures a different aspect of the diversity structure of any population of interest.

C.1.2.1 Simpson's index

One of the oldest diversity indices incorporating both species richness and evenness is Simpson's index, which measures the probability that two randomly selected individuals from a population are from the same species [203]. For a finite population (or sample) X with a set of possible species S, Simpson's index is calculated as follows:

$$L(X) = \frac{\sum_{s \in S} n_s(n_s - 1)}{N(N-1)} \tag{C.2}$$

As the population size tends to infinity, this simplifies to

$$L^*(X) = \frac{\sum_{s \in S} n_s^2}{N^2} = \sum_{s \in S} p_s^2 \tag{C.3}$$

As originally formulated, Simpson's index can be considered a *dominance index* (also known as a concentration index [203]), measuring the degree to which a population is dominated by a small number of species; as such, a higher Simpson's index indicates a less-diverse population. Conversely, the complement $1 - L(X)$ of Simpson's index, known as the Gini-Simpson index [193], represents the "probability of interspecific encounter" [201], and is also widely used as a diversity index in the literature.

C.1.2.2 The Berger-Parker index

The Berger-Parker index is a very simple measure of diversity, given by the relative frequency p_x of the most abundant species x in the population [202, 204]. Like Simpson's index, the Berger-Parker index is a dominance index, measuring the degree to which the population is dominated by the single largest species. Despite its simplicity, the Berger-Parker is a true diversity index, responding to both the richness and the evenness of a population, and can be used to distinguish different diversity structures in real populations [204]. As it focuses only on the most common species in each population, it also has the particular advantage of being much less vulnerable to sampling bias than other diversity metrics, especially compared to the species richness itself [202].

C.1.2.3 Entropic diversity indices

In information theory, the Shannon entropy $H(Y)$ of a random variable Y provides a measure of the unpredictability of that variable, and therefore of the degree to which its value can be predicted in advance [205]. A variable which can take only a single value has zero entropy (it is perfectly predictable), while one which can take all values in its state space with equal probability has maximum entropy (it is maximally unpredictable) for that state space. All else being equal, a variable which can take a larger number of possible values has greater entropy than one which can take fewer values; hence, the entropy of a variable increases with both the richness and evenness of its state space [205].

Although the concept of Shannon entropy was developed in the context of electronic communication, it can be extended quite naturally to ecology, where the process of sampling individuals from a population represents an information source and the species identity of each sampled individual represents an output. In this case, the Shannon entropy represents "the amount of uncertainty that exists regarding the species of an individual selected at random from the population" [201]. The Shannon entropy of a population X is given by

$$H(X) = -\sum_{s \in S} p_s \cdot \log p_s \tag{C.4}$$

Any base of logarithm can be used, though bases 2 and e are most common; in these bases, the units of Shannon entropy are the bit ("binary digit") and nat ("natural unit"), respectively.

A generalisation of the Shannon entropy, also used as a diversity measure, is the Rényi entropy [206]:

$$H_q(X) = \frac{1}{1-q} \log \left[\sum_{s \in S} p_i^q \right] \tag{C.5}$$

where q is denoted the *order* of the entropy measure. The Rényi entropy is undefined at $q = 1$, but reduces to the Shannon entropy (with the same base of logarithm) in the limit as $q \to 1$ [206]. As p_i is always between 0 and 1, raising the relative frequencies to a power greater than 1 downweights rarer species (with smaller p_i) relative to more common ones; as a result, higher-order Rényi entropies put more emphasis on the most common species compared to Shannon entropy when computing population diversity, while Rényi entropies with order less than 1 put more weight on rarer species; at $q = 0$, H_0 is simply the logarithm of the species richness $|S|$.

C.1.3 Effective species richness and true diversity

Appendix C.1.2 discusses several commonly-used diversity indices, including Simpson's index, the Gini-Simpson index, Shannon entropy (& other Rényi entropies), and the Berger-Parker index; many other diversity indices are possible. These indices differ importantly in their forms, numeric ranges and biological interpretation, as well as their response behaviours to different changes in population composition [193, 201]. For example, Simpson's index and the Berger-Parker index range from 0 to 1, with lower values indicating greater population diversity, while Shannon- and Rényi-entropy measures range from 0 to infinity, with higher values indicating greater diversity (in log-scale). The use of different diversity indices can therefore yield importantly different results when used to compare different populations: comparing two populations using only one such measure will capture only a small part of the diversity structure of those populations, while comparing them using two or more raw indices will often yield results that are difficult to accurately interpret [193, 201].

Fortunately, different diversity indices can be transformed into a common framework by considering, for each index, the number D of *equally-common* species that would be needed to produce the same diversity value using that index. This transformation gives an estimation, for each index, of the *effective species richness* of the population: if we corrected for differences in species abundance while holding the diversity index constant, how rich would the resulting population be? This effective

richness can itself be considered a diversity measure; one that is comparable across diversity indices in a way the raw indices are not [193].

For many important diversity indices, the equation for the effective richness D takes a common form, known as the Hill number or "true diversity" for some parameter q [192, 193]:

$$ {}^{q}D(X) = \left(\sum_{s \in S} p_s^q \right)^{\frac{1}{1-q}} \tag{C.6} $$

Indices whose effective richnesses take this form (Table C.1) include species richness ($q = 0$), Rényi entropy in base e ($q = q$), Shannon entropy in base e ($q \to 1$), Simpson's index ($q = 2$) and the Berger-Parker index ($q \to \infty$) among others [182, 192, 193, 201]. As with the Rényi entropy, the parameter q, known here as the *diversity order*, describes the degree to which rare species are downweighted compared to common ones when calculating the Hill number: at one extreme ($q = 0$, the species richness), all species are given equal weight regardless of their frequency, while at the other ($q \to \infty$, the reciprocal Berger-Parker index) only the most common species in the sample is considered. As a result of this downweighting, higher-order indices are also less sensitive to undersampling than lower-order ones; this is particularly important to keep in mind in cases, like immune repertoires, where the number of rare species is extremely large and undersampling of rare species is pervasive [30].

The use of Hill numbers in diversity estimation enables many different, widely-used diversity measurements to be considered and compared in a common framework, giving a description of population diversity that is easy to interpret biologically and compare across different populations [193]. Typically, the limit at $q \to 1$ is substituted for the (undefined) value at $q = 1$, giving a function that is both continuous and monotonically decreasing:

$$ {}^{q}D(X) = \begin{cases} \left(\sum_{s \in S} p_s^q \right)^{\frac{1}{1-q}} & q \neq 1 \\ \exp\left(-\sum_{s \in S} p_s \cdot \ln p_s \right) & q = 1 \end{cases} \tag{C.7} $$

This equation can then be easily used to construct diversity profiles (or *spectra*) spanning many different orders of diversity [182]; since each value of q captures a different aspect of a population's diversity structure, these profiles can be much more informative than any single metric when analysing and comparing the diversity structure of populations.

Table C.1: Summary of effective-richness measures for some common diversity indices (adapted and expanded from [193]).

Diversity index, $f(X)$	Effective species richness, ${}^qD(X)$	Diversity order, q
Species richness	${}^0D(X) = f(X)$	0
Shannon entropy (base e)	$\lim_{q\to 1}\ {}^qD(X) = \exp f(X)$	1
Simpson index	${}^2D(X) = \frac{1}{f(X)}$	2
Gini-Simpson index	${}^2D(X) = \frac{1}{1-f(X)}$	2
Renyi entropy (base e, order q)	${}^qD(X) = \exp f(X)$	q
Berger-Parker index	$\lim_{q\to\infty}\ {}^qD(X) = \frac{1}{f(X)}$	∞

C.2 Diversity in structured populations

Section C.1 discusses methods for analysing the diversity of a single, unitary population. In many cases, however, we are interested in groups of related populations, and the relative amount of variability within and between the populations in each group. In order to extend the mathematical framework of diversity measurement, and especially that of true diversities and Hill spectra, to this more complex case, some additional clarification of terminology is needed.

C.2.1 Terminology

In a *structured* population C, the individuals in C can be partitioned into some number M of disjoint unitary subpopulations $X_1, X_2, \ldots$, such that:

- Each subpopulation X has size N_X, with the total size of the whole population given by $N = \sum_{X\in C} N_X$.
- Each individual in a subpopulation X is assigned to a species s drawn from a set of possible species S_X, with the total species set for the population given by $S = \bigcup_{X\in C} S_X$.
- Each subpopulation can be assigned a relative statistical weight w_X; this could be equal for all populations ($w_X = 1/M$), proportional to each population's relative size $w_X = \frac{N_X}{N}$, or proportional to some other measure of each population's "importance" to the system.
- The relative frequency of a species s in a subpopulation X is given by $p_{X,s} = \frac{n_{X,s}}{N_X}$, where $n_{X,s}$ is the number of individuals in X belonging to s.

What is the diversity of C? There are several possible answers, depending on which features of the makeup of C are most salient:

1. The **gamma diversity** of C is the *total* diversity of the population when its subpopulations are pooled according to their weights; it represents the species diversity across the whole population, ignoring the subpopulation membership of individuals.

2. The **alpha diversity** of C is the diversity arising from differences in species identity among individuals *within* each subpopulation, and is given by a weighted *average* of the unitary diversities of those subpopulations; the appropriate weighting function depends on the order of diversity under consideration. In some sense, the alpha diversity of C can be thought of as the expected diversity of a single population drawn from C.

3. The **beta diversity** of C is the diversity arising from variability in species composition *among* the subpopulations in C; it is lowest when all subpopulations have identical species compositions and highest when they have no species in common.

The alpha and beta diversities of a population are independent; two different structured populations can have identical alpha and very different beta diversities, or vice versa, depending on the exact species compositions of their subpopulations. The alpha and beta diversity of a structured population also completely determine its gamma diversity [207]:

$$D_\gamma(C) = D_\alpha(C) \times D_\beta(C) \tag{C.8}$$

and therefore

$$D_\beta(C) = \frac{D_\gamma(C)}{D_\alpha(C)} \tag{C.9}$$

This (Equation C.9) is typically the easiest way of computing the beta diversity of a given collection of populations.

In the rest of this section, I will primarily consider only subpopulations with equal sizes $N_X = \frac{N}{M}$ and equal weights $w = \frac{1}{M}$, as this is how they are used in the repertoire-diversity methods discussed in Section 2.3.6.4 and Chapter 4. Equal-sized and -weighted populations can be produced by downsampling each subpopulation to the same number of individuals (in the IgSeq case, by downsampling each repertoire to the same number of unique sequences) prior to calculating the diversity of the population.

C.2.2 Calculating alpha, beta, and gamma diversity

The alpha, beta and gamma diversities of structured populations can be calculated analogously to the unitary true diversities discussed in Appendix C.1.3. In this framework, the alpha diversity represents the effective number of equally-abundant species present in the "average" subpopulation drawn from

C, while the gamma diversity represents the effective number of species present in the population as a whole (ignoring subpopulation membership). The beta diversity also represents an effective number of groupings, but in this case the unit is subpopulations rather than species: given an alpha diversity value for C, the beta diversity gives the number of equally-weighted subpopulations, with no species in common, that would give rise to the gamma diversity of C.

Under this framework, the diversities of order q for a structured population C are given [207] by

$$ {}^{q}D_{\gamma}(C) = \left[\sum_{s \in S} \left(\frac{\sum_{X \in C} w_X p_{X,s}}{\sum_{X \in C} w_X} \right)^q \right]^{\frac{1}{1-q}} = \left[\frac{\sum_{s \in S} \left(\sum_{X \in C} w_X p_{X,s} \right)^q}{\left(\sum_{X \in C} w_X \right)^q} \right]^{\frac{1}{1-q}} \tag{C.10} $$

$$ {}^{q}D_{\alpha}(C) = \left[\frac{\sum_{X \in C} w_X^q \left(\sum_{s \in S_X} p_{X,s}^q \right)}{\sum_{X \in C} w_X^q} \right]^{\frac{1}{1-q}} = \left[\frac{\sum_{X \in C} \sum_{s \in S_X} (w_X p_{X,s})^q}{\sum_{X \in C} w_X^q} \right]^{\frac{1}{1-q}} \tag{C.11} $$

$$ {}^{q}D_{\beta}(C) = \frac{{}^{q}D_{\gamma}(C)}{{}^{q}D_{\alpha}(C)} = \left[\frac{\sum_{s \in S} \left(\sum_{X \in C} w_X p_{X,s} \right)^q}{\sum_{X \in C} \sum_{s \in S_X} (w_X p_{X,s})^q} \times \frac{\sum_{X \in C} w_X^q}{\left(\sum_{X \in C} w_X \right)^q} \right]^{\frac{1}{1-q}} \tag{C.12} $$

When all the subpopulations are equally-weighted (i.e. $w_X = w = \frac{1}{M}$ for all subpopulations), these formulae simplify considerably:

$$ {}^{q}D_{\gamma}(C) = \left[\frac{\sum_{s \in S} \left(\sum_{X \in C} w p_{X,s} \right)^q}{\left(\sum_{X \in C} w \right)^q} \right]^{\frac{1}{1-q}} = \left[\frac{w^q \sum_{s \in S} \left(\sum_{X \in C} p_{X,s} \right)^q}{M^q w^q} \right]^{\frac{1}{1-q}} = \left[\frac{\sum_{s \in S} \left(\sum_{X \in C} p_{X,s} \right)^q}{M^q} \right]^{\frac{1}{1-q}} \tag{C.13} $$

$$ {}^{q}D_{\alpha}(C) = \left[\frac{\sum_{X \in C} \sum_{s \in S_X} (w p_{X,s})^q}{\sum_{X \in C} w^q} \right]^{\frac{1}{1-q}} = \left[\frac{w^q \sum_{X \in C} \sum_{s \in S_X} (p_{X,s})^q}{M w^q} \right]^{\frac{1}{1-q}} = \left[\frac{\sum_{X \in C} \sum_{s \in S_X} (p_{X,s})^q}{M} \right]^{\frac{1}{1-q}} \tag{C.14} $$

$$ {}^{q}D_{\beta}(C) = \frac{{}^{q}D_{\gamma}(C)}{{}^{q}D_{\alpha}(C)} = \left[\frac{\sum_{s\in S}\left(\sum_{X\in C} p_{X,s}\right)^{q}}{\sum_{X\in C}\sum_{s\in S_X}(p_{X,s})^{q}} \times \frac{M}{M^{q}} \right]^{\frac{1}{1-q}} = \left[\frac{\sum_{s\in S}\left(\sum_{X\in C} p_{X,s}\right)^{q}}{M^{q-1}\sum_{X\in C}\sum_{s\in S_X}(p_{X,s})^{q}} \right]^{\frac{1}{1-q}} \tag{C.15} $$

These equations are valid for all values of $q \in \mathbb{R}$ except 1, providing a spectrum of alpha-, beta- or gamma-diversity measures analogous to the diversity spectra provided for unitary populations in Appendix C.1.3. As in the unitary case, a special case needs to be made for $q = 1$ in order to make these functions continuous[1] [207]:

$$ \begin{aligned} {}^{1}D_{\gamma}(C) &= \lim_{q\to 1} {}^{q}D_{\gamma}(C) = \exp\left[-\sum_{s\in S}\left(\left[\sum_{X\in C} w p_{X,s}\right] \cdot \ln\left[\sum_{X\in C} w p_{X,s}\right] \right) \right] \\ &= \exp\left(-\sum_{s\in S} p_s \cdot \ln p_s \right) = {}^{1}D(C) \end{aligned} \tag{C.16} $$

$$ \begin{aligned} {}^{1}D_{\alpha}(C) &= \lim_{q\to 1} {}^{q}D_{\alpha}(C) = \exp\left[-\sum_{X\in C} w \sum_{s\in S_X} (p_{X,s} \cdot \ln p_{X,s}) \right] \\ &= \exp\left[\frac{1}{M}\sum_{X\in C}\left(-\sum_{s\in S_X} (p_{X,s} \cdot \ln p_{X,s}) \right) \right] = \exp\left[\frac{1}{M}\sum_{X\in C} \ln {}^{1}D(X) \right] \end{aligned} \tag{C.17} $$

$$ {}^{1}D_{\beta}(C) = \frac{{}^{1}D_{\gamma}(C)}{{}^{1}D_{\alpha}(C)} = \frac{\exp\left[-\sum_{s\in S} p_s \cdot \ln p_s\right]}{\exp\left[\frac{1}{M}\sum_{X\in C}\left(-\sum_{s\in S_X}(p_{X,s} \cdot \ln p_{X,s})\right)\right]} = \frac{{}^{1}D(C)}{\exp\left[\frac{1}{M}\sum_{X\in C} \ln {}^{1}D(X)\right]} \tag{C.18} $$

C.2.3 Rescaling beta diversity

As discussed in Appendix C.2.2, while alpha and gamma diversity are expressed in terms of an effective number of species (in an average subpopulation and the entire structured population, respectively), beta diversity is expressed in terms of an effective number of subpopulations. Since the effective number of subpopulations is determined in part by the actual number of subpopulations, this means that the beta diversity, unlike alpha and gamma diversity, is directly dependent on the number of subpopulations M. If two different structured populations contain different numbers of subpopulations,

[1] Note that, when $w = \frac{1}{M}$ and $N_X = \frac{N}{M}$ for all populations, $\sum_{X\in C} w p_{X,s} = \sum_{X\in C} \frac{1}{M}\frac{n_{X,s}}{N_X} = \frac{1}{M}\frac{M\sum_{X\in C} n_{X,s}}{N} = \frac{n_s}{N} = p_s$

it is therefore not possible to compare their beta diversity values directly; rather, the beta diversity spectra of the populations must be *rescaled* to a common range before such a comparison is performed.

