Dynamics of the T cell receptor repertoire of a small vertebrate in response to infection

Dissertation
in fulfillment of the requirements for the degree
Doctor rerum naturalium
of the Faculty of Mathematics and Natural Sciences
at the Christian Albrechts University of Kiel

Submitted by
Artemis Efstratiou

Max Planck Institute for Evolutionary Biology
Germany
2024

First examiner:	Prof. Dr. Tobias L. Lenz
Second examiner:	Prof. Dr. Olivia Roth
Date of oral examination:	17.12.2024

Summary

The adaptive immune system emerged 500 million years ago together with the first vertebrates, and stands as one of the most intricate biological systems known to humans. It is based on a dynamic and diverse collection of novel cell types called lymphocytes, each of which expresses on its surface an antigen recognition receptor of unique specificity. T lymphocytes -or T cells-, in particular, lie at the epicenter of the adaptive immunity; they perform immune surveillance by means of engagement with peptides presented by Major Histocompatibility Complex (MHC) molecules of other host cells. This interaction is mediated by the T cell receptor (TCR), which upon binding a foreign-derived antigen, prompts the activation of the T cell. Subsequent clonal expansion of the antigen-triggered T cell marks the initiation of all downstream cellular and humoral immune responses. Thus, the ability of a vertebrate host to recognize and successfully combat a potential pathogen is linked to both the peptide-presenting capacity of its germline-encoded MHC molecules, as well as the diversity of its available repertoire of somatically rearranged TCRs.

Over the past decade, an avalanche of in-depth TCR repertoire studies, fueled by the unparalleled resolution of Next Generation Sequencing (NGS) technologies, has helped uncover many of the mechanisms determining the TCR repertoire and its diversity. The way the TCR repertoire responds to infection, specifically, has been a topic of intensive research, with the majority of our current understanding stemming from human and murine empirical data. In this thesis, I employ the three-spined stickleback (*Gasterosteus aculeatus*) as an experimental model to further advance our knowledge regarding the factors that shape the TCR repertoire during infection. As a teleost, the adaptive immune system of the stickleback bears remarkable similarities to that of mammals, including an extensive set of germline TCR genes shown to contribute to a diverse pre-immune repertoire. Furthermore, this species offers the critical advantage of encompassing a substantial level of natural genetic variation – a crucial component missing from previous murine studies.

First, I focus my attention at the MHC genes of the three-spined stickleback. In **Chapter 1**, my co-authors and I sought to develop an accurate and reproducible NGS-based protocol that can be employed for genotyping multi-locus gene families, such as the MHC, across different species. Thus far, efforts to reliably characterize the individual MHC diversity in non-model species have been obstructed by the inability to target each MHC locus individually, leading to complexities pertaining to the differentiation between genuine alleles and artifacts during data analysis. Here, we performed systematic comparison of three popular NGS genotyping pipelines by using MHC sequencing data derived from stickleback genomic DNA and cDNA samples, independently genotyped by an established non-NGS method. Based on an exhaustive breakdown of genotyping errors, we underscore the necessity for template-specific optimization of adjustable pipeline parameters. Furthermore, we show that optimized genotyping of paired gDNA and cDNA samples allows for accurate expression analysis of individual alleles and haplotypes, which will assist researchers in future efforts to reliably genotype MHC, and other multigene families, using cDNA samples.

Expression of a highly diverse set of MHC alleles in the three-spined stickleback, as confirmed in **Chapter 1**, is strongly indicative of a fully functional adaptive immune arm. I thus turn my attention to the TCR repertoire of the three-spined stickleback. A complex network of genetic and environmental factors are known to affect the TCR repertoire, prominent among which is the host's immunological status. Previous infection studies have primarily examined the TCR repertoire's response to antigenically simple pathogens, such as viruses. To build upon these studies, **Chapter 2** of this thesis is dedicated to thoroughly investigating

the effect of a controlled macroparasitic infection on the TCR repertoire. Towards this end, I used a stickleback-specific helminth, *Schistocephalus solidus*, widely presumed to exert immunomodulatory effects on the host's immune system. Importantly, the use of lab-reared, MHC-genotyped sibships allowed us to simultaneously account for genetic variation at the individual level. In doing so, I illuminate the significant extend to which genetic background beyond the MHC can dictate TCR repertoire diversity, a factor largely unacknowledged in past studies. I postulate upon this experimental finding by discussing the possible genetic and epigenetic mechanisms partaking in shaping the TCR repertoire. Moreover, infection with a fast-growing *S. solidus* strain lead to what I presume is parasite-induced suppression of the private TCR repertoire compartment. Our results also revealed a set of MHC-unrestricted, infection-non-specific TCR sequences of distinct properties, strongly indicative of specialized biological functions. I interpret this as indirect evidence of the presence of unconventional T cells in the stickleback, and consider the implication regarding their role in the evolution of the vertebrate adaptive immune system.

In **Chapter 3**, the last of this thesis, I characterize the TCR repertoire of the stickleback under semi-natural conditions, achieved via exposure of lab-bred sibships to their natal lake environment. By emulating its natural living conditions, our goal was to better capture the stickleback's immunological status in the wild, which often involves co-infection with multiple macro-parasite species. Such natural context is often overlooked in laboratory-based single infection studies and can bias conclusions regarding the TCR repertoire dynamics. Here, eight weeks of lake exposure facilitated high individual macro-parasitic burdens, which coincided with a robust activation of the TCR repertoire via a substantial increase of clonally expanded TCR clonotypes. I further confirm the role of individual genetic variation in the TCR repertoire, in agreement with results from **Chapter 2**. Moreover, I detect strong sex-based differences in the diversity of the TCR repertoire in sticklebacks, whose sexual maturation was prompted by our experimental setup. Intriguingly, comparison of TCR repertoires among control sticklebacks of common genetic background differing in age revealed strikingly similar changes to those described in mammals, a result that points towards the existence of common foundational principles governing the TCR repertoire structure throughout a vertebrate's lifetime.

Zusammenfassung

Das adaptive Immunsystem entstand vor 500 Millionen Jahren zusammen mit den ersten Wirbeltieren und ist eines der kompliziertesten biologischen Systeme, die der Mensch kennt. Es basiert auf einer dynamischen und vielfältigen Ansammlung neuartiger Zelltypen, den Lymphozyten, von denen jeder auf seiner Oberfläche einen Antigenerkennungsrezeptor mit einzigartiger Spezifität aufweist. Insbesondere die T-Lymphozyten - oder T-Zellen - stehen im Zentrum der adaptiven Immunität; sie führen die Immunüberwachung durch, indem sie mit Peptiden interagieren, die von Molekülen des Major Histocompatibility Complex (MHC) anderer Wirtszellen präsentiert werden. Diese Interaktion wird durch den T-Zell-Rezeptor (TCR) vermittelt, der nach Bindung eines fremden Antigens die Aktivierung der T-Zelle auslöst. Die anschließende klonale Expansion der durch das Antigen ausgelösten T-Zelle markiert den Beginn aller nachgeschalteten zellulären und humoralen Immunantworten. Die Fähigkeit eines Wirbeltierwirts, einen potenziellen Krankheitserreger zu erkennen und erfolgreich zu bekämpfen, hängt also sowohl von der Peptidpräsentationskapazität seiner keimbahnkodierten MHC-Moleküle als auch von der Vielfalt seines verfügbaren Repertoires an somatisch umstrukturierten TCRs ab.

In den letzten zehn Jahren hat eine Lawine eingehender Studien des TCR-Repertoires, angetrieben durch die beispiellose Auflösung der Next Generation Sequencing (NGS)-Technologien, dazu beigetragen, viele der Mechanismen aufzudecken, die das TCR-Repertoire und seine Vielfalt bestimmen. Insbesondere die Art und Weise, wie das TCR-Repertoire auf eine Infektion reagiert, ist ein Thema intensiver Forschung, wobei der Großteil unseres derzeitigen Verständnisses aus empirischen Daten von Menschen und Mäusen stammt. In dieser Arbeit verwende ich den Dreistachligen Stichling (*Gasterosteus aculeatus*) als experimentelles Modell, um unser Wissen über die Faktoren, die das TCR-Repertoire während einer Infektion formen, zu erweitern. Als Teleost weist das adaptive Immunsystem des Stichlings bemerkenswerte Ähnlichkeiten mit dem von Säugetieren auf, einschließlich eines umfangreichen Satzes von Keimbahn-TCR-Genen, die zu einem vielfältigen präimmunen Repertoire beitragen. Darüber hinaus bietet diese Spezies den entscheidenden Vorteil, dass sie ein erhebliches Maß an natürlicher genetischer Variation aufweist - eine entscheidende Komponente, die in früheren Studien an Mäusen fehlte.

Zunächst konzentriere ich meine Aufmerksamkeit auf die MHC-Gene des Dreistachligen Stichlings. Im ersten Kapitel haben meine Co-Autoren und ich versucht, ein genaues und reproduzierbares NGS-basiertes Protokoll zu entwickeln, das für die Genotypisierung von Multi-Locus-Genfamilien, wie dem MHC, bei verschiedenen Arten eingesetzt werden kann. Bisher wurden die Bemühungen um eine zuverlässige Charakterisierung der individuellen MHC-Vielfalt in Nicht-Modellarten dadurch behindert, dass nicht jeder MHC-Locus einzeln amplifiziert werden konnte, was zu komplexen Problemen bei der Unterscheidung zwischen echten Allelen und Artefakten während der Datenanalyse führte. Hier haben wir einen systematischen Vergleich von drei gängigen NGS-Pipelines zur Genotypisierung durchgeführt, indem wir MHC-Sequenzierungsdaten von genomischer Stichling-DNA und cDNA-Proben verwendet haben, die separat auch mit einer etablierten Nicht-NGS-Methode genotypisiert wurden. Auf der Grundlage einer umfassenden Aufschlüsselung der Genotypisierungsfehler unterstreichen wir die Notwendigkeit einer ausgangsmaterial-spezifischen Optimierung der einstellbaren Pipeline-Parameter. Darüber hinaus zeigen wir, dass eine optimierte Genotypisierung von gepaarten gDNA- und cDNA-Proben eine genaue Expressionsanalyse einzelner Allele und Haplotypen ermöglicht, was bei künftigen Bemühungen um eine zuverlässige Genotypisierung von MHC- und anderen Multigenfamilien unter Verwendung von cDNA-

Proben helfen wird.

Die Expression einer Vielzahl von MHC-Allelen beim Dreistachligen Stichling, wie im ersten Kapitel bestätigt, ist ein deutlicher Hinweis auf einen voll funktionsfähigen adaptiven Teil des Immunsystems. Ich wende mich daher im zweiten und dritten Kapitel dem TCR-Repertoire des Dreistachligen Stichlings zu. Es ist bekannt, dass ein komplexes Netzwerk genetischer und umweltbedingter Faktoren das TCR-Repertoire beeinflusst, wobei der immunologische Status des Wirts eine wichtige Rolle spielt. Frühere Infektionsstudien haben in erster Linie die Reaktion des TCR-Repertoires auf antigenisch einfache Erreger, wie z. B. Viren, untersucht. Um auf diesen Studien aufzubauen, widmet sich das zweite Kapitel dieser Arbeit der gründlichen Untersuchung der Auswirkungen einer kontrollierten makroparasitären Infektion auf das TCR-Repertoire. Zu diesem Zweck haben wir einen stichlings-spezifischen Helminthen, *Schistocephalus solidus*, verwendet, von dem allgemein angenommen wird, dass er immunmodulatorische Effekte auf das Immunsystem des Wirts ausübt. Durch die Verwendung von im Labor gezüchteten, MHC-genotypisierten Stichlingsfamilien konnten wir gleichzeitig die genetische Variation auf individueller Ebene berücksichtigen. Auf diese Weise konnten wir aufzeigen, in welch hohem Maße der genetische Hintergrund über den MHC hinaus die Vielfalt des TCR-Repertoires bestimmen kann - ein Faktor, der in früheren Studien weitgehend unberücksichtigt blieb. Ich gehe auf diese experimentellen Ergebnisse ein, indem ich die möglichen genetischen und epigenetischen Mechanismen diskutiere, die an der Gestaltung des TCR-Repertoires beteiligt sind.

Darüber hinaus führte die Infektion mit einem schnell wachsenden *S. solidus*-Stamm zu einer, wie wir vermuten, parasitenbedingten Unterdrückung des privaten TCR-Repertoire-Kompartiments. Unsere Ergebnisse zeigten auch eine Reihe von MHC-unbeschränkten, nicht-infektionsspezifischen TCR-Sequenzen mit unterschiedlichen Eigenschaften, die stark auf spezialisierte biologische Funktionen hinweisen. Wir interpretieren dies als indirekten Beweis für das Vorhandensein unkonventioneller T-Zellen im Stichling und betrachten die Implikationen hinsichtlich ihrer Rolle in der Evolution des adaptiven Immunsystems der Wirbeltiere.

Im dritten und letzten Kapitel dieser Arbeit charakterisieren wir das TCR-Repertoire des Stichlings unter halbnatürlichen Bedingungen, indem wir im Labor gezüchtete Stichlinge der natürlichen Umgebung ihres ursprünglichen Sees aussetzen. Unser Ziel war es, durch die Nachahmung der natürlichen Lebensbedingungen den immunologischen Status des Stichlings in freier Wildbahn besser zu erfassen, der häufig eine Koinfektion mit mehreren Makroparasitenarten beinhaltet. Ein solcher natürlicher Kontext wird in laborgestützten Einzelinfektionsstudien oft übersehen und kann zu falschen Schlussfolgerungen hinsichtlich der Dynamik des TCR-Repertoires führen. Hier ermöglichte eine achtwöchige See-Exposition eine hohe individuelle Makroparasiten-Belastung, die mit einer starken Aktivierung des TCR-Repertoires durch eine erhebliche Zunahme klonal erweiterter TCR-Klonotypen einherging. Wir bestätigen außerdem die Rolle individueller genetischer Variation im TCR-Repertoire, was mit den Ergebnissen des zweiten Kapitels übereinstimmt. Darüber hinaus stellen wir starke geschlechtsspezifische Unterschiede in der Vielfalt des TCR-Repertoires bei Stichlingen fest, deren sexuelle Reifung durch unseren Versuchsaufbau ausgelöst wurde. Interessanterweise zeigte der Vergleich des TCR-Repertoires von Kontroll-Stichlingen mit gleichem genetischen Hintergrund und unterschiedlichem Alter auffallend ähnliche Veränderungen wie bei Säugetieren, ein Ergebnis, das auf die Existenz gemeinsamer Grundprinzipien hinweist, die die Struktur des TCR-Repertoires während des gesamten Lebens eines Wirbeltiers bestimmen.

Fish magic, Paul Klee (1925)

"We shall not cease from exploration
And the end of all our exploring
Will be to arrive where we started
And know the place for the first time."

- T.S. Eliot, *Four Quartets* (1941)

Contents

Glossary

Chapter 1	
Individual MHC diversity	Total number of unique MHC alleles present in an individual.
Major Histocompatibility Complex	(MHC). The highly polymorphic region of the vertebrate genome that encodes MHC cell surface proteins that present antigenic peptides.
MHC allele	A unique sequence variant indiscriminately amplified by a given pair of locus-nonspecific MHC-I or MHC-II primers.
MHC-I	Class I MHC genes encoding for MHC proteins expressed on all nucleated cells that present intracellular antigenic peptides.
MHC-II	Class II MHC genes encoding for MHC proteins expressed on antigen presenting cells that present extracellular antigenic peptides.
Chapters 2, 3	
Clone size	Copy numbers of a given clonotype in a TCRβ repertoire. Here equivalent to number of UMIs, i.e. input cDNA molecules.
Clonotype	Functional unit of the TCR repertoire. Here defined as a unique amino acid CDR3 sequence of a TCR, with its corresponding V gene.
Complementarity-Determining Region 3	(CDR3). The most polymorphic region of each TCR chain, which directly contacts the peptide-MHC complex.
Generation probability	(P_{gen}). Probability for a given TCR sequence to be generated during V(D)J recombination.
Germline genes	Variants of variable (V), diversity (D) and joining (J) genes, representing the building blocks of recombined variable regions of a TCR.
Private clonotype	A TCR sequence that occurs exclusively in the repertoire of a single individual.
Public clonotype	A TCR sequence that occurs more than n times in a given TCR repertoire dataset of multiple individuals.
TCRβ	The β chain of the T cell receptor, a heterodimer consisting of one α and one β chain.
TCRβ repertoire	Collection of all TCRs present in an individual at a given point in time.
TCRβ1/TCRβ2 repertoire	TCRβ repertoire stemming from the TRB1 or TRB2 genomic locus of the three-spined stickleback.
TCRβ repertoire diversity	Number of clonotypes present in a TCRβ repertoire of a certain size.
TCRβ repertoire size	Total number of TCRβ sequences comprising a given repertoire.
Singleton	A clonotype of clone size equal to one in a given TCRβ repertoire.
Unique Molecular Identifiers	(UMIs). Random barcode sequences added to each input template molecule during PCR-based high-throughput sequencing library preparation protocols, used for sequencing error correction and amplification bias elimination.

General Introduction
The great gamble of adaptive immunity

The adaptive immune system of vertebrates

A (very) brief history of immunology

Since antiquity, humans have repeatedly and independently observed that prior exposure to a pathogen can "safeguard" an individual from succeeding exposures. In 430 BC, the historian Thucydides noted that those who had survived the Athenian plague were subsequently able to tend to the ill without fear of reinfection (Poole and Holladay, 1979). For most of our history as a species, however, the existence of pathogens has eluded us. Invention of the microscope paved the path to the unveiling of a whole new world of microorganisms previously invisible to the naked eye, when in the 17th century Anthonie van Leeuwenhoek directly observed bacteria from his own mouth (Leewenhoeck, 1684). A century later, in 1798, Edward Jenner's ground-breaking work on the cowpox vaccine proved that vaccinated individuals were immune to subsequent challenges with smallpox. As a result, cowpox vaccination rapidly replaced the common yet far more dangerous practice of 'variolation'– the use of smallpox itself to induce immunity.

Even though Jenner's experiments helped popularize the concept of 'immunity', they offered no scientific understanding as to how it develops. As it stood, the dominant theory was that infectious diseases were caused by a type of 'bad air' (or miasma) stemming from rotting organic matter. It would take a hundred more years, until the pioneering work of Louis Pasteur, followed by Robert Koch in the late 19th century, for the realization that bacteria can be the causative agents of disease. This stood as direct and definitive evidence for the long-hypothesized "germ theory of disease", which had, up until then, been largely met with contempt by both the scientific and medical communities. The answer to the etiology of infectious disease, however, was not immediately linked to the notion that the host could be equipped with tools against invaders; Koch and Pasteur themselves seemed to lean towards the idea of a defenseless host. It wasn't until Metchnikoff discovered the intracellular killing of microbes through phagocytosis, followed by Behring and Ehrlich, who identified neutralizing antibodies against bacterial toxins (Kaufmann, 2008), that the role of host immunity was finally recognized as an important defense mechanism against pathogens. These seminal works marked the birth of the science of immunology, and set in motion a series of further remarkable scientific discoveries in the field (Kaufmann, 2019), which, over the course of a century, helped weave an intricate picture of arguably one of the most complex biological systems known to humans: the vertebrate immune system.

Innate immunity preceded the adaptive immune system

Every living organism is equipped with an immune system - this universality underscores how absolutely critical it is for survival to be able to defend against pathogens, *i.e.*, disease-inducing micro- and macro-organisms (Janeway and Medzhitov, 2002). It is a hostile world out there – it has been estimated that 40% of all species on Earth follow a parasitic life strategy, or in other words, survive in or on another host organism by using them for nourishment (Dobson et al., 2008). Not all parasites are pathogenic, *i.e.*, cause disease to their hosts; yet, a host's fitness can be drastically reduced by being parasitized (Begon, 2009). For simplicity, and because all pathogens are parasitic in nature, in the context of this thesis the terms "parasite" and "pathogen" will be used interchangeably. This vast array of threats contributes to what Darwin described as a competitive "struggle for existence"- it thus follows logically that evolution has selected for hosts with efficient lines of defense. How are these immune defenses organized and how do they manage to protect the host?

Conventionally, the immune system is broadly divided into two functional arms: the innate immune system and the adaptive one. All species, from prokaryotes to fungi, from plants to invertebrates to vertebrates, possess an innate immune system (Janeway and Medzhitov, 2002). Innate immunity is multi-layered and complex, with a large number of cell types, such as macrophages, natural killer cells and neutrophils, interacting together to combat invading pathogens. Initiation of innate immune responses depends on germline–encoded receptors that have evolved to recognize highly conserved molecular patterns found in pathogens- but not in host cells. These receptors, termed pattern recognition receptors, and are constitutively expressed on the surface of innate immune cells, awaiting the infective agent (Janeway and Medzhitov, 2002). For this reason, effector cells can become quickly activated and lead to rapid innate responses, often within hours following invasion (Marshall et al., 2018). On the other hand, the sharing of these conserved features by many pathogens renders innate immune responses largely non-specific, as responds in the same way to all germs and foreign substances – often resulting in collateral damage, *i.e.*, harm to the host's own tissues, via *e.g.*, oxidative stress. It also means that, some exceptions notwithstanding, innate responses are not improved upon with subsequent encounters with the same pathogen – or, as immunologists say, they have no "memory" (Marshall et al., 2018). Nevertheless, innate immunity is very efficient: most infections are believed to successfully eliminated in the early stages of infection by means of the innate immune system (Parham, 2003). Thus, innate immunity is fast, non-specific, efficient, and represents the only mode of immunity available to the vast majority of species on Earth.

Innate immunity with its various layers stood as the only immune system for the most part of life on Earth, since the appearance of single cell microorganisms more than 3.5 billion years ago. This changed with the emergence of vertebrates, 500 million years ago, which was accompanied by the arrival of adaptive immunity (Pancer and Cooper, 2006). As opposed to the limited number of pattern recognition receptors of the innate immune system, the adaptive immune system is characterized by an extraordinarily large set of somatically diversified antigen receptors, which form an anticipatory form of defense against never-before-encountered threats. These new kind of receptors are expressed on a novel cell type, called lymphocyte, with each lymphocyte expressing a single such receptor. Lymphocytes are divided into B cells and T cells, a functional dichotomy present in jawed and jawless vertebrates alike (Boehm et al., 2012). Antigen receptors on the B cell surface are termed B cell receptors (BCRs). Upon antigen encounter, B cells undergo clonal expansion and differentiation into plasma cells, which secrete a soluble form of their BCRs knows as immunoglobulins or antibodies. B cells are therefore the effectors of humoral immune responses (Murphy and Weaver,

2016). In contrast to B cells, the antigen receptors of T cells (TCRs) remain cell-surface bound after antigen-mediated stimulation. Similar to B cells, activation of T cells prompts their clonal proliferation and differentiation into effector cells. Cytotoxic T cells, in particular, can destroy foreign and infected host cells, manifesting cell-mediated adaptive immunity (Andersen et al., 2006). After primary pathogenic exposure, multiple days to weeks pass before the adaptive immune system can mount a response sufficient for clearing the infection. Subsequent encounters, however, produce markedly enhanced, both in speed and potency, secondary immune responses (Murphy and Weaver, 2016). This is due to the formation of immunological memory, the critical feature of adaptive immunity that Thucydides and Jenner unknowingly observed more than 2000 years apart. Memory is one of the of the two pillars of the adaptive immune system – the second being specificity: both cell-mediated and humoral immune responses act against individual antigens derived from specific microbes, and as such, can precisely target a given pathogen at the species or even strain level.

MHC and its centrality in vertebrate adaptive immunity

How are the host's B and T cells activated upon pathogenic intrusion? The adaptive immune system is centered around the process of antigen presentation and recognition, as facilitated by Major Histocompatibility Complex (MHC) proteins (Flajnik and Kasahara, 2010). The MHC molecules are tasked with binding and presenting short peptides at the surface of host cells, where they are inspected by circulating T cells (Murphy and Weaver, 2016). Depending on the source of the bound peptide, MHC molecules are classified either as class I or class II. In case of infection, the presentation of fragments of pathogen-derived peptides by the MHC alerts the immune system as to the presence of an intruder: if a cell is infected with an intracellular pathogen, such as a virus or bacterium, proteasome-mediated degradation of associated proteins in the cytosol is followed by the loading of pathogen-derived peptides onto MHC class I molecules, found on all nucleated host cells. Analogously, MHC class II molecules, expressed only by certain immune cells called antigen presenting cells (APCs), bind and present peptides from the extracellular space, following phagocytosis of foreign proteins or pathogens by the APC (Murphy and Weaver, 2016). Successful interaction of the formulated peptide-MHC molecule (pMHC) by a patrolling T cell via its matching unique TCR triggers downstream adaptive immune responses. These responses diverge: pMHC-I complexes are recognized by T-cells carrying a cluster of differentiation 8 (CD8) co-receptor; engagement of both TCR and CD8 will result in T cell activation and initiation of cell-mediated immune responses. pMHC-II complexes are recognized by T-cells with a CD_4 co-receptor, which will trigger humoral responses via B cell activation and subsequent antibody production. With the exception of jawless fish, the adaptive immune system of vertebrates is based on this triad of MHC-TCR-BCR (Boehm et al., 2012).

The host's chances of successfully combating infectious agents is directly dependent upon its ability to bind and present a wide array of foreign-derived peptides by its MHC molecules. Each MHC has a limited binding capability, determined by the amino acid sequence of its binding grove (Brown et al., 1993). It is, therefore, advantageous to the host to express a set of MHC molecules with a wide array of binding specificities; this necessity is reflected both at the population level, where the MHC is recognized as the most polymorphic region in the vertebrate genome (Piertney and Oliver, 2006), as well as at the individual level, where typically vertebrate species possess multiple MHC-I and MHC-II loci (Babik, 2010). A number of different theories have been put forward to explain, in evolutionary terms, the rise and maintenance of such high polymorphism at the MHC. While the debate is still not settled, it

is widely accepted that host-parasite interactions have had a key role to play via pathogen-mediated balancing selection (Spurgin and Richardson, 2010).

The T cell receptor (TCR) repertoire and its diversity

These lines are being written during the 40th anniversary of the identification and cloning of the T cell receptor (Hedrick et al., 1984; Yanagi et al., 1984), the molecule at the very heart of adaptive immunity, nicknamed "the Holy Grail" of immunology (Shi et al., 2024). This scientific milestone allowed for the thorough elucidation of the TCR structure: it is a heterodimeric molecule, typically consisting of one α and one βglycoprotein chain, each of which is composed of two extracellular domains. The Variable (V) region, facing the extracellular space, and the Constant (C) region, adjacent to the cell membrane, which also includes a transmembrane region and the cytoplasmic tail. Combined, the two chains form the peptide-MHC complex binding region (Figure 1). The $\alpha\beta$ T cell receptor (TCR) is present in one of the two main lineages of T cells in jawed vertebrates, the other being $\gamma\delta$ T cells (Murphy and Weaver, 2016). Within this thesis, I focus on $\alpha\beta$ T cells, which constitute the overwhelming majority of T cells in humans, mice, and presumably all jawed vertebrates (Holderness et al., 2013).

The engagement of a foreign peptide-MHC complex by a TCR of matching specificity is the key requirement for adaptive immune response initiation. How does the host ensure that the pMHC will encounter a T cell bearing the "right" TCR? The answer came with the discovery of somatic recombination, a mechanism initially described for immunoglobulins (Hozumi and Tonegawa, 1976). This process, known as VDJ recombination, offers an elegant solution to the "generation of diversity problem"; its premise is that, unlike innate immune receptors, α and ß TCR chains are not fully encoded in the genome, are instead organized in arrays of gene segments. The numerous variable (V), joining (J), and -for the β chain locus-, diversity (D) gene segments in the germline state provide the building block for somatic recombination (Figure 1). This process takes place in the thymus, where precursor T cells migrate after hematopoiesis. Here, the TCR germline genes of these precursor T cell, termed thymocytes, are rearranged in an almost random fashion, but in strict order; first D-to-J, then V-to-DJ rearrangements of the β locus, followed by V-to-J recombination of the TCR α locus. These reactions are performed by a group of enzymes, key amongst which are RAG1 and RAG2 recombinases (Roth, 2014).

The resulting diversity of the somatically rearranged α and β TCR genes stems both from the combination of various segments, as well as from the random insertion and/or removal of nucleotides at their junctions (Figure 1). In order for each T cell to bear a TCR of single specificity, the mechanism of allelic exclusion halts VDJ recombination as soon as rearranged TCR β and α genes from one of the two chromosomal loci are produced (Brady et al., 2010). These recombination events result in the assembly of somatically diversified α and β chains, the coupling of which produces an $\alpha\beta$TCR of unique antigen specificity, expressed on the surface of a developing T cell. This specificity is determined by three regions on each chain: Complementarity Determining Region (CDR) 1 and CDR2, both fully encoded by the corresponding V gene segments, and CDR3, which encompasses parts of the V, (D), and J gene segments (Figure 1). The CDR3 is the most diverse region of the TCR, directly interacting with the peptide of the peptide-MHC complex (Birnbaum et al., 2014).