The *minimum* beta diversity of a structured population obtains when all subpopulations have identical species composition (i.e. $p_{X,s} = p_s$ for all species s and subpopulations X). In this case, the beta diversity for the structured population is given by:

$$ {}^{q}D_{\beta\,\min}(C) = \left[\frac{\sum_{s\in S}\left(\sum_{X\in C} p_{X,s}\right)^{q}}{M^{q-1}\sum_{X\in C}\sum_{s\in S_X}(p_{X,s})^{q}} \right]^{\frac{1}{1-q}} = \left[\frac{\sum_{s\in S}(Mp_s)^{q}}{M^{q-1}\sum_{s\in S_X} Mp_s^{q}} \right]^{\frac{1}{1-q}} = \left[\frac{M^{q}\sum_{s\in S} p_s^{q}}{M^{q}\sum_{s\in S_X} p_s^{q}} \right]^{\frac{1}{1-q}} = 1 \quad \text{(C.19)} $$

The *maximum* beta diversity, meanwhile, obtains when there is no overlap in species between populations. In this case, for any given species, $n_{X,s}$ is equal to n_s for one subpopulation and 0 for all others, and therefore $(\sum_{X\in C} p_{X,s})^{q} = \sum_{X\in C}(p_{X,s})^{q} = p_s^{q}$ and $\sum_{X\in C}\sum_{s\in S_X}(p_{X,s})^{q} = \sum_{s\in S} p_s^{q}$. The beta diversity is therefore given by:

$$ {}^{q}D_{\beta\,\max}(C) = \left[\frac{\sum_{s\in S}\left(\sum_{X\in C} p_{X,s}\right)^{q}}{M^{q-1}\sum_{X\in C}\sum_{s\in S_X}(p_{X,s})^{q}} \right]^{\frac{1}{1-q}} = \left[\frac{\sum_{s\in S} p_s^{q}}{M^{q-1}\sum_{s\in S} p_s^{q}} \right]^{\frac{1}{1-q}} = \left[\frac{1}{M^{q-1}} \right]^{\frac{1}{1-q}} = M \qquad \text{(C.20)} $$

The same identities hold for the special case when $q = 1$:

$$ \begin{aligned} {}^{1}D_{\beta\,\min}(C) &= \frac{\exp\left[-\sum_{s\in S} p_s \cdot \ln p_s\right]}{\exp\left[\frac{1}{M}\sum_{X\in C}\left(-\sum_{s\in S_X}(p_{X,s}\cdot \ln p_{X,s})\right)\right]} = \frac{\exp\left[-\sum_{s\in S} p_s \cdot \ln p_s\right]}{\exp\left[\frac{1}{M}\sum_{X\in C}\left(-\sum_{s\in S}(p_s \cdot \ln p_s)\right)\right]} \\ &= \frac{\exp\left[-\sum_{s\in S} p_s \cdot \ln p_s\right]}{\exp\left[\frac{M}{M}\left(-\sum_{s\in S} p_s \cdot \ln p_s\right)\right]} = 1 \end{aligned} \qquad \text{(C.21)} $$

$$
\begin{aligned}
{}^{1}D_{\beta\max}(C) &= \frac{\exp\left[-\sum_{s\in S} p_s \cdot \ln p_s\right]}{\exp\left[\frac{1}{M}\sum_{X\in C}\left(-\sum_{s\in S_X}(p_{X,s}\cdot \ln p_{X,s})\right)\right]} \\
&= \exp\left[-\sum_{s\in S} p_s \cdot \ln p_s - \frac{1}{M}\sum_{X\in C}\left(-\sum_{s\in S_X}(p_{X,s}\cdot \ln p_{X,s})\right)\right] \\
&= \exp\left[H(C) + \frac{1}{M}\sum_{X\in C}\left(\sum_{s\in S_X}\frac{n_{X,s}}{N_X}\cdot \ln\frac{n_{X,s}}{N_X}\right)\right] \\
&= \exp\left[H(C) + \frac{1}{M}\sum_{X\in C}\left(\sum_{s\in S_X}\frac{Mn_{X,s}}{N}\cdot \ln\frac{Mn_{X,s}}{N}\right)\right] \\
&= \exp\left[H(C) + \sum_{X\in C}\left(\sum_{s\in S_X}\frac{n_{X,s}}{N}\cdot \ln\frac{Mn_{X,s}}{N}\right)\right] \\
&= \exp\left[H(C) + \left(\sum_{s\in S}\frac{n_s}{N}\cdot \ln\frac{Mn_s}{N}\right)\right] = \exp\left[H(C) + \left(\sum_{s\in S} p_s \cdot \ln(Mp_s)\right)\right] \\
&= \exp\left[H(C) + \left(\sum_{s\in S} p_s \cdot [\ln M + \ln p_s]\right)\right] \\
&= \exp\left[H(C) + \left(\sum_{s\in S} p_s \cdot \ln p_s\right) + \ln M \cdot \sum_{s\in S} p_s\right] = \exp\left[H(C) - H(C) + \ln M\right] \\
&= \exp(\ln M) = M
\end{aligned}
\tag{C.22}
$$

The beta diversity for a structured population with M subpopulations therefore ranges between 1 (identical composition) and M (maximally-divergent composition). The beta diversities of such a population can thus be transformed onto a new scale from 0 to 1 as follows:

$$
{}^{q}D_{\beta\,\mathrm{rescaled}}(C) = \frac{{}^{q}D_{\beta}(C) - {}^{q}D_{\beta\,\min}(C)}{{}^{q}D_{\beta\,\max}(C) - {}^{q}D_{\beta\,\min}(C)} = \frac{{}^{q}D_{\beta}(C) - 1}{M - 1} \tag{C.23}
$$

By transforming the beta-diversity spectra of different structured populations onto this common scale, the inter-subpopulation variability of those populations can be meaningfully compared, even if they differ in the number of subpopulations they contain.

Appendix D

Supplementary figures

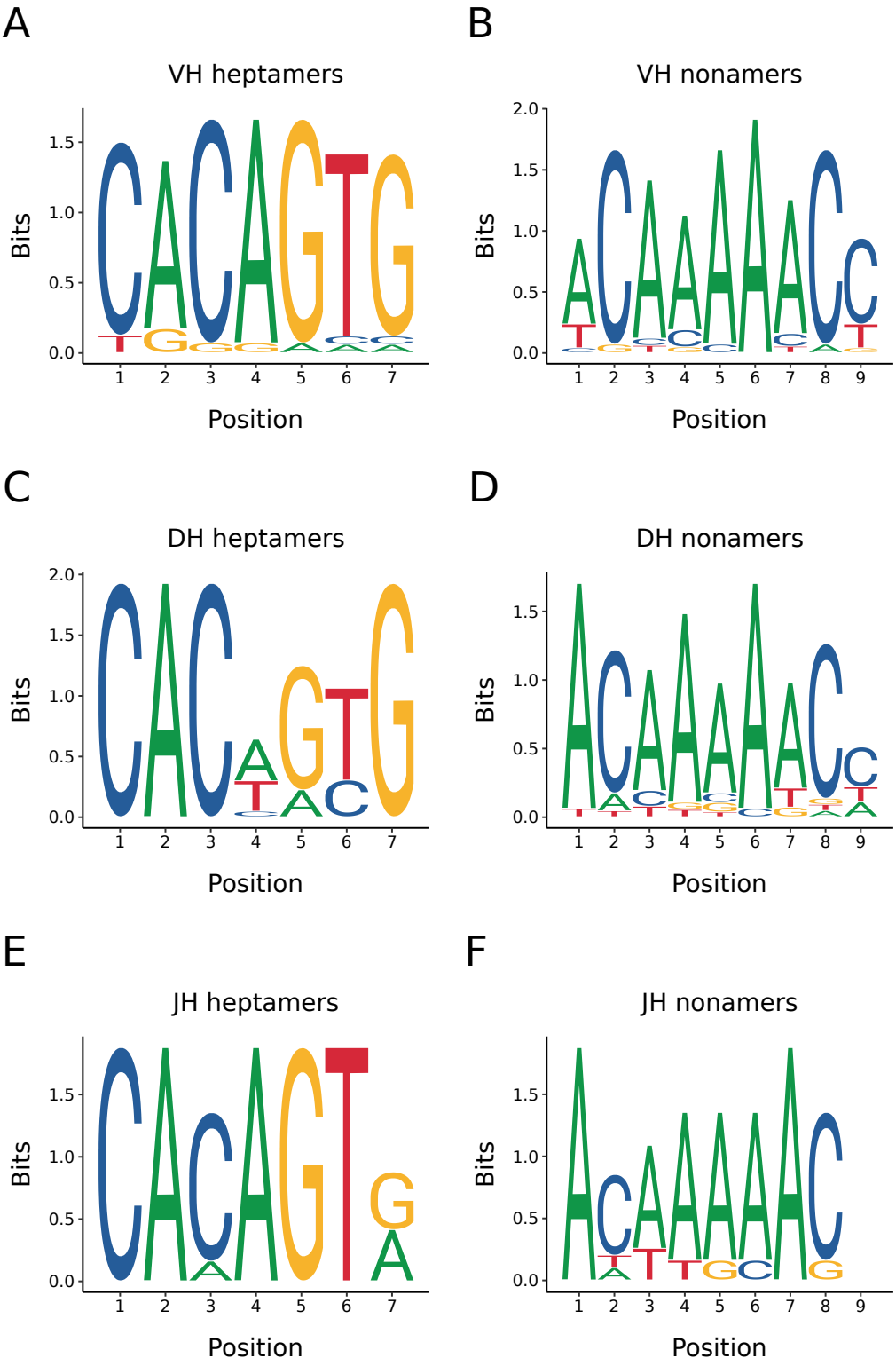


Figure D.1: *N. furzeri* recombination signal sequences by segment type: Sequence composition of conserved heptamer (A,C,E) and nonamer (B,D,F) sequences from *N. furzeri* heavy-chain RSSs associated with VH (A,B), DH (C,D) or JH (E,F) gene segments.

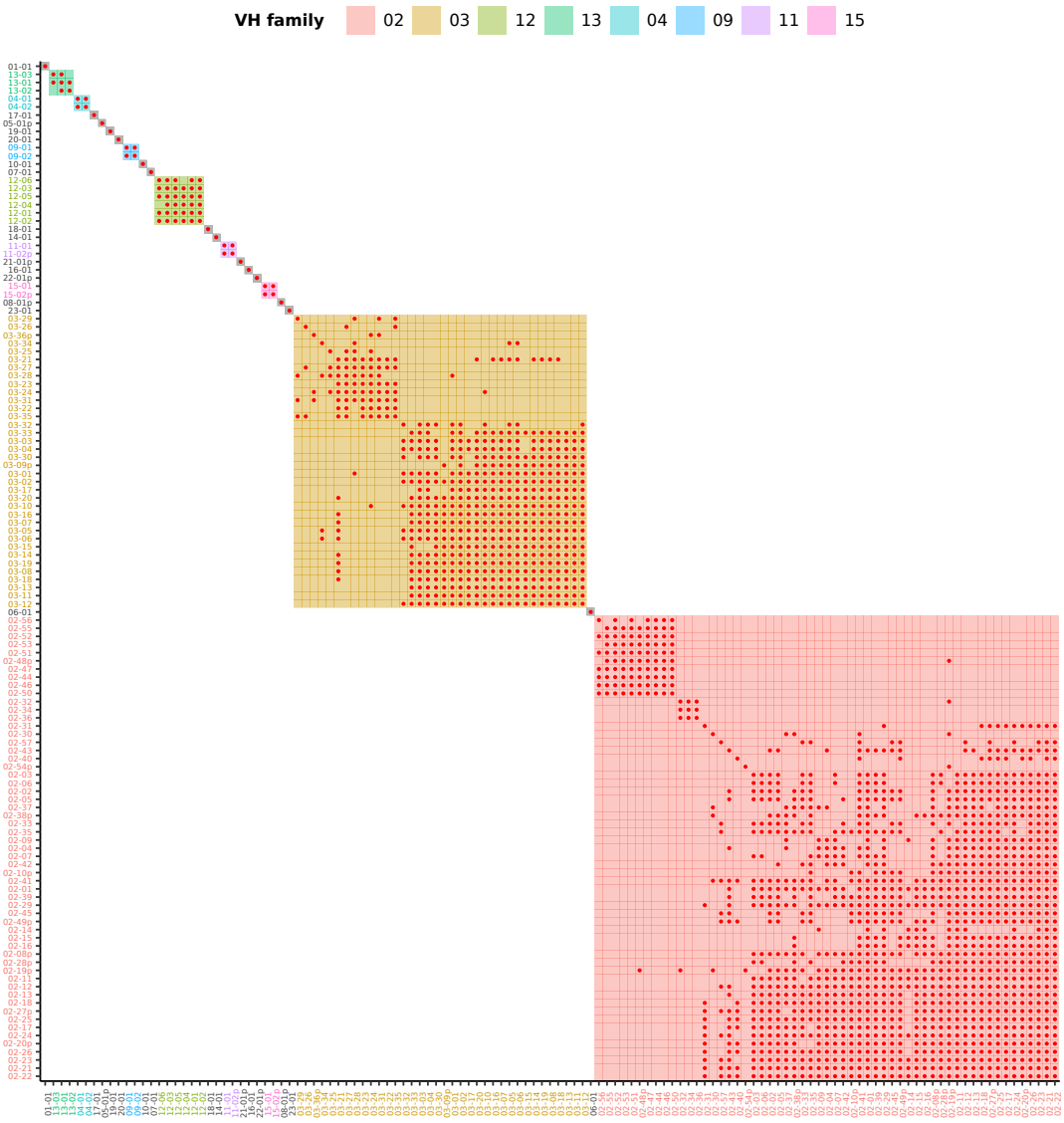


Figure D.2: Heatmap of VH families in the *X. maculatus IGH* locus: Heatmap of family relationships among *X. maculatus* VH segments, with coloured shading indicating families and red dots indicating pairwise nucleotide sequence identity of at least 80%. VH families containing multiple segments are uniquely coloured, while single-segment families are in grey.

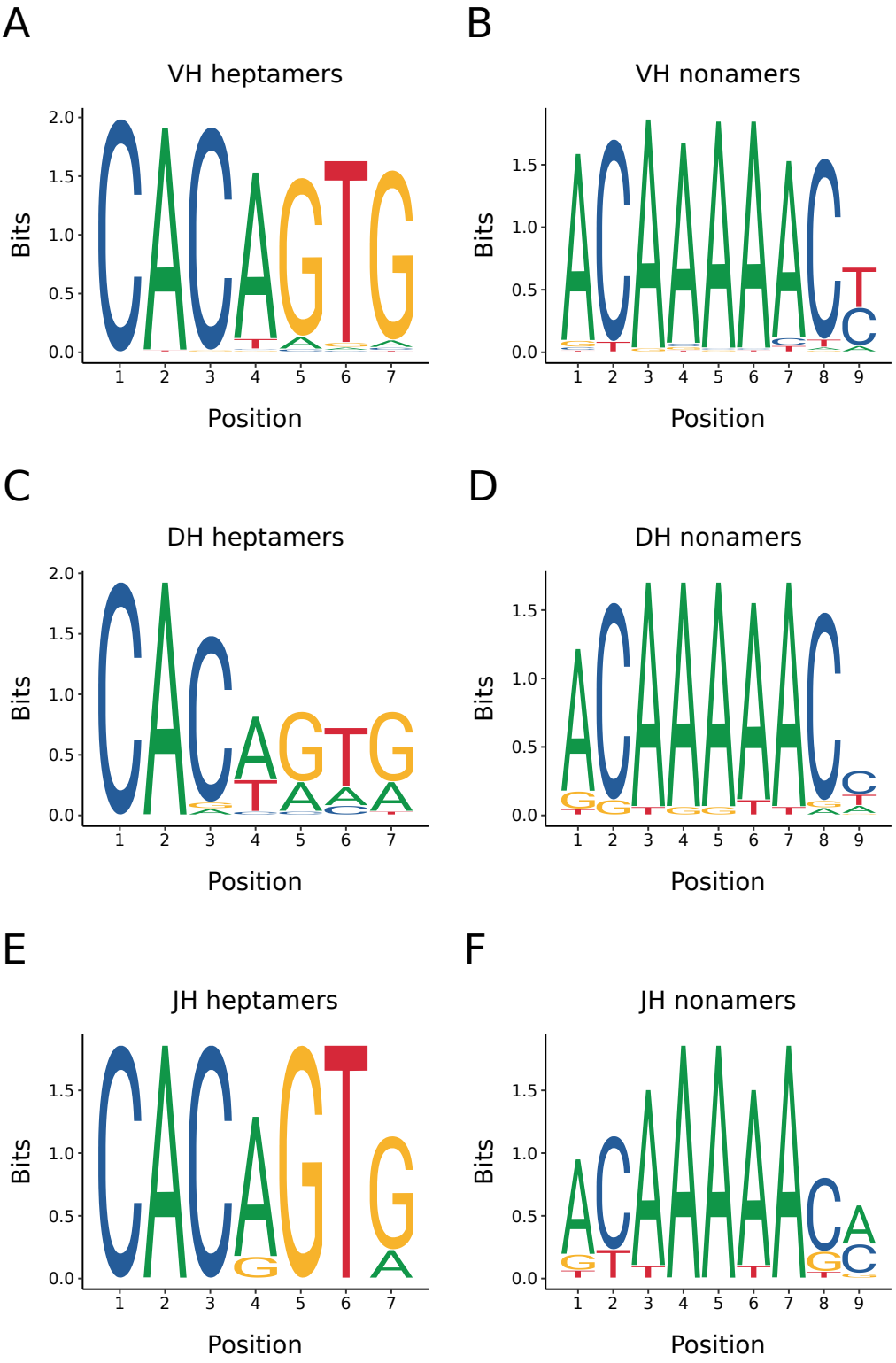


Figure D.3: ***X. maculatus*** **recombination signal sequences by segment type:** Sequence composition of conserved heptamer (A,C,E) and nonamer (B,D,F) sequences from *X. maculatus* heavy-chain RSSs associated with VH (A,B), DH (C,D) or JH (E,F) gene segments.

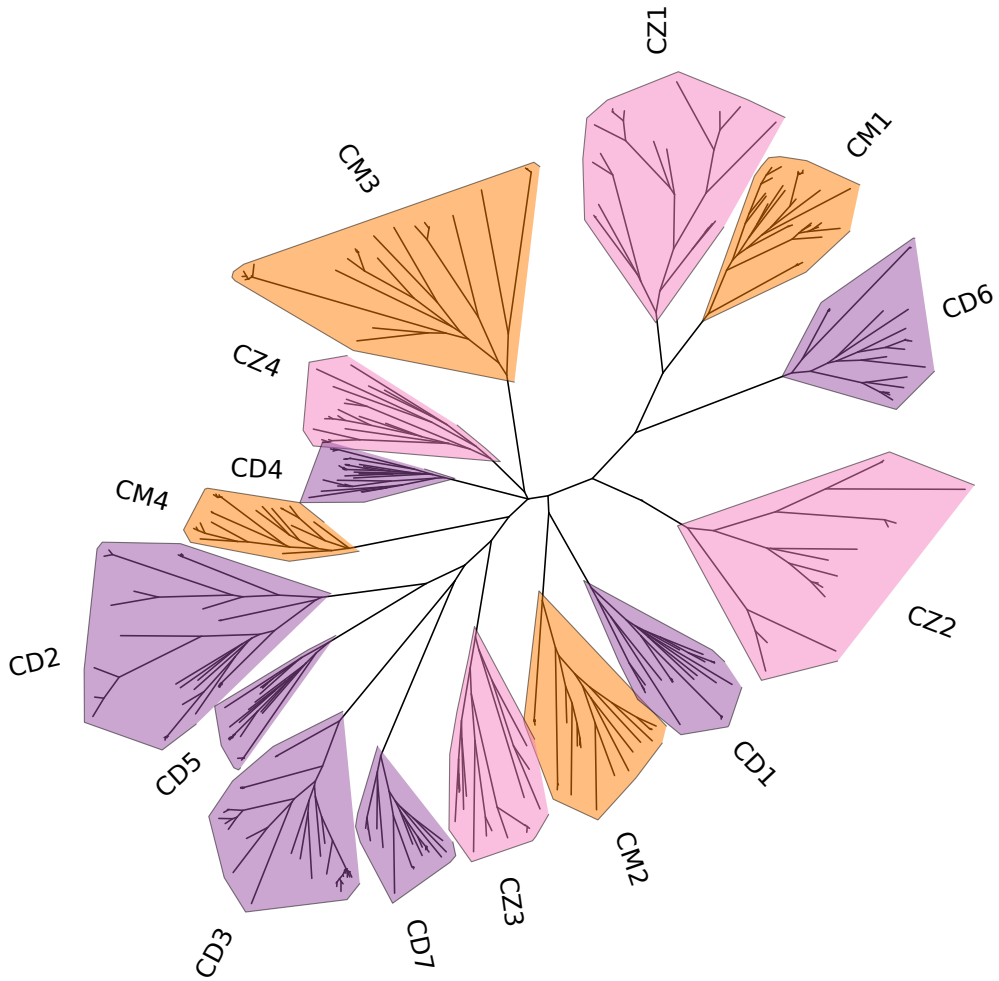


Figure D.4: Constant-region CH exons in the Atherinomorpha: Unrooted phylogram of CH exons from thirteen fish species from the Atherinomorpha (Table 3.6), constructed using PRANK and RAxML. Each exon type is clustered separately in the tree topology, indicating that the types of the identified exons have all been correctly annotated.

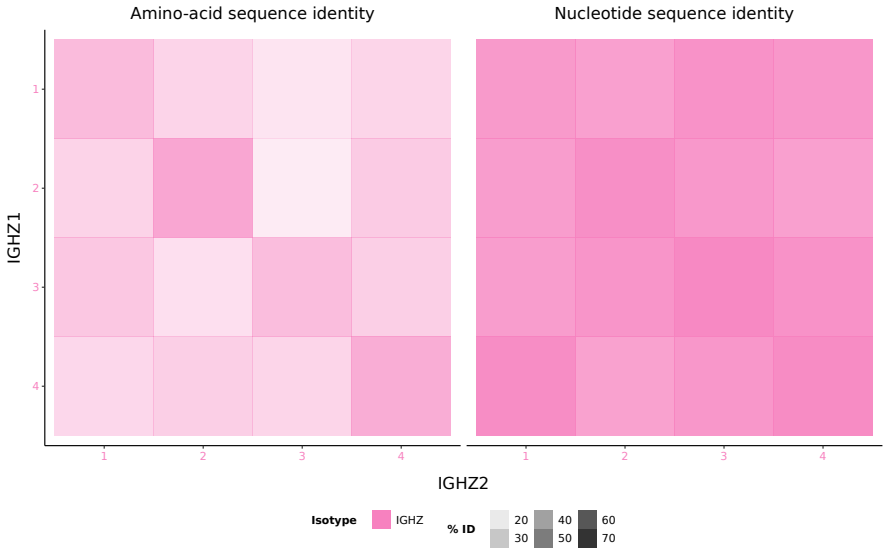


Figure D.5: Sequence similarity between *IGHZ* constant-regions in *X. maculatus*: Heatmap of percentage sequence identity between amino-acid (right) and nucleotide (left) sequences of C_ζ exons from the two *X. maculatus IGHZ* constant regions, calculated using pairwise Needleman-Wunsch global alignments.

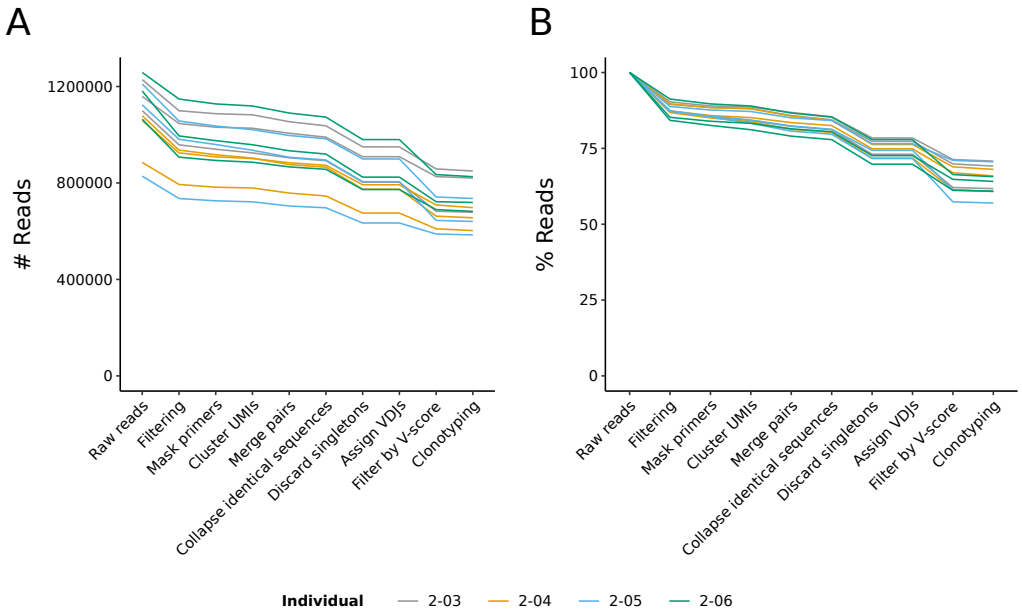


Figure D.6: Read survival during complete pre-processing of the IgSeq pilot dataset: Line graphs of absolute (A) and relative (B) read survival during pre-processing of the IgSeq pilot dataset, up to and including clonotyping.

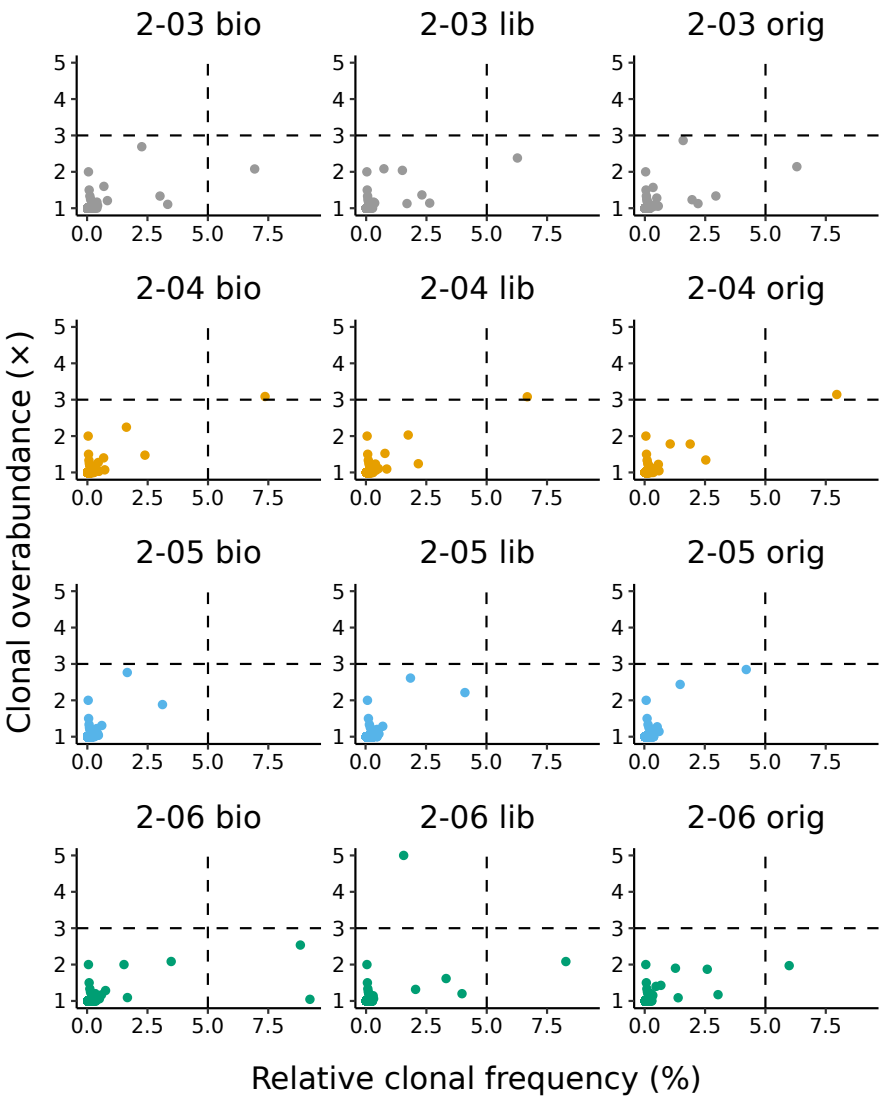


Figure D.7: Clonal expansions in *N. furzeri* pilot replicates: Scatter plots of clonal abundance for each replicate in the IgSeq pilot dataset, measured in terms of the proportion of unique sequences in the repertoire (x-axis) and the abundance relative to the next-largest clone (y-axis). Thresholds for identifying clonal expansions (5 % and 3-fold for the x- and y-axis, respectively) suggested by Rosenfeld *et al.* [191].

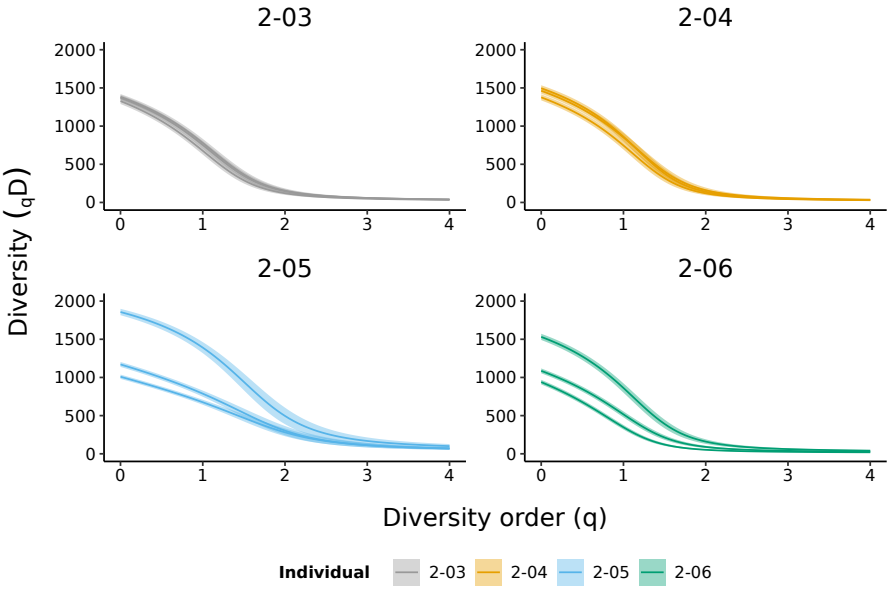


Figure D.8: Per-replicate clonal-diversity spectra for the IgSeq pilot dataset: Hill diversity spectra of clone sizes (as measured by number of unique sequences per clone) for each replicate in the IgSeq pilot dataset, grouped by source individual.

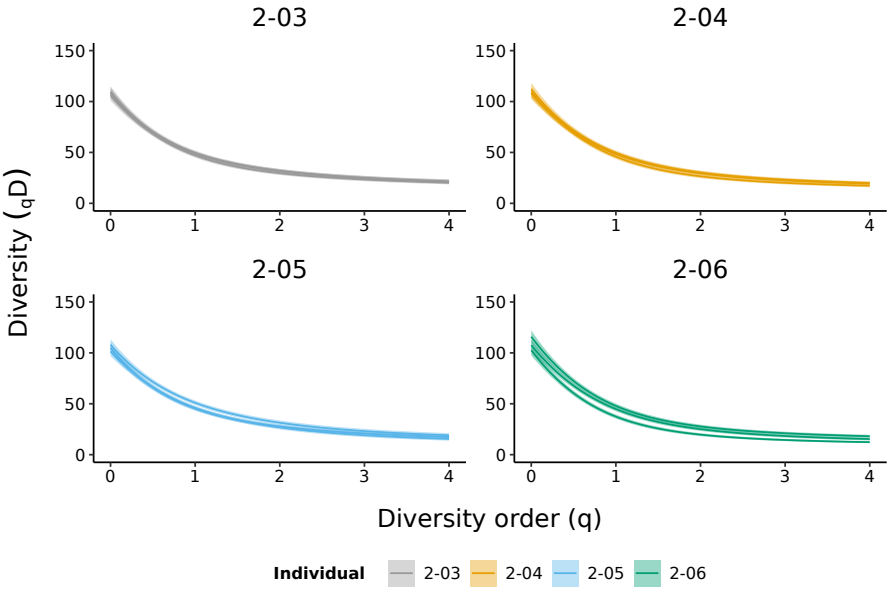


Figure D.9: Per-replicate VJ-diversity spectra for the IgSeq pilot dataset: Hill diversity spectra of VJ usage (as measured by number of unique sequences per V/J combination) for each replicate in the IgSeq pilot dataset, grouped by source individual.

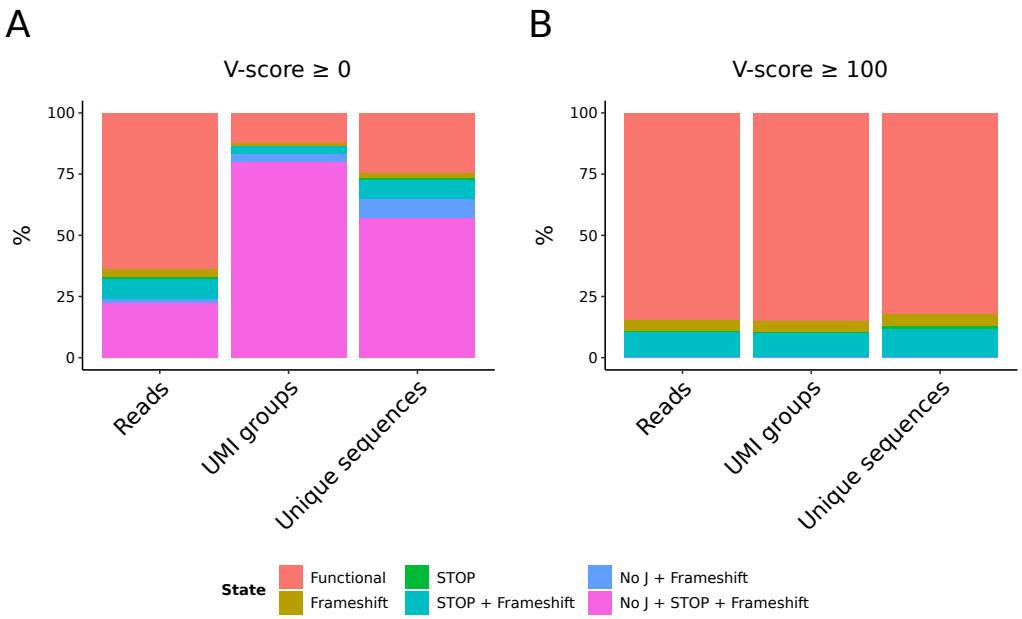


Figure D.10: Functional composition and V-score filtering in the IgSeq ageing dataset: Proportion of input reads, UMI groups and unique sequences in the IgSeq ageing dataset belonging to different (non)functional categories, before (A) and after (B) filtering on V-alignment score.