En masse, VDJ recombination generates of a vast diversity of TCRs, which is a key prerequisite for vertebrates to combat the innumerable threats with which are faced throughout their lifetime. In humans, this recombinatorial apparatus has the capacity to generate approximately $2x10^{19}$ unique $\alpha\beta$ TCRs (Dupic et al., 2019)- a number comparable to that of stars in the observable universe (Miles et al., 2011). Two problem arise: the first pertains to the fact that

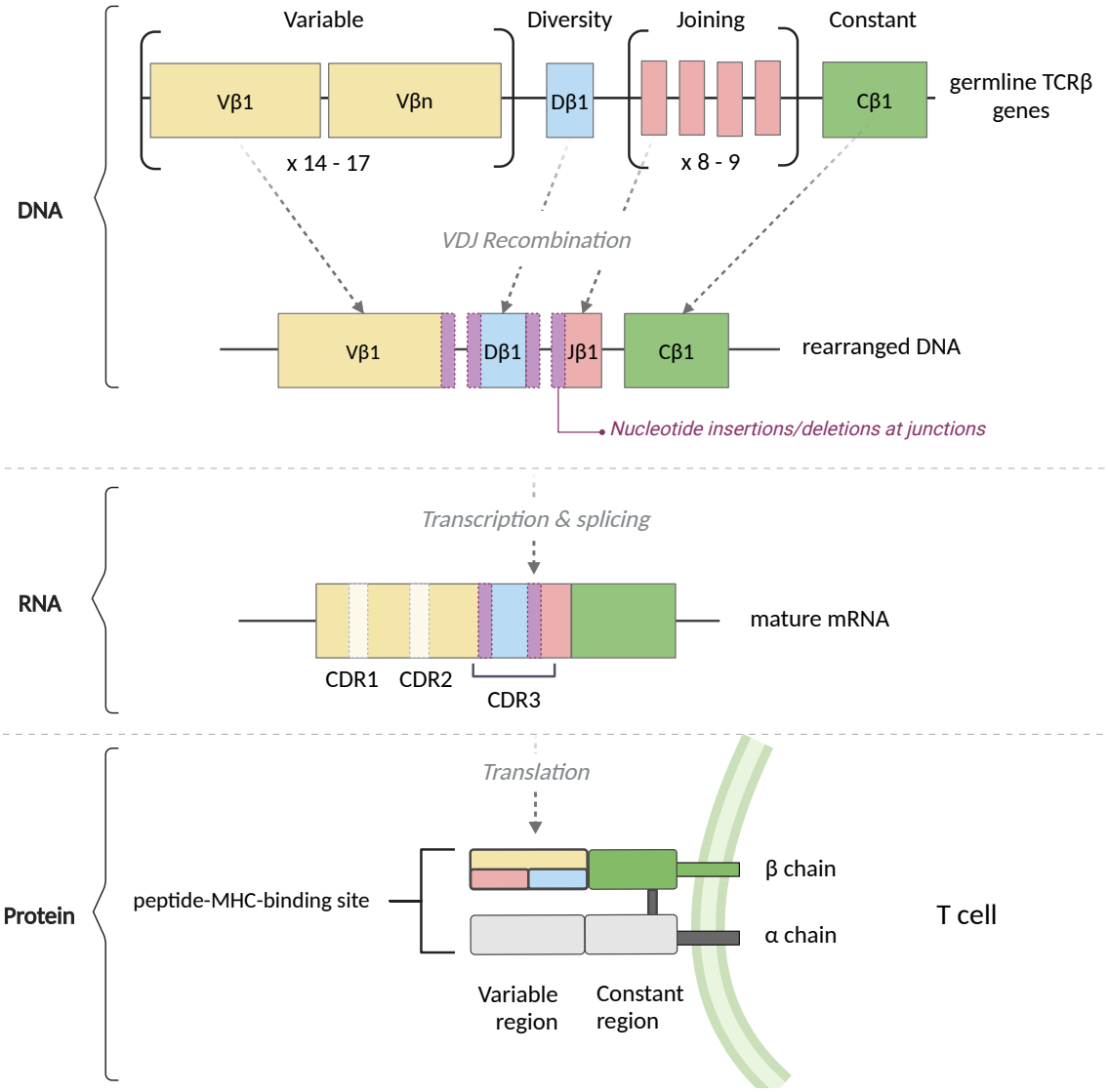


Figure 1: Generation process of the three-spined stickleback T cell receptor β chain through VDJ recombination. Created with BioRender.com.

some of the produced TCRs will not be useful, *i.e.*, they will not be able to interact with any self MHC molecule. The second is that, by chance, some TCRs will be autoreactive, *i.e.*, triggered by self peptide-MHC complexes. How is this diverse collection of receptors exclusively directed against pathogens and not the host?

Upon successful VDJ recombination, thymocytes go through two sequential 'quality control' steps in a process known as thymic selection, whose outcome is determined by their TCR specificity and strength of signaling. During the first step, which occurs in the thymus cortex, the newly generated TCRs are tested for their ability to recognize self-MHC molecules; thymocytes with TCRs that cannot bind to the host's MHC molecules undergo apoptosis, termed 'death by neglect' (Ashby and Hogquist, 2023). Surviving thymocytes subsequently move to the medulla, where those bearing TCRs that recognize self-pMHC with high affinity either undergo programmed cell death ('clonal deletion'), or are redirected into a regulatory T cell (Treg) phenotype (Ashby and Hogquist, 2023). Thus, thymic selection achieves its double goal: positive selection ensures TCR functionality through MHC-restriction (Zinkernagel & Doherty,

1997), while negative selection bestows immune self-tolerance upon TCRs. An estimated 1-3% of pre-selection thymocytes (Goldrath and Bevan, 1999) exit the thymus as fully mature T cells expressing unique TCRs that are both functional and safe. In the periphery, this diverse pool of T cells conducts self surveillance by 'scanning' the host's MHC molecules, and initiate an immune response only in case of detection of non-self peptides, *e.g.*, when the inspected host cell is infected. This post-selection TCR pool of diverse specificities, commonly referred to as a "repertoire", represents only a fraction of the maximal theoretical capacity. The dichotomous nature of the TCR repertoire was first defined by theoretical immunologist Niels Jerne, a prominent contributor to the theory of clonal selection of lymphocytes:

> "*When an antigen confronts the immune system, it impinges upon a repertoire of available lymphocytes(...). We must distinguish between the potential repertoire of specificities that could arise given the genetic constitution of the zygote from which the animal develops, and the available repertoire embodied in the cells that can respond to antigens at a given moment in the life of the animal.*"
>
> - Jerne (1972)

At any given point, therefore, an individual expresses only a tiny slice of that immense theoretical diversity – this realized diversity is conservatively estimated at $2x10^7$ unique T cell specificities, *i.e.*, unique TCRs (Arstila et al., 1999), born by a total population of approximately $4\text{-}5x10^{11}$ T cells in an adult human (Jenkins et al., 2010; Sender et al., 2023).

The vital role of Next Generation Sequencing in studying adaptive immunity

Much like the invention of microscopy opened a window to the previously hidden world of microbes, recent technological improvements has enabled researchers to delve into new depths in their pursuit of better understanding the adaptive immune system. In particular, the advent of Next Generation Sequencing (NGS), also known as high-throughput sequencing (HTS) represents a promising development for the field of MHC research (Babik et al., 2009). MHC genes exhibit extreme polymorphism, often manifesting as hundreds or thousands of distinct variants for a single locus. In species where the MHC genomic region has undergone recent duplications, locus-specific amplification is often impossible, making it challenging to decipher individual and population level diversity – as is commonly the case in non-model species (Babik, 2010; Lighten et al., 2014). Furthermore, there can be significant within-species variation in the number of MHC genes per individual (Kelley et al., 2005; Gaigher et al., 2016; Roved et al., 2022). This unpredictability adds an extra layer of complexity in reliably characterizing MHC diversity. NGS has the capacity to facilitate tremendous progress in this regard, as it requires no prior knowledge of allelic variants, unlike other traditional genotyping approaches (Argüello et al., 1998; Lenz et al., 2009a), and offers nucleotide-level resolution of sequences.

NGS has also ushered a modern era of adaptive immune receptor (*i.e.*, TCR and BCR) repertoire studies (Hou et al., 2016); its first successful application managed to yield hundreds of thousands of TCR sequences in a single experiment (Robins et al., 2009). Comparing this efficiency to that of studies employing earlier sequencing strategies, such as Sanger sequencing, whose typical output falls within the range of hundreds or thousands sequences at a time (Li et al., 2013), underlines the sharp technical improvement that NGS embodied. NGS outperforms other traditional methods, not only due to its significantly higher resolution, but

because it is considerably less labor-intensive; for example, Sanger sequencing relies on bacterial cloning of rearranged TCR sequences, a process that is both time-consuming and costly (Hou et al., 2016). Another popular technique of TCR repertoire analysis has been CDR3 spectratyping, an immunoscoping approach for analyzing the distribution of the length of CDR3 sequences, used to infer total repertoire diversity (*e.g.*, Ahmed et al. (2009)), and infection-status-driven disturbances, *e.g.*, Boudinot et al. (2001, 2004). This technique, however, fails to provide single-clone sequence resolution and information regarding the clonal size of each sequence. Despite their numerous shortcomings, traditional techniques facilitated the early characterization of the TCR repertoire and offered a first glimpse into the degree of its diversity (Six et al., 2013). In NGS, researchers have a tool powerful enough to capture a broad picture of the TCR repertoire. What has been revealed is a near unprecedented level of biological diversity, which in turn has spurred the development of novel bioinformatics and statistical tools for its analysis (Shugay et al., 2015; Rosati et al., 2017; Davis and Boyd, 2019; Mhanna et al., 2024). Previously hard to tackle questions, especially ones pertaining to quantitative features of the TCR repertoire, such as its true size, diversity, and inter-individual overlap, have now been placed firmly within reach.

What shapes the TCR repertoire? Knowns and unknowns

Following thymic selection, surviving T cells are released into the circulation. Some of these T cells remain in the periphery, while others take residency in various host tissues and organs, but all contribute to forming a pre-immune (or "naive") TCR repertoire. Upon exposure to a pathogen, activated T cells undergo clonal expansion and differentiation into mature effector lymphocytes with cytotoxic and pro-inflammatory functions, or into plasma cells that secrete antibodies. Thus, at the acute stage of infection, this population of expanded and differentiated T cells is represented in the T cell pool at a greater frequency compared to naive, *i.e.*, non-antigen-driven T cells (Catron et al., 2004; Hataye et al., 2006; Moon et al., 2007). Upon pathogen elimination, a few effector T cells are retained for long periods as memory T cells, ensuring protection from future re-invasion (Murphy and Weaver, 2016), while the majority of expanded T cells undergoes apoptosis (Zhang et al., 2005b). Replenishment of T cell by the thymus and homeostatic proliferation of naïve T cells are crucial mechanisms of maintenance of a diverse pool of T cell specificities over the course of an individual's lifetime (Murray et al., 2003; Takada and Jameson, 2009). The TCR repertoire therefore represents a highly dynamic process of T cell turnover via generation, expansion, and programmed cell death.

What shapes an individual's TCR repertoire at a given point in time? A plethora of NGS-enabled efforts to examine the TCR repertoire has informed us of both environmental and intrinsic factors that are crucial in shaping its size, shape, and composition. As an assembly of antigen-binding receptors, the TCR repertoire harbors information on both past and ongoing immune responses (DeWitt et al., 2018). Thus, a focal research point has been how the TCR repertoire responds to antigenic challenge. Such studies have been particularly informative regarding changes in its diversity; overall, we observe drops in diversity after exposure to various pathogen-derived antigens, especially from viruses, due to expansion of specific T cell clones (Zhu et al., 2013; Costa et al., 2014; Krishna et al., 2020). The overwhelming majority of TCR repertoire research under conditions of immune challenge has focused on antigenically simple pathogens, as well as immunizations with model allergens (Ballesteros-Tato et al., 2016; Hondowicz et al., 2016). As such, even though our understanding has been greatly advanced with regards to the response of the TCR repertoire to a few dominant epitopes, there is a significant gap in our knowledge of TCR repertoire changes induced by more epitope-rich antigenic

stimuli, such as those found in macroparasitic infections (Brown et al., 2021). Helminths are large parasitic worms whose complex life cycles are characterized by the production and secretion of a variety of stage-specific molecules (Hewitson et al., 2009; Harnett, 2014; Berger et al., 2021). As a result, a unique feature of helminth infection is the diversity of available immunogenic epitopes.

Beyond conditions of eternal stimuli, TCR repertoire studies have also tried to elucidate the ways in which intrinsic factors, such as sex (Britanova et al., 2016; Migalska et al., 2019; Krishna et al., 2020), age (Qi et al., 2014; Britanova et al., 2014, 2016; Yoshida et al., 2017; Egorov et al., 2018; Minervina et al., 2021), and gut microbiome (Ivanov et al., 2022; Nagashima et al., 2023) affect the TCR repertoire in healthy individuals. These factors can impact various functional aspects, such as its diversity, clonality (*i.e.*, clonal composition), and overlap between individuals. A big question that remains largely unexplored pertains to the role of genetics in shaping the TCR repertoire. Thus far, efforts to identify relevant genetic factors have been limited to specific immune genes, most notably those of the MHC region, due to their close functional association with T cells (DeWitt et al., 2018; Krishna et al., 2020; Migalska et al., 2019; Brown et al., 2024). An effect that has not received nearly as much attention from researchers is the one exerted by the genetic background as a whole, not just the MHC, on the TCR repertoire (Russell et al., 2022). The assessment of the role of individual genetic variation has been hindered by the fact that most TCR repertoire studies are restricted to humans. Besides often lacking information regarding the subjects' genetic background, conclusions drawn from observational studies can be compromised due to external factors influencing the TCR repertoire composition. For example, there is a large number of asymptomatic chronic infections that cause robust T cell responses (Simon et al., 2015), such as CMV infection (Krishna et al., 2020), which could be part of a subject's unknown clinical history. Alternatively, researchers have used experimental approaches by employing model species. Such approaches offer control over the age, sex, and subjects' exposure to antigenic stimuli, and can be used for designing balanced studies with regards to other TCR repertoire-relevant factors, such as the individuals' MHC background. The standardized format of experimental studies has thus been very successful in uncovering molecular mechanisms of TCR repertoire generation and maintenance through genetically engineered strains (*e.g.*, Gilfillan et al. (1993); Li et al. (2010); Perera et al. (2013)). Nevertheless, their low (or non-existent) level of genetic variation (Collins et al., 2015; Watson et al., 2019) is a bad proxy for the species-specific genetic diversity existing in the wild, and renders them an unsuitable medium through which to consider questions pertaining to the role of natural genetic variation in shaping the TCR repertoire.

The above points emphasize the need to go beyond the established experimental models and start exploring the TCR repertoire -specifically-, and the adaptive immune system -broadly-, within the context of natural genetic variability, in order to draw more robust conclusions to questions regarding their functionality. Incorporating genetically variable species into research would also facilitate addressing queries at the intersection of immunology and evolution.

The three-spined stickleback as a model vertebrate host

This thesis is based on experimental work carried out using the three-spined stickleback (*Gasterosteus aculeatus* L.), a small teleost species that has a wide geographic distribution across the Northern hemisphere. Owing to the fact that sticklebacks have repeatedly colonized a broad spectrum of freshwater habitats from an ancestral marine population (McKinnon and Rundle, 2002), they exist as morphologically and ecologically distinct ecotypes (Reusch et al.,

2001). Their rapid ecological radiation and rich natural variation have thus been extensively utilized in evolutionary biology (Bell and Foster, 1994), while their elaborate reproductive behavior has sparked rigorous research in the field of behavioral ecology (FitzGerald and Wootton, 1986; Lachance and FitzGerald, 1992), aided by their relatively short life span, and easy breeding and maintenance in the lab. The stickleback has also emerged as a model system for the study of host-parasite interactions, with an extensive literature regarding parasite-driven local adaptation (Scharsack et al., 2007a; Eizaguirre et al., 2012b; Feulner et al., 2015; Robertson et al., 2016). This research has been supported by extensive background knowledge regarding the stickleback's diverse range of parasites (Kalbe et al., 2002; Wegner et al., 2003b). Crucially, there is substantial evidence for the role of the three-spined stickleback's individual MHC diversity influencing resistance to sympatric parasite species (Wegner et al., 2003b,a; Kurtz et al., 2004; Wegner et al., 2008; Kalbe and Kurtz, 2006; Eizaguirre et al., 2012a). The above are strongly indicative of MHC-mediated adaptive immune responses, including T cell responses, affecting the stickleback's immunocompetence against parasites (Scharsack and Kalbe, 2014; Haase et al., 2016; Robertson et al., 2016; Piecyk et al., 2019a).

As a teleost, the adaptive immune system of the stickleback displays significant similarities with the mammalian immune system (Sunyer, 2013; Bjørgen and Koppang, 2021; Zapata, 2024), which extent beyond its functional and polymorphic MHC-I and MHC-II loci (Figure 2). Specifically, the stickleback's thymus is structured into zones that are equivalent to the cortex and medulla of the mammalian thymus, where lymphocytes reside (Wölfing, 2012). In sticklebacks, expression of the recombination activating gene 1 (RAG-1), which is key for VDJ recombination, as well as FOXP3, a specific marker of regulatory T-cells, have been detected (Wölfing, 2012). Furthermore, we know from studies in other species that teleosts possess most of the components of the mammalian innate immune system (Magnadóttir, 2006), in addition to elements of a functional adaptive immune system, including various immunoglobulin classes, and major T cell subsets (Flajnik and Kasahara, 2010). Although teleosts lack a bone marrow, their head kidney is considered its functional ortholog (Zapata and Amemiya, 2000). Therefore, due to the homology of immune organs and cell types to its mammalian counterparts, the stickleback constitutes an excellent model species for exploring questions at the intersection of evolution and immunology.

The stickleback-*Schistocephalus solidus* infection system

Yet another advantage of the stickleback as an experimental model is that it can be used for controlled infections with different parasite species (Wegner et al., 2003a), arguably the best known among which is the helminth *Schistocephalus solidus*. *S. solidus* is a naturally occurring cestode parasite of the stickleback, which serves as its obligatory second intermediate host during its complex three-host life cycle, as depicted in Figure 3. In the wild, the parasite is trophically transmitted to the stickleback via ingestion of infected copepods, the parasite's first intermediate hosts (Barber and Scharsack, 2010). Once inside the stickleback, *S. solidus* penetrates the intestinal wall and establishes infection in the body cavity in the form of plerocercoid larvae, where it can grow to reach significant body sizes (Barber and Scharsack, 2010). Subsequently, the parasite is transmitted to its definitive host, typically a piscivorous bird, through consumption of the infected stickleback. In the bird's gut, the parasite undergoes sexual maturation and reproduction (Smyth, 1946; Bråten, 1966). The *S. solidus* life cycle is completed when the eggs are released into a water body through bird feces, where they hatch into free-living coracidia that are consumed by copepods (Wedekind, 1997; Barber and Scharsack, 2010).

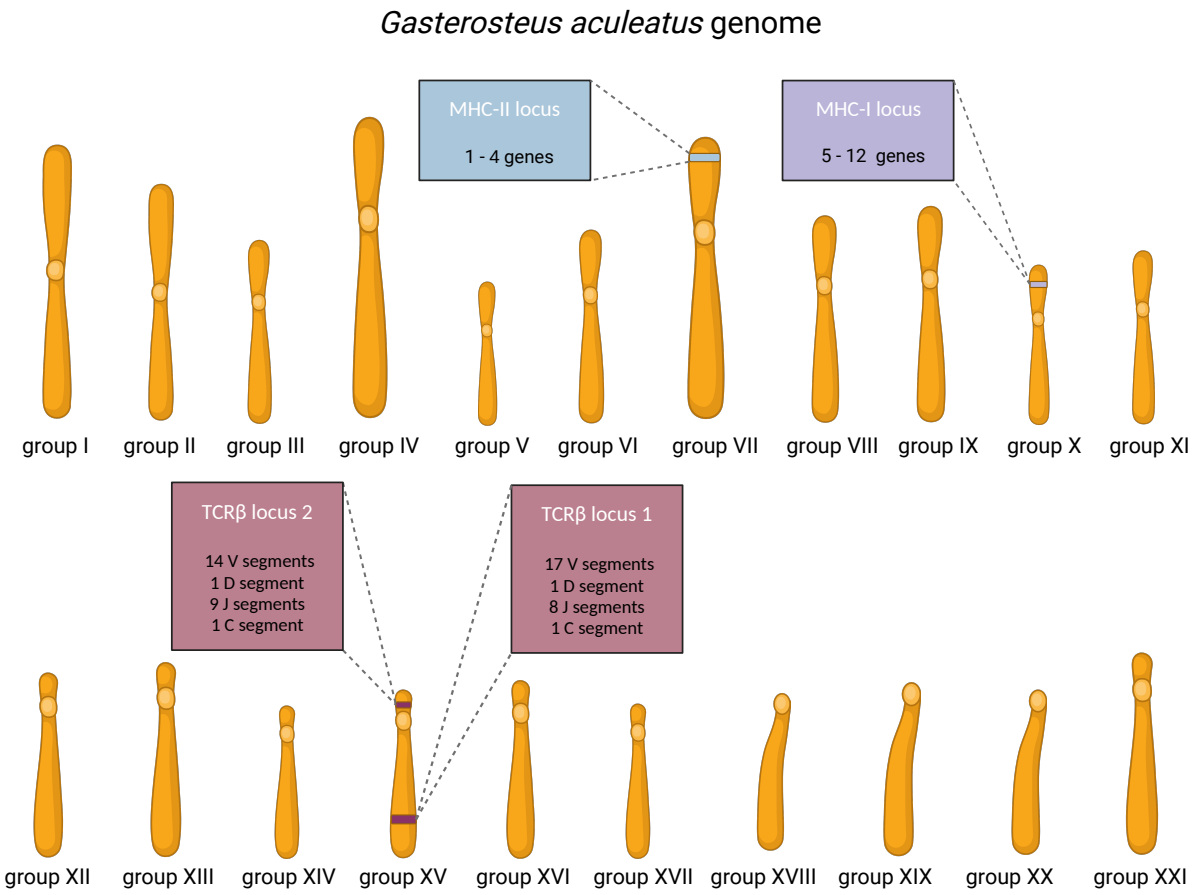


Figure 2: Schematic representation of the linkage groups (chromosomes) of the three-spined stickleback (*Gasterosteus aculeatus*). Highlighted in different colors are the genomic loci of interest within the context of this thesis (MHC class I, MHC class II, TCRβ locus 1 and TCRβ locus 2). Numbers of involved genes are provided within the corresponding boxes. Genomic positions and relative distances are in approximation. Created with BioRender.com.

In a laboratory setting, the definite host of *S. solidus* can be replaced by an *in vitro* breeding system emulating the conditions of the bird gut, thus enabling controlled infections under standardized conditions (Smyth, 1946, 1954; Wedekind, 1997). This cemented the stickleback-*S. solidus* infection system as an established model of helminth infection of vertebrates (Arme and Owen, 1967), and allowed researchers to investigate a plethora of questions at the intersection of ecology, evolution and immunology (Barber and Scharsack, 2010). The fact that it is a specialist parasite to the stickleback, *i.e.*, only grows optimally inside this species, while exhibiting sub-optimal growth in closely related species (Orr et al., 1969), ensures that a strong interplay of selection forces can potentially shape both host and pathogen over large-scale evolutionary times. Indeed, a series of studies has shown that *S. solidus* strains originating from distinct geographic locations differ in their growth rate depending on the host strain (Kalbe et al., 2016; Piecyk et al., 2019b, 2021). *S. solidus* growth within the stickleback has been suggested to be partially controlled by the adaptive immune system, as individuals who exhibit median MHC diversity can better control its growth following repeated exposures (Kurtz et al., 2004). However, once inside the stickleback's body cavity, *S. solidus* total elimination is rare (Scharsack et al., 2007b), if not elicited via external immune stimulation soon after infection (Wedekind and Little, 2004). Conversely, *S. solidus* can be cleared by the immune system during infection in other teleost species (Orr et al., 1969). This stands as a strong indication

of *S. solidus* exerting immunomodulatory effects on its specific host, further supported by a number of *in vitro* and *in vivo* studies (Franke et al., 2014; Hammerschmidt and Kurtz, 2005; Scharsack et al., 2004, 2007a; Piecyk et al., 2019a; Berger et al., 2021; Robertson et al., 2016).

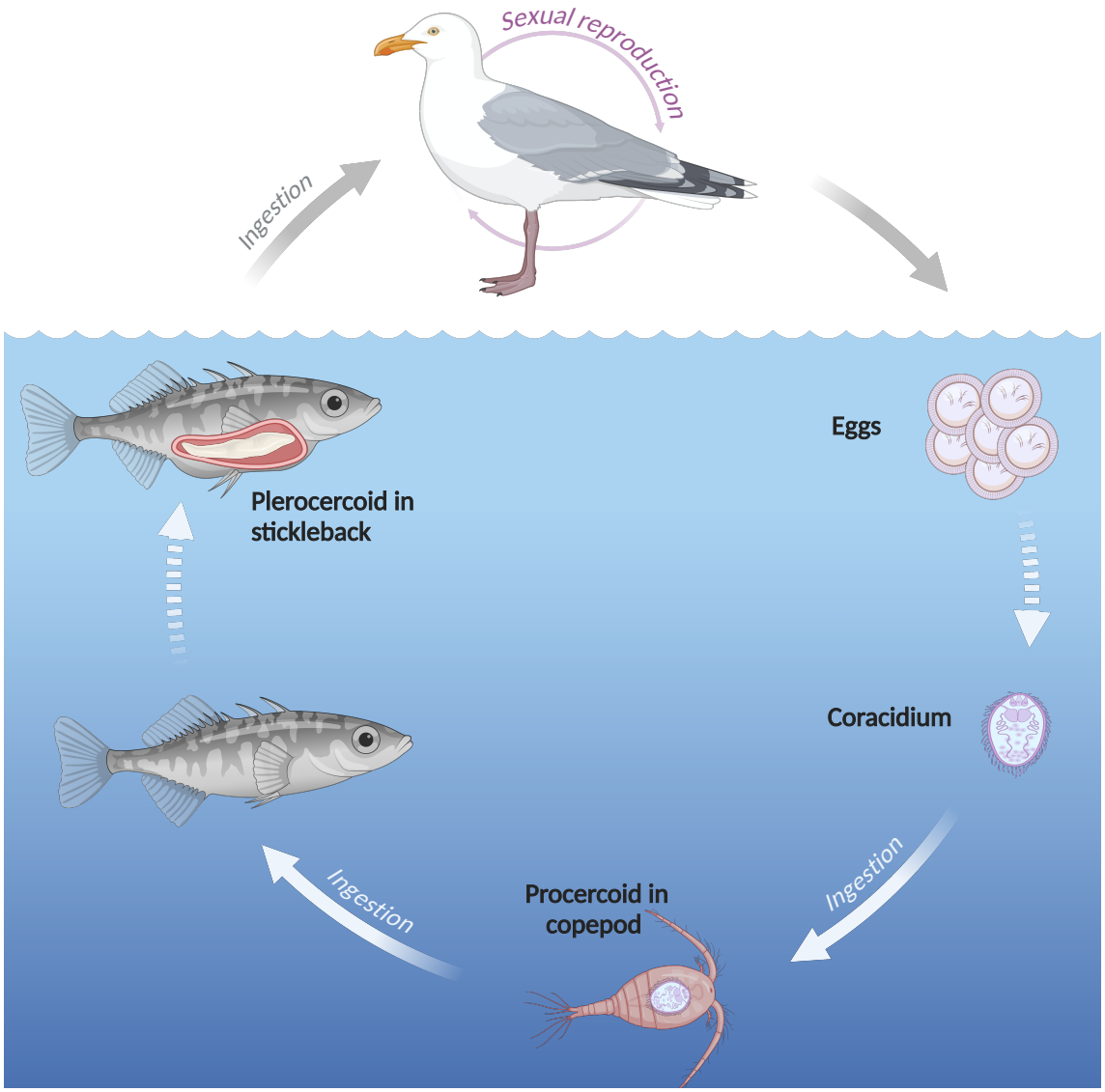


Figure 3: The complex life cycle of the three-spined stickleback-specific parasite, *Schistocephalus solidus*. This cestode has a three-host life cycle, within which the three-spined stickleback serves as second intermediate host. Created with BioRender.com.

Previous work on the three-spined stickleback TCR repertoire

The sequence of the whole genome of the three-spined stickleback (Jones et al., 2012), was a significant advancement towards expanding our focus beyond its well-studied MHC system. As such, previous efforts were able to assemble a library of V, D, and J germline gene segments present in the stickleback through mapping to reference genome (Wölfing, 2012). Based on gene segment sequence information, the subsequent design of stickleback-specific

primers facilitated sequencing efforts which revealed that the stickleback possesses a diverse TCR repertoire (Wölfing, 2012), mirroring that of other teleosts (Covacu et al., 2016; Giorgetti et al., 2021; Castro et al., 2011). Intriguingly, the stickleback possesses two independent TRB loci on same chromosome (Figure 2), in contrast to other teleosts species, in a configuration closely resembling that of the two separate TRG loci of the Atlantic salmon (Yazawa et al., 2008). The above efforts, even though restricted by their low-throughput sequencing capacity, provide valuable initial insights into the TCR repertoire of immunologically naïve sticklebacks.

Thesis outline

The aim of this thesis is to advance our understanding of the factors shaping the vertebrate T cell receptor (TCR) repertoire during infection. Towards this end, I use the three-spined stickleback, *Gasterosteus aculeatus*, a prominent species in eco-evolutionary research, whose natural genetic variation can be directed towards elucidating the effects of genetic background on the TCR repertoire. I employ advanced NGS methodology to explore the expressed individual MHC diversity of the stickleback, a genetic factor implicated in shaping the TCR repertoire. Then, I thoroughly investigate the intra- and inter-individual TCR repertoire dynamics after a single controlled *Schistocephalus solidus* infection, as well as after co-infection with multiple macro-parasites under semi-natural conditions. By comparing with control, *i.e.*, immunologi-cally naive, individuals, I examine how well factors such as infection status, individual genetic variation, and sex can explain the inter-individual variance in the TCR repertoire composition during response to antigenically complex pathogens. As outlined below, my thesis consists of three main projects. All projects have been conducted in cooperation with colleagues, whose individual contributions are provided in full in the corresponding section of the thesis.

Chapter 1

Template-specific optimization of NGS genotyping pipelines reveals allele-specific variation in MHC gene expression

Reliable genotyping is a crucial first step toward furthering our understanding of the drivers of the exceptional genetic variability in the MHC. Next Generation Sequencing (NGS) genotyping approaches, despite their broad advantages, face challenges due to simultaneous co-amplification of multiple loci, which can cause overestimation of individual diversity. Here, my co-authors and I used stickleback MHC-II samples of known allelic composition to develop a highly accurate and reproducible NGS-based protocol which can be used to genotype MHC, and other multigene families, and is not restricted to genomic DNA. We show that use of cDNA can help retrieve valuable information regarding MHC gene expression status, as well as allele expression patterns. By applying this newly validated protocol to stickleback MHC-I samples, we reveal a higher-than-previously-reported expressed MHC-I allelic diversity.

Chapter 2

Effect of genetic background and helminth infection on the TCR repertoire of the three-spined stickleback

This project examines the dynamics of TCR diversity in three-spined sticklebacks following controlled experimental infection with the presumably immunosuppressive parasite, *S. solidus*. By analyzing the TCR repertoire of lab-reared individuals belonging to genetically diverse sibships, we found that genetic background significantly influences inter-individual TCR reper-

toire variation and biases V-J gene usage, indicating genetically controlled VDJ recombination mechanisms. Parasite infection did not significantly alter the repertoire diversity or its architecture; however, infection with a virulent *S. solidus* strain increased TCR repertoire overlap, likely due to suppression of private T cell responses. We also identified a 'public' TCR compartment comprising sequences of unique characteristics, which aligns with findings in other vertebrates, and hints at the presence of "innate-like" T cells in the stickleback.

Chapter 3

Exposure to natural parasitic fauna leads to TCR repertoire activation in the three-spined stickleback

For the last part of my Ph.D., I attempt to answer similar questions regarding the dynamics of the TCR repertoire under semi-natural conditions, which better emulate the naturally encountered variation in the abiotic and biotic environment of the stickleback. Lab-bred naïve individuals were placed in cage enclosures and exposed to the local parasite fauna of their natal lake. After eight weeks of exposure, significant T cell activation was observed, with a marked increase in clonally expanded TCR sequences, decreased repertoire diversity, and altered repertoire structure, all linked to high macroparasitic burdens. Notably, sex differences were found in the TCR diversity among exposed and control fish. Lastly, age-related changes in TCR repertoire diversity and composition in lab-kept control fish paralleled patterns seen in mammals.

Chapter 1

Template-specific optimization of NGS genotyping pipelines reveals allele-specific variation in MHC gene expression

1.1 Abstract

Using high-throughput sequencing for precise genotyping of multi-locus gene families, such as the Major Histocompatibility Complex (MHC), remains challenging, due to the complexity of the data and difficulties in distinguishing genuine from erroneous variants. Several dedicated genotyping pipelines for data from high-throughput sequencing, such as next-generation sequencing (NGS), have been developed to tackle the ensuing risk of artificially inflated diversity. Here, we thoroughly assess three such multi-locus genotyping pipelines for NGS data, the DOC method, AmpliSAS and ACACIA, using MHC class IIβ datasets of three-spined stickleback gDNA, cDNA, and "artificial" plasmid samples with known allelic diversity. We show that genotyping of gDNA and plasmid samples at optimal pipeline parameters was highly accurate and reproducible across methods. However, for cDNA data, gDNA-optimal parameter configuration yielded decreased overall genotyping precision and consistency between pipelines. Further adjustments of key clustering parameters were required to account for higher error rates and larger variation in sequencing depth per allele, highlighting the importance of template-specific pipeline optimization for reliable genotyping of multi-locus gene families. Through accurate paired gDNA-cDNA typing and MHC-II haplotype inference, we show that MHC-II allele-specific expression levels correlate negatively with allele number across haplotypes. Lastly, sibship-assisted cDNA-typing of MHC-I revealed novel variants linked in haplotype blocks, and a higher-than-previously-reported individual MHC-I allelic diversity. In conclusion, we provide novel genotyping protocols for the three-spined stickleback MHC-I and -II genes, and evaluate the performance of popular NGS-genotyping pipelines. We also show that fine-tuned genotyping of paired gDNA-cDNA samples facilitates amplification bias-corrected MHC allele expression analysis.

1.2 Introduction

The ability of the host's immune system to recognize and specifically target pathogens is a significant contributor to its overall fitness. Thus, pathogen-mediated selection is regarded as a major driver of vertebrate evolution, with immunity-related genes being hotspots of positive selection (Hillier et al., 2004; Bustamante et al., 2005; Kosiol et al., 2008; Barreiro and Quintana-Murci, 2010). In vertebrates, pathogen-specific immune responses rely on the presentation of short peptides by classical major histocompatibility complex (MHC) molecules, categorized into class I or II depending on the peptide origin -intracellular or extracellular (Murphy and Weaver, 2016).These molecules are coded by genes of the MHC region, which exhibits the highest genetic variation in the jawed vertebrate genome (Piertney and Oliver, 2006). MHC gene numbers can differ not only between, but also within species, where individual allele number variation has been repeatedly observed (Kelley et al., 2005; Gaigher et al., 2016; Roved et al., 2022).

Reliable genotyping is a crucial step toward furthering our understanding of the drivers of this extensive polymorphism, which, despite the pivotal role of MHC in adaptive immunity, remain elusive (Radwan et al., 2020). Early genotyping methods, such as Reference Strand-mediated Conformation Analysis (RSCA, Argüello et al. (1998)) helped decipher MHC polymorphism in humans and model species, where loci can be addressed separately. In most non-model species, however, recent gene duplication and conversion prohibit the designing of locus-specific MHC primers (Babik, 2010; Lighten et al., 2014). Unlike traditional methods, the increased resolution offered by next-generation sequencing (NGS) can tackle the resulting complexity of numerous indiscriminately amplified allelic variants (Babik et al., 2009) - hereafter referred to as "alleles". However, NGS technologies exhibit inherent error rates (Schirmer et al., 2015; Manley et al., 2016; Schirmer et al., 2016; Stoler and Nekrutenko, 2021) , and are susceptible to cross-sample contamination (Esling et al., 2015; Schnell et al., 2015; Eisenhofer et al., 2019) Together with known multi-locus genotyping issues, such as preferential allele amplification and formation of random base-mismatch errors and PCR-associated chimeras (Kanagawa, 2003; Lenz and Becker, 2008; Babik, 2010), the increased chances of observing artifacts alongside genuine alleles in NGS genotyping makes their separation a formidable task (Galan et al., 2010; Lighten et al., 2014; Biedrzycka et al., 2017).

To counter the ensuing risk of allelic diversity overestimation (Sommer et al., 2013; Biedrzycka et al., 2017), multiple bioinformatic workflows have been developed for the processing and genotyping of sequencing data (*i.e.*, genotyping pipelines), each one representing a different approach towards tackling artifacts (Lighten et al., 2014; Biedrzycka et al., 2017). So far, efforts to assess pipelines at MHC genotyping have largely depended on allele-calling repeatability between duplicate samples (Lighten et al., 2014; Biedrzycka et al., 2017; Rekdal et al., 2018). However, this can yield misleading results, given the non-random nature of PCR/sequencing artifacts (Lighten et al., 2014; Biedrzycka et al., 2017). Other studies have utilized datasets exclusively containing genomic DNA samples of a priori known genotypes (Sebastian et al., 2016; Gillingham et al., 2021) , missing out on the importance of cDNA-typing in MHC studies for verification of the allele expression status (Babik, 2010) . As such, a rigorous and unbiased pipeline evaluation using independently validated genotyping results from different sample templates has been lacking.

Here, we thoroughly evaluate three popular genotyping pipelines: the 'Degree of Change' (DOC) criterion method (Lighten et al., 2014) , the Amplicon Sequence Assignment (AmpliSAS) pipeline (Sebastian et al., 2016) , and the Allele CAlling procedure for Illumina Amplicon sequencing data (ACACIA) pipeline (Gillingham et al., 2021) . We use NGS data derived from

genomic DNA and cDNA of the three-spined stickleback, *Gasterosteus aculeatus* L., a species exhibiting significant natural variation in the individual number of MHC alleles (Lenz et al., 2009b) . We utilize two sources of reliable information for estimating genotyping performance: an established RSCA stickleback MHC-II typing protocol, and a comprehensive plasmid library of previously isolated MHC-II sequence variants, collected over the years by cloning various stickleback populations (Lenz and Becker, 2008; Lenz et al., 2009b) . Systematic identification of contaminants, substitution errors, PCR chimeras, and allelic drop-outs illuminates template-specific genotyping challenges. We show that optimized typing of paired gDNA and cDNA samples facilitates MHC-II allele expression analysis, hereby providing evidence of allele-specific MHC-II expression patterns. Lastly, the best-performing pipeline was incorporated into our analysis of expressed MHC-I diversity in sticklebacks, presented here for the first time using NGS technology.

1.3 Materials and methods

1.3.1 Sampling for MHC-II genotyping dataset

We generated two sets of data to serve as a benchmark for accurate pipeline evaluation at MHC-II genotyping: (i) biological samples of sticklebacks with known MHC-II genotypes, and (ii) artificial samples created from a plasmid library of isolated MHC-II alleles. Offspring of wild three-spined sticklebacks caught from the lake Großer Plöner See were bred and maintained in the aquaria rooms of the Department of Evolutionary Ecology at the Max Planck Institute for Evolutionary Biology in Plön, Germany. Stickleback breeding and husbandry was approved by the Ministry of Agriculture, Environment, and Rural Areas of the State of Schleswig-Holstein, Germany. All experiments were performed in accordance with Directive 2010/63/EU.

The fish selection aimed to include multiple different MHC-II haplotypes in our dataset (Table A.1, Supporting Information), *i.e.*, MHC-II allele combinations segregating in stable haplotype blocks due to strong linkage disequilibrium (Reusch et al., 2004; Lenz et al., 2009b) , and, as the number of alleles varies between haplotypes, for the biological samples to capture the whole range of individual MHC-II diversity (two to seven alleles) (Lenz et al., 2009b) (Table A.2, Supporting Information). 30 sticklebacks were selected for isolation of tail tissue DNA with the DNeasy Blood and Tissue kit (Qiagen), as well as for cDNA synthesis from spleen total RNA. An additional 10 and 12 sticklebacks were used only for DNA and RNA extraction, respectively. In brief, the spleen was mechanically homogenized in the TissueLyser II (Qiagen) using TRIzol (Invitrogen) and the RNeasy Micro Kit (Qiagen) following the manufacturer's instructions. RNA samples were transcribed to cDNA using the High-Capacity cDNA Reverse Transcription Kit (Thermo Fischer), and then purified with AMPure XP Beads (Beckman Coulter), following the manufacturer's instructions. Out of the 30 cDNA and gDNA samples taken from the same individuals, 24 were heterozygous. Samples derived from these 24 heterozygous individuals constitute the "paired" subset used for MHC-II allele expression analysis (see Table 1.1 for a summary of all datasets used in this study).

Additionally, a set of 18 artificial samples (hereafter 'plasmid samples') of random allele combinations were created from our comprehensive plasmid library of isolated MHC-II alleles from the local stickleback population. (Lenz et al., 2009b) Specifically, 35 different alleles from the plasmid library were used to produce three random combinations each of 3, 4, 5, 6, 8, and 10 alleles per sample, with each sample containing 0.01ng of each allele (Table A.3, Supporting Information). Overall, our MHC-II dataset thus consisted of 40 gDNA, 42 cDNA, and 18 plasmid samples, amounting to 100 samples in total (Table 1.1). To ensure a fair comparison

Table 1.1: Overview of all sample datasets used in this study.

Target	Dataset name	Total samples	Composition	Use
MHC-II	'MHC-II genotyping set'	100	- 40 genomic DNA - 42 cDNA - 18 plasmid	Evaluation of pipeline performance at MHC-II genotyping
	'"Paired" MHC-II subset'	48	- 24 gDNA - 24 cDNA	MHC-II allelic expression analysis
MHC-I	'MHC-I primer subset'	32	- 32 cDNA	Comparison of MHC-I primer pairs
	'MHC-I genotyping dataset'	185	- 185 cDNA	MHC-I genotyping & haplotype inference

of pipeline performance, gDNA and cDNA samples were balanced in terms of allele number distribution (Fig. A.1, Supporting Information).

1.3.2 Traditional approach: RSCA genotyping of MHC-II

Reliable evaluation of NGS genotyping pipelines was based on a priori knowledge of MHC-II genotypes of all individuals gained via an established RSCA protocol (Lenz et al., 2009b). This protocol (see Supporting Materials and Methods) targets all currently known stickleback MHC-II loci using established primers which amplify a 247-bp fragment of exon 2 of the MHC-IIβ genes.

1.3.3 NGS approach: Illumina library preparation and sequencing of MHC-II

After independent RSCA genotyping of our MHCI-II dataset, we developed an NGS-based genotyping protocol, based on a two-step PCR approach for NGS library preparation, as described by Cruaud et al. (2017) . To allow for the calculation of between-run genotyping repeatability, duplicates of each sample were processed separately (2x100 amplicons in total). The first PCR primers contained Illumina sequencing primers, a "heterogeneity spacer" of 0-3 bp for optimal Illumina base-calling (Fadrosh et al., 2014) , and the same MHC-IIβ-targeting primer pair used for RSCA typing. The resulting 4 forward and 4 reverse primers (Table A.4, Supporting Information) allowed 16 different primer combinations, with each being used in 6-7 different samples per library. A detailed description of the two-step PCR protocol is provided in the Supporting Materials and Methods. Following the final PCR step, one-step purification was performed separately for each sample using AMPure XP Beads. Samples were pooled together in an equimolar manner and sequenced with Illumina MiSeq sequencing at 2×150 bp read lengths using a Micro reagent kit. Generated sequences were de-multiplexed automatically and used in downstream bioinformatics analyses.

1.3.4 Genotyping pipelines and parameter permutations

Three different bioinformatics pipelines were tested with our Illumina MHC-II dataset: the DOC method, AmpliSAS, and ACACIA. While all three pipelines assume that erroneous se-

quences ('artifacts') occur less frequently than genuine variants, each one employs different clustering, filtering, and allele-calling strategies to combat their occurrence. A comprehensive overview of these steps, alongside each pipeline's user-adjustable parameters, are presented in Table 1.2. The pipeline evaluation was done by using the three sample subsets (gDNA, cDNA, and plasmid) separately, following common parameter optimization approaches, as described below. To comply with the DOC method parameter requirements, and in order to facilitate a fair comparison, a limit of 10 expected alleles for a single amplicon was imposed for all three pipelines, in accordance with the maximum number of alleles present in our plasmid samples.

The DOC method

The DOC method was one of the first pipelines to incorporate a similarity-based clustering step, during which variants differing slightly from a dominant sequence are clustered with said 'parental' sequence, if their frequency (read depth relative to total amplicon depth) is below a user-specified 'minimum dominant frequency' threshold (Table 1.2). Then, the DOC method estimates the number of true alleles in each amplicon based on the DOC (the second derivative) in the cumulative sequencing depth between depth-ranked variants. The variant with the highest DOC value is considered the last true allele. For downstream analysis, merging and quality filtering were done with the AmpliMerge and AmpliClean tools on the AmpliSAT web server (minimum average Phred score: 30, 247-253 bp sequence length). We used the automatic DOC method available in the AmpliLEGACY online tool (Biedrzycka et al., 2017) .We assessed the pipeline using a 'minimum dominant frequency' threshold of 2%, in accordance with the authors' suggestion, then proceeded to test values of 10%, 15%, 25%, 35%, 45%, and 55%, covering a range of clustering thresholds recommended and commonly employed for MHC studies (Sebastian et al., 2016; Gillingham et al., 2021).

AmpliSAS

Merging and quality filtering were done with AmpliMerge and AmpliClean, same as above. AmpliSAS implements a 'minimum dominant frequency' clustering threshold, similar in function to that of the DOC method. Following clustering, artifacts are eliminated by removing low-frequency clusters according to a number of user-selected filtering parameters, including the 'minimum per amplicon frequency' threshold; all remaining variants are considered to be true alleles (Table 1.2). We tested for the optimal 'minimum dominant frequency' threshold from a set of values ranging from 2% to 55% (default 25%) for each subset, for a given 'minimum amplicon frequency' filtering threshold (*i.e.*, 1%). The rest of the clustering parameters were used at Illumina-default values (1% substitution error rate, 0.001% indel error rate), with the chimera and singleton removal options selected. For our cDNA samples specifically, a substitution error rate of 2% was also tested, following careful evaluation of the artifacts included in the results. For each sample subset, as optimal was considered the clustering threshold at which most high-frequency errors were eliminated while all true variants retained in the final results. Next, we proceeded to test a series of 'minimum per amplicon frequency' values (0% to 5% at 0.5% intervals, default 3%) for each subset, while employing its optimal clustering threshold.

ACACIA

ACACIA utilizes an entropy-informed clustering step, followed by multiple filtering steps (Table 1.2). Comparable merging and quality control values were used as with AmpliClean: a read

overlap of 47-53bp, calculated based on the expected sequence lengths, a minimum Phred score of 30 for at least 90% of the sequence, and removal of chimeras and singletons. ACACIA also removes unspecific PCR priming products by blasting against a set of reference sequences, for which we provided a list of all known stickleback MHC-IIβ exon 2 variants. Its final filtering step incorporates two adjustable parameters: the 'absolute number of reads' (abs_nor) and the 'lowest proportion of reads' (low_por), both referring to the coverage that any variant needs to be present within an amplicon in order to be considered true. We followed the authors' recommendation by first testing various 'absolute number of reads' values (0 to 50, at 10-read intervals, with low_por set at 0), and then 'lowest proportion of reads' values, from 0% to 10% with 0.5% intervals, with the optimal abs_nor for each subset.

1.3.5 Evaluation of pipelines at MHC-II genotyping

Genotyping errors are divided into type I, which refers to the inclusion of erroneous variants in genotypes, and type II, *i.e.*, genuine alleles mistakenly excluded from genotypes. When genotyping accuracy is calculated to include both error types (*e.g.*, Gillingham et al. (2021)), no clear distinction is drawn, so valuable information regarding genotyping performance is lost. Here, we have defined genotyping accuracy, the measure of overall pipeline performance, as the average of two different metrics, specificity and sensitivity. Specificity is defined solely as the proportion of variants correctly identified as true alleles out of the total allele calls across all amplicons, thus acting as a metric of type I error occurrence. To account for type II errors (here forth referred to as "drop-outs"), we have introduced a new metric, sensitivity, defined as the inverse proportion of drop-outs over the total allele calls. Our specificity and sensitivity metrics follow the logic behind the 'false discovery' and 'false negative' rates utilized by Biedrzycka et al. (2017) . Lastly, we estimate the repeatability between duplicate samples as the ratio of all variants, genuine or not, called in both duplicates, as per previous studies (Lighten et al., 2014; Biedrzycka et al., 2017) . For a given parameter configuration (*i.e.*, genotyping run), pipeline evaluation was done on the basis of the above metrics, calculated according to the following formulas:

$$\text{Specificity}(\%) = [1 - \frac{\text{erroneous sequences}}{\text{total allele calls}}] \times 100$$

$$\text{Sensitivity}(\%) = [1 - \frac{\text{drop-outs}}{\text{total allele calls}}] \times 100$$

$$\text{Accuracy}(\%) = \frac{\text{Specificity}(\%) + \text{Sensitivity}(\%)}{2}$$

$$\text{Repeatability}(\%) = [1 - \frac{\text{different cells among duplicates}}{\text{total cells among duplicates}}] \times 100$$

For our cDNA subset, allelic drop-outs were distinguished from non-expressed alleles provided they could be i) successfully detected and ii) determined to be genuine alleles (*i.e.*, not contaminants) at lower filtering thresholds. Type I errors comprise contaminants, polymerase-generated and NGS random base-mismatch errors (here forth referred to as substitution errors), and chimeras. Contaminants are relatively low-depth variants, which often occur as the products of 'tag-jumping', *i.e.*, sequences mis-assigned to samples during demultiplexing. In order to distinguish between substitution errors and contaminants, we have taken advantage of the heterogeneity spacers in our protocol: contaminants usually bear different spacers than

Table 1.2: Detailed overview of the pre-processing, clustering, filtering and allele-calling approaches employed by each of the three genotyping pipelines included in our evaluation. Key user-adjustable parameters for each step are shown, alongside their function, default values and range of values tested here.

	Genotyping pipeline		
Steps	**Degree of Change (DOC)**	**AmpliSAS**	**ACACIA**
0. Pre-processing	Not included in core pipeline	Not included in core pipeline. AmpliMerge (for paired-end read merging) and AmpliClean (for quality filtering) are online tools available as part of AmpliSAT web server	Optional trimming of low-quality read ends, and paired-end read merging. Primer trimming and quality filtering. Singleton and chimera removal. Removal of unrelated sequences by blasting against user-provided reference library, retaining sequences of $E \leq 10^{-10}$ hits.
Key user-adjustable parameters (default, values tested)	N/A	AmpliMerge: minimum and maximum overlapping length. AmpliClean: minimum average Phred quality score (sequences with lower score are removed)	Quality control parameters p and q (default p=90, q=30): percentage p of sequence bases that have to achieve reach quality threshold q for that sequence to pass the quality filter
1. Sequence clustering Aim: Removal of erroneous sequences & addition of their coverages to true ones	Within-amplicon pairwise global alignment of sequences. Clustering of sequences differing by 1–3 bp & are below clustering threshold with parental sequences	Step-wise clustering of within-amplicon sequences. Pairwise global alignments between dominant sequence and others within cluster for error detection based on platform-specific error rates and clustering thresholds. Creation of 'consensus sequence' with most frequent nucleotide in each position	Global alignment of remaining sequences & clustering based on automated configuration of the Oligotyping tool: concatenation of high-information nucleotide positions and clustering of divergent variants using entropy information, filtering out artefacts
Key user-adjustable parameters (default, values tested)	**Minimum dominant frequency** (default 2%, range tested 2-55%): Putative erroneous sequences with same or higher frequency respect to the parental variant are assumed to be true alleles	**Substitution error rate** (Illumina default 1%, range tested 1-2%), Indel error rate (Illumina default 0.001%), **Minimum dominant frequency** (default 25%, range tested 2-55%): same as for DOC. Optional singleton removal	N/A

Detailed overview of the pre-processing, clustering, filtering and allele-calling approaches employed by each of the three genotyping pipelines included in our evaluation. Key user-adjustable parameters for each step are shown, alongside their function, default values and range of values tested here.

2. Variant filtering Aim: Separation of artefactual variants from genuine alleles	Variants within amplicon ranked by decreasing sequencing depth. Variant with highest DOC value (second derivative in cumulative depth) retained as last true allele, remaining variants discarded	Removal of non-clustered low-depth sequences and small artifactual clusters according to a number of user-defined filtering thresholds	Optional removal of allele variants identified in a single amplicon, removal of artefacts according to user-defined filtering thresholds
Key user-adjustable parameters (default, range tested): function	N/A	**Minimum per amplicon frequency** (default 3%, range tested 0-5%): Sequences with lower amplicon frequency are discarded Optional removal of chimeric, non-coding, and frame-shift sequences	**Absolute number of reads** (default 10, range tested 0-50): minimum number of sequences that need to support an allele; otherwise, it is considered an artefact **Lowest proportion of reads** (default 0%, range tested 0-10%): minimum proportion of reads within individual amplicon for an allele to be retained
3. Allele-calling Aim: Post-filtering screening of putative alleles	N/A	N/A	Putative alleles dropped if frequency less than 1/10th of next higher-ranking allele, or if 1-bp difference from more frequent allele in amplicon Amplicons with fewer than 50 sequences are eliminated
Key user-adjustable parameters	N/A	N/A	N/A

User-adjustable parameters of each pipeline which were tested in our evaluation are marked **in bold**.

true alleles within a sample, and can thus be manually identified upon inspection. Rare contaminants misassigned to a sample with identical spacers were still regarded as such, provided they could be traced back to another sample within the same column/row in the multiplexing matrix. Unlike contaminants, substitution errors emerge during PCR or NGS amplification from a within-sample parental variant and thus bear the correct spacer combination. These artifacts were easily identified as differing by 1-2 bases from the 'true' within-amplicon parental variant of higher coverage. Finally, chimeras were identified by eye when co-occurring with both parental sequences within a sample. Following their identification, we quantified the occurrence of all three error types among the total allele calls across all amplicons for all genotyping runs.

1.3.6 Assessment of genotyping error repeatability and manual allele screening

To evaluate the relationship between genotyping accuracy and repeatability, we assessed the occurrence of duplicates sharing recurring genotyping errors, *i.e.*, drop-outs and artifacts, across our gDNA, cDNA and plasmid subsets for each genotyping pipeline at optimal configuration. Next, for all pipelines, a final step of manual screening was employed, by taking the consensus genotype (*i.e.*, alleles called in both duplicates) for each sample. Post-screening, genotyping accuracy, specificity, and sensitivity were calculated as described above.

1.3.7 Pairwise comparison of genotyping agreement between pipelines

To assess the level of genotyping consistency between pipelines, we performed pairwise comparisons of results from all pipelines at their optimal parameter configuration. Agreement rates were calculated at both the allele level, *i.e.*, the proportion of common allele calls out of the total calls made by both pipelines, and at the genotype level, where agreement required the exact same alleles being called by the two pipelines for the same sample.

1.3.8 MHC-II allele amplification efficiency and expression analysis

Stochastic processes and biased amplification efficiency can result in uneven allele amplification during PCR of multilocus MHC genes (Burri et al., 2014; Sommer et al., 2013). Using sequencing data from our "paired" subset of gDNA-cDNA samples, we assessed the amplification efficiencies of MHC-II alleles. Specifically, we selected 24 heterozygous individuals whose two haplotypes shared no alleles, to ensure each allele is present in a single copy within an individual. The 24 individuals corresponded to 17 different genotypes consisting of 14 haplotypes and 29 alleles. Analyses were performed separately for gDNA and cDNA samples including both duplicates using the R script by Sommer et al. (2013), whose technique uses observed frequencies in a dataset to assign an efficiency value to each allele through a maximum likelihood optimization approach. Next, we analyzed the allele expression levels by subtracting the gDNA from the cDNA amplification efficiency values of each allele, thereby accounting for biases occurring due to random processes and/or biased amplification efficiency during PCR, as well as the allele number variation across samples. Lastly, to examine haplotype expression levels, we calculated the average allele expression level of all alleles present in each haplotype.

1.3.9 Illumina library preparation and sequencing of MHC-I

Following careful comparison of two sets of stickleback-specific MHC-I primers (see Supporting Materials and Methods), our newly designed MHC-I primers (forward 5'-GAGTCCCAAACTTCCCAGAGT-3', reverse 5'-AAACAACGCTTCAACCAAACTGGA-3'), targeting a 227-bp fragment of exon 2, were employed in the subsequent analysis of expressed MHC-I diversity using our complete MHC-I dataset. Briefly, total RNA was extracted using the RNeasy Mini Kit (Qiagen) from whole-body tissue of 185 sticklebacks belonging to 10 lab-bred families (each represented by 6-24 offspring, on average 18.5 offspring/family, SD = 6.3). RNA samples were transcribed to cDNA using the High-Capacity cDNA Reverse Transcription Kit (Thermo Fischer), and the cDNA was purified with AMPure XP Beads. Next, each cDNA sample was processed in duplicate via independent PCRs and sequencing runs, in order to allow for the calculation of between-run genotyping repeatability. A two-step PCR protocol, employing heterogeneity spacer-linked primers during the 1st PCR (Table A.5, Supporting Information), was followed for Illumina library preparation, as described in the Supporting Materials and Methods. The two duplicate libraries were sequenced separately with a MiSeq Reagent Kit v2 Micro, and reads demultiplexed automatically.