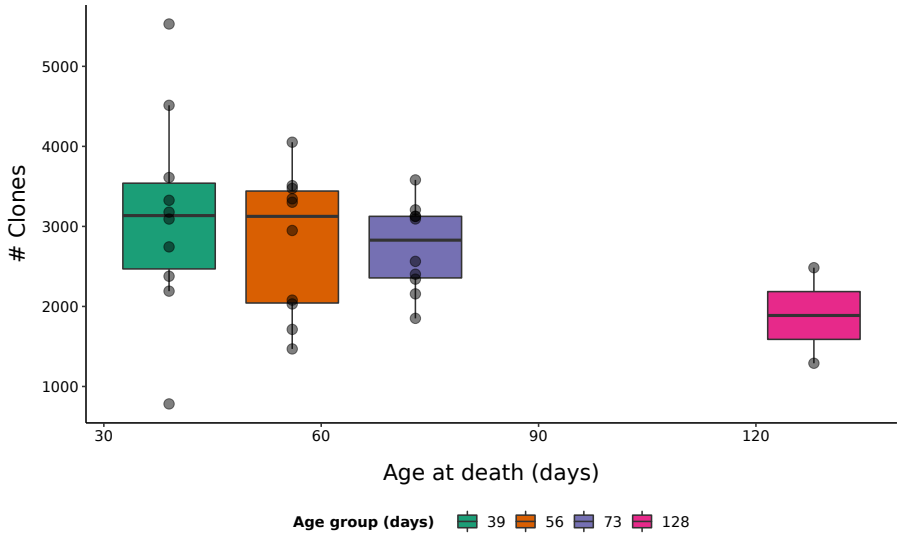


Figure D.11: Number of clones in the IgSeq ageing dataset: Boxplots of clonal counts for each individual in the IgSeq ageing dataset, grouped by age at death. The apparent decline in clonal count with age is not significant (Kruskal-Wallis analysis of variance, $p = 0.45$).

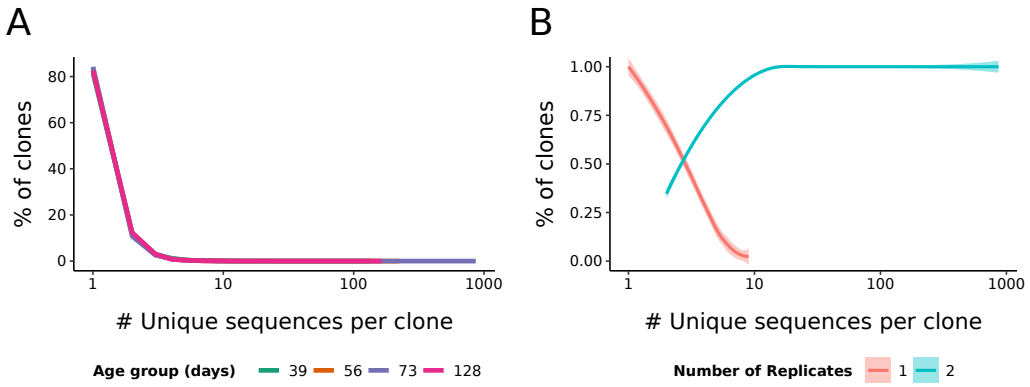


Figure D.12: Clone size and cross-replicate reproducibility in the IgSeq ageing dataset: (A) Proportion of clones of different sizes for each individual in the IgSeq ageing dataset, measured in unique sequences per clone. (B) LOESS-smoothed curves [188] showing proportion of clones of each size found across one or both replicates of the appropriate individual.

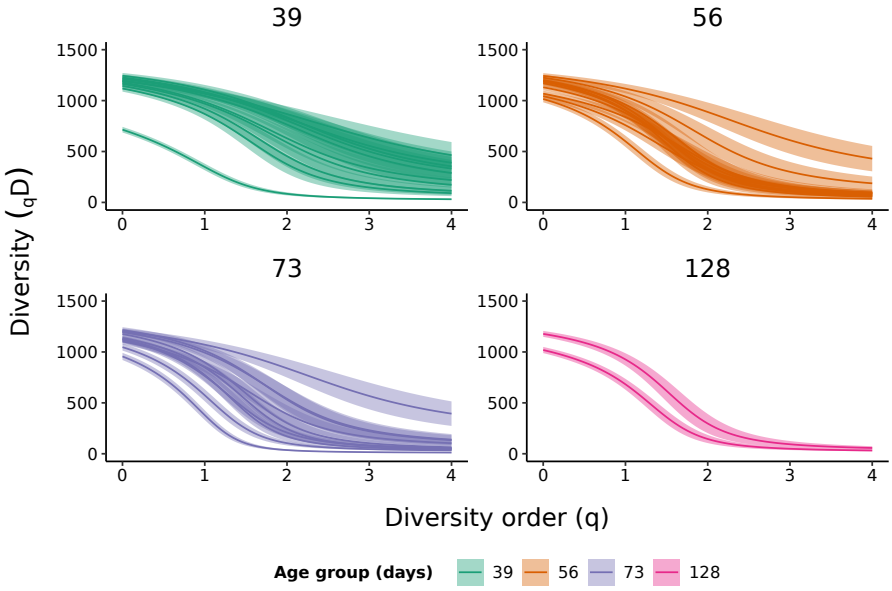


Figure D.13: Per-individual clonal-diversity spectra in the IgSeq ageing dataset: Hill diversity spectra of clone sizes (as measured by number of unique sequences per clone) for each individual in the IgSeq ageing dataset, grouped by age at death.

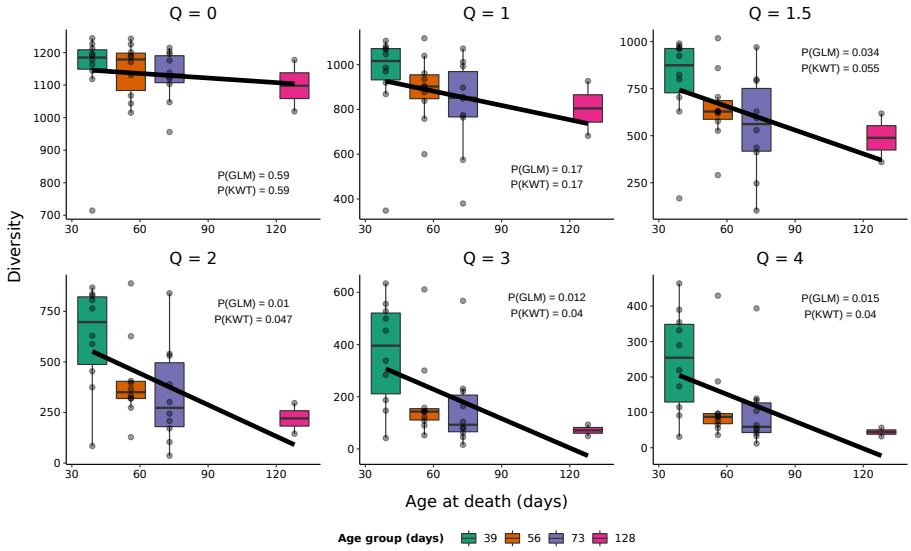


Figure D.14: Comparing clonal alpha-diversities between age groups in the IgSeq ageing dataset (linear fit): Boxplots of Hill diversity values for the antibody repertoires of individuals of each age group in the IgSeq ageing dataset at a sample of diversity orders, overlaid with the predictions of the best-fit linear model at each order. Annotated p-values indicate the statistical significance of the estimated age effect on diversity under the linear model ($P(GLM)$) and a Kruskal-Wallis test ($P(KWT)$) for each diversity order.

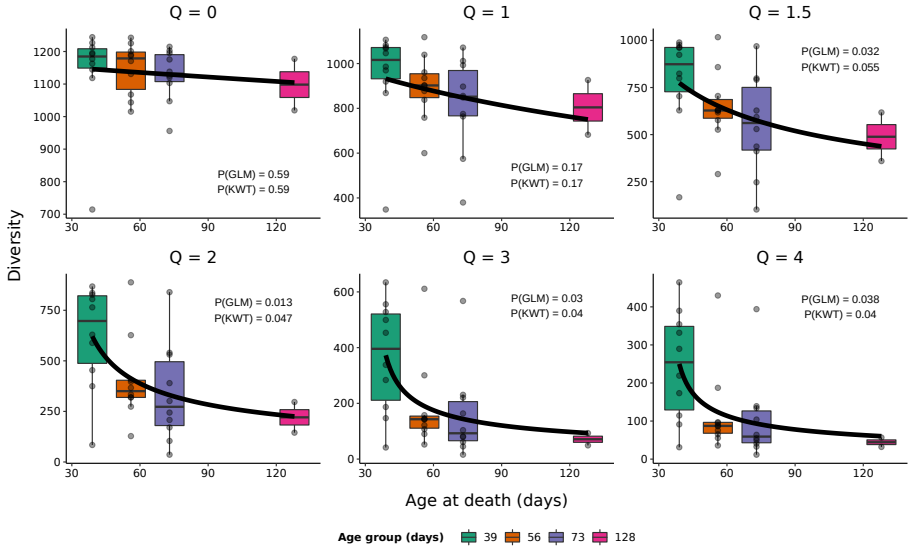


Figure D.15: Comparing clonal alpha-diversities between age groups in the IgSeq ageing dataset (inverse-Gaussian fit): Boxplots of Hill diversity values for the antibody repertoires of individuals of each age group in the IgSeq ageing dataset at a sample of diversity orders, overlaid with the predictions at each order of the best-fit inverse-Gaussian-family GLM under an inverse-squared link function. Annotated p-values indicate the statistical significance of the estimated age effect on diversity under the GLM ($P(GLM)$) and a Kruskal-Wallis test ($P(KWT)$) for each diversity order.

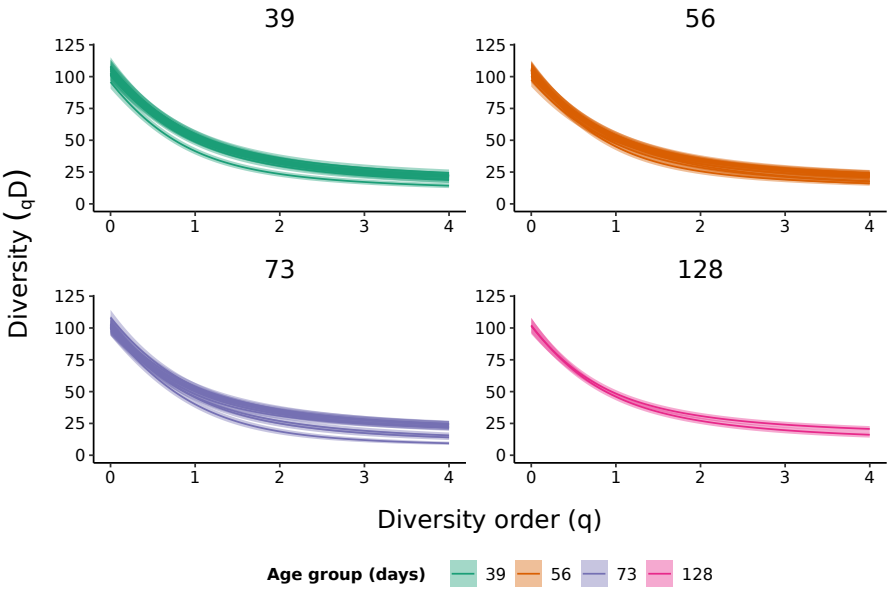


Figure D.16: Per-individual VJ-diversity spectra for the IgSeq ageing dataset: Hill diversity spectra of VJ usage (as measured by number of unique sequences per V/J combination) for each individual in the IgSeq ageing dataset, grouped by age at death.

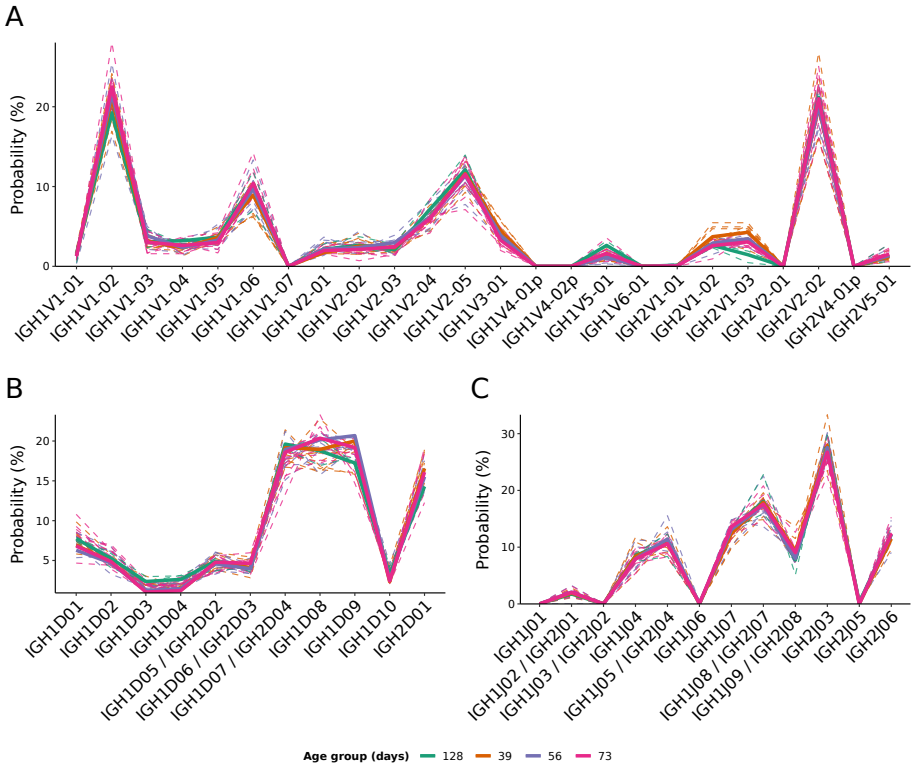


Figure D.17: Generative segment-choice distributions in the IgSeq ageing dataset: Probability distributions of segment choice for (A) VH-, (B) DH- and (C) JH-segments during VDJ recombination in adult male turquoise killifish of different ages, inferred from the IgSeq ageing dataset using IGoR. Thin dashed lines represent the distributions inferred for individual killifish, while the thick solid lines represent those inferred from pooled data from all individuals in each age group. DH and JH segments with identical sequences (which cannot be distinguished in the repertoire data even in principle) are collapsed together.

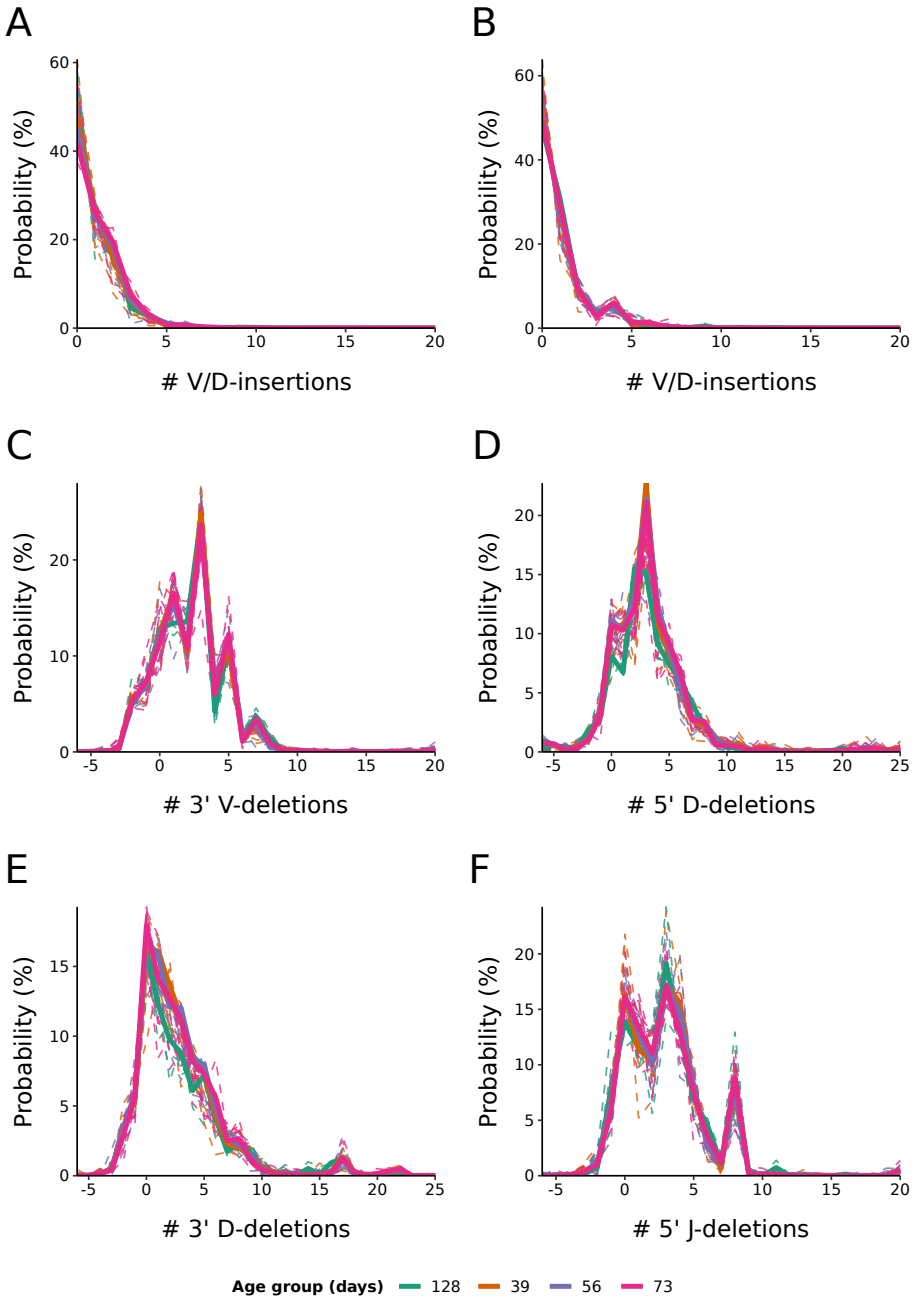


Figure D.18: Generative insertion/deletion distributions in the IgSeq ageing dataset: Probability distributions of the number of N-insertions (A-B) or P-insertions/deletions (C-F) following VDJ recombination in adult male turquoise killifish of different ages, inferred from the IgSeq ageing dataset using IGoR. P-insertions are modelled as negative deletions. Thin dashed lines represent the distributions inferred for individual killifish, while the thick solid lines represent those inferred from pooled data from all individuals in each age group.