1.3.10 Sibship-assisted NGS cDNA-typing of expressed MHC-I genes

AmpliSAS, having yielded the best results with our MHC-II cDNA samples, was employed for MHC-I typing, with a 25% clustering threshold, so as to minimize similar true alleles of varying read depths getting clustered together (Biedrzycka et al., 2017) , and a low per amplicon frequency threshold of 0.5%, to account for the higher expected MHC-I allelic diversity. AmpliSAS-called MHC-I variants were only regarded as true alleles if they passed manual scrutiny (carried correct spacer combination, and not determined to be 1-2 base substitution errors from common parental variants within amplicon, a tactic successfully employed in previous studies (*e.g.*, Lighten et al. (2014); Zagalska-Neubauer et al. (2010)). This step is important for the elimination of the few remaining, low-depth erroneous variants that can pass the error correction filters, as shown in our pipeline evaluation for MHC-II. Consensus genotypes were subsequently inferred from the genotypes of both duplicates; only alleles present in both duplicates were called. In cases where seemingly correct calls were only present in the second library, said alleles were regarded as true on the assumption that the higher sequencing depth of the second library allows for more sensitive allele detection, and provided that no contaminants were found within the sample. Similarly, ambiguous, low-depth calls with seemingly correct spacers only present in first library samples alongside other contaminants were regarded as 'putative artifacts' (as contaminants bearing the correct spacer combination are rare but possible). For final genotype verification, we used allele segregation patterns within families in a manner similar to a previous study by Gaigher et al. (2016): true alleles were confirmed on the bases of agreement within sibships, and within-amplicon variant depths. High-depth variants shared by at least one sibling were regarded as genuine, while low-depth 'putative artifacts' were disregarded if they were found in a single individual within a sibship. During this process, we excluded six individuals mistakenly included in sibships (also based on MHC-II genotyping). We thus obtained high-confidence consensus genotypes for all 179 remaining individuals.

1.3.11 MHC-I haplotype and linkage inference

The observed patterns of allelic segregation within the 10 sibships (each represented by 6-24 offspring, on average 17.9 offspring/family) were used to infer potential MHC-I haplotypes. Inferred parental genotypes that could have resulted in the observed offspring genotypes were further validated with the R package MHCtools (Roved et al., 2022). Furthermore, the pattern of MHC-I allelic segregation within sibships was used as the basis for the estimation of linkage of the corresponding loci (see Gaigher et al. (2016, 2018)). We chose 8 sibships with parental genotypes inferred to be heterozygous, as the occurrence of recombination is observable in individuals possessing two different haplotypes. The occurrence of only four different allele combinations within siblings would suggest that the loci are tightly linked. By assessing the rate of recombinant offspring genotypes, we can estimate the extent to which the loci are linked.

1.3.12 Polymorphism and pattern of positive selection of expressed MHC-I genes

All 48 confirmed MHC-I alleles detected in this study were named following the established nomenclature (Klein et al., 1993) and deposited to NCBI (GenBank accession numbers PP097231-PP097278). Previously published sequences retained their given numerical suffix, with an 'X' as the locus designating letter, since alleles could not be assigned to specific loci (*e.g.*, Gaac-UX*08). Sequence diversity, including the number of polymorphic sites, the average number of nucleotide differences (k) and the average number of pairwise differences per base pair (π), was estimated with DnaSP 6 (Rozas et al., 2017). The amino acid p-distance was obtained with MEGA 11 (Tamura et al., 2021). The phylogenetic relationships between MHC-I alleles were assessed with a Neighbor-Net network using SplitsTree4 (Huson and Bryant, 2006). MHC-I sequences retrieved from the three-spined stickleback reference genome (Jones et al., 2012) were also included in the network for comparison purposes (Table A.6). To infer codon-specific footprints of positive selection, we used the maximum-likelihood site models in CodeML implemented in PAML v4.7 (Yang, 2007). The likelihood ratio test was carried out by comparing models M7 (neutral) with M8 (positive selection). When the best-fit model was M8, the Bayes Empirical Bayes (BEB) method was employed to detect sites evolving under positive selection (Zhang et al., 2005a). Input trees for the CodeML analysis were produced with MrBayes 3 (Ronquist and Huelsenbeck, 2003) using the K80+G substitution model. To guarantee that signals of selection were not affected by the tree topology, the CodeML analysis was repeated with six different trees picked at random from the posterior distribution of tree topologies.

1.3.13 Statistical analyses and software tools

All statistical tests were performed in R 4.1.2 (Team, 2021). Plots were generated using the R packages ggplot2 (Wickham, 2016) and ggstatsplot (Patil, 2021), and the software Graphpad Prism 6 for Windows (GraphPad Software, 2012).

1.4 Results

Evaluation of pipelines for NGS genotyping of MHC-II

Both duplicates were successfully sequenced for 99 out of 100 samples in our MHC-II dataset, achieving an average final sample coverage of over 15,000 reads for all three groups -gDNA, cDNA, and plasmid (Table A.7, Supporting Information). We proceeded to evaluate the DOC method, AmpliSAS, and ACACIA with each group separately, based on a priori knowledge of MHC-II genotypes obtained through RSCA (see Supporting Results, Fig. A.2).

The DOC method

Optimal results were obtained using a 35% clustering threshold for gDNA and plasmid samples, and 2% for cDNA samples (Fig. A.3, Supporting Information); the DOC method reached high genotyping specificity across groups (96.3-99%), performing best with the gDNA subset (95.5% sensitivity, 95% repeatability). Genotyping sensitivity of cDNA and plasmid samples was markedly lower; 81.5% and 88.8%, respectively. Drop-out occurrence was positively correlated with increasing allele numbers in both cDNA and plasmid samples (Spearman's R = 0.61, P <0.0001 and R = 0.43, P <0.0083, respectively). Despite the frequent drop-outs, repeatability for cDNA and plasmid samples remained high (93.3% and 99%, respectively). Notably, chimeras were included as genuine alleles in genotypes obtained with the DOC method for cDNA and plasmid -but not gDNA- samples (Fig. A.3, Supporting Information). Overall, the DOC method was most accurate for gDNA samples (97.3%), followed by plasmid (93.9%), and cDNA samples (88.9%).

AmpliSAS

With AmpliSAS, a 35% clustering threshold yielded the best results for gDNA and plasmid samples, and 45% for cDNA samples. These values were employed for subsequent benchmarking of the 'minimum per amplicon frequency' threshold. At low filtering thresholds, a high number of artifacts were included in the gDNA results (Fig. 1.1A), the majority of which were contaminants. The number of contaminants was negatively correlated with sample depth (Spearman's correlation coefficient: R = -0.50, P <0.0001). With gradually increasing cut-off values, the presence of artifacts progressively declined, until the pipeline reached optimal performance. Notably, at optimal configurations, contaminants still represented 1.4 to 2.6% of total allele calls in final genotypes across pipelines. Beyond the optimal point, further threshold increases resulted in allele drop-outs. With gDNA, AmpliSAS reached 98.5% accuracy (97% specificity, 100% sensitivity), and 98.8% repeatability at a 3% filtering threshold, averaging 0.13 cumulative genotyping errors per sample (Fig. 1.1A), corresponding to nine contaminants and one substitution error (Fig. 1.2A). Regarding the plasmid samples, AmpliSAS optimally achieved 99.3% accuracy at a 1.5% threshold, yielding fewer drop-outs than the DOC method (Fig. A.4, Supporting Information).

Notably, cDNA-typing exhibited worse sensitivity at gDNA-optimal clustering and filtering configurations (Fig. 1.1C), as drop-outs occurred more frequently compared to the gDNA subset. Lowering of the filtering threshold improved sensitivity, but allowed for more erroneous sequences; accuracy peaked at 96.5% with 0.30 total errors per sample at a 1.5% filtering threshold at gDNA-optimal clustering configurations (Fig. 1.1C). Implementation of the optimal 45% clustering threshold, in combination with further cDNA-tailored adjustment of the substitution error rate to 2%, led to fewer average per sample errors (0.20), with an overall

improved optimal accuracy of 97.5% (97.7% specificity, 97.4% sensitivity), and 96.8% repeatability (Fig. 1.1E). This corresponded to a total of seven contaminants, one substitution error, and nine drop-outs, all occurring in samples containing six or seven alleles (Fig. 1.2B).

ACACIA

With ACACIA, 'absolute number of reads' thresholds of 20 reads for gDNA and cDNA, and 30 reads for plasmid samples yielded the best results, and were employed for benchmarking of the downstream filtering threshold ('percentage of reads'). Similar to AmpliSAS, at low filtering thresholds a high number of artifacts were included in gDNA (Fig. 1.1B) and cDNA results (Fig. 1.1D). For gDNA, ACACIA reached a peak of 98.3% accuracy (97.8% specificity and 98.7% sensitivity), and 98.4% repeatability, averaging 0.14 cumulative errors per sample (Fig. 1.1B), performing at similar levels as AmpliSAS. For plasmid samples, ACACIA achieved 99.1% accuracy (98.2% specificity, 100% sensitivity) at optimal settings (Fig. A.4, Supporting Information). Finally, for cDNA samples, ACACIA yielded an optimum of 92.4% accuracy, with an average of 0.69 total errors per sample (Fig. 1.1D).

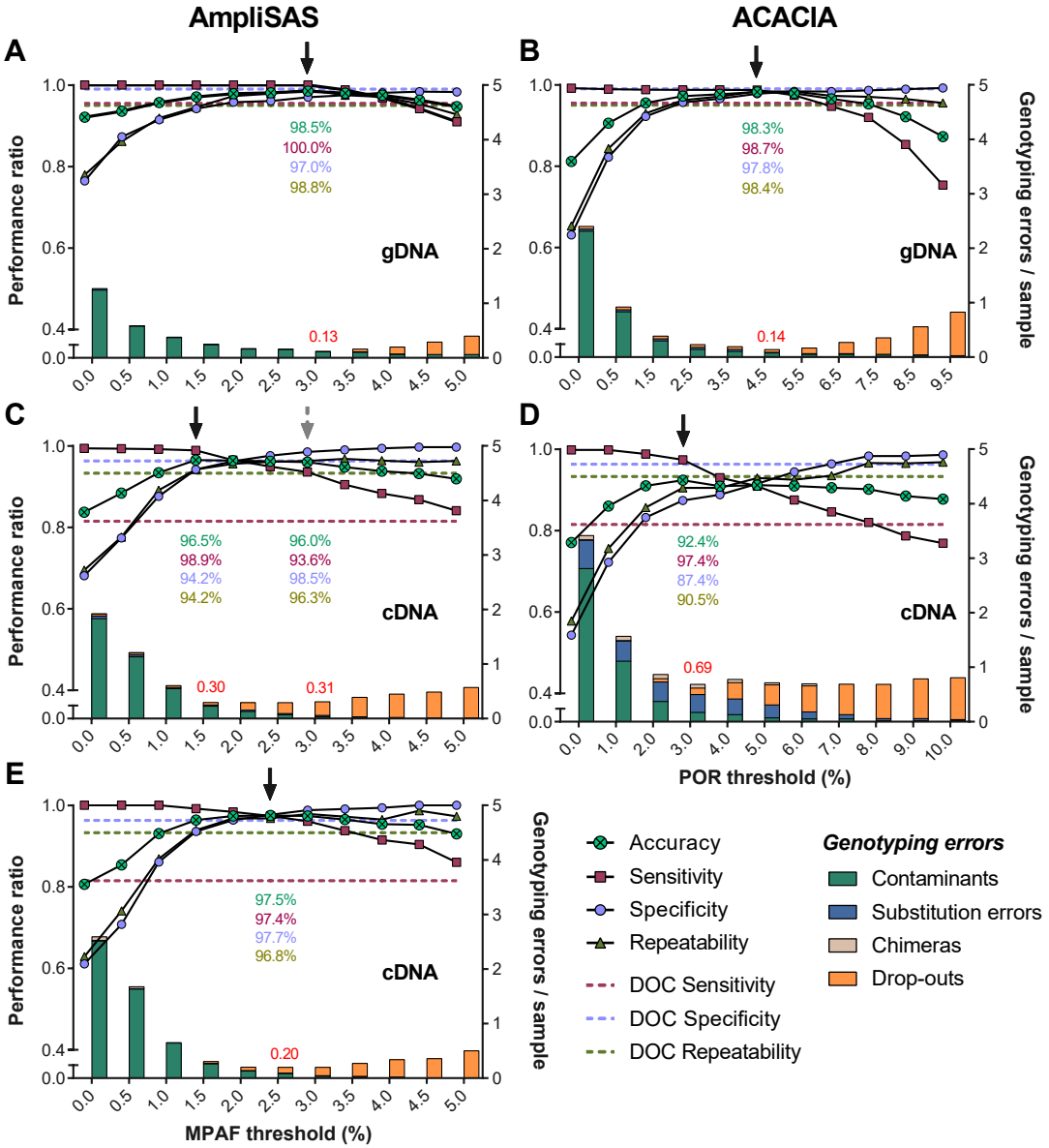


Figure 1.1: Evaluation of next-generation sequencing bioinformatics pipelines at genotyping three-spined stickleback MHC-IIβ exon 2 using samples of genomic DNA and cDNA. Performance of AmpliSAS with the **(A)** gDNA and **(C)** cDNA datasets at different filtering threshold settings with minimum dominant frequency parameter set at 35% and substitution rate at 1%. Performance of AmpliSAS with the **(E)** cDNA dataset at different filtering threshold settings with minimum dominant frequency at 45% and substitution rate at 2%. Performance of ACACIA with the **(B)** gDNA and **(D)** cDNA datasets at different filtering threshold settings with "abs_nor" parameter set at 20. The left y axis depicts the ratio of performance parameter (sensitivity, specificity, repeatability), and the right y axis the average per sample number of genotyping errors. Optimal filtering thresholds for a given configuration are pinpointed with a black arrow; corresponding performance percentages are shown in matching colors, and average number of genotyping errors per sample in red. The grey arrow in (C) pinpoints the filtering threshold that yielded optimal gDNA results. The DOC method's optimum performance for each dataset is depicted as horizontal dotted lines for comparison. POR: percentage of reads filtering threshold of ACACIA. MPAF: Minimum per amplicon frequency filtering threshold of AmpliSAS. Note the truncated y-axis scale.

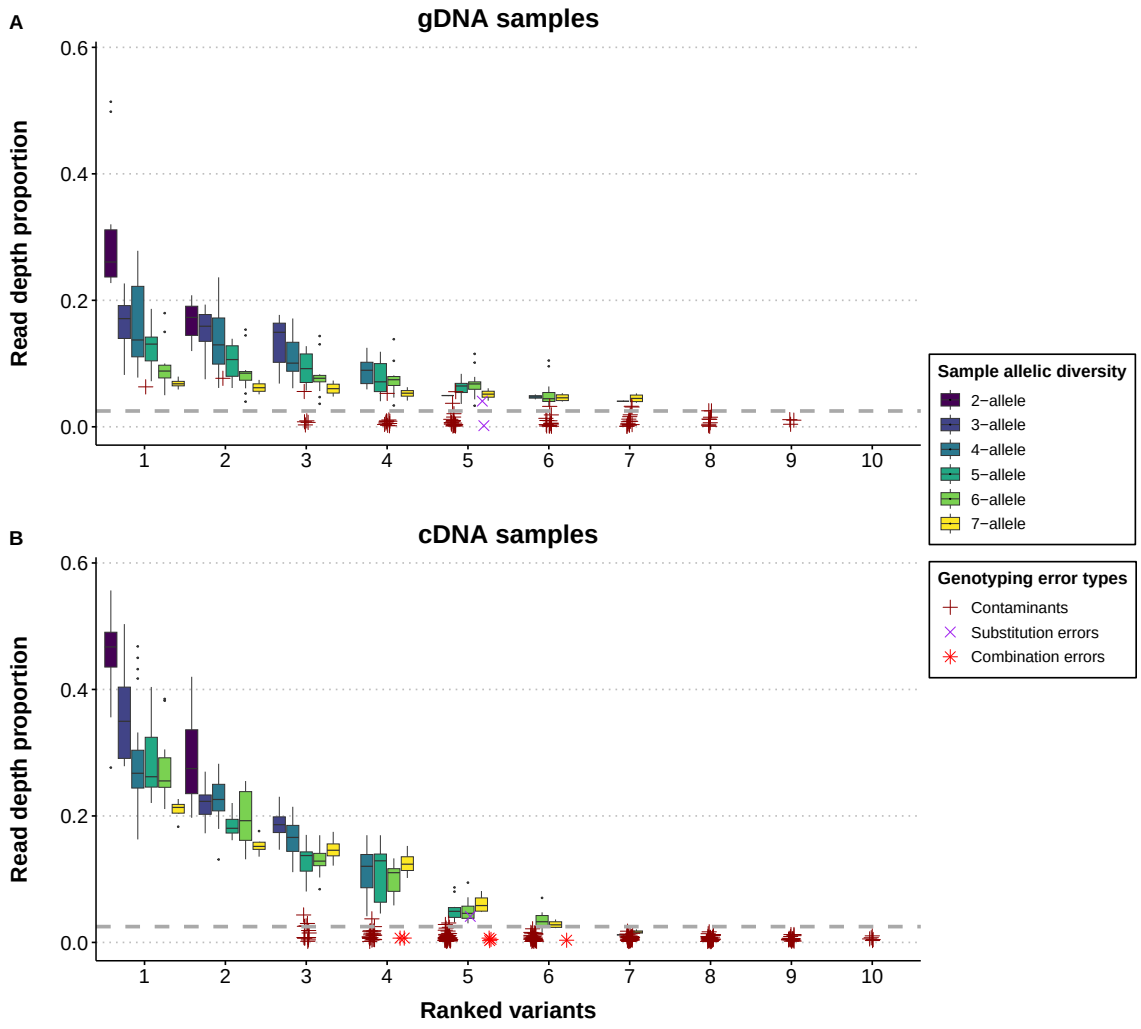


Figure 1.2: Relationship between variant read depth and allele number in a sample. Shown are the within-amplicon read depth proportions of all detected sequence variants (true alleles and genotyping errors) in all the **A)** gDNA and **B)** cDNA duplicate samples of our MHC-II dataset. Boxes represent true alleles, and dots represent genotyping errors. All variants are ranked from highest to lowest depth with regards to the total sample depth. Samples are grouped and color-coded according to the number of true alleles. Shown is the proportion of read depth of each variant following AmpliSAS genotyping using template-specific optimal clustering and filtering settings (see main text for details). Horizontal dashed lines represent the optimal minimum per amplicon frequency filtering threshold for each subset. Combination errors refer to chimeric artifacts in the cDNA subset which also bear 2 bp substitutions from the parental variants.

1.4.1 Assessment of genotyping error repeatability and manual allele screening

At optimal pipeline configurations, the occurrence of gDNA and plasmid duplicates sharing genotyping errors was low (1-2 samples per subset); conversely, more cDNA duplicates included recurring errors, mostly drop-outs (Fig. A.5, Supporting Information). Manual screening, whereby only duplicate-shared alleles were retained, improved genotyping specificity across sample subsets via elimination of non-recurring artifacts - especially in cDNA samples (Fig. A.5, A.6, Supporting Information). Post-screening sensitivity was largely unaffected for gDNA and plasmid samples, whereas drops were observed both for cDNA samples across pipelines, and for gDNA samples called by the DOC method, due to the presence of non-recurring drop-outs (Fig. A.6, Supporting Information). Overall, manual scrutiny improved the genotyping accuracy of AmpliSAS and ACACIA for all sample subsets (Table 1.3).

Table 1.3: Genotyping accuracy of the three pipelines at optimal parameter configurations before and after manual screening of final allele calls.

	gDNA		cDNA		Plasmid	
	Before	After	Before	After	Before	After
DOC	97.3	94.8	88.9	88.5	93.9	94.3
AmpliSAS	98.5	99.1	97.5	98.2	99.3	99.5
ACACIA	98.3	99.0	92.4	96.5	99.1	99.1

1.4.2 Pairwise comparison of genotyping agreement between pipelines

Assessment of the genotyping consistency across pipelines revealed that at the allele level, the highest agreement rates were achieved between AmpliSAS and ACACIA across all sample sets, ranging from 92.7% for the cDNA to 98.9% for the plasmid samples (Fig. 1.3). The agreement rates between the DOC method and the other two pipelines were lower, with most disagreements occurring due to the DOC method excluding genuine alleles in one or both duplicates. On the genotype level, however, the DOC method and AmpliSAS produced the highest agreement rates for both gDNA and cDNA samples, while for the plasmid samples, the results of ACACIA and AmpliSAS were the closest ones. Notably, the lowest pipeline agreement rates were observed with the cDNA dataset at both the allele and the genotype levels (Fig. 1.3).

1.4.3 MHC-II allele-specific expression analysis

Having observed decreased cDNA-typing sensitivity across all three pipelines, we subsequently examined the variation in sequencing depth between alleles from gDNA and cDNA samples. We used 24 duplicated pairs derived from heterozygous individuals; our "paired" dataset included 17 different genotypes, corresponding to 14 haplotypes and 29 alleles. cDNA alleles exhibited a significantly higher variance of within-amplicon read depth proportions compared to corresponding gDNA alleles (Levene's test, P = 0.0001236, Fig.A.7, Supporting Information). Specifically, alleles DXB*01 and DXB*27 show significantly (P<0.01) higher proportions, whereas DXB*28 and DXB*31 significantly lower ones (Fig. A.8A, Supporting Information). Indeed, DXB*28 and DXB*31 were the most frequently dropped out in our pipeline evaluation

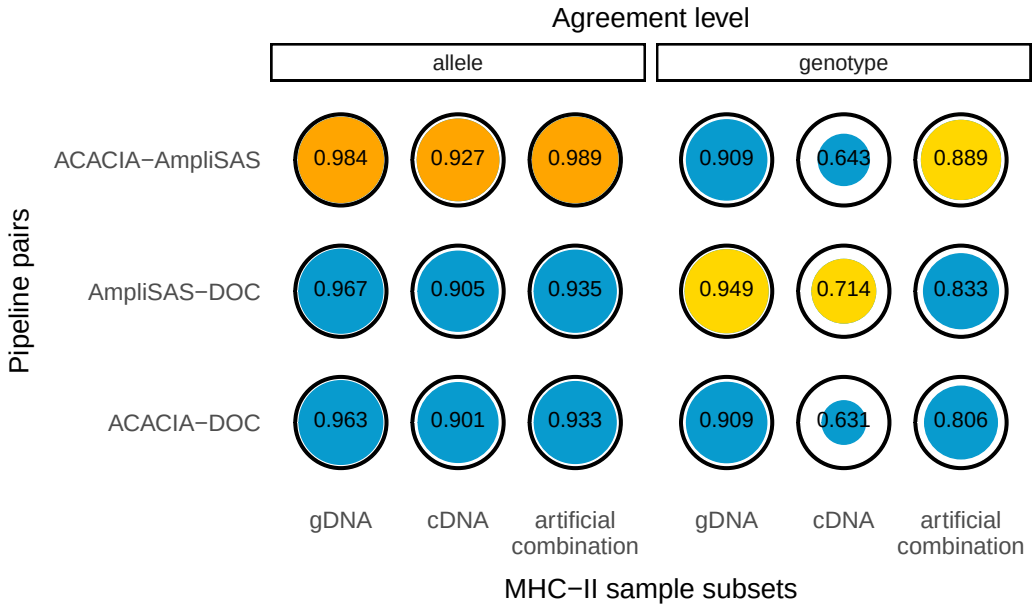


Figure 1.3: Pairwise agreement between the ACACIA, AmpliSAS and the DOC method at MHC-II genotyping. Pairwise genotyping agreement rates for each of the gDNA, cDNA and plasmid sample subsets in our MHC-II genotyping dataset were calculated at optimal parameter configuration and at both the allele and the genotype levels, with diameters of coloured circles being proportional to the corresponding agreement rates. The highest agreement rates for each dataset are highlighted in orange (allele level) and yellow (genotype level).

of cDNA samples (corresponding to all nine allele drop-outs at optimal AmpliSAS configuration). Calculation of allele-specific gDNA and cDNA amplification efficiencies confirmed the low variation in gDNA samples, with a ratio of maximum to minimum amplification efficiency of 3.1 (1.62 for DXB*23 to 0.53 for DXB*07, Fig. 1.4A). All gDNA amplification efficiencies were below two, confirming that all alleles in our dataset were present in single copies within individuals. In contrast, the cDNA dataset was characterized by a five-times higher ratio of 15.6 (3.90 for DXB*23 to 0.25 for DXB*31 of haplotype W, Fig. 1.4A). Even after the exclusion of outlier allele DXB*23, cDNA amplification efficiencies maintained a ratio of 12.7, four-times higher than that of the gDNA ones. Remarkably, three out of the five alleles with the highest expression levels, and four out of the five alleles with the lowest, belong to three haplotypes in our dataset containing an above-average three or four loci (II-O and II-G, and II-W, Fig. 1.4B). Haplotype II-W, in particular, contains both a highly expressed (DXB*27) and two lowly expressed alleles (DXB*28 and DXB*31). Allele DXB*01, exhibiting the third highest expression level, belongs to the only single-locus haplotype in our dataset, II-F (Fig. 1.4B, A.8B). The three alleles in our dataset which are shared by two different haplotypes seem to retain their expression pattern across haplotypes; DXB*28 and DXB*31 show suppressed, and DXB*27 enhanced expression levels (Fig. A.9, Supporting Information). At the haplotype level, there is no correlation between the number of alleles within haplotypes and their average gDNA amplification efficiency (Fig. 1.4C). Conversely, we observe a negative trend between haplotype allele number and average MHC-II expression level (Spearman's R = -0.48, P = 0.08); the correlation becomes significant (Spearman's R = -0.71, P = 0.007) after excluding haplotype II-V, expression of which is driven by the outlier allele DXB*23 (Fig. 1.4D).

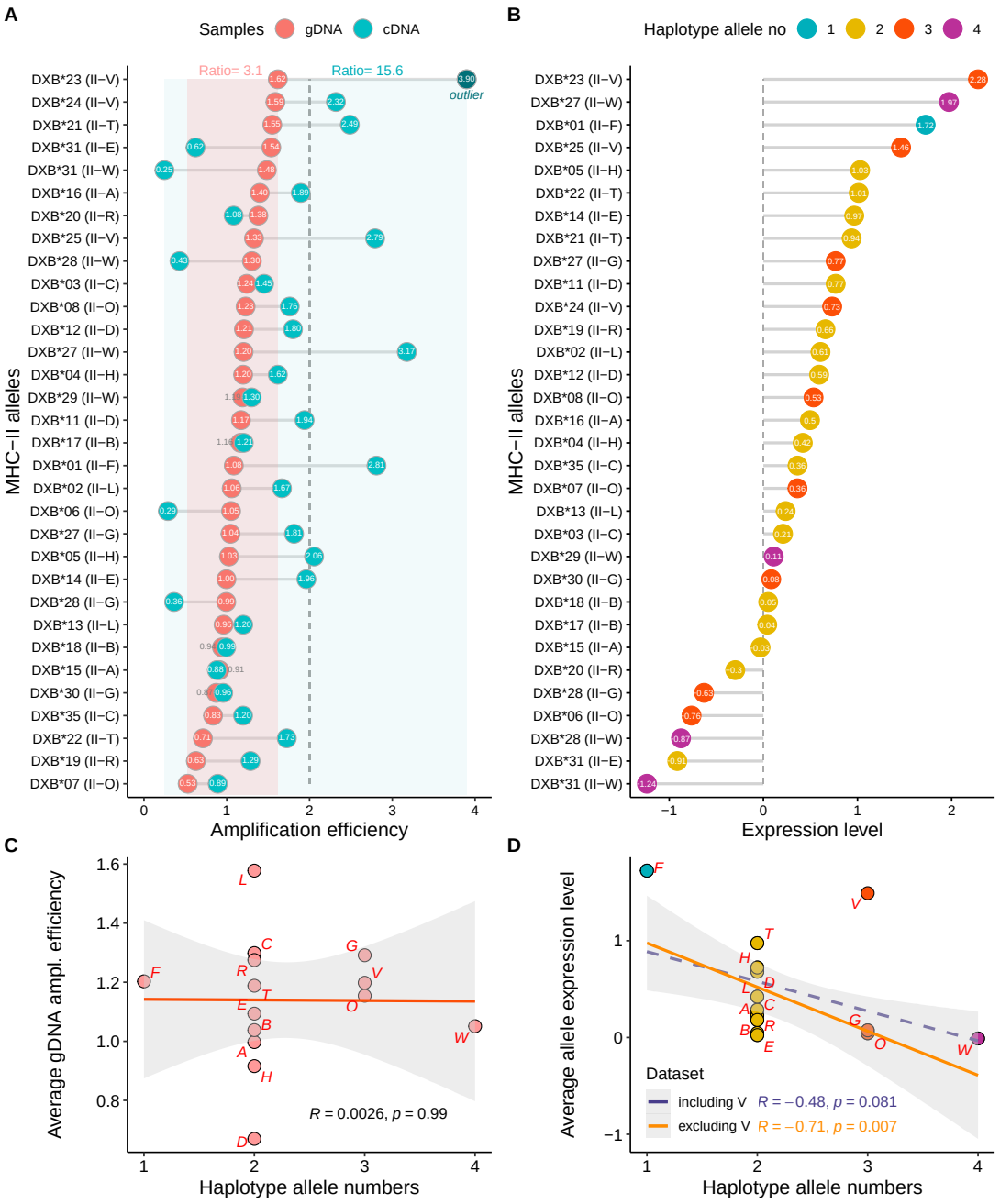


Figure 1.4: Expression analysis of MHC-II alleles and haplotypes. A) Relative allele amplification efficiencies of gDNA and cDNA samples in our MHC-II paired sample subset. The name of the corresponding haplotype for each allele is mentioned in parenthesis. The dashed vertical line represents an efficiency of two, which in the gDNA samples would represent duplicated or homozygous alleles. Allele Gaac-DXB*23 is marked as an outlier among cDNA amplification efficiency values. Red and blue shaded areas represent the range in allele amplification efficiencies of gDNA and cDNA samples, respectively. **B)** Expression level of each allele, calculated as the difference between gDNA and cDNA amplification efficiency values, and grouped according to number of alleles within the corresponding haplotype where each allele is found. **C)** Relationship between number of MHC-II alleles per haplotype and average amplification efficiency of corresponding gDNA alleles. 95% confidence intervals are indicated by the shaded area. **D)** Relationship between number of MHC-II alleles per haplotype and average haplotype expression level, calculated as the mean difference between the cDNA and gDNA amplification efficiency values of all alleles within each haplotype. 95% confidence intervals are indicated by the shaded area.