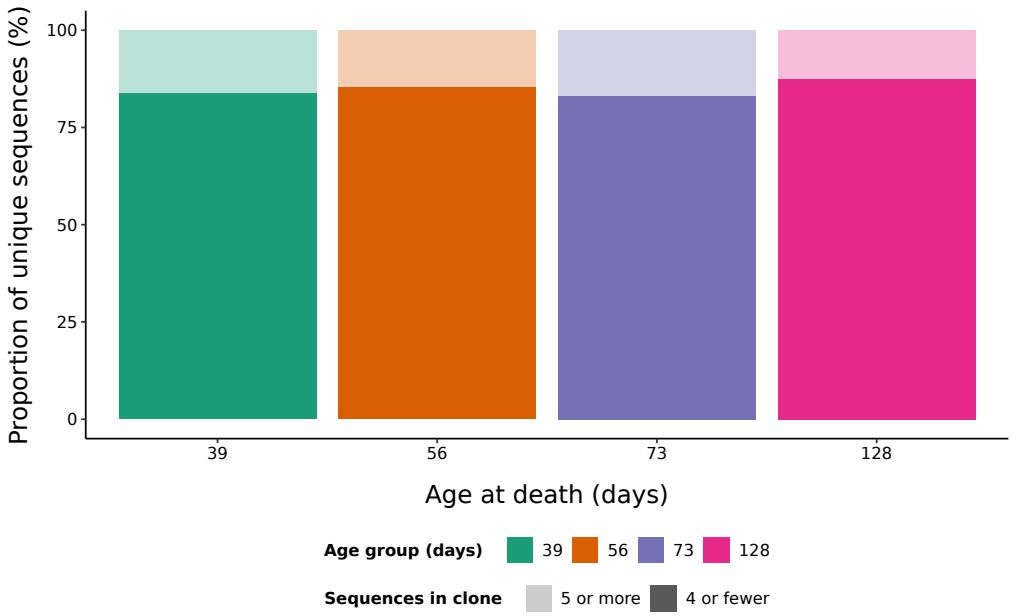


Figure D.19: Proportion of unique sequences in large vs small clones in the IgSeq ageing dataset: Stacked barplots showing the mean proportion of unique sequences in large (5 or more unique sequences, top, pale) vs small (4 or fewer, bottom, dark) clones in each age group in the IgSeq ageing dataset. The proportion of sequences in non-abundant clones does not change significantly with age (Kruskal-Wallis analysis of variance, $p = 0.4$).

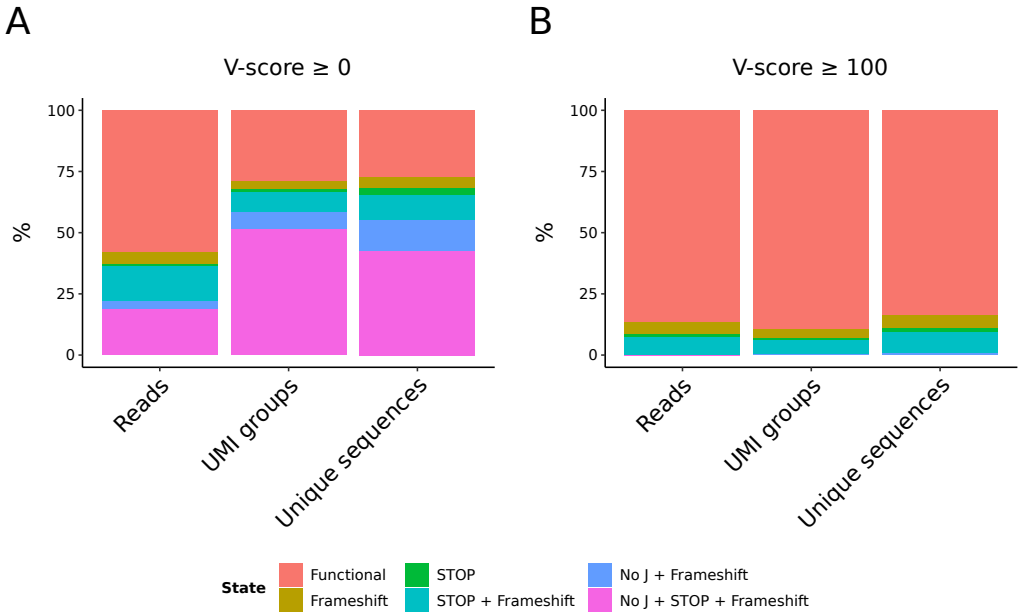


Figure D.20: Functional composition and V-score filtering in the IgSeq gut dataset: Proportion of input reads, UMI groups and unique sequences in the IgSeq gut dataset belonging to different (non)functional categories, before (A) and after (B) filtering on V-alignment score.

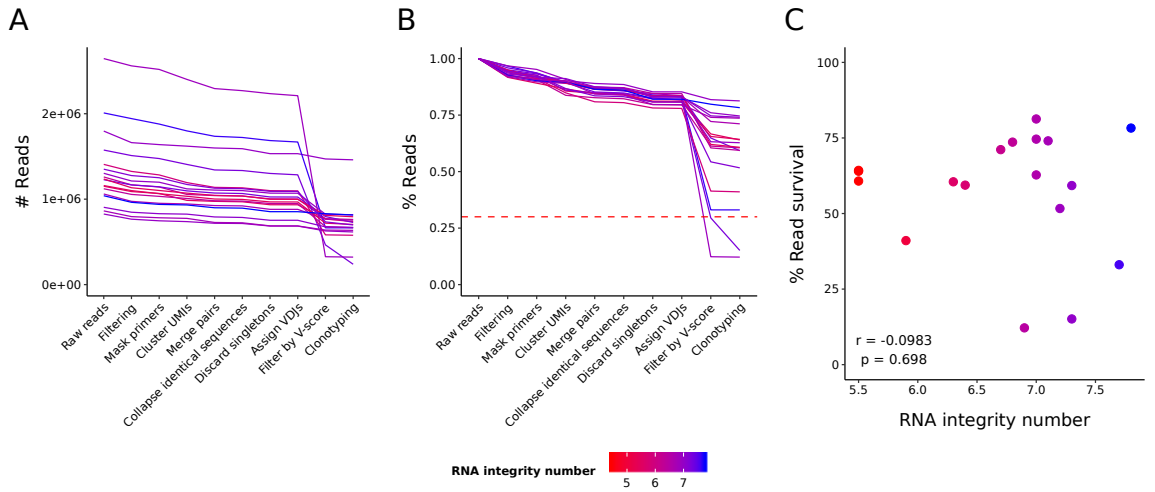


Figure D.21: Relationship between RNA integrity and read survival in the IgSeq gut dataset: (A-B) Absolute (A) and relative (B) read survival during pre-processing of the IgSeq gut dataset, up to and including clonotyping, coloured by the RNA integrity number of each input sample. The dotted red line in (B) indicates the 30% read-survival cutoff, below which samples were discarded prior to downstream analysis. (C) Scatterplot of RNA integrity number vs percentage read survival, up to and including clonotyping.

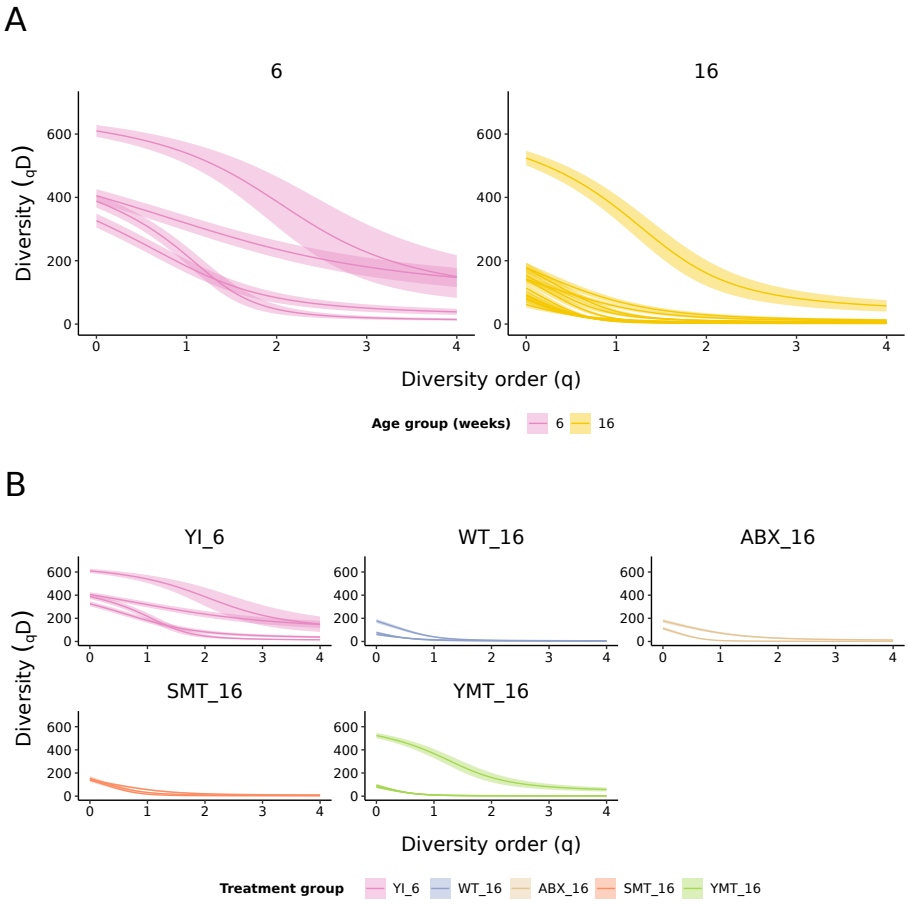


Figure D.22: Per-individual clonal-diversity spectra for the IgSeq gut dataset: Hill diversity spectra of clone sizes (as measured by number of unique sequences per clone) for each individual in the IgSeq gut dataset, grouped by (A) age at death and (B) treatment group.

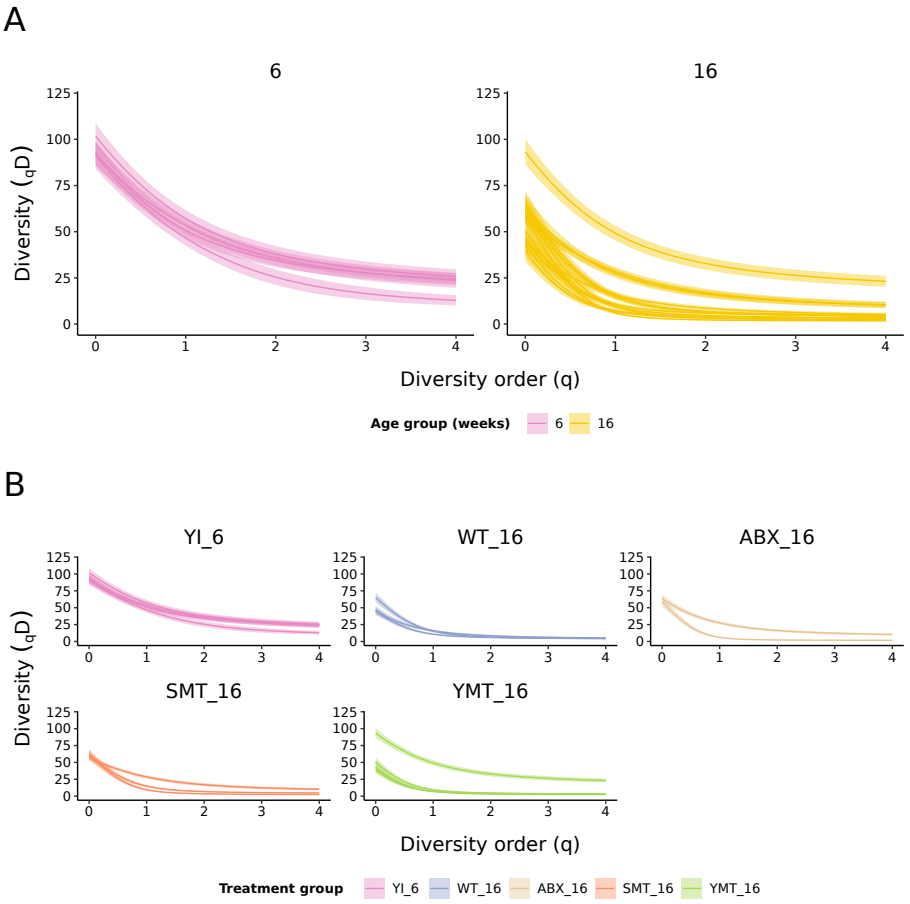


Figure D.23: Per-individual VJ-diversity spectra for the IgSeq gut dataset: Hill diversity spectra of VJ usage (as measured by number of unique sequences per V/J combination) for each individual in the IgSeq gut dataset, grouped by (A) age at death and (B) treatment group.

Appendix E

Supplementary tables

Table E.1: Software versions used in computational analyses

Program	Version
Basemount	0.15.96.2154
BLAST	2.7.1
Bowtie 2	2.2.6
CD-HIT-EST	4.6.8
Change-O	0.4.5
EMBOSS (FUZZNUC)	6.6.0
FigTree	1.4.2
HMMER	3.2
IgBLAST	1.7.0
IGoR	1.3.0
IGV	2.3.68
IMGT/DomainGapAlign	4.9.2
PRANK	v.170427
pRESTO	0.5.10
Primer3	2.3.6
Python 2	2.7.14
Python 3	3.6.4
QuorUM	1.0.0
R	3.4.1/3.6.2
RAxML	8.2.12
RepeatMasker	4.0.6
SAMtools	1.9
sed	4.2.2
seqtk	1.3
Snakemake	5.3.0
SPAdes	3.6.1
SSPACE	3.0
STAR	2.5.2b
Trimmomatic	0.32
VSEARCH	2.8.0

Table E.2: RNA-sequencing datasets used for *IGH* locus characterisation

Species	*N. furzeri*	*X. maculatus*
Tissues	Gut	Various[a]
BioProject Accession	PRJNA379208	PRJNA420092
SRA Run Accessions	SRR5344350 SRR5344343 SRR5344344 SRR5344345 SRR5344346 SRR5344347 SRR5344348 SRR5344349 SRR5344350	SRR6327069 SRR6327070 SRR6327071 SRR6327072 SRR6327073 SRR6327074 SRR6327075 SRR6327076 SRR6327077 SRR6327078 SRR6327079 SRR6327080 SRR6327081 SRR6327082 SRR6327083 SRR6327084 SRR6327085 SRR6327086 SRR6327087 SRR6327088 SRR6327089 SRR6327090 SRR6327091 SRR6327092 SRR6327093 SRR6327094
Source	[78]	Citation not given

[a] Tissues used for *X. maculatus* RNA-sequencing included brain, heart, liver, gut, skin or whole fish; see BioProject entry for details.