1.4.4 Sibship-assisted NGS typing of expressed MHC-I genes

Using our newly designed primer pair, which better captures the MHC-I diversity (see Supporting Information), we proceeded to sequence a duplicated cDNA library of a total of 185 individuals from 10 sibships. Both duplicates were successfully acquired for 179 out of the 185 individuals, with an average depth of 12,385±5,892 reads (Table A.8, Supporting Information). Despite duplicates differing in depth by an average 10,637 ± 9,086 reads, the mean difference in relative coverage of the same alleles between the two duplicates was only 1.1% (SD: 1.7%). On a par with our previous evaluation, 97.6% of total allele calls were identical between our 179 duplicates, confirming good genotyping repeatability. For six individuals, only a single duplicate was sequenced sufficiently, however, their high sequencing depth (14,853 ± 7,349 reads) allowed for unambiguous genotyping. AmpliSAS genotypes included some manually-identified erroneous sequences; of the total 56 variants called, 48 were attributed to true alleles after visual inspection. The remaining variants predominantly represented single-base substitution errors and contaminants. Indeed, the lowest relative depth of an assumed genuine allele was 0.006, while the highest among substitution errors and contaminants was 0.016 and 0.018, respectively. Of the total 48 unique MHC-I allelic variants, 19 have previously been published, while 18 have been previously detected in our laboratory using the old primer pair via cloning (unpublished data, Tobias Lenz), or by others via Illumina sequencing (data summarized in Table A.9, Supporting Information). The number of expressed MHC-I alleles per individual ranges between 5 and 23, with a mean of 12.8 (SD = 4) (Fig. A.10A, Supporting Information).

1.4.5 MHC-I haplotype and linkage inference

Segregation analysis of the MHC-I alleles within sibships for each of the 10 families (6-24 offspring per family) confirmed the anticipated pattern of Mendelian inheritance. As expected for parental fish with distinct genotypes, four different genotypes were detected in seven of the ten families (Table A.10, Supporting Information); an example is shown in Fig. 1.5A. In two families, only two offspring genotypes were detected, indicative of a homozygous state for one of the parents. Lastly, the offspring of one family exhibited three different genotypes, which is consistent with parents sharing a heterozygous genotype. By observing which alleles appear to be inherited together within a kinship we were able to decipher ten distinct MHC I haplotypes, with each containing 5-12 alleles (Table 1.4). Haplotypes frequently shared alleles; 14 out of the 48 alleles (29.2%) were found in more than one haplotype, with the most shared allele being present in six haplotypes (Fig. A.10B, Supporting Information). For comparison, allele sharing across MHC-II haplotypes from the same individuals was limited to only two out of the 28 alleles (7.1%), both shared between two haplotypes from a total of 13 (Fig. A.10B, Supporting Information). Of the 328 observed gametes (164 offspring) in the eight sibships with heterozygous parents, five (1.52%) appeared to be parental recombinants (example shown in Fig. 1.5B), suggesting that the linkage group encompassing the expressed MHC class I loci of the three-spined stickleback spans >1 cM.

Table 1.4: Ten three-spined stickleback MHC-I haplotypes and corresponding alleles inferred from allele segregation analysis of NGS data from ten sibships.

Haplotype	Sibships	Allele number	Alleles
I-A	2	5	Gaac-UX*18, 44, 51, 55, 56
I-B	1	6	Gaac-UX*10, 19, 27, 52, 53, 54
I-C	8	7	Gaac-UX*08, 10, 17, 18, 19, 27, 52
I-D	3	8	Gaac-UX*18, 19, 27, 40, 44, 46, 51, 57
I-E	2	12	Gaac-UX*18, 19, 44, 47, 51, 62, 63, 64, 65, 66, 67, 74
I-F	3	9	Gaac-UX*08, 10, 17, 58, 59, 60, 61, 66, 73
I-G	4	12	Gaac-UX*13, 14, 15, 16, 18, 21, 24, 27, 41, 46, 50, 68
I-H	1	7-9	Gaac-UX*11, (18), 22, 26, (27), 42, 69, 70, 71
I-I	1	6	Gaac-UX*08, 09, 19, 20, 27, 52
I-J	2	5-6	Gaac-UX*11, 22, 26, 27, 52, (72)

Alleles whose presence in haplotypes could not be confidently determined are shown inside parentheses.

A

Family 4	Inferred genotype of parent 1													Inferred genotype of parent 2												
Haplotypes	I-A					I-D								I-C							I-I					
Alleles (UX*)	**18**	**44**	**51**	55	56	**18**	19	27	40	**44**	46	**51**	57	**08**	10	17	18	**19**	**27**	**52**	**08**	09	**19**	20	**27**	**52**
	X	X	X	X	X									X	X	X	X	X	X	X						
	X	X	X	X	X									X	X	X	X	X	X	X						
						X	X	X	X	X	X	X	X								X	X	X	X	X	X
						X	X	X	X	X	X	X	X								X	X	X	X	X	X
						X	X	X	X	X	X	X	X								X	X	X	X	X	X
	X	X	X	X	X																X	X	X	X	X	X
	X	X	X	X	X																X	X	X	X	X	X
	X	X	X	X	X																X	X	X	X	X	X
	X	X	X	X	X																X	X	X	X	X	X
	X	X	X	X	X																X	X	X	X	X	X
	X	X	X	X	X																X	X	X	X	X	X
Offspring genotypes	X	X	X	X	X																X	X	X	X	X	X
	X	X	X	X	X																X	X	X	X	X	X
	X	X	X	X	X																X	X	X	X	X	X
	X	X	X	X	X																X	X	X	X	X	X
	X	X	X	X	X																X	X	X	X	X	X
						X	X	X	X	X	X	X	X	X	X	X	X	X	X	X						
						X	X	X	X	X	X	X	X	X	X	X	X	X	X	X						
						X	X	X	X	X	X	X	X	X	X	X	X	X	X	X						
						X	X	X	X	X	X	X	X	X	X	X	X	X	X	X						
						X	X	X	X	X	X	X	X	X	X	X	X	X	X	X						
						X	X	X	X	X	X	X	X	X	X	X	X	X	X	X						
						X	X	X	X	X	X	X	X	X	X	X	X	X	X	X						
						X	X	X	X	X	X	X	X	X	X	X	X	X	X	X						

B

Family 5	Inferred genotype of parent 1												Inferred genotype of parent 2														
Haplotypes	I-A					I-C							I-F									I-I					
Alleles (UX*)	**18**	44	51	55	56	08	10	17	**18**	19	27	52	**08**	10	17	58	59	60	61	66	73	**08**	09	19	20	27	52
	X	X	X	X	X								X	X	X	X	X	X	X	X	X						
	X	X	X	X	X								X	X	X	X	X	X	X	X	X						
	X	X	X	X	X								X	X	X	X	X	X	X	X	X						
	X	X	X	X	X								X	X	X	X	X	X	X	X	X						
						X	X	X	X	X	X	X										X	X	X	X	X	X
						X	X	X	X	X	X	X										X	X	X	X	X	X
						X	X	X	X	X	X	X										X	X	X	X	X	X
						X	X	X	X	X	X	X										X	X	X	X	X	X
						X	X	X	X	X	X	X										X	X	X	X	X	X
Offspring genotypes						X	X	X	X	X	X	X										X	X	X	X	X	X
						X	X	X	X	X	X	X										X	X	X	X	X	X
						X	X	X	X	X	X	X										X	X	X	X	X	X
						X	X	X	X	X	X	X										X	X	X	X	X	X
						X	X	X	X	X	X	X										X	X	X	X	X	X
	X	X	X	X	X																	X	X	X	X	X	X
	X	X	X	X	X																	X	X	X	X	X	X
						X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X						
						X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X						
	X	X	X			X	X	X	X	X	X	X										X	X	X	X	X	X
	X			X	X	X	X	X	X	X	X	X										X	X	X	X	X	X

Figure 1.5: Inference of genetic linkage between MHC-I loci of the three-spined stickleback using sibship data. A) Shown are two examples in which the inferred parental genotypes consist of different haplotypes, with shared alleles between parents indicated in bold. The presence and absence of individual alleles in the offspring is designated by grey and white backgrounds, respectively. **A)** Family 4, as an example of a sibship with only four genotypic combinations observed in the offspring. **B)** Family 5, as an example of a family in which recombination has been detected in two offspring individuals. Black backgrounds indicate recombinant haplotypes, with diagonal-filled backgrounds designating alleles whose presence or absence cannot be determined in recombinant haplotypes due to them being present on the haplotype from the other parent. Note that the order of the allele names in this table is purely numerical and does not reflect the order of the MHC-I gene copies on the chromosome, which is unknown.

1.4.6 MHC-I allelic diversity and signatures of selection

Sequence analysis of the 48 confirmed MHC-I alleles revealed a high level of genetic diversity (Table 1.5). Pairwise comparisons between alleles revealed a wide range of sequence divergence, with alleles differing by one to 45 nucleotides (mean difference: 23.43 ± 6.39 nucleotides) (shown as p-distances in Fig. A.11, Supporting Information). However, the phylogenetic networks revealed no clear genetic clusters, as neither the alleles from our European individuals nor the ones from the Alaskan stickleback genome formed distinct clusters (Fig. 1.6). Furthermore, evidence of positive selection was found at 21 amino acid sites (35% of the entire sequence), most of which are residues associated with the peptide binding domain, based on human structure (Bjorkman et al., 1987; Wallny et al., 2006) (Fig. A.12, Supporting Information).

Table 1.5: Genetic diversity at the MHC-I genes of the three-spined stickleback.

	Number of alleles	Number of sites	S	k	π (S.D.)	AA distance (S.E.)
MHC-I	48	182	83	23.43	0.129 (0.005)	0.285 (0.038)

S, number of polymorphic sites; k, average number of nucleotide differences; π, average number of pairwise differences per base pair; AA distance, amino acid pairwise p-distance.

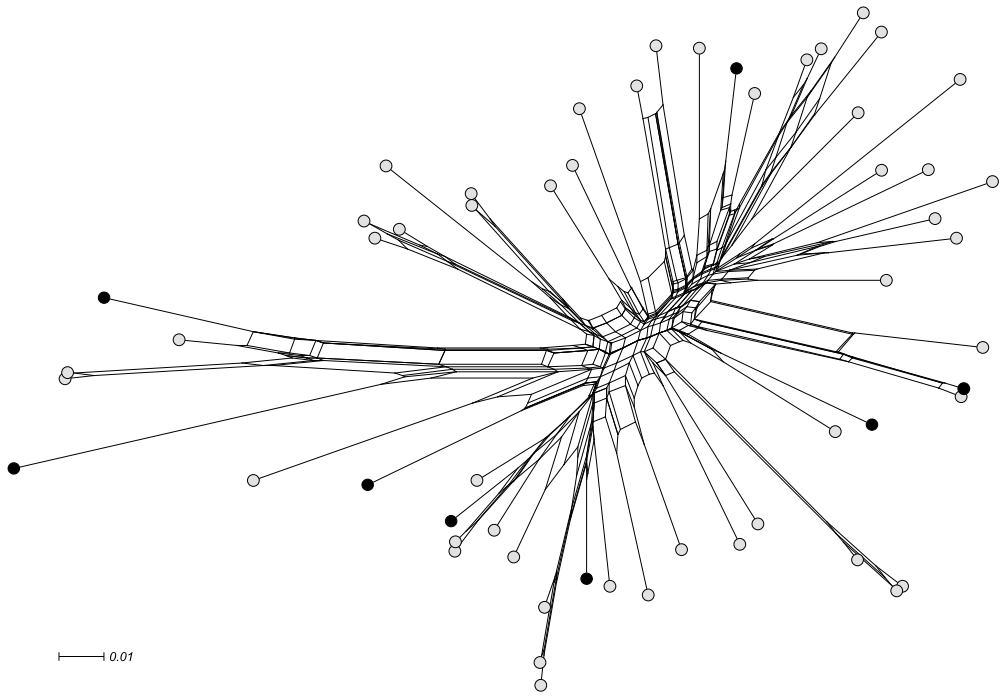


Figure 1.6: Phylogenetic network of MHC-I nucleotide sequences. Shown are the 48 alleles detected in this study (grey dots), and the eight MHC-I sequence variants from the stickleback reference genome (black dots).

1.5 Discussion

The present study illustrates how template-informed optimization of pipeline settings enables highly reliable genotyping of complex multigenic families, such as the MHC, even at the face of increased allelic complexity, where traditional methods, such as RSCA, fail. Furthermore, we show that optimized genotyping of paired gDNA and cDNA samples allows for accurate expression analysis of individual alleles and haplotypes. To ensure our results, including outcomes of the genotyping pipeline evaluation, can be transferable to other MHC systems, in addition to three-spined sticklebacks, we have utilized plasmid samples containing 3-10 alleles each; this range covers the allelic diversity that researchers often encounter in multi-locus genotyping datasets from many non-model species; from cartilaginous fish (Gaigher et al., 2023)and other teleosts (Lighten et al., 2014), to mammals (Sommer et al., 2013). Nevertheless, this range might not cover the extreme diversity observed in some species, *e.g.*, some passerine birds. We in turn discuss our results and hope to provide take-away messages regarding accurate MHC genotyping which can applied by researchers working with a plethora of diverse species in downstream evolutionary ecology analyses.

1.5.1 Pipeline clustering and filtering adjustability is key for genotyping error correction

At optimal parameter configuration, the DOC method, AmpliSAS, and ACACIA performed satisfactorily at distinguishing true alleles from artifacts in our gDNA and plasmid samples. A common characteristic of all three pipelines is the incorporation of a clustering step, whereby divergent variants join their putative parental sequences. Indeed, increasing the main clustering threshold of AmpliSAS and the DOC method (*i.e.,* minimum dominant frequency) from its default value facilitated the clustering of a few remaining high-depth 1-2 bp substitution errors, thus enhancing genotyping specificity. Such artifacts are probably formed during early PCR cycles and become abundant in the final product (Kanagawa, 2003). Compared with AmpliSAS and the DOC method, ACACIA performed equally well in genotyping the gDNA and plasmid samples, despite lacking adjustability in its automated, entropy-based clustering tool (Gillingham et al., 2021).

For AmpliSAS and ACACIA, persistent artifacts beyond clustering algorithm capacities, mostly contaminants, can subsequently be eliminated via fine-tuning of the filtering threshold, *i.e.*, 'minimum per amplicon frequency' and 'lowest proportion of reads', respectively. Overall, pipeline optimization was highly effective in eliminating artifacts from gDNA genotypes. Regardless of the pipeline, no chimeras were included in the genotypes of gDNA samples; this could be partly due to the intentionally few PCR cycles used in our protocol (Lenz and Becker, 2008). Accordingly, cDNA-typing results included chimeras, as cDNA samples underwent extra PCR cycles. AmpliSAS proved better than ACACIA at chimera removal from the cDNA subset, despite both pipelines incorporating dedicated chimera-removing steps employing modified versions of external software. On the other hand, the DOC method, lacking a dedicated removal tool, was more prone to their inclusion, evident in both the plasmid and cDNA subsets. Furthermore, the DOC method was particularly vulnerable to drop-outs, a shortcoming stemming from the existence of alleles with differential amplification efficiencies within amplicons (Biedrzycka et al., 2017).

Despite its limitations, the DOC method outperformed the other two pipelines at gDNA sample genotyping for most filtering thresholds tested. However, AmpliSAS and ACACIA fine-tuning can yield strikingly improved results. At optimal settings, these pipelines achieved an

accuracy of over 97% with gDNA samples, in agreement with previous evaluations (Biedrzycka et al., 2017; Gillingham et al., 2021). We thus join the AmpliSAS and ACACIA authors in encouraging researchers to dedicate time to discover the optimal configuration for their datasets, as these will vary depending on factors such as the per-sample allelic diversity and sequence diversity of target amplicons. Otherwise, the DOC method might offer more reliable results for gDNA samples, at least for MHC systems not expected to exceed an individual diversity of five or six alleles. Previous studies have highlighted the risk of relying on allele repeatability to validate genotyping pipelines (Biedrzycka et al., 2017; Gillingham et al., 2021). Here, due to the relative low frequency of within-duplicate recurring errors, accuracy and repeatability were on a par for our gDNA samples. This assumption lies behind the practice of eliminating alleles not shared by duplicates, a final step often employed in genotyping studies of species of unknown MHC complexity (*e.g.*, Rekdal et al. (2018)). The degree to which this manual screening approach can improve genotyping performance depends on the balance between non-recurring artifacts (elimination of which improves specificity) and non-recurring drop-outs, which decrease sensitivity. As such, we show that AmpliSAS and ACACIA genotyping of gDNA samples was further improved. In agreement with previous findings (Biedrzycka et al., 2017), post-screening results of duplicates lacking non-recurring errors, such as our plasmid samples, remained mostly the same.

1.5.2 Reliable cDNA-typing requires template-specific pipeline optimization

We show that, besides the sequencing platform and study system, optimal pipeline configuration also depends upon the sample template (*i.e.*, genomic DNA, cDNA, plasmid DNA), with cDNA samples, in particular, posing unique challenges. As artifacts in our dataset frequently exceeded 1% read depths, distinguishing them from low-frequency alleles in high-diversity samples became complicated. This was not a serious issue with our gDNA and plasmid samples, as the variation in amplification efficiency rates was quite limited in our MHC system (ratio of maximum to minimum amplification efficiency of 3.1), compared to others (ratio of 12.6 reported by Sommer et al. (2013) , 21.0 to 22.7 reported by Biedrzycka et al. (2017) . However, due to wider variation in our cDNA samples, the inclusion of low-depth alleles required lower filtering thresholds, in turn allowing high-depth artifacts into the final allele calls. Hence, our cDNA dataset violated the fundamental assumption that artifacts are less common than real alleles (Babik et al., 2009; Galan et al., 2010). The DOC method exhibited the lowest sensitivity rate as a consequence of gaps in depth between high- and low-depth alleles.

cDNA datasets posed a challenge for AmpliSAS and ACACIA, too, as evidenced by their overall reduced performance when used at gDNA-optimal configurations. A good indication that variability in allele expression levels could explain the observed sub-par genotyping is that the same alleles, namely DXB*28 and DXB*31, tend to be dropped out. We also hypothesize that the stochasticity of RNA sub-sampling, coupled with reverse transcription, an additional step where errors can be introduced (Arezi and Hogrefe, 2007)further complicates artifact distinction. The need for additional PCR cycles for adequate amplification of target RNA-derived sequences can also lead to more artifacts (Lenz and Becker, 2008). Such was the case here for MHC-II, the expression of which is presumably limited to antigen-presenting cells of the teleost immune system as it is in mammals (Yamaguchi and Dijkstra, 2019). Further adjustment of the AmpliSAS clustering parameters proved valuable for the elimination of rare errors exceeding the default Illumina error rate, and of artifacts of exceptionally high frequency, both of which were only found in the cDNA dataset. ACACIA, with its automated oligotyping clus-

tering tool, was more prone to their inclusion.

All three pipelines were developed and benchmarked using gDNA samples, as do, to our knowledge, all other pipelines for NGS amplicon data that have been published so far (*e.g.*, Pavey et al. (2013); Sommer et al. (2013); Stutz and Bolnick (2014)). Nevertheless, pipeline parameter flexibility can aid decisively in improving cDNA-typing performance; AmpliSAS, incorporating the most adjustable clustering settings, produced the best results for cDNA samples -at optimal settings performing on a par with the gDNA subset. Discarding alleles not present in both duplicates can further improve cDNA-typing, provided non-recurring artifacts outnumber recurring errors, which was the case with AmpliSAS and ACACIA in our study. Still, as our cDNA duplicates were characterized by relatively high recurring error rates, we would deter researchers from relying solely on repeatability as a proxy for accuracy in cDNA-typing. Instead, we encourage researchers to i) adjust the clustering and filtering thresholds, and ii) perform careful manual inspection of pipeline-generated genotyping results for identification of artifacts. Besides pipeline optimization, the inclusion of replicate samples, unsaturated multiplexing design with 'empty' index combinations, and, if possible, pedigree data are valuable tools toward ensuring high reliability of cDNA-typing.

1.5.3 Contaminants are an underestimated hurdle in MHC genotyping

Besides the well-documented genotyping errors of PCR/sequencing artifacts and allele dropouts, in our study, we also encountered contaminant sequences. By definition, contaminants are identical to real alleles but don't originate from within-amplicon variants- a fundamental assumption on which stepwise threshold clustering is based (Stutz and Bolnick, 2014); therefore, genotyping pipelines have limited capacity to remove them. Contaminants are generated from misassignment of sequences to wrong samples due to tag-jumping (Eisenhofer et al., 2019; Esling et al., 2015; Schnell et al., 2015), a phenomenon linked to sizable degrees of cross-contamination (Schnell et al., 2015; Zinger et al., 2019; Carøe and Bohmann, 2020)across various library set-ups (Bohmann et al., 2022)and multiplexing designs (Esling et al., 2015). Yet, the possibility of contaminants has been largely overlooked in MHC studies, as covered by Lighten et al. (2014). In some studies, contaminants were detectable only at low frequencies, thereby not jeopardizing genotyping reliability (Gaigher et al., 2016; Biedrzycka et al., 2017). Still, the need to improve the underlying methods to minimize the impact of tag-jumping remains (Rodriguez-Martinez et al., 2022).

Here, as we have employed a saturated dual-indexing design vulnerable to tag-jumping (Esling et al., 2015), we took advantage of heterogeneity spacer combinations to detect contaminants. We primarily observed contaminants at sub-optimal filtering thresholds, where they represented the majority of errors across sample types. Adjustment of genotyping parameters was partially successful. In agreement with (van der Valk et al., 2020), low-depth samples in our dataset received the greatest proportion of contaminants, further highlighting the need for sufficient sequencing depth for genotyping reliability (Biedrzycka et al., 2017). From our experience, minimizing the storage time of indexed libraries prior to sequencing (Illumina, 2017)helped curtail, but not completely eradicate, contamination. We recognize that our multiplexing design doesn't allow for the definitive disentanglement of contamination sources, but hypothesize that remaining low-level contamination could be the result of mixed-clusters on the sequencing platform (Kircher et al., 2011), or even low rates of 2nd PCR primer cross-contamination (Bohmann et al., 2022).

1.5.4 Genotyping discrepancies between pipelines calls for caution in literature comparison

As researchers have historically used a number of different genotyping methods, awareness of their levels of consistency is vital. Here, we demonstrate that the three pipelines exhibit generally high agreement levels for gDNA samples (>96%), as per previous comparisons (Biedrzycka et al., 2017). However, in our cDNA subset, pipelines generated demonstrably different results, with identical genotypes called in as few as 63% of samples. Insufficient sequencing depth has been implicated in increased genotyping differences between pipelines (Biedrzycka et al., 2017), but this is not the case in our study, where an above 2,000-read coverage was achieved for 92% of amplicons. Attention has previously been drawn to the potential risks of comparing gDNA results across studies employing different genotyping strategies (Rekdal et al., 2018). Rather, our study shows that caution is required when attempting to compare cDNA-typing data obtained via different approaches.

1.5.5 MHC-II allele expression negatively correlates with haplotype allele numbers

MHC-I gene expression in birds has been studied via NGS typing of paired gDNA-cDNA samples with multiple primer sets (Drews et al., 2017; Drews and Westerdahl, 2019). Here, using the same PCR primers ensured sharing of amplification biases, allowing us to draw indirect- yet robust- comparisons regarding MHC-II allele expression levels. We were able to show ubiquitous MHC-II alleles expression, but also observed higher variance in the allele depth proportion of cDNA samples. This could potentially stem from differential regulation of the expression of specific MHC genotypes, haplotypes, and/or alleles at the transcription level. Studies have proposed the presence of genetic elements, such as promoters, influencing MHC expression regulation (reviewed by Handunnetthi et al. (2010); van den Elsen (2011)). In sticklebacks, genetic control of MHC-II gene expression has been speculated, as individuals of high allelic diversity exhibit lower MHC-II expression levels (Wegner et al., 2006). Given that low individual MHC diversity has been linked to poorer immune outcomes in sticklebacks (Kurtz et al., 2004; Wegner et al., 2003a), the observed up-regulation of MHC-II expression in individuals of low allelic diversity has been proposed as a compensatory mechanism (Wegner et al., 2006). In further support of this observation, our results show that lower allele numbers per haplotype correlated with higher allele expression. In humans, there is mounting evidence for haplotype-specific coordinated expression of MHC loci (Petersdorf and O'hUigin, 2019), with studies indicating a role for transcriptional regulation and DNA methylation (Vandiedonck et al., 2011), as well as clustered polymorphisms in specific MHC regions to exert gene regulatory effects (Lam et al., 2017). Here, we show that alleles of the same haplotype markedly vary in expression levels, an observation hinting at the existence of within-haplotype differential allele expression regulation. In further support of this observation, three alleles found in two different haplotypes seemingly retain their expression patterns. Further studies are needed to elucidate the basis of differential MHC allele expression regulation within haplotypes.

1.5.6 Sibship-assisted MHC-I typing reveals high individual expressed diversity

As already exhibited for MHC-II, careful optimization of genotyping parameters, although essential, does not guarantee perfect genotyping. For our MHC-I dataset, which consists of

cDNA samples of unknown composition, AmpliSAS genotypes were therefore manually scrutinized. In the absence of an independent typing method for genotype validation, we relied on the repeatability between duplicate samples for accurate allele-calling, as well as pedigree information, shown to be a useful tool (Gaigher et al., 2016; Rekdal et al., 2018). Given the importance of sufficient sequencing depth (Sommer et al., 2013)and careful primer design (Gillingham et al., 2021)in enabling the distinction of poorly amplified alleles from artifacts, we evaluated two primer pairs targeting MHC-I and ensured very deep coverage of samples. Expecting a higher individual diversity (Schaschl and Wegner, 2006; Jäger et al., 2007), we employed lower clustering and filtering thresholds than the ones found optimal for MHC-II. As fewer PCR cycles sufficed for adequate MHC-I amplification, presumably due to ubiquitous MHC-I expression (Yamaguchi and Dijkstra, 2019), we anticipated artifacts of lower depth in our MHC-I dataset compared to MHC-II. Lastly, as our lab-bred sticklebacks were kept under constant conditions, no significant variation was anticipated in MHC gene expression due to *e.g.*, infection (Hansen and La Patra, 2002; Peatman et al., 2008). However, as this assumption can be easily violated in wild populations, we encourage awareness of the possibility of greater variance in allelic expression levels when cDNA-typing individuals in non-controlled settings.

Our final results, involving a total of 48 alleles, offer us a first glimpse into the expressed MHC-I allelic diversity of the three-spined stickleback using NGS technologies. Interestingly, the 11 previously undiscovered alleles were insufficiently amplified by the old primer pair or only found in a single MHC-I haplotype. The successful detection of these novel variants highlights the advantage of deep sequencing in spotting under-expressed or less preferentially amplified alleles (Sommer et al., 2013). All newly detected alleles had sequence similarity of over 92% with U-lineage MHC-I genes of the Alaskan reference genome (Grimholt et al., 2015), and two matched those of a local European genome (Thorburn et al., 2023). Individual expressed MHC-I allelic diversity ranged between five and 24 expressed alleles, exceeding previous estimations based on non-NGS genotyping methods (Schaschl and Wegner, 2006; Jäger et al., 2007). These results place the stickleback on the higher end of the number of expressed MHC-I loci among other teleost taxa, which range from one in Atlantic salmon (Grimholt et al., 2002) to 17 in the Atlantic cod (Persson et al., 1999).

1.5.7 MHC-I alleles are highly polymorphic and often shared between haplotypes

Sequence analysis revealed typical features of functional MHC genes, including high polymorphism, high sequence divergence between alleles, and evidence of positive selection on specific amino acid sites involved in peptide binding. These findings are in line with prior efforts to characterize MHC-I genes in sticklebacks (Schaschl and Wegner, 2006). Furthermore, as neither the alleles from our individuals nor the ones from the Alaskan stickleback genome formed significantly supported phylogenetic clusters, we hypothesize that the allelic variation at the MHC-I genes predates the divergence of the Pacific and the Atlantic clade, indicating continuous balancing selection for more than 44 thousand years (Fang et al., 2018). Inferred MHC-I haplotypes from our analysis carried between five and 12 expressed alleles; these results suggest the expression of all 12 known MHC-I loci in the stickleback Sato et al. (1998). Same as MHC II (Reusch and Langefors, 2005), MHC I genes are believed to be generated by repeated and recent gene duplications (Schaschl and Wegner, 2007). The MHC-I inter-locus recombination rate of 1.52% we report here is over five times higher than previous estimations (Schaschl and Wegner, 2006), likely due to the increased sensitivity of our NGS-based genotyping. Hence, our results suggest that the linkage group encompassing the expressed

MHC class I loci of the three-spined stickleback spans more than 1 cM. Additionally, we observe increased allele sharing among MHC-I haplotypes compared to MHC-II, possibly the result of more frequent recombinations within a longer MHC-I genomic region. Further NGS studies, ideally incorporating gDNA data, are needed to elucidate the differences in molecular mechanisms of evolution of the MHC-I genomic region in the three-spined stickleback.

1.6 Conclusion

To the best of our knowledge, this study stands as a pioneering effort to present an independent and thorough evaluation of genotyping pipelines that is not restricted to gDNA samples. By employing different sample templates, and constructing our analysis on an in-depth breakdown of genotyping errors, we highlight the need for template-specific optimization of adjustable pipeline parameters. In doing so, we hope to assist researchers in future efforts to reliably genotype MHC, and other multigene families, using cDNA samples. Valuable information is to be gained regarding MHC gene expression status, in general, as well as explicit allele expression patterns, as showcased here with our MHC-I and MHC-II data, respectively.

Chapter 2

Effect of genetic background and helminth infection on the TCR repertoire of the three-spined stickleback

2.1 Abstract

Vertebrate adaptive immunity is controlled by a set of somatically diversified receptors on the surface of T cells. The cumulative diversity of this T cell receptor (TCR) repertoire is consequential for the outcome of antigenic challenges, yet our understanding of its genetic determinants remains poor. Here, we set out to investigated the dynamics of the TCR repertoire after experimental infection with *Schistocephalus solidus*, a parasite known for its immunosuppressive affects, while also controlling for the effects of genetic background on inter-individual TCR repertoire variation. To do so, we employ advanced molecular and bioinformatics approaches to analyze the TCRβ repertoire of lab-reared three-spined sticklebacks from genetically diverse sibships. We provide the first direct experimental evidence of significant background genetic effects on the variation of TCRβ repertoire diversity among individuals. Also evident are substantial family-based biases in the usage of V-J gene segments, reflecting genetically controlled variance in the mechanisms controlling the VDJ recombination machinery. Individuals infected with a fast-growing *S. solidus* strain displayed enhanced TCRβ repertoire overlap, presumably due to parasite-induced suppression of private T cell responses. Furthermore, we detect public clonotypes of high abundance present in the repertoires of fish across treatment groups. In-depth analyses of these MHC-unrestricted, infection-non-specific TCR sequences reveal distinct properties, such as high generation probabilities and post-generation selection pressures, signifying a biological function. Combined with a limited VJ gene usage profile and formation of tight TCRβ network clusters based on sequence similarity, these results strongly suggest the existence of a group of natural killer-like T cells in the stickleback, in accordance with other teleost species.

2.2 Introduction

The adaptive immune system of jawed vertebrates is centered around the presentation of non-self peptides by Major Histocompatibility Complex (MHC) molecules, and the subsequent interaction of the resulting pMHC-complex with a specialized receptor on the T cell surface; activation of the latter is the starting point of all downstream cellular and humoral immune responses (Murphy and Weaver, 2016). T cell receptors (TCRs) are heterodimeric molecules produced during lymphocyte development in the thymus via VDJ recombination, which is the somatic rearrangement of germline V (variable) and J (joining) segments to compose the variable domain of the α chain, and the rearrangement of V, D (diversity), and J segments for the β chain – which jointly form an $\alpha\beta$ TCR molecule (Alt et al., 1992). This semi-stochastic process results in the generation of a pool of diverse TCRs, the total number of which varies widely among species: there are at least 2×10^7 distinct $\alpha\beta$ TCRs in adult humans (Arstila et al., 1999), 2×106 in mice (Casrouge et al., 2000), and at least one order of magnitude less in zebrafish (Covacu et al., 2016). The term 'TCR repertoire' describes the entirety of an individual's T cells with their corresponding TCR at a given point in time. As each TCR has a limited binding specificity, the diversity of the TCR repertoire is believed to be a crucial factor in an individual's capacity to defend against pathogens (Nikolich-Žugich et al., 2004; Turner et al., 2009).

The advent of next generation sequencing (NGS) technologies has propelled an unprecedented "deep dive" into the vertebrate TCR repertoire (Benichou et al., 2012). A consistent finding emerging from the literature is that there is substantial intra-species variation in the diversity of the TCR repertoire among individuals (Robins et al., 2010; Elhanati et al., 2018; Krishna et al., 2020). Numerous studies attempting to explain this variation have identified several determinants of TCR repertoire diversity: some are genetic, *i.e.*, sex (Britanova et al., 2016; Migalska et al., 2019; Krishna et al., 2020), while others are environmental, such as age (Qi et al., 2014; Britanova et al., 2014, 2016), acute infection status, and the presence or absence of chronic infections, like CMV (Krishna et al., 2020). A crucial observation is that considerable levels of variation are maintained even when other key factors are accounted for; for example, two age- and sex-matched individuals can differ in their diversity by a factor of two (Yoshida et al., 2017). Special emphasis has been given to uncovering the role of individual MHC diversity in shaping the TCR repertoire, motivated by the long-theorized association between the number of expressed MHC alleles and its size, on the basis of constrains imposed by thymic selection (Nowak et al., 1992; Borghans et al., 2003; Woelfing et al., 2008). So far, however, studies in humans and mice have produced contradictory results (Krishna et al., 2020; Brown et al., 2024).

An obstacle in the above attempts to comprehensively unravel the factors shaping the TCR repertoire stems from the fact that the overwhelming majority of research is limited to humans and mice; this undermines the assessment of the -likely significant- role of individual genetic variation (Gragert, 2022; Omer et al., 2022; Corcoran and Karlsson Hedestam, 2024). Studies in humans are -by default- observational and typically lack information on the subjects' genetic background (*e.g.*, Emerson et al. (2017))– and yet, rare studies that have recruited subjects of known kinship offer valuable insights (Zvyagin et al., 2014). On the other hand, experimental approaches using murine models - while useful for uncovering molecular mechanisms through genetic manipulations and production of knock-out strains - are unsuitable for examining effects of natural genetic variation due to the genetic invariability of inbred strains (Watson et al., 2019; Corcoran and Karlsson Hedestam, 2024). Studies in non-model species encompassing natural genetic diversity are hard to come by, but can advance our understanding - as demonstrated in a study on the MHC diversity effect on the TCR repertoire of bank

voles (Migalska et al., 2019).

In order to bridge the gap in our knowledge regarding the role of genetic background in the shaping the individual TCR repertoire, we have conducted our research in the three-spined stickleback, *Gasterosteus aculeatus*. This is a vertebrate species well-established in eco-evolutionary research, that combines the benefits of controlled experimental studies while also incorporating the natural genetic variation of an out-bred population. Importantly, its adaptive immune system shares important features with its mammalian counterpart, including extensive natural MHC variation (**Chapter 1**), which can be reliably genotyped through established techniques (Lenz et al., 2009a; Efstratiou et al., 2024). Furthermore, it possesses a functional thymus and complete VDJ recombination machinery capable of producing a diverse T cell repertoire (Wölfing, 2012). This is therefore an ideal system for us to elucidate crucial questions regarding non-MHC genetic background effects on the TCR repertoire. To take things one step further, we employed the well-studied experimental *Schistocephalus solidus* parasite model to examine the effects of infection on the TCR repertoire. *S. solidus* is a cestode highly adapted to the host's immune system and believed to exert immunomodulatory effects on its host (Scharsack et al., 2004; Piecyk et al., 2019a; Berger et al., 2021). Towards this end, we use two parasite strains of divergent eco-evolutionary histories and virulence levels (Weber et al., 2017; Piecyk et al., 2019b; Franke et al., 2014).

The aim of this study is three-fold. First, we characterize basic aspects of the TCR repertoire of the three-spined stickleback, by employing advanced NGS methodologies for an unbiased and in-depth analysis. By including fish from multiple families in our study, we expect to uncover novel insights regarding the role of genetic background in shaping this part of the adaptive immune system. Secondly, we investigate how infection with a helminth parasite postulated to have immunomodulatory ability shapes the TCRβ repertoire of the stickleback. Lastly, this experiments stands -to the best of our knowledge- as a pioneering attempt at separately examining the TCR repertoires stemming from two independent genomic loci in a species where two such loci exist (Wölfing, 2012).

2.3 Materials and Methods

2.3.1 Experimental fish and parasites

Three-spined sticklebacks were caught during April-May 2019 from the lake Großer Plöner See in Northern Germany (54° 9'21.16"N, 10°25'50.14"E), and kept under standardized summer conditions (18°C, 16/8 hours light/dark regime) at the Max Planck Institute for Evolutionary Biology (Plön, Germany). All fish were genotyped at the MHC class IIB loci according to (Lenz et al., 2009a), and subsequently, male-female pairings were chosen so that resulting F1 sibships have genotypes segregating across a range of MHC-II allele numbers reflecting the MHC-II variability in natural populations (Eizaguirre et al., 2011). During summer, multiple clutches from each parental couple were produced. Hatched F1 offspring of each family were kept under constant summer conditions to limit sexual maturation. To avoid exposure to eukaryotic parasites, fish were fed ad libitum with frozen insect larvae and supplied with ultra-filtrated water. In March 2020, sibships from six F1 families (families A-F) were selected for the experiment. Each family consisted of individuals of various genotypes corresponding to varying levels of individual MHC-II allelic diversity (Table B.1), as deciphered through RSCA genotyping.

Two S*chistocephalus solidus* strains were used in this study, each from a different population: the GOT strain was obtained from Gotland, Sweden, and the NO strain from Lake Skog-

seidvatnet, Norway. Worms were reared in the laboratory by employing a modified *in vitro* system (Jakobsen et al., 2012)which simulates the conditions in a bird's gut, were the parasites naturally reproduce (Smyth, 1954). The protocol followed is described in detail in (Kalbe et al., 2016). In brief, worm eggs were washed and stored at 4℃ until use. In order to obtain coracidia, the copepod-infecting larval stage of *S. solidus*, eggs were incubated for 3 weeks at 20°C in the dark, before light stimulation for hatching. *Macrocyclops albidus* copepods, first intermediate hosts in the *S. solidus* life cycle, were cultured in the laboratory. Medium sized copepods were separated into 24-well microtiter plates with 1.5 mL tap water, starved for a day, and fed a single *S. solidus* coracidium. Subsequently, copepods were kept at 18°C and fed with live *Artemia salina naupliae* and *Paramecium caudatum*. Two weeks post exposure, copepods microscopically confirmed to be infected with *S. solidus* procercoids were used for controlled stickleback infections.

2.3.2 Stickleback infections and dissections

Experimental infections with laboratory-cultured *Schistocephalus solidus* were completed on two days within one week in March 2020, when the sticklebacks were approximately 7 months old. On each day, 10 sticklebacks from each of the six families were exposed to *S. solidus*, while 5 fish from each of the families were used as controls. As one parasite strain was used per day, sticklebacks were processed in two batches (GOT and NO). Each batch consisted of 90 individuals: 60 exposed and 30 control. Overall, 30 individuals per family were used, amounting to 180 fish in total. An overview of the experimental design is provided in Figure 2.1. Exposure-group sticklebacks were distributed into 1L tanks and received a single *S. solidus*-infected copepod, whereas control fish were sham-exposed, *i.e.*, received one uninfected copepod. All controls were treated identically, but processed on different days, as part of either the GOT or NO experimental batches. All sticklebacks were returned to their original 16L tanks after two days, and confirmation that the copepods were ingested was provided by microscopically examining the filtered 1L tank water.

Dissections took place eight weeks post *S. solidus* exposure, following the same order as described above. The fish were sacrificed with an overdose of the anesthetic MS-222, weighed, and their standard length was measured. *S. solidus* plerocercoids were recovered from the abdominal cavity of infected fish; additionally, the weighs of the spleen, head kidney and liver were recorded. We counted 42 infected fish, 78 exposed but uninfected, and 58 controls (Table B.2). Two individuals belonging to the sham-exposed group died during the course of the experiment and were excluded from downstream analysis. Experimental procedures were approved by the animal ethics committee of the local ministry (Ministerium für Energiewende, Landwirtschaft, Umwelt, Natur und Digitalisierung) of the state of Schleswig Holstein (approval no. V 241 - 870612018 (23-3118)).

The parasite index (PI) (% PI = (parasite weight/ (fish + parasite weight)) * 100) was calculated as a proxy for each parasite's virulence (Arme and Owen, 1967). The body condition of all fish was evaluated on the basis of two indices. The condition factor (CF), calculated by the formula 100 × fish weight/fish lengthb (Frischknecht, 1993). Exponent b represents a stickleback population-specific constant and was computed from all control fish in our experiment (b=2.565). Secondly, the hepatosomatic index (HSI) was used as a measure of each individual's nutritional status (Chellappa et al., 1995). Lastly, host immunological activation was assessed through the splenosomatic index (SSI) and the head kidney index (HKI), following the formula (organ weight/fish weight-worm weight) *100 (Bolger and Connolly, 1989). Spleens were stored in RNAlater at -20℃ for further use.

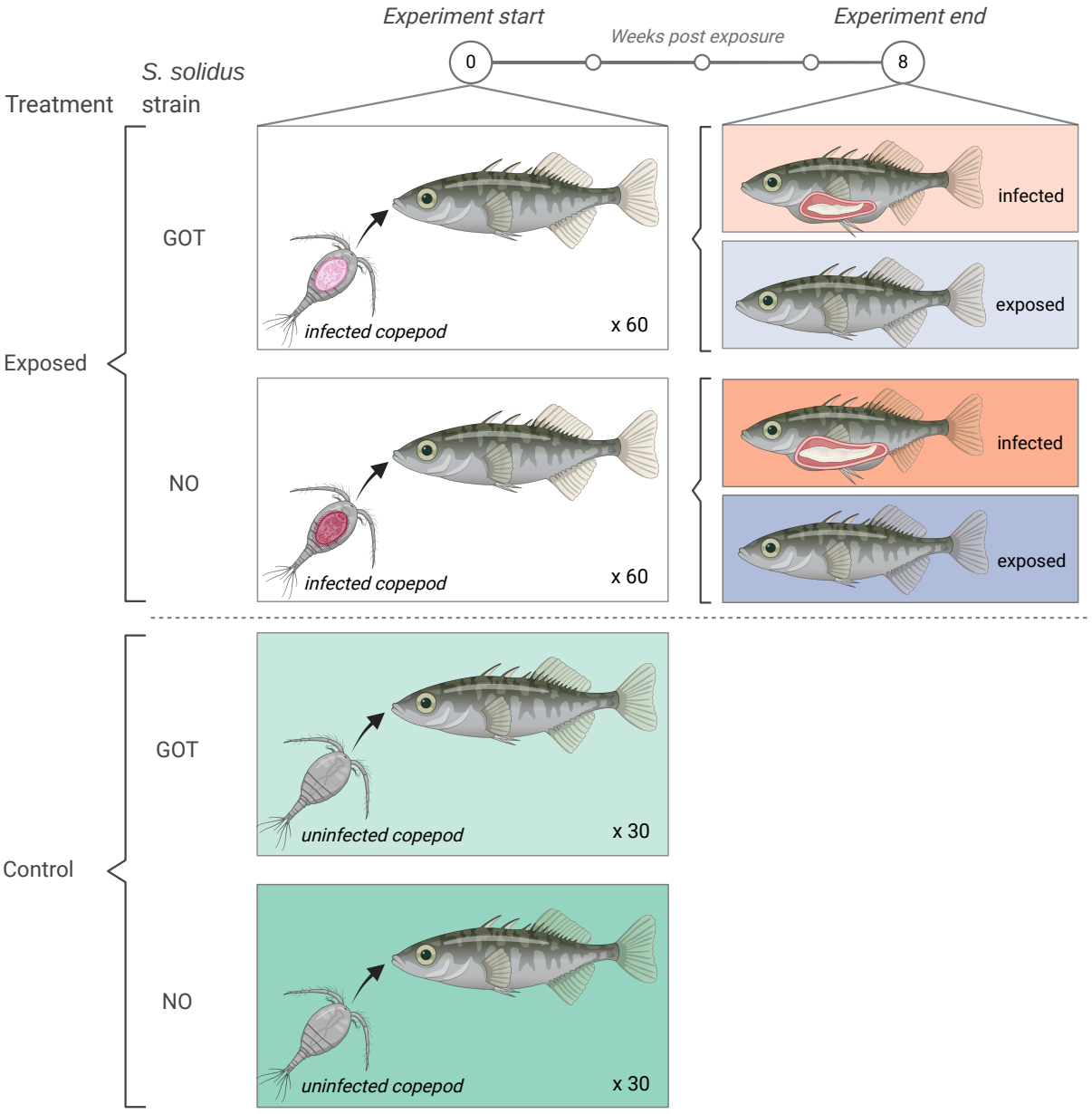


Figure 2.1: Schematic diagram of the experimental design and timeline. Control individuals underwent identical treatment but were handled on separate days, as part of the *S. solidus* GOT strain or the *S. solidus* NO strain batches (see Materials and Methods); are thus referred to as GOT and NO control, respectively. Fish numbers per family are presented in Table B.2. Created with BioRender.com.

2.3.3 TCRβ-targeting 5' RACE and library preparation for Illumina sequencing

A total of 114 fish from the 180 individuals used in the infection experiment were selected for TCRβ sequencing, so as to have a balanced representation of each family in every treatment group (Table B.3). Our TCR sequencing efforts focused on the β chain, the more diverse of the two chains comprising the TCR, due to its CDR3 including a D segment, likely to offer a better picture of the extent of the stickleback TCR diversity. We also chose the spleen, an immunologically important organ with high T cell density (Bjørgen and Koppang, 2021). Total RNA was extracted from the spleens and cDNA synthesis performed using a 5'RACE protocol with unique molecular identifiers (UMI) coupled with a SmartNNNa 5'-template switch adapter and a reverse primer for the constant regions of the TCRβ chain, following a published protocol (Egorov et al., 2015). Two aliquots of the single strand cDNA of each individual were used as technical replicates to perform a two-step PCR amplification with a forward primer targeting the template switch adapter, and reverse primers specific for each constant region of the two TCRβ loci. In the second PCR, the two loci were targeted separately using primers which included Illumina sequencing adaptors and heterogeneity spacers (listed in Table B.4), following (Fadrosh et al., 2014). A schematic overview of the protocol is given in Figure B.1, and a detailed description is provided in the Supplementary Materials and Methods. The libraries were subjected to 2x250 bp paired-end sequencing on the Illumina NovaSeq 6000 platform.

2.3.4 Bioinformatics pipeline for TCRβ repertoire extraction

Raw TCR sequencing data were processed with the MiXCR software (version 4.5.0) (Bolotin et al., 2015). The pipeline groups raw sequencing reads by unique molecular identifiers (UMI), performs UMI-based error correction, and extracts individual repertoires. Due to the range in sequencing depth across sample time points, a moderate threshold of six reads per UMI group (*e.g.*, MIG size of six) was chosen for both TCRβ1 and TCRβ2 repertoires; thus, MIGs supported by fewer than 6 reads were discarded. This allows for the analysis of large TCR repertoires at the expense of more stringent error correction (*e.g.*, MIG size > 10). As a reference database for repertoire extraction we used an expanded version of the three-spined stickleback reference genome-based TCRβ gene library previously assembled (Wölfing, 2012). Because the stickleback reference genome (Jones et al., 2012) is derived from an individual belonging to a relatively distant Canadian population, we sought to further confirm the validity of our TCRβ gene library by mapping our raw sequencing reads to a recently published local stickleback genome (Thorburn et al., 2023). By using the BWA-MEM aligning algorithm and the software IGV (Robinson et al., 2017) we were able to find previously unidentified loci in both RB1 and TRB2 loci. This led to the incorporation of two TRBV1, one TRBV2, and one TRBJ2 additional gene segments. In total, our updated TCRβ gene library consisted of 19 V, 1 D, and 8 J TRB1 gene segments, and 14 V, 1D, and 9 J TRB2 gene segments (Table B.5).

Our analysis focused on the CDR3 region, and exclusively on functional sequences, *i.e.*, nucleotide sequences without stop codons or frame shifts, which made up approximately 70% of the total sequences of each individual. We also filtered out any sequences consisting of gene segments from both loci (*e.g.*, a sequence made of adjoined TRV1-TRJ1-TRC2 segments), as the distance between the two loci on the chromosome renders any organic recombination between them highly improbable - they thus likely represent technical artifacts stemming from cross-contamination (Wölfing, 2012). Because of considerable variation in the size (*i.e.*, total UMIs or starting cDNA molecules) of the resulting functional TCRβ1 and TCRβ2 repertoires across individuals (Supplementary Results), a down-sampling step was implemented in or-

der to avoid sampling-related biases, as described in detail in Supplementary Materials and Methods (Izraelson et al., 2018; Egorov et al., 2018, 2015; Britanova et al., 2014).

2.3.5 Clonotype sharing and inter-repertoire overlap analysis

Initial analysis of repertoire diversity was performed on unique CDR3 nucleotide sequence variants together with their corresponding V, D (if appropriate), and J segments; each such rearrangement corresponds to a unique starting TCRβ molecule. Conversely, all subsequent analyses were performed based on unique CDR3 amino-acid sequences coupled with their corresponding V gene segment, referred to as 'clonotypes'. This definition better captures each sequence's biological function, given that both the CDR3 region and parts of the V gene (CDR1 and CDR2) of the TCR engage with the peptide-MHC complex (Birnbaum et al., 2014). The inter-individual repertoire overlap is dependent on the number and size of study samples (Bradley and Thomas, 2019). In order to avoid biases that could potentially stem from an unbalanced design, we used a subset of 96 samples, where individuals are equally distributed between the three infection status groups and two *S. solidus* strains. We used VDJtools to calculate the extent of pairwise overlap for various fractions of the top most abundant clonotypes in each repertoire, as well as the full, size-standardized repertoire. Sharing level for each clonotype was decided based on the extent of its sharing across the entire dataset of TCRβ1 and TCRβ2 repertoires: clonotypes detected in only a single individual were deemed as private, those detected in 2 to 12 individuals were relatively private, while those clonotypes present in 13 to 48 different individuals were relatively public. Lastly, as public were regarded clonotypes found in the majority (>48) of individuals in our dataset.

2.3.6 Generation probability inference and synthetic repertoire simulation

We used the software IGOR (Marcou et al., 2018) to infer two independent three-spined stickleback-specific models of VDJ recombination, one for each TRB genomic locus. IGOR is a probabilistic model that is trained on non-productive TCR rearrangements; by learning the recombination statistics (*i.e.*, germline gene choice, nucleotide deletions at the V-D and D-J junctions, and N nucleotide insertions) from these unselected sequences, it can then be used to calculate the generation probability of any given TCR sequence. 5,000 of an individual's non-productive sequences suffice for IGOR to learn an accurate TRB model (Marcou et al., 2018). Due to the small repertoire sizes in our study, acquiring such individual-specific models was not feasible. Nevertheless, since variability in recombination statistics across individuals is generally low so that a universal model captures reasonably well the statistics of each individual (Elhanati et al., 2018; Sethna et al., 2020), we proceeded to pool together all non-functional TCRβ1 (n=86,690) and TCRβ2 (n=37,893) unique nucleotide sequences across all individual repertoires in our dataset. We kept sequences found in multiple individuals because they represent independent recombination events. The inferred 'population'-specific TRB1 and TRB2 recombination models had sufficiently low error rates (0.005 and 0.003, respectively). We then used these models to estimate the generation probability of each TCRβ1 and TCRβ2 unique nt sequence in our dataset.

Next, we wanted to study pre-selection repertoires, *i.e.*, repertoires that have not undergone any form of selection, either in the thymus, or in the periphery. We used the software OLGA (Sethna et al., 2019) to generate artificial TCR repertoires, which represent the output of VDJ-recombination and, as such, reflect its inherent biases. For this part of the analyses, we

generated an equal number of TCRβ1 and TCRβ2 repertoires as our balanced dataset (n=96), with repertoire size matching the size of our down-sampled observed repertoires (2395 and 860 sequences for the TCRβ1 and TCRβ2 repertoires, respectively). The simulated datasets were used to calculate the level of sharing of each artificially generated clonotype among all simulated TCRβ1 and TCRβ2 samples; afterwards, we compared the distribution of clonotype sharing levels against the observed repertoires. For the calculation of the average pairwise overlap of simulated TCRβ1 repertoires, we generated a set of 36 artificial repertoires of a total of 10,000 sequences each; each repertoire corresponding to a control sample in our observed dataset. We performed 5,000 iterations of sub-sampling 2395 sequences from each sample, then calculated the average percentage of overlapping clonotypes for all 496 sample pairs in each iteration. Final mean and standard deviation values correspond to the mean of all 5,000 iterations. We repeated this analysis for simulated TCRβ2 repertoires, by sub-sampling to 860 sequences in each iteration. Lastly, we used the software Sonnia to estimate the Q selection factor of each unique clonotype, with a slight modification to allow for specific V–J combinations (Sethna et al., 2020). The Q selection factor is a proxy of the sequence-specific selection pressure exerted upon each clonotype that can best explain the difference in its observed frequency in experimental TCR repertoires from its frequency in artificial repertoires, generated based on the inferred IGOR recombination models.

2.3.7 Similarity network analysis

TCR repertoire similarity network objects were created in R with the NAIR package (Yang et al., 2023), through which we obtained the number of clusters and clustered nodes, as well as node degree and betweenness. To generate a network of sequence similarity, functional nt sequences where collapsed into one node when their amino acid CDR3 sequence and V segments were identical, in accordance with our definition of a clonotype. A Levenshtein distance of 1, denoting sequences differing by at most one amino acid substitution/ insertion/ deletion, was selected as the sequence similarity cut-off used in cluster formation. Thus, edges only connected those nodes that differed by maximum 1 Levenshtein distance, resulting in an undirected graph. For visualization, network objects and node metadata files were exported as mat and csv files, respectively, and imported into Cytoscape (Shannon et al., 2003).

2.3.8 Detection of differentially expanded clonotypes

Clonotypes that were differentially expanded between infection status groups in a statistically significant manner were identified using EdgeR, a software for differential gene expression analysis (M. D. Robinson et al., 2010). In each sample, TCRβ1 clonotypes were filtered by UMI counts via thefilterByExpr function with the following parameters: min.count = 4, min.total.count = 4, large.n = 0, and min.prop = 0. For TCRβ2 clonotypes min.count and min.total.count were set to 3, as an adjustment to the overall smaller size of the repertoire. EdgeR automatically adjusts any differential expression analysis for varying sequencing depths as represented by differing library sizes (*i.e.*, total UMI counts per sample). No count normalization method was applied, as we did not expect the majority of clonotypes to be equally expanded between sample pairs. Each infection status group was tested against all others using a quasi-likelihood (QL) negative binomial generalized log-linear model followed by the F-test. Final clonotypes were filtered for an FDR <0.05 and a logFC $\geq$ 2.5. EdgeR was used as implemented in the TCRgrapher R package (https://github.com/KseniaMIPT/tcrgrapher).