Table E.3: Co-ordinate table of constant-region exons in the *N. furzeri IGH* locus

Name	Isotype	Start	End	Length	Strand
IGH1M-1	M	130848	131144	297	+
IGH1M-2	M	131971	132312	342	+
IGH1M-3	M	132394	132705	312	+
IGH1M-4	M	132816	133288	473	+
IGH1M-TM1	M	134262	134413	152	+
IGH1M-TM2	M	138431	138819	389	+
IGH1D-1	D	139381	139689	309	+
IGH1D-2A	D	139774	140064	291	+
IGH1D-3A	D	140178	140489	312	+
IGH1D-4A	D	140572	140853	282	+
IGH1D-2B	D	145613	145909	297	+
IGH1D-3B	D	146000	146311	312	+
IGH1D-4B	D	146398	146676	279	+
IGH1D-5	D	146795	147124	330	+
IGH1D-6	D	147210	147527	318	+
IGH1D-7	D	147598	147885	288	+
IGH1D-TM1	D	148016	148164	149	+
IGH1D-TM2	D	148323	148504	182	+
IGH2D-TM2	D	187624	187803	180	-
IGH2D-TM1	D	187963	188111	149	-
IGH2D-7	D	188658	188945	288	-
IGH2D-6	D	189016	189333	318	-
IGH2D-5	D	189419	189748	330	-
IGH2D-4B	D	189867	190145	279	-
IGH2D-3B	D	190232	190543	312	-
IGH2D-2B	D	190636	190932	297	-
IGH2D-4A	D	195644	195925	282	-
IGH2D-3A	D	196008	196319	312	-
IGH2D-2A	D	196433	196723	291	-
IGH2D-1	D	196808	197116	309	-
IGH2M-TM2	M	198315	198506	192	-
IGH2M-TM1	M	199834	199985	152	-
IGH2M-4	M	200953	201425	473	-
IGH2M-3	M	201536	201847	312	-
IGH2M-2	M	201929	202270	342	-
IGH2M-1	M	203549	203845	297	-

Name	Start	End	Length	Strand	RSS Start	Heptamer	Spacer Length	Nonamer	RSS End	RSS Length	Comment
IGH1V1-01	1252	1540	289	+	1541	CACAGTG	22	ACAAAAACC	1578	38	
IGH1V1-02	3365	3656	292	+	3657	CACAGTG	22	ACAAAAACC	3694	38	
IGH1V2-01	5907	6201	295	+	6202	CACAGAA	15	ACAAAAACT	6232	31	
IGH1V1-03	13690	13964	275	+	13965	CACAGTG	22	ACAAAAACC	14002	38	
IGH1V3-01	14862	15162	301	+	15163	CACAGTG	23	ACAAAAACC	15201	39	
IGH1V2-02	17433	17730	298	+	17731	CACAATG	23	ACAAAAACC	17769	39	
IGH1V4-01p	24566	24837	272	+	24838	CGCAGTG	22	CCACAAACC	24875	38	Nonsense mutation
IGH1V1-04	37305	37596	292	+	37597	CACAGTG	22	ACAAAAACC	37634	38	
IGH1V2-03	48845	49139	295	+	49140	CACAGTG	23	TCAAAAACT	49178	39	
IGH1V1-05	49909	50197	289	+	50198	CACAGTG	22	ACAAAAACC	50235	38	
IGH1V5-01	51710	51998	289	+	51999	CACAGTG	22	ACAAAAACT	52036	38	
IGH1V2-04	56322	56616	295	+	56617	CACAGTG	23	ACAAAAACC	56655	39	
IGH1V6-01	57465	57762	298	+	57763	CACAGTG	21	ACTAAATCT	57799	37	
IGH1V1-06	59678	59966	289	+	59967	CACAGTG	22	ACAAAAACC	60004	38	
IGH1V4-02p	68017	68288	272	+	68289	TGCAGTG	22	TCACAAACC	68326	38	Nonsense mutation
IGH1V2-05	69787	70084	298	+	70085	CACAGTG	23	ACAAAAACC	70123	39	
IGH1V1-07	155485	155763	279	+	155764	CACAGTG	22	TCAAAACCC	155801	38	
IGH2V2-02	282620	282914	295	-	282915	CACAGTG	23	ACAAAAACC	282953	39	
IGH2V4-01p	284404	284675	272	-	284676	TGCAGTG	22	TCACAAACC	284713	38	Nonsense mutation
IGH2V5-01	288808	289096	289	-	289097	CACAGTG	22	ACAGAAACT	289134	38	
IGH2V1-03	289977	290271	295	-	290272	CACAGTG	22	ACAAAAACC	290309	38	
IGH2V1-02	293835	294126	292	-	294127	CACAGTG	22	ACAAAAACC	294164	38	
IGH2V2-01	303780	304074	295	-	304075	CAGGGCC	24	AGCACAAAG	304114	40	
IGH2V1-01	304926	305204	279	-	305205	CACAGTG	22	TCAAAACCC	305242	38	

Table E.4: Co-ordinate table of VH segments in the *N. furzeri IGH* locus

Table E.5: Co-ordinate table of DH segments in the *N. furzeri IGH* locus

Name	Start	NT Sequence	End	Length	Strand
IGH1D01	25782	ATACGTACTTTCGTGGTATATAGAGA	25807	26	+
IGH1D02	76700	GATATCTGGGTGGGGG	76715	16	+
IGH1D03	77027	TGAAATGATTAC	77038	12	+
IGH1D04	77476	TCGCGTAGCGGC	77487	12	+
IGH1D05	78717	GAAACCACGGCAGC	78730	14	+
IGH1D06	79049	TTTATAGCGGCTAC	79062	14	+
IGH1D07	80417	CAGACTGGAGA	80427	11	+
IGH1D08	81362	TTCATGGCAGCCAC	81375	14	+
IGH1D09	82067	CAGACTGGAGC	82077	11	+
IGH1D10	84282	TGGGGTGGCAGC	84293	12	+
IGH2D04	263497	CAGACTGGAGA	263507	11	-
IGH2D03	270243	TTTATAGCGGCTAC	270256	14	-
IGH2D02	270878	GAAACCACGGCAGC	270891	14	-
IGH2D01	271749	GACTTTTACTAC	271760	12	-

Table E.6: Co-ordinate table of DH 5'-RSSs in the *N. furzeri IGH* locus

Name	5'-RSS Start	Nonamer	Spacer Length	Heptamer	5'-RSS End	Length
IGH1D01	25754	GGTTGTTGT	12	CACTGTG	25781	28
IGH1D02	76672	AGTTTTTGA	12	CACAGTG	76699	28
IGH1D03	76999	TGTTGTTGT	12	CACAGTG	77026	28
IGH1D04	77448	AGTTTTTGT	12	CACGGTG	77475	28
IGH1D05	78688	GATGTTTTT	13	CACAGTG	78716	29
IGH1D06	79021	TGTTTTTGT	12	CGCTGTG	79048	28
IGH1D07	80389	AGTTTTGGT	12	CACAGTG	80416	28
IGH1D08	81334	TGTTTTTGT	12	CGCTGTG	81361	28
IGH1D09	82039	AGTTTTGGT	12	CACAGTG	82066	28
IGH1D10	84254	TCATTCATT	12	CACTGTG	84281	28
IGH2D04	263469	AGTTTTGGT	12	CACAGTG	263496	28
IGH2D03	270215	TGTTTTTGT	12	CGCTGTG	270242	28
IGH2D02	270850	TGTTTTTGT	12	CACAGTG	270877	28
IGH2D01	271721	AGTTTTTAT	12	CATGGTG	271748	28

Table E.7: Co-ordinate table of DH 3'-RSSs in the *N. furzeri IGH* locus

Name	3'-RSS Start	Heptamer	Spacer Length	Nonamer	3'-RSS End	Length
IGH1D01	25808	CACAGTG	12	ACAAAAACC	25835	28
IGH1D02	76716	CACAGTG	12	ACAAAAACC	76743	28
IGH1D03	77039	CACTGTG	11	AATATAACC	77065	27
IGH1D04	77488	CACAGCG	12	ACATAAAAC	77515	28
IGH1D05	78731	CACAGCG	12	ACAAAAGCC	78758	28
IGH1D06	79063	CACTGTG	12	ACAAGATCC	79090	28
IGH1D07	80428	CACAACG	12	ACAAAAACC	80455	28
IGH1D08	81376	CACTGTG	12	ACAAAATCC	81403	28
IGH1D09	82078	CACAATG	12	ACAAAAACC	82105	28
IGH1D10	84294	CACAGTG	12	ACAAAAACC	84321	28
IGH2D04	263508	CACAACG	12	ACAAAAACC	263535	28
IGH2D03	270257	CACTGTG	12	ACAAGATCC	270284	28
IGH2D02	270892	CACAGCG	12	ACAAAAGCC	270919	28
IGH2D01	271761	CACAATG	12	ACAAAAACC	271788	28

Name	Start	NT Sequence	AA Sequence	End	Length	Strand
IGH1J01	26187	GTGCTTTAGACAACTGGGGAAAAGGAACGGAGGTTACTGTTCAACCTG	ALDNWGKGTEVTVQP	26234	48	+
IGH1J02	128176	ATGACTACTTTGACTACTGGGGAAAAGGAACAATGGTGACGGTCACATCAG	DYFDYWGKGTMVTVTS	128226	51	+
IGH1J03	128354	ACCGTGGGGTAAAGGGACAACAGTCACGGTCAAAACAG	PWGKGTTVTVKT	128391	38	+
IGH1J04	128533	ACGGTGCTCTTGACTACTGGGGTAAAGGGACCGCAGTCACTGTAACATCAG	GALDYWGKGTAVTVTS	128583	51	+
IGH1J05	128887	ACAACGCTTTTGACTACTGGGGAAAAGGAACAACGGTCACCGTCACTTCAG	NAFDYWGKGTTVTVTS	128937	51	+
IGH1J06	129346	CTACGATGCTTTTGACTACTGGGGGAAAAGGACGATGGTCACGTCACTTCAG	YDAFDYWGKRTMVTSLQ	129397	52	+
IGH1J07	129635	TTAACTGGGCTTTCGACTACTGGGGAAAAGGGACGATGGTAACGGTGACTTCAG	NWAFDYWGKGTMVTVTS	129688	54	+
IGH1J08	129965	TTACCACGCAGCTTTGGACTACTGGGGAAAAGGGACGACGGTCACCGTCACCTCAG	YHXALDYWGKGTTVTVTS	130020	56	+
IGH1J09	130612	TCTACGCTGCTTTTGACTACTGGGGTAAAGGTACAACGGTAACCGTTTCATCAG	YAAFDYWGKGTTVTVSS	130665	54	+
IGH2J08	204031	TCTACGCTGCTTTTGACTACTGGGGTAAAGGTACAACGGTAACCGTTTCATCAG	YAAFDYWGKGTTVTVSS	204084	54	-
IGH2J07	204673	TTACCACGCAGCTTTGGACTACTGGGGAAAAGGGACGACGGTCACCGTCACCTCAG	YHXALDYWGKGTTVTVTS	204728	56	-
IGH2J06	205005	ATAACTGGGCTTTCGACTACTGGGGAAAAGGGACGATGGTAACGGTGACTTCAG	NWAFDYWGKGTMVTVTS	205058	54	-
IGH2J05	205296	CTACGATGCTTTTGACTACTGGGGGAAAAGGACGATGGTCACGTCACTTCAG	YDAFDYWGKRTMVTSLQ	205347	52	-
IGH2J04	205756	ACAACGCTTTTGACTACTGGGGAAAAGGAACAACGGTCACCGTCACTTCAG	NAFDYWGKGTTVTVTS	205806	51	-
IGH2J03	206111	ATGGTGCTTTTGACTACTGGGGTAAAGGGACCGCAGTCACTGTAACATCAG	GAFDYWGKGTAVTVTS	206161	51	-
IGH2J02	206303	ACCGTGGGGTAAAGGGACAACAGTCACGGTCAAAACAG	PWGKGTTVTVKT	206340	38	-
IGH2J01	206466	ATGACTACTTTGACTACTGGGGAAAAGGAACAATGGTGACGGTCACATCAG	DYFDYWGKGTMVTVTS	206516	51	-

Table E.8: Co-ordinate table of JH segments in the *N. furzeri IGH* locus

Name	RSS Start	Nonamer	Spacer Length	Heptamer	RSS End	RSS Length
IGH1J01	26196	TGTTTTTGT	23	CACTGTG	26186	39
IGH1J02	128188	AGTGTTTGT	23	CACTGTG	128175	39
IGH1J03	128353	TGTTTATTT	23	CACTGTG	128353	39
IGH1J04	128545	GGTTTTTGT	23	CACTGTG	128532	39
IGH1J05	128899	GGTTTTAGT	23	TACTGTG	128886	39
IGH1J06	129360	TCTTCTTGT	22	TACTTTG	129345	38
IGH1J07	129650	AGTTTTTGT	23	TACTGTG	129634	39
IGH1J08	129983	AGTTTTAGT	22	TACTGTG	129964	38
IGH1J09	130628	CGTTTTTAT	22	CACTGTG	130611	38
IGH2J08	204047	CGTTTTTAT	22	CACTGTG	204030	38
IGH2J07	204691	AGTTTTAGT	22	TACTGTG	204672	38
IGH2J06	205020	AGTTTTTGT	23	TACTGTG	205004	39
IGH2J05	205310	TCTTCTTGT	22	TACTTTG	205295	38
IGH2J04	205768	GGTTTTAGT	23	TACTGTG	205755	39
IGH2J03	206123	GGTTTTTGT	23	CACTGTG	206110	39
IGH2J02	206302	TGTTTATTT	23	CACTGTG	206302	39
IGH2J01	206478	AGTGTTTGT	23	CACTGTG	206465	39

Table E.9: Co-ordinate table of JH RSSs in the *N. furzeri IGH* locus

Table E.10: Co-ordinate table of constant-region exons in the *X. maculatus IGH* locus

Name	Isotype	Start	End	Length	Strand
IGHZ1-1	Z	3380	3667	288	+
IGHZ1-2	Z	3814	4098	285	+
IGHZ1-3	Z	4195	4497	303	+
IGHZ1-4	Z	4934	5263	330	+
IGHZ1-S	Z	5264	5459	196	+
IGHZ1-TM1	Z	6345	6490	146	+
IGHZ1-TM2	Z	6645	7043	399	+
IGHZ2-1	Z	256059	256337	279	+
IGHZ2-2	Z	256453	256734	282	+
IGHZ2-3	Z	256893	257171	279	+
IGHZ2-4	Z	257319	257636	318	+
IGHZ2-S	Z	257637	257850	214	+
IGHZ2-TM1	Z	258059	258213	155	+
IGHZ2-TM2	Z	258410	258629	220	+
IGHM-1	M	279664	279960	297	+
IGHM-2	M	280880	281224	345	+
IGHM-3	M	281321	281629	309	+
IGHM-4	M	281789	282291	503	+
IGHM-TM1	M	282910	283034	125	+
IGHM-TM2	M	285028	285740	713	+
IGHD-1	D	285902	286219	318	+
IGHD-2A	D	286310	286597	288	+
IGHD-3A	D	286814	287128	315	+
IGHD-4A	D	287250	287534	285	+
IGHD-2B	D	288876	289166	291	+
IGHD-3B	D	289262	289576	315	+
IGHD-4B	D	289680	289964	285	+
IGHD-5	D	290052	290381	330	+
IGHD-6	D	290472	290789	318	+
IGHD-7	D	290865	291152	288	+
IGHD-TM1	D	291286	291434	149	+
IGHD-TM2	D	291541	291642	102	+

Name	Start	End	Length	Strand	RSS Start	Heptamer	Spacer Length	Nonamer	RSS End	RSS Length	Comment
IGHV01-01	1159	1450	292	+	1451	CACAGTG	23	GTAAAAACC	1489	39	
IGHV02-01	10534	10825	292	+	10826	CACAGTG	23	ACAAAACCC	10864	39	
IGHV02-02	11961	12261	301	+	12262	CACTGTG	23	ACAAAAACT	12300	39	
IGHV02-03	13319	13616	298	+	13617	CACAGTG	23	ACACAAACT	13655	39	
IGHV03-01	15440	15734	295	+	15735	CACAGTG	22	ACAAAAACT	15772	38	
IGHV02-04	16618	16908	291	+	16909	CACAGTG	23	ACAAAAACC	16947	39	
IGHV02-05	17522	17822	301	+	17823	CACTGTG	22	ACAAAAACT	17860	38	
IGHV02-06	18881	19178	298	+	19179	CACAGTG	23	ACACAAACT	19217	39	
IGHV03-02	21000	21294	295	+	21295	CACAGTG	22	ACAAAAACT	21332	38	
IGHV02-07	22179	22467	289	+	22468	CACAGTG	23	ACAAAAACC	22506	39	
IGHV02-08p	24234	24514	281	+	24515	CACAGTG	23	ACAAAAACT	24553	39	Frameshift
IGHV04-01	25359	25659	301	+	25660	CACAGTG	23	ACAAAAACT	25698	39	
IGHV04-02	27066	27366	301	+	27367	CACAGTG	23	ACAAAAACA	27405	39	
IGHV02-09	28669	28958	290	+	28959	CACAGTG	23	ACAAAAACC	28997	39	
IGHV02-10p	30460	30741	282	+	30742	CACAATG	23	ACAAAACTC	30780	39	Frameshift
IGHV02-11	32395	32681	287	+	32682	CACAGTG	23	ACAAAAACC	32720	39	
IGHV03-03	33663	33957	295	+	33958	CACTGTG	22	ACAAAAACT	33995	38	
IGHV02-12	35012	35299	288	+	35300	CACAGTG	23	ACAAAAACC	35338	39	
IGHV03-04	36281	36575	295	+	36576	CACTGTG	22	ACAAAAACT	36613	38	
IGHV02-13	37639	37931	293	+	37932	CACAGTG	23	ACAAAAACT	37970	39	
IGHV02-14	39019	39311	293	+	39312	CACAGTG	23	ACAAAAACT	39350	39	
IGHV03-05	41008	41302	295	+	41303	CACAGTG	22	ACAAAAACT	41340	38	
IGHV02-15	42660	42952	293	+	42953	CACAGTG	23	ACAAAAACT	42991	39	
IGHV03-06	45081	45375	295	+	45376	CACAGTG	22	ACAAAAACT	45413	38	
IGHV02-16	46732	47024	293	+	47025	CACAGTG	23	ACAAAAACT	47063	39	