2.3.9 Statistical analyses and data visualization

All statistical analyses were performed with R (Team, 2021). Infection rates, calculated as the proportion of infected out of all exposed fish were used in generalized mixed effects models (GLMMs) with binomial error structure using the lme4 R package (Bates et al., 2024). Independent variables included stickleback family, parasite strain, and their interaction. Sex was included in all statistical analyses as random term, unless otherwise noted. Post hoc tests included Tukey's all-pair comparisons with R package multcomp (Hothorn et al., 2024) or Type III Wald $\chi 2$-tests using Anova() from car (Fox et al., 2024). Next, parasite index (PI) was used as response variable in a GLMM with family, parasite strain, and individual MHC-II diversity level as fixed effects; tank, which was confounded with fish family and treatment group (GOT, NO), was used as random effect.

Three infection status groups were defined for further analyses: the control group, the exposed group, which included sticklebacks exposed to *S. solidus* but uninfected at the time of dissection, and the infected group. Unless otherwise stated, comparisons were restricted within the GOT and NO batches, in order to control for batch effects. Body condition (CF, HSI), and immunological indices (SSI, HKI) were used as independent variables in GLMMs to examine differences between groups. Biologically relevant interaction terms were fitted, and random factors included sex (to account for potential sex-specific differences) and fish tank. Best fitting models were selected with likelihood ratio tests and the Akaike information criterion by sequentially dropping non-significant terms. Significance of remaining components was determined by Wald F-tests using the anova() function from the 'stats' core R package. The effect of individual MHC-II allelic diversity on the TCRβ1 and TCRβ2 repertoire diversity was explored while controlling for other variables, by expressing the diversity as residuals of a GLMM with family and infection status as fixed effects. Tank was used as random effect. Based on previous findings of MHC-specific effects on the TCR repertoire (Migalska et al., 2019; Brown et al., 2024), we tried fitting both linear and quadratic polynomial on these residuals to test the effect of individual MHC-II diversity on each repertoire's diversity.

V and J gene usage frequencies were obtained for all TCRβ1 and TCRβ2 samples with the VDJtools software. Weighted usage of each V and J gene from the two TRB genomic loci was analyzed on the basis of host family, infection status, and strain by GLMMs. Differences in weighted gene usage frequencies by parasite strain were analyzed separately in infected individuals. The V and J gene usage profiles of pre-selected TCR repertoires were examined in a similar manner, by using non-functional TCRβ1 and TCRβ2 clonotypes from each individual.

For inter-individual repertoire overlap analysis, we first tested the homogeneity in dispersion between different families and infection status groups using ANOVA. Non-parametric permutational multivariate analyses of variance (PERMANOVA) were then calculated on Euclidian distance matrices using function adonis() from R package vegan (Oksanen et al., 2024). 10,000 permutations were run for each test, which family, infection status, and tank as fixed effects. Significant results were followed up with post-hoc Tukey's multiple comparisons of means of dispersion to identify which groups differed significantly. Non-metric multidimensional scaling (NMDS) on Euclidian distances and two dimensions was employed for visualization of inter-individual repertoire overlap with function metaMDS()).

The R package 'ggplot2' (Wickham, 2016) was used for generating figures. Color palettes used in plots and figures were from ColorBrewer (Neuwirth, 2022). Heatmaps were created using the package ComplexHeatmap (Gu et al., 2016), and R packages ggmap and rnaturalearth were used for drawing the map of the sampling site (Kahle and Wickham, 2013; Massicotte et al., 2023). The software Cytoscape (Shannon et al., 2003) was used for visualization of TCR repertoire similarity networks.

2.4 Results

2.4.1 Results of *S. solidus* experimental infection

We performed controlled single infections of lab-bred three-spined sticklebacks from Northern Germany with two allopatric strains of *S. solidus*- a fast-growth strain from Norway (NO) and a slow-growth strain from Sweden (GOT) (Figure 2.2A). Of the 120 individuals exposed in total, 42 (35%) harbored a parasite at the time of dissection; 18 fish were infected with the GOT strain and 24 with the NO strain – corresponding to an average infection rate of 30% for the GOT and 40% for the NO strain (Table B.2). We tested whether the infection rate was fish family- or parasite strain-dependent using binomial logistic regression (results presented in Table B.6). Infection rates did not differ significantly between parasite strains (X21 = 1.99, p = 0.158) or fish families (X21 = 7.81, p = 0.167), despite the observed variation (Figure 2.2B). Individual MHC-II allelic diversity also did not have a significant effect on the infection rate (X23 = 1.86, p = 0.602). As expected, the parasite index of infected individuals, a proxy of parasite virulence, differed significantly between the two parasite strains (X21 = 84.77, p = <0.0001), with NO *S. solidus* growing considerably bigger (Figure 2.2C). Neither fish family nor individual MHC-II allelic diversity had a significant effect on parasite index (Table B.7).

Next, we examined the effects of infection status, parasite strain, their interaction, as well as fish family on the various stickleback body and immunological condition indices. Interestingly, out of all variables in our GLMM models, only family had a significant effect on the stickleback condition factor, whereas infection status had a major effect on the HSI, the HKI, and the SSI (Table B.8). Given the pronounced difference in growth between the two parasite strains, and the fact that either strain or the interaction between infection status and strain had significant effects on the HSI, HKI and SSI (Table B.8), we proceeded to analyze the differences in the three indices between infection status for each strain separately. No differences were observed between the control and exposed fish of the two strains for any of the 3 indices (Figure 2.2D, Table B.9). Furthermore, NO-infected fish had significantly lower HSI, and higher HKI and SSI compared to their control and exposed counterparts (Figure 2.2D, Table B.9). GOT-infected fish, on the other hand, had a higher SSI compared to the control. Between the two strains, NO-infected fish had significantly higher HKI and SSI, and lower HSI compared to GOT-infected fish (Figure B.2, Table B.10). Lastly, unlike other indices which are not sensitive to the parasite index, the HKI of both NO- and GOT-infected individuals is significantly affected by parasite size (Table B.11).

2.4.2 TCR repertoire diversity and gene usage

We focused our TCR sequencing efforts on a dataset of 114 individuals (Table B.3). We yielded samples ranging in size between 2,000 to 8,300 TCRβ cDNA molecules (TRBMs) in the TCRβ1 repertoire, and 860-4,200 TRBMs in the TCRβ2 repertoire. Samples of NO-infected individuals were on average smaller, while those from GOT-infected fish were larger than their control and exposed counterparts (Figure B.3A, B). According to conservative estimate, these numbers correspond to a total of 10,000 unique TCRβ nt sequences in the spleen (Supplementary Results). Following normalization of sample sizes (Supplementary Materials and Methods, Supplementary Results, Figure B.4A, B), we fitted GLMMs to predict the observed diversity (*i.e.*, number of unique nt sequences) of both repertoires. We found that family and infection status provide substantial explanatory power, with a strong family effect on both TCRβ1 and TCRβ2, and infection status affecting the TCRβ1 repertoire only (Table 1, conditional $R2_{TCR\beta_1}$=0.36, $R2_{TCR\beta_2}$=0.32,). Post-hoc testing revealed significant differences in TCRβ1

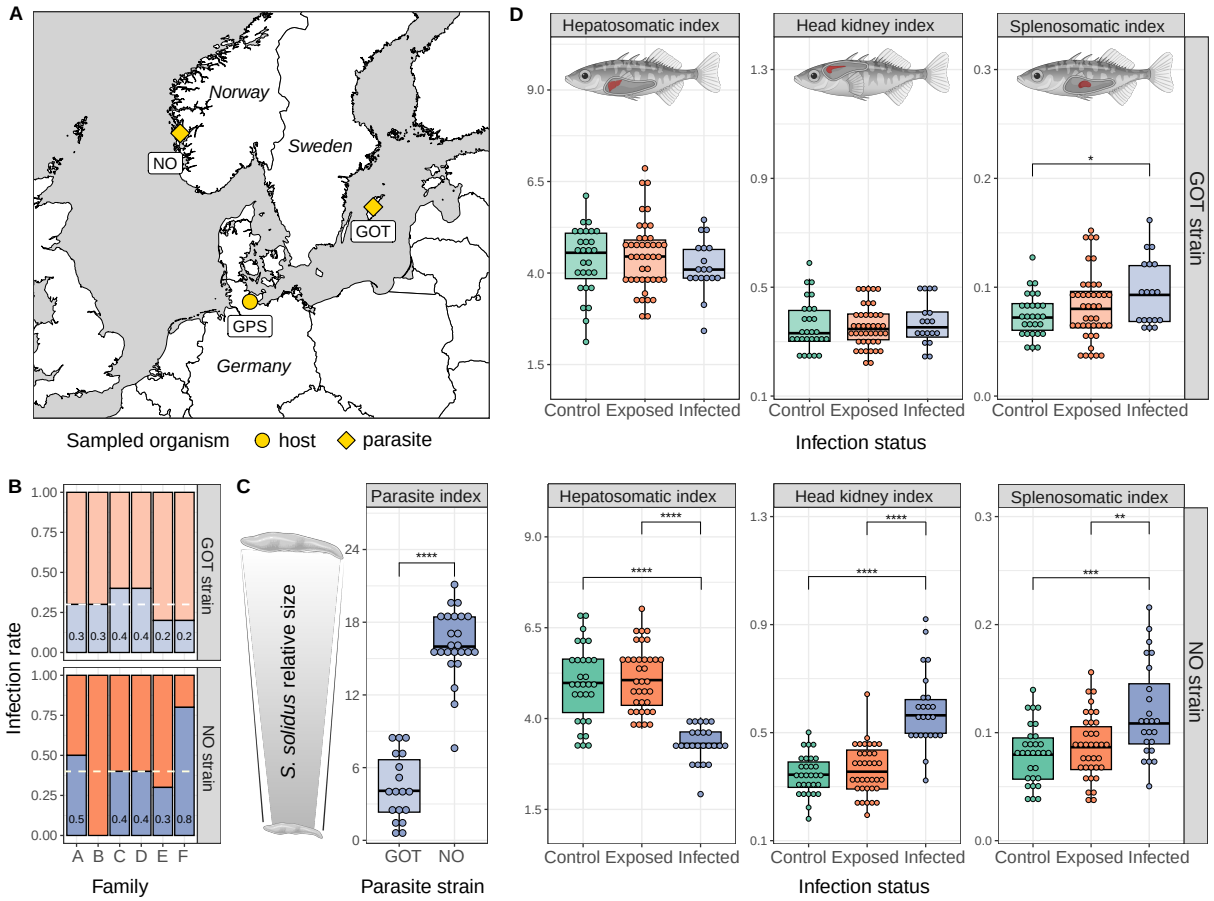


Figure 2.2: Infection with virulent *S. solidus* strain leads to deteriorated condition in three-spined sticklebacks. A) Map of the sampling spots for the host and the parasites. Sticklebacks originated from the Großer Plöner See (GPS) lake in Northern Germany, and *S. solidus* parasites were captured from Gotland, Sweden (GOT), and lake Skogseidvatnet, Norway (NO). **B)** Stickleback infection rates eight weeks post exposure to *S. solidus* shown by parasite strain (GOT and NO), and stickleback family (A to F). Horizontal dashed lines indicate across-family average infection rates for each strain. **C)** Comparison of the parasite index of the two *S. solidus* strains, a measure of relative growth inside their hosts. **D)** Comparison of hepatosomatic index, head kidney index, and splenosomatic index between the three experimental groups of sticklebacks (control, exposed but uninfected, and infected) and by parasite strain. Asterisks indicate P value significance levels as inferred from GLMM analysis described in Tables S9, S10 (**** $p < 10^{-5}$, *** $p < 10^{-4}$, ** $p < 10^{-3}$, * $p < 0.05$). Stickleback and *S. solidus* icons created with BioRender.com.

(Figure 2.3A) and TCRβ2 diversity (Figure B.5A) between various family pairs. Overall, family effects alone explain 22.9% and 28.5% of the TCRβ1 and TCRβ2 diversity variance observed across all individuals, respectively. When the cumulative TCRβ diversity is considered, family accounts for 29.9% of the total variance, while infection status for only 0.9%. The inclusion of individual MHC-II diversity or MHC-II genotype slightly improved our GLM models (with MHC-II diversity: $R2_{TCR\beta1}$=0.38, $R2_{TCR\beta2}$=0.33, with MHC genotype: $R2_{TCR\beta1}$=0.41, $R2_{TCR\beta2}$=0.40). Nevertheless, their effects were non-significant (F<1.3, P>0.27 in all cases), as also shown via residual regression analysis (Figure B.6). Neither MHC-II diversity nor MHC-II genotype had a substantial impact on TCR diversity within the same family.

As expected due to the power law distribution of clone sizes in both repertoires (Figure B.4C, D), increased family diversity stems from a higher percentage of T cells represented by single-cell clonotypes ('singletons'), as well as expanded clonotypes occupying less repertoire space (Figures 2.3B, B.5B). Interestingly, infected individuals have a higher average TCRβ1 repertoire diversity compared to exposed fish (Figure 2.3C); this difference was retained when control and exposed fish were grouped together (Tukey's post-hoc test, P=0.0347). We see a lower percentage of repertoire space occupied by expanded clonotypes in the NO-infected individuals (Figure 2.3D). Neither infection status- nor parasite strain-based differences are observed in the diversity of the TCRβ2 repertoire (Figure B.5C). Overall, we find a strong and statistically significant positive correlation (Spearman's rho= 0.60, P= 5.1super^{-12}) between intra-individual diversity of the two repertoires (Figure B.7).

Table 1. Family and infection status effects on the average observed diversity of TCRβ1 and TCRβ2 repertoires, from 10 independent size normalization iterations. For each repertoire, the entire dataset of successfully sequenced individuals was used (N=112). Results from Type III F tests with Satterthwaite's method of generalized linear mixed effect models (GLMMs). Tank was used as random effect. Statistically significant factors are marked in bold.

Observed TCR diversity						
Repertoire	Explanatory	Df	Sum Sq	F value	Pr(>F)	
TCRβ1	**Host family**	**5**	**279981**	**4.6025**	**0.0087**	**
	Infection status	**2**	**85989**	**3.5338**	**0.0410**	*
TCRβ2	**Host family**	**5**	**112869**	**6.8136**	**0.0021**	**
	Infection status	2	541	0.0817	0.9218	

Next, we examined the V and J gene usage frequency of the TCRβ1 and TCRβ2 repertoires (genes listed in Table B.5). Strong biases are observed within all gene groups, with specific V-J combinations being heavily favored in each repertoire (Figure B.8, Figure B.9). GLMM analyses revealed a significant family effect on gene usage frequency for various V and J gene segments of both loci (Figure B.8). Notably, infection status also had a significant effect on the frequency of nine gene segments (Figure B.10). Of those genes, only V1.17 varied significantly between infected individuals based on parasite strain (Figure B.10); NO-infected individuals used V1.17 at higher frequencies (F=6.73, P= 0.03098). In order to characterize the initial repertoire formed by VDJ recombination itself, we performed separate analysis of non-functional nt sequences; we find that family-based differences are even more prominent in the pre-thymic selection repertoires, particularly when it comes to V gene selection in both loci (Figure B.11).

2.4.3 Clonotype sharing and inter-individual TCR repertoire overlap

Our sequencing efforts yielded 74,795 unique TCRβ1 nt sequences, corresponding to 63,307 clonotypes, *i.e.*, distinct CDR3 aa sequences with their corresponding V gene. The overwhelming majority of clonotypes (80%) belong to a single individual. However, hundreds of sequences were highly shared among individuals; 275 sequences (0.43%) were shared by more than half of the 96 samples in our balanced dataset. Notably, we found ten clonotypes that were shared by all 96 fish (0.016% of all clonotypes). Thus, clonotypes have a wide range of sharing levels. We defined a clonotype as "private" if it appeared in only one individual, as "relatively private" if it was shared by two to 12 fish, as "relatively public" if shared by 13–48 fish, and as "public" if shared more than half the fish (n>48). Public clonotypes are on average more abundant than private ones within each repertoire (Figure B.12A). Furthermore, the average TCRβ1 repertoire comprises of 33% private, 25% relatively private, 29% relatively public, and 13% public clonotypes (Figure B.12B).

To investigate the factors affecting the degree of clonotype sharing between individuals, we performed distance-matrix-based analysis of inter-individual TCRβ1 repertoire overlap. We employed a balanced sample subset (Materials and Methods) and focused on each sample's top 500 most abundant clonotypes. Families exhibited homogeneous variance in dispersion (ANOVA: $F_{5,90}$= 1.1557, p= 0.3373), and differed moderately in their composition of overlapping clonotypes (PERMANOVA: $R2$ = 0.05699, $F_{5,89}$ = 1.0879, p <0.0001), as visualized with non-metric multidimensional scaling (NMDS) (Figure B.12C). Overall, the average within- and between-family repertoire overlap of control individuals did not differ significantly (p= 0.266, Table B.12). Accordingly, sharing an MHC-II haplotype did not increase the average repertoire overlap of control samples (p= 0.349, Table B.13).

Grouping by infection status revealed significant heterogeneity in dispersion (ANOVA: $F_{2,93}$ = 6.0905, p = 0.00327). The infected group, in particular, displayed substantially smaller inter-repertoire distances than the control and exposed groups (Table B.14; Figure 2.4A). Post-hoc testing showed this clustering was driven by NO-infected individuals (Table B.14). Indeed, NO-infected repertoires had the highest average number of overlapping clonotypes (n=148), far exceeding that of GOT-infected repertoires (n=103) (Figure 2.4B, B.12D). As NMDS analysis did not reveal distinct clustering centers, we deduce that no compositional difference in overlapping clonotypes exists between groups. In other words, the high clustering of NO-infected individuals is based on non-group-exclusive clonotypes. As confirmation, we identified all 42 clonotypes shared exclusively between NO-infected samples (Figure 2.4C). Removal of these clonotypes resulted in same clustering patterns, allowing us to reject the hypothesis that they play any role in clustering. Indeed, the majority of overlapping clonotypes within each treat-

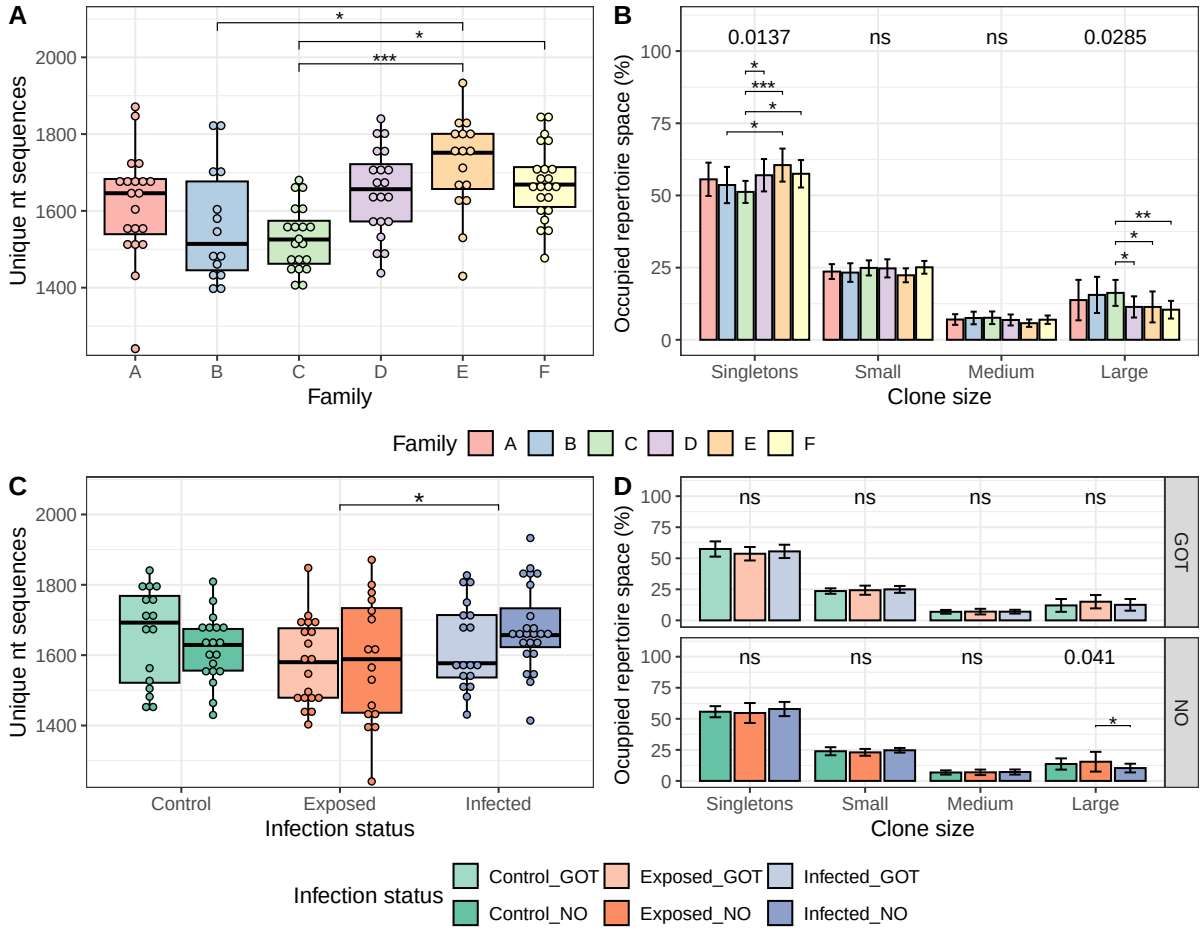


Figure 2.3: TCRβ1 repertoire diversity and clonality are genetic background-dependent. Mean per individual observed diversity (unique nt sequences) obtained from down-sampled repertoires of 2,395 T cells was compared between stickleback families **(A)** and infection status grouped by parasite strain **(C)**. Where GLMM showed a statistically significant effect, Tukey's post-hoc tests were performed with asterisks signifying P value significance levels *** P < 10^{-4}, ** P < 10^{-3}, * P < 0.05). **B, D)** Clonal space homeostasis showing the average percent of clonal space occupied by clones of a given size. Singletons, small, medium and large clones were defined as constituting <0.05, <0.2, <0.3 or >= 0.3% of analyzed T cells, respectively. In terms of starting cDNA molecules, these frequencies correspond to 1, 2-4,5-7 and >=8 UMIs, respectively. Comparisons were made between families **(B)** and infection status by parasite strain **(D)**. For each size type, GLMMs were fitted (P values on top of each category, ns non-significant); for statistically significant cases, they were followed by Tukey's post-hoc tests, with asterisks signifying P value significance levels after Holm correction (***P < 10^{-4}, ** P < 10^{-3}, * P < 0.05). Tank was incorporated in all models as random effect. Non-significant comparisons are omitted for visual clarity.

ment group stems from the public compartment, from which NO-infected repertoires draw more clonotypes on average (Figure 2.4D).

It follows logically that NO-infected repertoires have lower percentage of private clonotypes; indeed, both singletons and expanded private clonotypes are fewer in NO-infected repertoires (Figure B.12E). This pattern of private clonotype suppression is prevalent across different repertoire fractions (Figure B.12F). When clonotype abundance is accounted for, the difference becomes even starker (Figure 2.4E). Next, we asked whether the total abundance of private clonotypes in a repertoire is a function of parasite growth within infected individuals. Statistical modeling showed that the private clonotype compartment in infected individuals is better predicted by the parasite index than the parasite strain itself (GLMM with family included as main effect, tank as random effect; PI: F=7.99, p=0.03; strain: F=4.6, p=0.11). Indeed, there is a very strong negative correlation (Spearman's R= -0.67, p=<0.0001) between parasite index and the space occupied by private clonotypes in repertoires of infected individuals (Figure 2.4F).

We have extended the above analysis to our TCRβ2 repertoires. As compared to TCRβ1, TCRβ2 repertoires have limited clonotype sharing; 1.8% of the total 33,334 clonotypes are either public (n=28) or relatively public (n=564), together comprising less than a quarter of the average individual's repertoire (Figure B.13A). These highly shared TCRβ2 clonotypes are generally more abundant compared to private ones, mirroring their TCRβ1 counterparts (Figure B.13B). Neither family nor MHC-II haplotype sharing affected the extend of repertoire overlap among control individuals (Tables S12, S13). TCRβ2 repertoires of NO-infected individuals on average tended to show more pairwise overlapping clonotypes (Figure B.13C), of which relatively public clonotypes constituted the majority (Figure B.13D). However, the difference was statistically non-significant (Table B.14). TCRβ2 repertoires of NO-infected individuals displayed mild private clonotype suppression and simultaneous expansion of public/relatively public clonotypes (Figure B.13E, F). We conclude that infection with NO *S. solidus* leads to a prominent contraction of the private TCRβ1 clonotype compartment relative to the public one, whereas such a contraction is less evident in the TCRβ2 repertoire.

2.4.4 Detection of differentially expanded clonotypes

Besides the widespread presence of abundant, non-infection-specific public clonotypes, we also found highly expanded TCRβ1 and TCRβ2 clonotypes in the private and relatively private compartments (Figure B.12A, B.13B). To test whether there exist clonotypes that are preferentially expanded among individuals of the same infection status, we employed EdgeR, a tool effectively employed in TCR repertoire studies (Pogorelyy et al., 2018; Minervina et al., 2021, Meinhardt et al.,2022). Overall, we detected 184 TCRβ1 and 59 TCRβ2 differentially expanded (DE) clonotypes across all three infection status groups (Figure 2.5A), nearly all of which are either private (60%) or relatively private (40%) (Figure 2.5B). 48 TCRβ1 and 20 TCRβ2 DE clonotypes were specific to *S. solidus*-infected individuals, and can potentially include *S. solidus*-specific responses. Notably, the average number of both DE TCRβ1 and TCRβ2 clonotypes per individual was significantly lower in NO-infected individuals compared to both the control, as well as the GOT-infected ones (Figure 2.5C). This agrees with our previous observations of suppressed private clonotypes following NO *S. solidus* infection. Interestingly, we found three DE clonotypes shared by control and seven DE clonotypes shared by exposed individuals in both treatments, while a single common DE clonotype was detected between GOT- and NO-infected fish (Figure 2.5D).

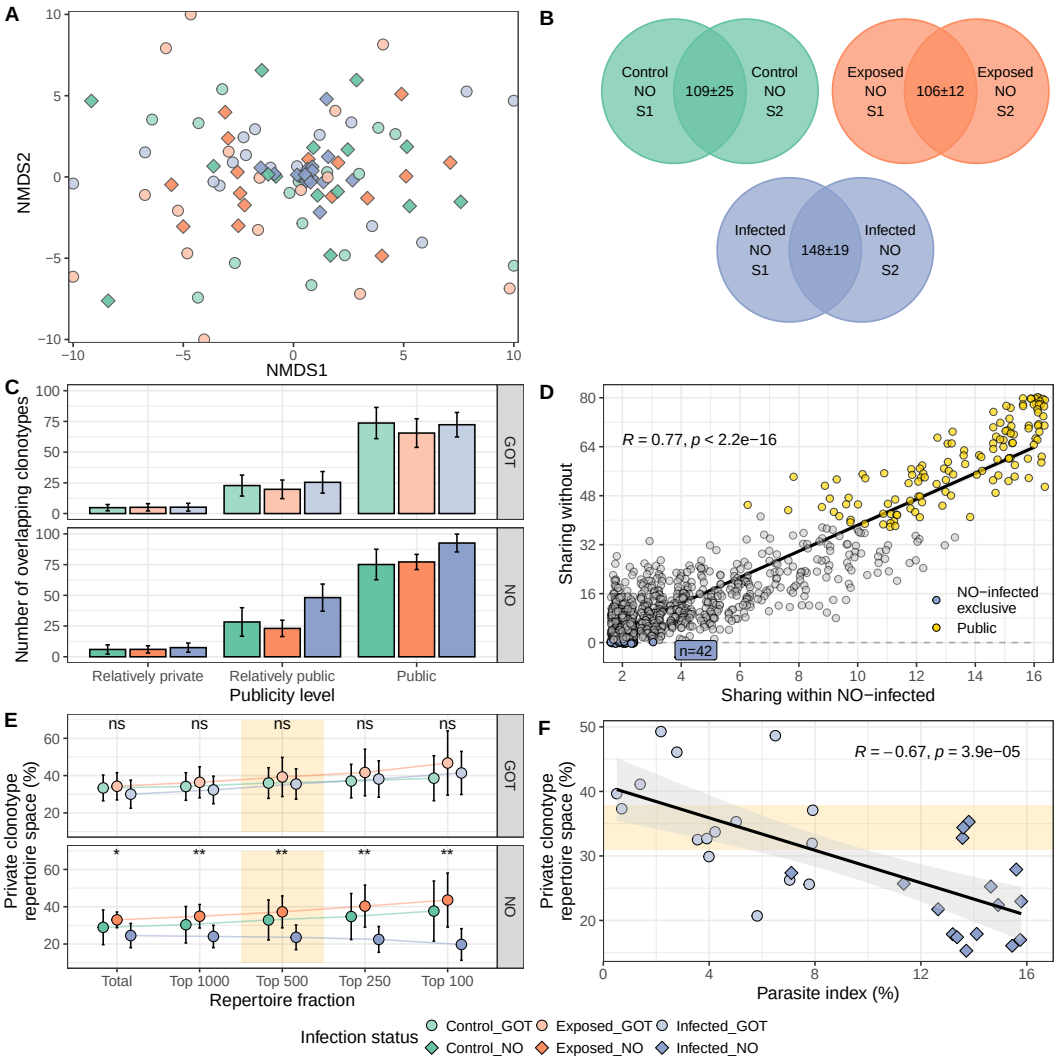


Figure 2.4: Diminished private clonotypes drive the increased TCRβ1 repertoire overlap between individuals infected with the virulent *S. solidus* strain. (A) Clustering of the TCRβ1 repertoires of infected individuals confirmed with non-metric multi-dimensional scaling (NMDS) based on the D index, calculated as the number of shared clonotypes between any two samples, normalized for their individual diversities; Euclidean distance between samples reflects the distance between repertoires. Repertoires down-sampled to the top 500 most abundant clonotypes were used. Outlier samples were moved inwards to the plot periphery for better visualization. **(B)** Average number of pairwise overlapping clonotypes between any two samples (S1/S2) out of the top 500 clonotypes by infection status within the NO treatment group. **(C)** Number of overlapping clonotypes by sharing level, grouped by infection status and parasite strain. **(D)** Scatterplot showing a strong correlation (Spearman's R=0.77, p<2.2e-16) between sharing level among NO-infected repertoires (n=16 samples) and sharing level among the rest of the samples (n=80). All clonotypes present in at least two NO-infected samples were used (n=1,001). Each dot represents a single clonotype. Gold color designates public clonotypes (n=128), and blue clonotypes exclusive to NO-infected samples (n=42). **(E)** Analysis of the repertoire space occupied (*i.e.*, cumulative frequency) of private clonotypes by infection status and parasite strain for different fractions of the TCRβ1 repertoire. P values for each fraction refer to the statistical significance of the effect of infection status, obtained from GLMM models with family and infection status as main effects, and tank as random effect. Yellow shading highlights repertoire fraction (top 500 clonotypes) analyzed in panels A-D and F. **(F)** Scatterplot showing a strong negative correlation (Spearman's R=-0.67, p=<0.0001) between the parasite index and the proportion of homeostatic space occupied by private clonotypes in the top 500 clonotypes of TCRβ1 repertoires of infected individuals. 95% confidence intervals are indicated by the grey shaded area. Horizontal yellow area marks the 95% confidence interval of private repertoire space for all control individuals (n=36).