Table E.11: Co-ordinate table of VH segments in the *X. maculatus IGH* locus, part 1

Name	Start	End	Length	Strand	RSS Start	Heptamer	Spacer Length	Nonamer	RSS End	RSS Length	Comment
IGHV03-07	48618	48912	295	+	48913	CACAGTG	22	ACAAAAACT	48950	38	
IGHV02-17	50323	50611	289	+	50612	CACAGTG	23	ACAAAAACC	50650	39	
IGHV03-08	51890	52184	295	+	52185	CACAGTG	22	ACAAAAACT	52222	38	
IGHV03-09p	53026	53274	249	+	53275						3'-truncated, no RSS
IGHV02-18	54462	54747	286	+	54748	CACAGTG	23	ACAAAAACC	54786	39	
IGHV02-19p	55729	55866	138	+	55867	CACAGTG	23	ACAAAAACC	55905	39	3'-truncated
IGHV03-10	57371	57662	292	+	57663	CACAGTG	22	ACAAAAACT	57700	38	
IGHV02-20p	58698	58986	289	+	58987	CACAGTG	23	ATAAAAACC	59025	39	Nonsense mutation
IGHV03-11	59940	60234	295	+	60235	CACAGTG	22	ACAAAAACT	60272	38	
IGHV02-21	61249	61537	289	+	61538	CACAGTG	23	ATAAAAACC	61576	39	
IGHV03-12	62491	62785	295	+	62786	CACAGTG	22	ACAAAAACT	62823	38	
IGHV02-22	63801	64089	289	+	64090	CACAGTG	23	ATAAAAACC	64128	39	
IGHV03-13	65043	65337	295	+	65338	CACAGTG	22	ACAAAAACT	65375	38	
IGHV02-23	66354	66640	287	+	66641	CACAGTG	23	ACAAAAACT	66679	39	
IGHV03-14	68452	68743	292	+	68744	CACTATG	22	ACAAAACTC	68781	38	
IGHV02-24	70101	70389	289	+	70390	CACAGTG	23	ACAAAAACC	70428	39	
IGHV03-15	72206	72501	296	+	72502	CACAGTG	22	ACAAAAACT	72539	38	
IGHV02-25	73484	73772	289	+	73773	CACAGTG	23	ACAAAAACC	73811	39	
IGHV03-16	75799	76090	292	+	76091	CACAGTG	22	ACAAAAACT	76128	38	
IGHV03-17	77773	78067	295	+	78068	CACAGTG	22	ACAAAAACT	78105	38	
IGHV02-26	79001	79289	289	+	79290	CACAGTG	23	ACAAAAACC	79328	39	
IGHV03-18	80492	80784	293	+	80785	CACAGTG	22	ACAAAAACT	80822	38	
IGHV02-27p	81799	82082	284	+	82083	CACAGTG	23	ACAAAAACC	82121	39	Frameshift
IGHV03-19	83736	84030	295	+	84031	CACAGTG	22	ACAAAAACT	84068	38	
IGHV02-28p	85093	85381	289	+	85382	CACAGGG	23	GCAAAAACC	85420	39	Nonsense mutation

Table E.12: Co-ordinate table of VH segments in the *X. maculatus IGH* locus, part 2

Name	Start	End	Length	Strand	RSS Start	Heptamer	Spacer Length	Nonamer	RSS End	RSS Length	Comment
IGHV02-29	86225	86505	281	+	86506	CACAGTG	23	ATAAAAACC	86544	39	
IGHV03-20	87419	87713	295	+	87714	CACAGTG	22	ACAAAAACT	87751	38	
IGHV03-21	94532	94826	295	+	94827	CACAGTG	23	ACAAAAACC	94865	39	
IGHV03-22	96192	96489	298	+	96490	CACAGTG	23	ACAAAAACC	96528	39	
IGHV03-23	98068	98368	301	+	98369	CACAGTG	23	ACAAAAACC	98407	39	
IGHV03-24	99482	99779	298	+	99780	CACAGTG	23	ACAAAAACC	99818	39	
IGHV03-25	101639	101936	298	+	101937	CACAGTG	23	ACAAAAACC	101975	39	
IGHV05-01p	102818	103096	279	+	103097	CAGAAGC	0	ACAAAAACT	103112	16	Frameshift
IGHV03-26	104098	104389	292	+	104390	CACAGTG	23	ACAAAATCC	104428	39	
IGHV06-01	105551	105831	281	+	105832	CACAGTG	23	ACAAAAACC	105870	39	
IGHV03-27	107274	107571	298	+	107572	CACAGTG	23	ACAAAAACC	107610	39	
IGHV03-28	108775	109072	298	+	109073	CACAGAG	23	ACAAAAACC	109111	39	
IGHV03-29	110372	110672	301	+	110673	CACAGTG	23	ACAAAAACC	110711	39	
IGHV07-01	111565	111856	292	+	111857	CACAATG	23	ACAAAAACT	111895	39	
IGHV08-01p	113033	113330	298	+	113331	CACAGAG	23	CCAAGAACC	113369	39	Nonsense mutation
IGHV09-01	115512	115800	289	+	115801	CACAGTG	22	ACAAAAACT	115838	38	
IGHV10-01	117078	117379	302	+	117380	CACAGTG	22	ACATAAACT	117417	38	
IGHV11-01	119462	119760	299	+	119761	CACAGTG	23	ACAAAAACT	119799	39	
IGHV03-30	126125	126416	292	+	126417	CACAGTG	22	ACAAAAACC	126454	38	
IGHV03-31	127109	127400	292	+	127401	CACAGTG	23	GCAAAAACC	127439	39	
IGHV12-01	128489	128786	298	+	128787	CACAGTG	23	ACAAAAACC	128825	39	
IGHV02-30	135711	136000	290	+	136001	CACAGTG	22	ACAAAAACA	136038	38	
IGHV13-01	136757	137057	301	+	137058	CACAGTG	23	ACAAAAACT	137096	39	
IGHV02-31	138344	138637	294	+	138638	CACAGTG	23	ACAAAAATC	138676	39	
IGHV02-32	140024	140315	292	+	140316	CACTGTG	23	ACAAAAACT	140354	39	

Table E.13: Co-ordinate table of VH segments in the *X. maculatus IGH* locus, part 3

Name	Start	End	Length	Strand	RSS Start	Heptamer	Spacer Length	Nonamer	RSS End	RSS Length	Comment
IGHV02-33	142332	142620	289	+	142621	CACAGTG	23	ACAAAAACA	142659	39	
IGHV02-34	144334	144625	292	+	144626	CACAGTG	23	ACAAAAACT	144664	39	
IGHV02-35	145740	146031	292	+	146032	CACAGTG	23	ACAAAAAAT	146070	39	
IGHV02-36	146903	147194	292	+	147195	CACAGTG	23	ACAAAAACT	147233	39	
IGHV02-37	147839	148138	300	+	148139	CACAGTG	23	ACAAAAATC	148177	39	
IGHV02-38p	150504	150797	294	+	150798	CACAATA	23	ACAAAAACC	150836	39	Nonsense mutation
IGHV02-39	152249	152537	289	+	152538	CACAGTA	23	ACAAAAACC	152576	39	
IGHV14-01	154075	154374	300	+	154375	CACAGTG	23	ACAAAAAGT	154413	39	
IGHV02-40	155433	155709	277	+	155710	CACAGTG	23	ACAAAAACC	155748	39	
IGHV02-41	156583	156870	288	+	156871	CACAGTG	23	ACAAAAACC	156909	39	
IGHV02-42	163977	164269	293	+	164270	CACAGTG	23	ACAAAACCC	164308	39	
IGHV03-32	165416	165708	293	+	165709	CACAGTG	22	ACAAAAACA	165746	38	
IGHV02-43	166994	167293	300	+	167294	CACAATG	23	ACAGAAACT	167332	39	
IGHV12-02	169602	169900	299	+	169901	CACAGTG	23	ACAAAAACC	169939	39	
IGHV02-44	171452	171752	301	+	171753	CACTGTG	23	GCAAAAACT	171791	39	
IGHV02-45	173096	173384	289	+	173385	CTCAGTG	23	ACAAAAACC	173423	39	
IGHV02-46	174714	175009	296	+	175010	CACAGTG	23	ACAAAAACT	175048	39	
IGHV02-47	176396	176697	302	+	176698	CACAGTG	23	ACAAAAACT	176736	39	
IGHV12-03	178422	178719	298	+	178720	CACAGTG	23	ACAAAAACA	178758	39	
IGHV12-04	181245	181543	299	+	181544	CACAGTG	23	ACAAAAACC	181582	39	
IGHV02-48p	182977	183236	260	+	183237	CACAGGT	8	ACAAAAACT	183260	24	5'-truncated
IGHV02-49p	184323	184611	289	+	184612	CACAGTG	23	ACAAAAACC	184650	39	Nonsense mutation
IGHV02-50	185946	186244	299	+	186245	CACAGTG	23	ACAAAAACT	186283	39	
IGHV02-51	187624	187925	302	+	187926	CACAGTG	23	ACAAAAACT	187964	39	
IGHV12-05	190987	191284	298	+	191285	CACAGTG	23	ACAAAAACA	191323	39	

Table E.14: Co-ordinate table of VH segments in the *X. maculatus IGH* locus, part 4

Name	Start	End	Length	Strand	RSS Start	Heptamer	Spacer Length	Nonamer	RSS End	RSS Length	Comment
IGHV02-52	192570	192868	299	+	192869	CACAGTG	19	CTGAAAACC	192903	35	
IGHV12-06	193608	193906	299	+	193907	CACAGTG	23	ACAAAAACA	193945	39	
IGHV02-53	195271	195572	302	+	195573	CACAGTG	23	ACAAAAACC	195611	39	
IGHV15-01	204396	204693	298	+	204694	CACAATC	23	ACAAAAACT	204732	39	
IGHV13-02	206203	206503	301	+	206504	CACAGTG	23	ACAAAAACT	206542	39	
IGHV16-01	207726	208020	295	+	208021	CACAGTG	22	ACAAAAACT	208058	38	
IGHV13-03	208477	208777	301	+	208778	CACAGTA	23	ACAAAAACT	208816	39	
IGHV03-33	209921	210215	295	+	210216	CACGGTG	22	ACGAAAACT	210253	38	
IGHV17-01	211322	211625	304	+	211626	CACAGTA	23	ACAAAAACC	211664	39	
IGHV15-02p	214600	214860	261	+	214861						3’-truncated, no RSS
IGHV18-01	215671	215962	292	+	215963	CACACTG	23	ACAAAAACC	216001	39	
IGHV19-01	217874	218174	301	+	218175	CACAGTG	23	ACAAAAACT	218213	39	
IGHV03-34	219368	219668	301	+	219669	CACAGTG	23	ACAAAAACA	219707	39	
IGHV20-01	220329	220632	304	+	220633	CACAGTG	23	ACAAAAATT	220671	39	
IGHV02-54p	228547	228838	292	+	228839	CACACTG	23	ACAACCCCC	228877	39	Nonsense mutation
IGHV02-55	229963	230267	305	+	230268	CACAGCG	23	ACAAAAAAA	230306	39	
IGHV03-35	231630	231928	299	+	231929	CACAGTG	23	ACAAAAACC	231967	39	
IGHV21-01p	233069	233230	162	+	233231						Nonsense mutation, 3’-truncated, no RSS
IGHV22-01p	234954	235102	149	+	235103	CACAGTG	23	TCAAAAACT	235141	39	5’-truncated
IGHV02-56	236029	236330	302	+	236331	CACAGTG	23	ACAAATACT	236369	39	
IGHV03-36p	238122	238413	292	+	238414	CACAATG	23	ACAGAATCC	238452	39	Nonsense mutation
IGHV11-02p	240281	240579	299	+	240580	CACAGTG	24	ACAAAAACT	240619	40	Nonsense mutation
IGHV09-02	241878	242166	289	+	242167	CACAGTG	22	ACAAAAACT	242204	38	
IGHV23-01	243867	244164	298	+	244165	CACAGTG	23	ACAAAATCC	244203	39	
IGHV02-57	245524	245813	290	+	245814	CACCATA	22	ACAAAATCC	245851	38	

Table E.15: Co-ordinate table of VH segments in the *X. maculatus IGH* locus, part 5

Table E.16: Co-ordinate table of DH segments in the *X. maculatus IGH* locus

Name	Start	NT Sequence	End	Length	Strand
IGHDZ01	2243	GTGGGCAGGAGGCTATGC	2260	18	+
IGHDZ02	119768	AGG	119770	3	+
IGHDZ03	128794	ACTAAAGG	128801	8	+
IGHDZ04	129907	ATCGGG	129912	6	+
IGHDZ05	158017	ATATATGGGGG	158027	11	+
IGHDZ06	197791	ATATACTGGGGTGG	197804	14	+
IGHDZ07	222022	ATGGACTGGGGGG	222034	13	+
IGHDZ08	247941	GTGATTACGGCTACGGGGC	247959	19	+
IGHDZ09	249514	TTATGGGCTGGGGAG	249528	15	+
IGHDZ10	253752	TGGGTGGGGC	253761	10	+
IGHDM01	267392	TATACAGTGGCAAC	267405	14	+
IGHDM02	268498	CAGTATAGCAAC	268509	12	+
IGHDM03	268836	TACAATGGCAAC	268847	12	+
IGHDM04	269694	TAAACAGTGGCTAC	269707	14	+

Table E.17: Co-ordinate table of DH 5'-RSSs in the *X. maculatus IGH* locus

Name	5'-RSS Start	Nonamer	Spacer Length	Heptamer	5'-RSS End	Length
IGHDZ01	2215	GGTTTTTGT	12	CACTGTG	2242	28
IGHDZ02	119739	TGTATTACT	13	CACAGTG	119767	29
IGHDZ03	128766	TTTACTTCT	12	CACAGTG	128793	28
IGHDZ04	129879	GGTTTTTGT	12	CACAGTG	129906	28
IGHDZ05	157989	AGTTTTTGT	12	CACAGTG	158016	28
IGHDZ06	197763	GGTTTTTGC	12	TACTGTG	197790	28
IGHDZ07	221994	GGTTTTTGT	12	CGCTGTG	222021	28
IGHDZ08	247913	TGTTTTTGT	12	ATCTGTG	247940	28
IGHDZ09	249486	AGTTTTTGT	12	TGTGGTG	249513	28
IGHDZ10	253724	AGTTTTTGT	12	TGTAGTG	253751	28
IGHDM01	267364	AGTTTTTGT	12	TACAGTG	267391	28
IGHDM02	268470	TGTTTTTGT	12	CACAGTG	268497	28
IGHDM03	268808	AGTTTTTGC	12	TACTGTG	268835	28
IGHDM04	269666	CGTTTTTGT	12	CATTGTG	269693	28

Table E.18: Co-ordinate table of DH 3'-RSSs in the *X. maculatus IGH* locus

Name	3'-RSS Start	Heptamer	Spacer Length	Nonamer	3'-RSS End	Length
IGHDZ01	2261	CACTAAG	12	ACAAAAAGT	2288	28
IGHDZ02	119771	CAAAATG	13	ACAAAAACT	119799	29
IGHDZ03	128802	CAGAGAA	8	ACAAAAACC	128825	24
IGHDZ04	129913	CACAATG	12	TCAAAAACC	129940	28
IGHDZ05	158028	CACAGAG	12	ACAAAAACC	158055	28
IGHDZ06	197805	CACACAG	12	ACAAAAACC	197832	28
IGHDZ07	222035	CACAGAG	12	ACAAAAACC	222062	28
IGHDZ08	247960	CACAATA	12	ACAAAAACC	247987	28
IGHDZ09	249529	CACAATG	12	ACAAAAACC	249556	28
IGHDZ10	253762	CACAGTA	12	ACAAAAACC	253789	28
IGHDM01	267406	CACAGTG	12	GCAAAAACC	267433	28
IGHDM02	268510	CACAGTG	12	ACAGAAACC	268537	28
IGHDM03	268848	CACAGTG	12	ACAAAAACC	268875	28
IGHDM04	269708	CACTGTG	12	ACAAAATCA	269735	28

Name	Start	NT Sequence	AA Sequence	End	Length	Strand
IGHJZ01	2653	ATGCCTTAGATTACTGGGGTGAAGGGACCAGAGTCACAGTGACTTCAG	ALDYWGEGTRVTVTS	2700	48	+
IGHJZ02	120639	ATTACGCTCTTGACTACTGGGGAGCAGGAACCAAAGTTACTGTAAAGCCAG	YALDYWGAGTKVTVKP	120689	51	+
IGHJZ03	130376	ACTACGGCTTTGATTACTGGGGAGACGGAACTGAAGTTACTGTTGAACCAG	YGFDYWGDGTEVTVEP	130426	51	+
IGHJZ04	158408	AGATTTAGACTACTGGGGTAATGGAACAACAGTCACGGTTCTACCAG	DLDYWGNGTTVTVLP	158454	47	+
IGHJZ05	198186	ATTATGGTTTTGACTACTGGGGAGACGGAACCACAGTCACTGTTAGTCCAG	YGFDYWGDGTTVTVSP	198236	51	+
IGHJZ06	222417	ATGCTTTTGACGTCTGGGGTAAAGGAACCACAGTTACTGTTGTACCAG	AFDVWGKGTTVTVVP	222464	48	+
IGHJZ07	254130	ATGTTTTTGACTACTGGGGTAAAGGGACTGATGTCACAGTATCTCCAG	VFDYWGKGTDVTVSP	254177	48	+
IGHJM01	276014	ACGGCTACTTCGACTACTGGGGGAAAGGAACACAAGTCACAGTGACTTCTG	GYFDYWGKGTQVTVTS	276064	51	+
IGHJM02	276284	CCACTACTTTGACTACTGGGGAAAAGGAACCACGGTTACCGTCACTTCAG	HYFDYWGKGTTVTVTS	276333	50	+
IGHJM03	276654	ACAATGCTTTTGACTACTGGGGAAAAGGAACTACGGTAACAGTAACATCAG	NAFDYWGKGTTVTVTS	276704	51	+
IGHJM04	276999	ACTACGCTTTTGACTACTGGGGAAAAGGAACAATGGTCACTGTCACTTCAG	YAFDYWGKGTMVTVTS	277049	51	+
IGHJM05	277322	ACAACTGGGCTTTTGACTACTGGGGAGCAGGAACCATGGTAACAGTAACATCAG	NWAFDYWGAGTMVTVTS	277375	54	+
IGHJM06	277672	CTACGGTGCTTTTGACTACTGGGGTAAAGGGACTACAGTCACCGTCACTTCAG	YGAFDYWGKGTTVTVTS	277724	53	+
IGHJM07	278150	CTACGATGCTTTTGACTATTGGGGGAAAGGAACAACAGTCACCGTCATCACTTCAG	YDAFDYWGKGTTVTVITS	278205	56	+
IGHJM08	278606	TTACTACTACGCTTTTGACTATTGGGGAAAAGGGACAATGGTCACCGTCACTTCAG	YYYAFDYWGKGTMVTVTS	278661	56	+

Table E.19: Co-ordinate table of JH segments in the *X. maculatus IGH* locus

Name	RSS Start	Nonamer	Spacer Length	Heptamer	RSS End	RSS Length
IGHJZ01	2662	TGTTTTTGT	23	CACTGTG	2652	39
IGHJZ02	120651	TGTTTTTGT	23	CACTGTG	120638	39
IGHJZ03	130388	TGTTTTTGT	23	CACCGTG	130375	39
IGHJZ04	158416	GGTTTTTGT	23	CACTGTG	158407	39
IGHJZ05	198198	GGTTTTTGT	23	CACTGTG	198185	39
IGHJZ06	222426	TGTTTTTGT	23	CACTGTG	222416	39
IGHJZ07	254139	GGTTTTTGT	23	CACTGTG	254129	39
IGHJM01	276026	TGTATTTGT	23	CACTGTG	276013	39
IGHJM02	276295	TATTTTTGC	23	CACCGTG	276283	39
IGHJM03	276666	TGTTTTTGT	23	TACTGTG	276653	39
IGHJM04	277011	TGTTTTAGT	23	TACTGTG	276998	39
IGHJM05	277338	GGTTTTTGT	22	TACTGTG	277321	38
IGHJM06	277687	GCTTTTTAT	22	CACTGTG	277671	38
IGHJM07	278168	CCTTTTTAC	22	CACTGTG	278149	38
IGHJM08	278624	GCTTTTTAA	22	CACTGTG	278605	38

Table E.20: Co-ordinate table of JH RSSs in the *X. maculatus IGH* locus

Species	Scaffold(s)	Region	Isotype	Known Exons [1]	Complete?	Pseudo-exons	Comments
Nothobranchius orthonotus	scf33878	IGHM1	M	1,2,3,TM1	**No**	–	CM4 missing (missing sequence)
Nothobranchius orthonotus	scf33878	IGHD1	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Nothobranchius orthonotus	scf34438	IGHM2	M	1,2,3,4,TM1	Yes	–	
Nothobranchius orthonotus	scf34438, scf33917	IGHD2	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Nothobranchius orthonotus	scf33917	IGHD3	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Nothobranchius orthonotus	scf33917	IGHD4	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Nothobranchius orthonotus	scf9255, scf26119, scf33917	IGHD5	D	3,4,2,3,4,5,6,7,TM1	**No**	–	CD1 & CD2A missing (missing sequence)
Nothobranchius orthonotus	scf27951, scf33789	IGHM3	M	1,2,3,4,TM1	Yes	–	
Nothobranchius orthonotus	scf27951, 32033	IGHD6	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Nothobranchius orthonotus	scf32137, scf21286	IGHM4	M	1,2,3,4,TM1	Yes	–	
Nothobranchius furzeri	chr6 [2]	IGH1M	M	1,2,3,4,TM1	Yes	–	
Nothobranchius furzeri	chr6 [2]	IGH1D	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Nothobranchius furzeri	chr6 [2]	IGH2M	M	1,2,3,4,TM1	Yes	–	
Nothobranchius furzeri	chr6 [2]	IGH2D	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Aphyosemion australe	scf373	IGHM	M	1,2,3,4,TM1	Yes	–	
Aphyosemion australe	scf373	IGHD	D	1,2,3,4,5,6,7,TM1	Yes	–	
Callopanchax toddi	scf107	IGHZ1	Z	1,2,3,4,TM1	Yes	–	
Callopanchax toddi	scf107	IGHZ2	Z	1,2,3,4,TM1	Yes	–	
Callopanchax toddi	scf1209	IGHZ3	Z	1,2,3,4,TM1	Yes	–	
Callopanchax toddi	scf1209	IGHM1	M	1	**No**	–	Isolated CM1 exon
Callopanchax toddi	scf945	IGHZ4	Z	1,2,3,4,TM1	Yes	–	
Callopanchax toddi	scf945	IGHM2	M	1,2,3,4,TM1	Yes	–	
Callopanchax toddi	scf945	IGHD1	D	1,2,3,4,5,6,7,TM1	Yes	1,4,5	Frameshift mutations in CD1, CD4 & CD5
Callopanchax toddi	scf265	IGHM3	M	1,2,3,4,TM1	Yes	–	
Callopanchax toddi	scf265	IGHD2	D	1,5,7,TM1	**No**	–	CD2-4 & CD5-6 missing (not in sequence)

[1] Excluding TM2 and secretory exons.

[2] Expanded *IGH* locus sequence from Section 3.2.

Table E.21: *IGH* constant regions in cyprinidontiform fish, part 1

Species	Scaffold(s)	Region	Isotype	Known Exons [1]	Complete?	Pseudo-exons	Comments
Pachypanchax playfairii	scf547	IGHZ	Z	1,2,3,4,TM1	Yes	–	
Pachypanchax playfairii	scf125	IGHM1	M	1,2,3,4,TM1	Yes	–	
Pachypanchax playfairii	scf125	IGHD	D	1,2,3,4,5,6,7,TM1	Yes	–	
Pachypanchax playfairii	scf547	IGHM2	M	1	**No**	–	Isolated CM1 exon
Austrofundulus limnaeus	NW_013954375.1	IGHZ	Z	TM1	**No**	TM1	Isolated TM1 exon with frameshift mutation
Austrofundulus limnaeus	NW_013952673.1	IGHM	M	1,2,3,4,TM1	Yes	–	
Austrofundulus limnaeus	NW_013952673.1, NW_013956335.1	IGHD	D	1,2,3,4,5,6,7,TM1	Yes	–	
Kryptolebias marmoratus	NW_016094348.1	IGHZ1	Z	1,2,3,4,TM1	Yes	–	
Kryptolebias marmoratus	NW_016094348.1	IGHZ2	Z	1,4,TM1	**No**	–	CZ2 & CZ3 missing (not in sequence)
Kryptolebias marmoratus	NW_016094301.1	IGHM1	M	1,2,3,4,TM1	Yes	–	
Kryptolebias marmoratus	NW_016094301.1	IGHD1	D	1,2,3,4,5,6,7,TM1	Yes	–	
Kryptolebias marmoratus	NW_016094277.1	IGHM2	M	1,2,3,4,TM1	Yes	–	
Kryptolebias marmoratus	NW_016094277.1	IGHD2	D	1,2,3,4,5,6,TM1	**No**	–	CD7 missing (not in sequence)
Poecilia reticulata	NC_024338.1	IGHZ1	Z	1,2,3,4	**No**	–	TM1 missing (missing sequence)
Poecilia reticulata	NC_024338.1	IGHZ2	Z	1,2,3,4,TM1	Yes	–	
Poecilia reticulata	NC_024338.1	IGHM	M	1,2,3,4,TM1	Yes	–	
Poecilia reticulata	NC_024338.1	IGHD	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Poecilia formosa	NW_006800081.1	IGHZ1	Z	1,2,3,4,TM1	Yes	–	
Poecilia formosa	NW_006800081.1	IGHZ2	Z	1,2,3,4,TM1	Yes	–	
Poecilia formosa	NW_006800081.1	IGHZ3	Z	1,2,3,4,TM1	Yes	–	
Poecilia formosa	NW_006800081.1	IGHM	M	1,2,3,4,TM1	Yes	–	
Poecilia formosa	NW_006800081.1	IGHD	D	1,2,3,4,5,6,7,TM1	Yes	–	
Xiphophorus maculatus	NC_036458	IGHZ1	Z	1,2,3,4,TM1	Yes	–	
Xiphophorus maculatus	NC_036458	IGHZ2	Z	1,2,3,4,TM1	Yes	–	
Xiphophorus maculatus	NC_036458	IGHM	M	1,2,3,4,TM1	Yes	–	

[1] Excluding TM2 and secretory exons.

Table E.22: *IGH* constant regions in cyprinidontiform fish, part 2

Species	Scaffold(s)	Region	Isotype	Known Exons [1]	Complete?	Pseudo-exons	Comments
Xiphophorus maculatus	NC_036458	IGHD	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Fundulus heteroclitus	NW_012234561.1	IGHZ1	Z	1,2,3,4,TM1	Yes	–	
Fundulus heteroclitus	NW_012230737.1	IGHZ2	Z	4,TM1	**No**	–	CZ1 to CZ3 missing (missing sequence)
Fundulus heteroclitus	NW_012234542.1	IGHM	M	1,2,3,4,TM1	Yes	–	
Fundulus heteroclitus	NW_012234542.1	IGHD	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Cyprinodon variegatus	NW_015154250.1, NW_015151047.1	IGHZ	Z	1,2,3,4,TM1	Yes	–	
Cyprinodon variegatus	NW_015151047.1	IGHM	M	1,2,3,4,TM1	Yes	–	
Cyprinodon variegatus	NW_015151047.1	IGHD	D	1,2,3,4,2,3,4,5,6,7,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHM1	M	1,2,3,4,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHD1	D	1,2,3,4,6,7,TM1	Yes	7	Nonsense mutation in CD7
Oryzias latipes	NC_019866.2	IGHM2	M	1,2,3,4,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHD2	D	1,2,3,4,6,7,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHM3	M	1,2,3,4,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHD3	D	1,2,3,4,6,7,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHM4	M	1,2,3,4,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHD4	D	2,7,TM1	**No**	–	CD1 & CD3-6 missing (not in sequence)
Oryzias latipes	NC_019866.2	IGHM5	M	1,2,3,4,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHD5	D	1,2,3,4,6,7,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHM6	M	1,2,3,4,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHD6	D	1,2,3,4,6,7,TM1	Yes	–	
Oryzias latipes	NC_019866.2	IGHD7	D	1,2,3,6	**No**	–	CD4, CD5, CD7 and TM1 missing (not in sequence)

[1] Excluding TM2 and secretory exons.

Table E.23: *IGH* constant regions in cyprinidontiform fish, part 3

Table E.24: Turquoise killifish used in IgSeq pilot and ageing experiments. All fish are GRZ-AD strain and male.

Group	#	Fish ID[1]	Death weight (g)	Hatch date	Sacrifice date	Age (days)	Age (weeks)
1	1	4194	1.24	2016-05-09	2016-06-17	39	5.57
1	2	4107	1.39	2016-05-09	2016-06-17	39	5.57
1	3	4127	1.29	2016-05-09	2016-06-17	39	5.57
1	4	4204	1.35	2016-05-09	2016-06-17	39	5.57
1	5	4189	1.43	2016-05-09	2016-06-17	39	5.57
1	6	4160	0.68	2016-05-09	2016-06-17	39	5.57
1	7	4164	1.57	2016-05-09	2016-06-17	39	5.57
1	8	4171	1.40	2016-05-09	2016-06-17	39	5.57
1	9	4200	1.42	2016-05-09	2016-06-17	39	5.57
1	10	4131	1.27	2016-05-09	2016-06-17	39	5.57
2	1	4159	1.37	2016-05-09	2016-07-04	56	8.00
2	2	4179	1.47	2016-05-09	2016-07-04	56	8.00
2	3	4152	1.33	2016-05-09	2016-07-04	56	8.00
2	4	4132	1.35	2016-05-09	2016-07-04	56	8.00
2	5	4177	1.22	2016-05-09	2016-07-04	56	8.00
2	6	4158	1.51	2016-05-09	2016-07-04	56	8.00
2	7	4182	1.12	2016-05-09	2016-07-04	56	8.00
2	8	4202	1.54	2016-05-09	2016-07-04	56	8.00
2	9	4143	1.28	2016-05-09	2016-07-04	56	8.00
2	10	4201	1.55	2016-05-09	2016-07-04	56	8.00
3	1	4155	2.06	2016-05-09	2016-07-21	73	10.43
3	2	4193	1.92	2016-05-09	2016-07-21	73	10.43
3	3	4170	1.80	2016-05-09	2016-07-21	73	10.43
3	4	4135	1.65	2016-05-09	2016-07-21	73	10.43
3	5	4190	1.87	2016-05-09	2016-07-21	73	10.43
3	6	4099	1.94	2016-05-09	2016-07-21	73	10.43
3	7	4198	1.49	2016-05-09	2016-07-21	73	10.43
3	8	4024	1.73	2016-05-09	2016-07-21	73	10.43
3	9	4044	1.53	2016-05-09	2016-07-21	73	10.43
3	10	4117	1.57	2016-05-09	2016-07-21	73	10.43
4	1	4173	2.20	2016-05-09	2016-09-14	128	18.29
4	2	4197	2.40	2016-05-09	2016-09-14	128	18.29

[1] grz-AD_..._E

Fish ID	Age at death (weeks)	Treatment group	RNA Integrity Number (RIN)	Date of library prep	Sequenced?	Reason for exclusion
1271	16	ABX_16	7.7	2018-11-20	Yes	–
1274	16	ABX_16	7.3	2018-12-01	Yes	–
1309	16	ABX_16	6.9	2018-12-01	Yes	–
dash	16	ABX_16	6.8	2018-12-01	Yes	–
1015	16	SMT_16	7.0	2018-11-20	Yes	–
1298	16	SMT_16	5.5	2018-11-20	Yes	–
1301	16	SMT_16	5.9	2018-11-20	Yes	–
402	16	WT_16	7.0	2018-11-20	Yes	–
938	16	WT_16	7.8	2018-11-20	Yes	–
940	16	WT_16	7.2	2018-11-20	Yes	–
1403	6	YI_6	5.5	2018-11-20	Yes	–
1409	6	YI_6	6.4	2018-12-01	Yes	–
1412	6	YI_6	6.7	2018-11-20	Yes	–
1414	6	YI_6	5.5	2018-11-20	Yes	–
1009	16	YMT_16	7.0	2018-11-20	Yes	–
1026	16	YMT_16	6.3	2018-11-20	Yes	–
1305	16	YMT_16	7.3	2018-11-20	Yes	–
999	16	YMT_16	7.1	2018-12-01	Yes	–
400	16	WT_16	7.1	–	No	Not enough RNA for library prep
1005	16	SMT_16	4.4	–	No	RNA integrity too low

Table E.25: Turquoise killifish used in IgSeq gut experiment. All fish are GRZ-Bellemans strain and male.

Erklärung zur Dissertation

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt habe, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen – die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universtität zur Prüfung vorgelegen hat; dass sie – abgesehen von unten angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde.

Die Bestimungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Dr. Dario Valenzano betreut worden.

Nachfolgend genannte Teilpublikationen liegen vor:

1. Bradshaw, W. J. and Valenzano, D. R. Extreme genomic volatility characterises the evolution of the immunoglobulin heavy chain locus in teleost fishes. *bioRxiv*, 752063 (2019). [208]

William John Bradshaw
10. March 2020

Lebenslauf

Personal details

Full name:	William John Bradshaw
Date of birth:	29th September 1991
Place of birth:	Hastings, England, UK
Sex:	Male
Address:	Luxemburger Straße 296, 50937 Cologne
E-mail:	wjbradshaw1@gmail.com
Tel:	+49 160 90505608
Citizenship:	British

Academic record

2006 to 2008	General Certificate of Secondary Education (9 × A*, 1 × A)
	The Judd School, Tonbridge, England, UK
2008 to 2010	A-levels (4 × A*)
	The Judd School, Tonbridge, England, UK
2010 to 2013	BA (Hons) in Natural Sciences (Class I)
	University of Cambridge, England, UK
2013 to 2014	MPhil in Computational Biology (Pass with Distinction)
	University of Cambridge, England, UK
2014 to 2019	*Dr. rer. nat.* in Genetics (ongoing)
	University of Cologne, Germany