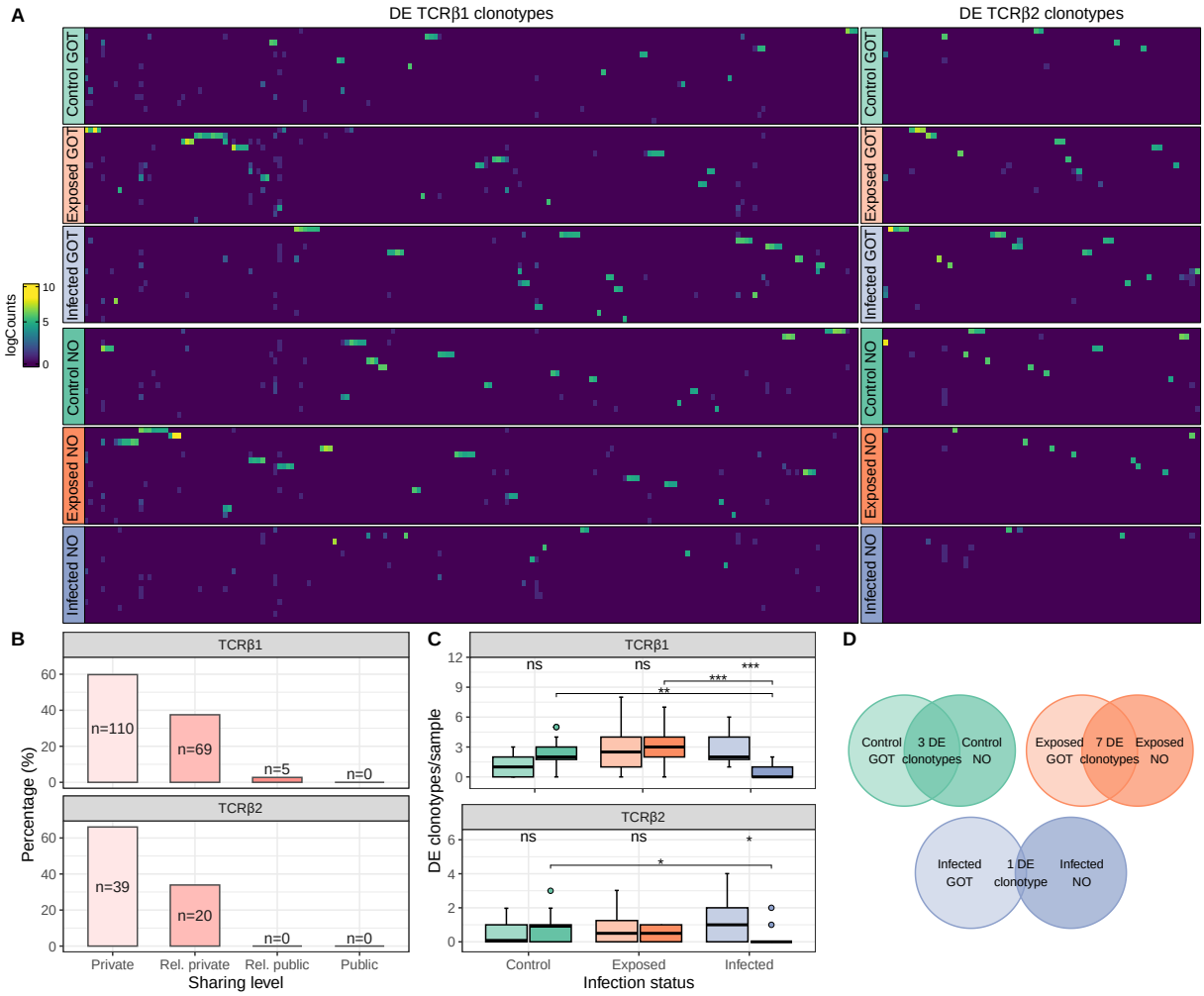


Figure 2.5: Differentially expanded clonotypes are mainly private and suppressed in NO *S. solidus*-infected sticklebacks. (A) Heatmaps show the clone size (in logarithmic scale) of all TCRβ1 (left) and TCRβ2 (right) clonotypes that were determined to be differentially expanded across infection status groups. Each row represents a single individual, and each column a DE clonotype. **(B)** Distribution of TCRβ1 and TCRβ2 DE clonotypes by sharing level. **(C)** Comparison of the number ofTCRβ1 andTCRβ2 DE clonotypes per sample by infection status and parasite strain. Upper asterisks refer to the statistical significance of the effect of strain, obtained from fitting GLMM models with family and strain as main effects, and tank as random effect. The effect of infection status was estimated for each repertoire by using GLMMs with family and infection status as main effects, and tank as random effect. When infection status had a significant effect, we performed Tukey's post-hoc tests, with asterisks above brackets asterisks signifying P value significance levels after Holm correction. Non-significant comparisons are omitted for visual clarity. **(D)** Overlap of DE clonotypes between same infection status groups treated by different parasite strain.

2.4.5 Generation probabilities and selection of private and public clonotypes

Both private and public clonotype compartments of the TCRβ1 and TCRβ2 repertoires contain abundant clonotypes, presumably resulting from expansion in response to an antigen (Figure 2.6A, Figure B.14A). We have further shown that the repertoires of NO-infected individuals have an increased cumulative abundance of public clonotypes, leading to repertoire convergence, but their private clonotypes, possibly including *S. solidus*-related ones, are constricted. We asked whether public clonotypes possess distinct characteristics that might explain this response to NO *S. solidus* infection. To answer, we compared public clonotypes to private expanded ones, including the previously detected DE clonotypes. We found that both TCRβ1 but not TCRβ2 public clonotypes possess a shorter average CDR3 amino-acid length (Figure B.15A, B); both have fewer non-template nucleotide (N) insertions (Figure B.15C). Furthermore, public clonotypes exhibit a high nucleotide sequence convergence, *i.e.*, multiple nucleotide sequences encoding for the same amino acid sequence (Figure B.13D). These features help explain the public clonotypes' higher generation probabilities, as inferred through a stickleback-trained computational model of VDJ recombination: on average, public TCRβ1 and TCRβ2 clonotypes are 3.6 and 4.2 times more likely to be produced than private expanded ones, respectively (Figure 2.6B, Figure B.14B). However, comparison of our observed dataset with simulated repertoires shows that the elevated likelihood of generation alone cannot explain the presence of public clonotypes in either repertoire (Figure B.15E). VDJ recombination biases can only support a small fraction of the levels of pairwise repertoire overlap observed between our control samples (Figure B.15F).

The existence of public repertoires can thus only be explained via a mechanism that selects for them, in the thymus or in the periphery- or both. In other words, public clonotypes must serve a biological function and their omnipresence is not a byproduct of recombination biases. The inference of a stickleback-specific selection model allowed us to assess the Q factor, a measure of the selection pressure for each clonotype that best explains its abundance in our observed dataset compared to artificial, pre-selection repertoires. We thus confirm that there is a selection bias towards public clonotypes compared to private expanded ones (Figure 2.6C, Figure B.14C). Biased selection heavily implies the existence of distinct structural properties of public TCRs that would allow for their preferential engagement during thymic selection. We therefore turn our attention to the V and J gene composition profile of public clonotypes. Interestingly, public TCRβ1 clonotypes exhibit restricted V and J gene usage, heavily skewed towards V1.17 and J1.7 (Figure 2.6D). Indeed, V1.17-J1.7 clonotypes represent the most frequent V-J combination in the relative public and public compartments of the TCRβ1 repertoire (Figure B.16A). Notably, V1.17-J1.7 clonotypes are driving the increased overlap between NO-infected repertoires, and constitute the most increased VJ combination of overlapping clonotypes compared to other infection groups (Figure 2.6E). Similarly, TCRβ2 public clonotypes are V2.1-J2.8-rich (Figure B.14D, B.16B), and drive the increased overlap between NO-infected TCRβ2 repertoires (Figure B.14E). As opposed to relatively private expanded clonotypes, 40% of which are only present in individuals with at least one MHC-II haplotype in common, none of the public clonotypes are MHC-II-restricted (Figure 2.6F, B.14F).

2.4.6 Network analysis of the TCRβ repertoire architecture

Next, we examined the organization of the stickleback TCRβ repertoires via network analysis, and investigated whether *S. solidus* infection impacts the network structure. We constructed TCRβ1 and TCRβ2 networks for each individual in our dataset using the top 500

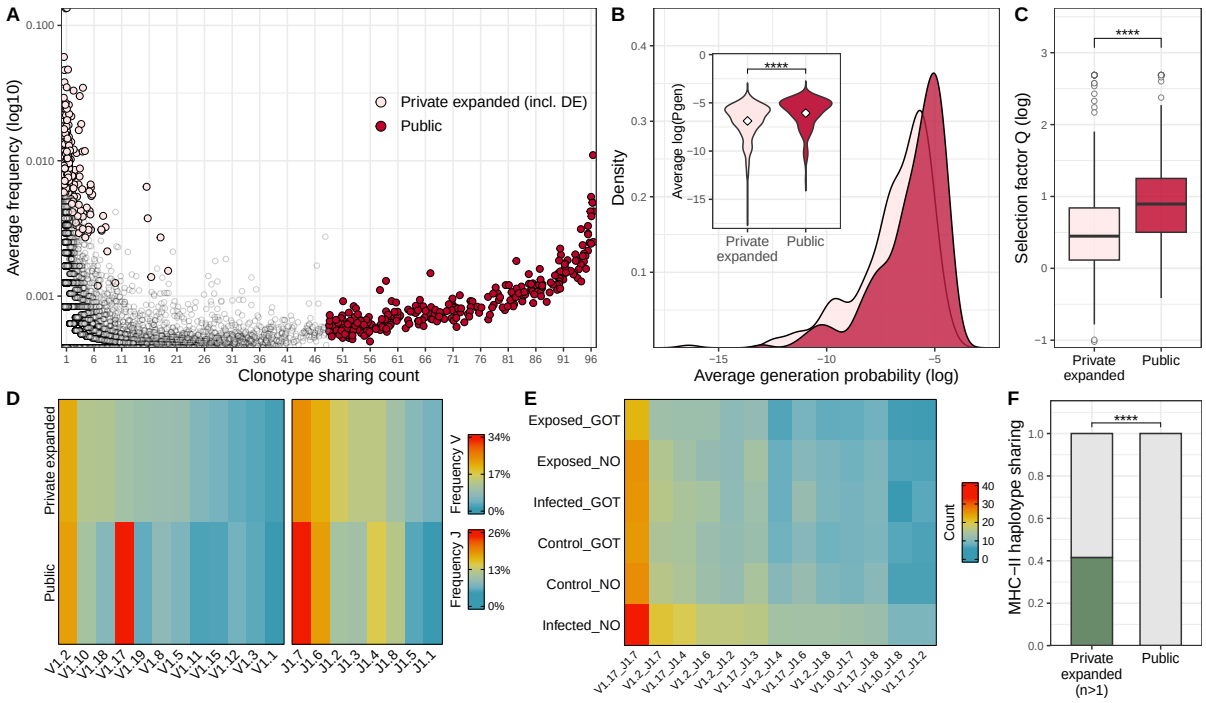


Figure 2.6: Public TCRβ1 clonotypes exhibit distinct properties that distinguish them from private expanded ones. **(A)** Relationship between clonotype average within-repertoire frequency (abundance) and sharing across different individuals. Each dot represents one clonotype. Both the private/relatively private and public compartments include expanded clonotypes, *i.e.*, with clone size >1. We focus on the 275 public (n>48) clonotypes (in red) and 275 mostly (relatively) private expanded clonotypes, including the 184 differentially expanded (DE) ones (in pink). **(B)** Distribution of generation probabilities Pgen for private expanded and public clonotypes, as calculated with the software IGOR. The Pgen value of each clonotype was calculated as an average of all nt sequences encoding each clonotype. Inset asterisks signify statistical significance as inferred from Wilcoxon rank-sum test (P= 5.93e-11). **(C)** Comparison of average Q values between private expanded and public clonotypes. Q is a proxy of the selection pressure of each clonotype in the thymus and the periphery, as calculated with the software Sonnia. Asterisks signify statistical significance as inferred from Wilcoxon rank-sum test (P= 7.24e-15). **(D)** Heatmaps of frequency of usage of V (left) and J (right) gene segments within the private expanded and the public clonotypes. **(E)** Heatmap of the average number of pairwise overlapping clonotypes by VJ gene usage within each of the six treatment groups. Shown are the most frequent VJ gene combinations, *i.e.*, those accounting on average for >2% of all overlapping clonotypes across treatment groups. **(F)** Proportion of relatively private expanded (n=77) and public (n=275) clonotypes that are shared exclusively by individuals sharing at least one MHC-II haplotype. Green color denotes sharing, grey non-sharing. Asterisks signify statistical significance as inferred from Chi-squared test (P= 4.37e-28).

TCRβ1 and top 250 TCRβ2 most abundant clonotypes from each repertoire, respectively. In each network, clonotypes are depicted as nodes, and are connected with each other based on their level of sequence similarity; *i.e.*, clonotypes differing by a maximum Levenshtein distance of 1 appear connected via an edge, while clonotypes separated by more than one amino acid difference from any other remain isolated. Any group of two or more nodes that are inter-connected by edges is termed a cluster.

We found that, in the TCRβ1 repertoire, an average of 194 ± 17 (mean ±SD) out of the 500 most frequent clonotypes per individual were clustered (Figure 2.7A). These TCRβ1 repertoire networks consisted of 248 ± 44 edges. On the other hand, networks constructed with an equal number of randomly selected TCRβ1 clonotypes were sparser, with 152 ± 17 nodes clustered, and 134 ± 23 edges (Figure 2.7B). An interesting observation is that network clusters are mostly formed by public and not private clonotypes; indeed, public TCRβ1 clonotypes were 6.9 times more connected on average that private ones (Figure 2.7C). We made similar observations by analyzing individuals' networks of top 250 TCRβ2 clonotypes; despite the noticeably lower connectivity levels, public clonotypes were still 3.7 times more connected on average than private ones (Figure 2.7C). Interestingly, we found no significant difference in the number of network edges in either repertoire from individuals infected with either strain of *S. solidus* compared to the controls (Figure B.17). We did, however, find that NO-infected individuals had TCRβ1 repertoires with decreased abundance inequality compared to the exposed group, as measured via the Gini index (z=-2.492, P= 0.034 from Tukey's post-hoc test, GLMM shown in Table B.15). No such differences were observed in the TCRβ2 repertoire (Table B.15).

In order to further examine the high sequence similarity exhibited by public TCRβ1 clonotypes, we analyzed the connectivity levels of a network of the 500 most widely shared clonotypes (all 275 public and 225 relatively public). The network included 252 clustered nodes (50.4%) and 391 edges (Figure B.18A). We compared this with a network formed by a set of 316 most abundant private clonotypes and the 184 differentially expanded ones (n=500 in total). This network was less connected, manifesting 109 clustered nodes (21.8%) and only 66 edges (Figure B.18B). It is worth noting that the vast majority of DE clonotypes (130 or 71%) were isolated, *i.e.*, not connected to any other nodes in the network (Figure B.18B). We further found that the biggest public network cluster, comprising 135 clonotypes, showed a strong bias towards gene V1.17 (50.4% usage frequency), as opposed to a frequency of 26.2% between unclustered public clonotypes and 11% among private ones (Figure B.19A). J gene usage of the big cluster public clonotypes was similarly biased towards J1.7 (37%, versus 19.4% among private clonotypes) (Figure B.19B). Lastly, analysis of the physicochemical properties of the amino acids making up each clonotype revealed that the public clonotypes forming the big cluster had considerably higher average polarity and lower hydropathy levels compared to private clonotypes, but didn't differ significantly from the unclustered public ones (Figure B.19C, D).

2.5 Discussion

In this study, we set out to comprehensively analyze the TCRβ repertoire of the three-spined stickleback via high-throughput NGS technology and an advanced error correction-inclusive bioinformatics approach. Our main goal was to examine the response of the stickleback's TCRβ repertoire to controlled infection with a naturally occurring parasite, the helminth *Schistocephalus solidus*, while also accounting for the role of host genetic background. We thus employed six lab-reared stickleback families and performed experimental exposure to two *S. solidus* strains of different virulence. The infection status of exposed fish was evaluated after a period of eight weeks, and their spleen TCRβ repertoires compared to their unexposed

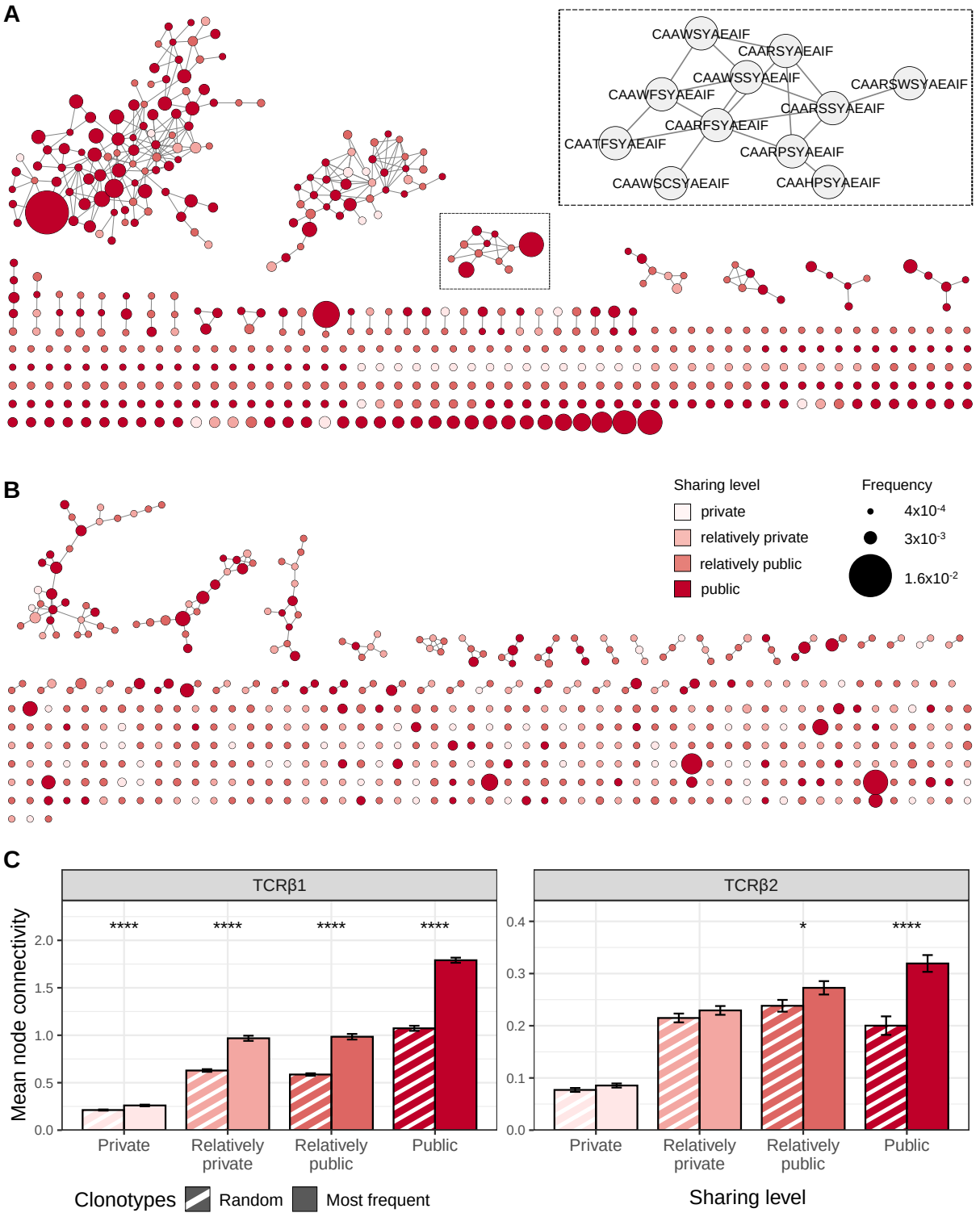


Figure 2.7: Stickleback TCRβ1 repertoires are organized in highly connected similarity networks around public clonotypes. (A) TCRβ1 repertoire similarity networks of the 500 most abundant clonotypes of a single stickleback. Clonotypes, represented by nodes, are connected with other nodes through edges, when differing by one amino acid at most, as presented in detail within the inset. The size of the nodes is proportional to their corresponding frequency, as shown in the scale of panel B. Nodes are color-coded according to the four sharing levels, from private to public (legend in panel B). **(B)** Networks consisting of 500 random TCRβ1 from the same stickleback. **(C)** Average number of edges per node (*i.e.*, node connectivity) by clonotype sharing level in top-500 (compact color) or random-500 (striped) networks of the TCRβ1 (left panel) and TCRβ2 repertoires (right panel). Error bars indicate standard error (SE) across all individuals (n=96). Asterisks signify Wilcoxon test P value significance levels (**** P< 10^{-5}, * P < 0.05).

counterparts kept under identical conditions in the lab.

2.5.1 Sticklebacks possess limited TCRβ repertoire diversity

Our sequencing protocol facilitates the independent targeting of both T cell receptor β chain loci (TRB1 and TRB2) present in the three-spined stickleback (Wölfing, 2012). It thus allowed us to perform, for the first time, an in-depth analysis of the TCRβ1 and TCRβ2 repertoires alike, in a vertebrate species where the two loci are independent. Previous TCR repertoire studies in the zebrafish, a teleost with a single TRB locus containing two constant regions, have focused solely on the TCRβ1 repertoire, *i.e.*, transcripts of the Cβ1 segment (Covacu et al., 2016), based on earlier reports that TCRβ2 transcripts are comparatively scarce (Meeker et al., 2010). Here, we show that the contribution of the TCRβ2 repertoire is not negligible, as it accounts for more than a third of stickleback's spleen repertoire diversity. We estimated the lower bound total TCRβ repertoire diversity in the stickleback spleen at 104 unique TCRs, following two different approaches (Chao, 1984; Giorgetti et al., 2021). his number is likely an underestimation, mainly due to the absence of information regarding the sequence diversity of the α chain - its joining with the β chain to form the TCR heterodimer noticeably increases the actual TCR repertoire diversity (Arstila et al., 1999; Casrouge et al., 2000). The relatively low sensitivity of the UMI-based 5'RACE methodology employed here, coupled with our data being the result of non-exhaustive sequencing efforts are additional factors that lead to diversity underestimation (Rosati et al., 2017; Barennes et al., 2021).

Nevertheless, this estimate can be used for insightful comparisons with other vertebrate species. For example, the 1000-fold difference in TCR repertoire size between humans and mice is thought to be a reflection of the area that each species needs to have 'patrolled' *i.e.*„ larger animals require more cells for immune surveillance (Langman and Cohn, 1987). Notably, this size difference is primarily driven by clonality, not diversity: a given TCR is expressed, on average, by 100 times more T cells in humans than in mice (Nikolich-Žugich et al., 2004). Here, however, the stickleback spleen TCRβ diversity was found to be two orders of magnitude lower than that of the mouse ($2x10^6$) (Casrouge et al., 2000) – despite the two species differing by approximately equal measure in body size. Similarly, the zebrafish's estimated $2x10^5$ T cells (Covacu et al., 2016) remain ten times fewer than those in mice ($3x10^8$), even after correcting for total body mass (Bousso et al., 1998). Teleost B cell receptor repertoires seem to be similarly restricted compared to mammals (Du Pasquier, 1982). It therefore appears that, for reasons extending beyond divergent body sizes, teleost species harbor intrinsically smaller adaptive immune receptor repertoires compared to higher vertebrates.

2.5.2 TCRβ diversity and gene usage are genetic background-dependent

A chief goal of the present study was to illuminate the long-overlooked factor of genetic background in shaping the TCRβ repertoire. By employing the three-spined stickleback, a model system that incorporates natural genetic variation, while simultaneously controlling for known determinants of the TCR repertoire diversity, we estimated that approximately 30% of the observed variance in total TCRβ repertoire diversity across individuals can be explained by the stickleback genetic background. In principle, the diversity of an individual's TCR repertoire at a given timepoint is determined by three separate processes: VDJ recombination, central selection in the thymus, and peripheral selection via clonal expansion or deletion in response to encounter with antigens. Here, peripheral TCR selective forces, measured in terms of changes in TCR diversity in *S. solidus*-infected individuals, were not a key

determinant of TCR repertoire diversity, as infection status accounted for less than 1% of the variance in its total diversity. We conclude that the main drivers of TCR repertoire diversity in our experiment were VDJ recombination and/or subsequent thymic selection, both processes dependent on the host genetic background.

VDJ recombination, the mechanism responsible for generating a massive number of unique TCRs (Alt et al., 1992), can potentially contribute to inter-individual diversity variation through differences in the frequency of usage of the various V, D, and J gene segments (Sethna et al., 2020). Here, in agreement with higher V-J segment usage similarity in monozygotic twins (Zvyagin et al., 2014), we observe significant differences in V-J usage among stickleback families. Family effects were even more pronounced in non-functional, unselected TCR repertoire, a strong indication of the role of genetic factors beyond the MHC in gene usage variation (Melenhorst et al., 2008). Such V-J usage variation could stem from polymorphisms in the TRB locus (Gragert, 2022; Omer et al., 2022; Corcoran and Karlsson Hedestam, 2024), or genetic variation in genes encoding key proteins, *e.g.*, RAG1/2 and TdT, of the recombination process (Schatz and Swanson, 2011; Lee et al., 2016; Troshchynsky et al., 2015; Russell et al., 2022)Furthermore, epigenetic effects could account for parts of the observed family effects in this experiment by means of altered chromatin accessibility at the TRB loci (Jaeger et al., 2013; Ji et al., 2019; Aburajab et al., 2023).

Besides VDJ recombination, inter-individual variation in the process of selection is also expected to contribute to the variance in TCR repertoire diversity in a population. During thymic selection, the strength of interaction with self-peptide–MHC complexes dictates the survival of a small proportion of T cells bearing 'appropriate' TCRs (Kyewski and Klein, 2006; Klein et al., 2014). Similar to peripheral selection, this process relies on the MHC-restricted antigen presentation, it is thus expected to primarily reflect the different selective pressures by divergent MHC genotypes. Surprisingly, computational approaches indicate that this step contributes a surprisingly limited amount to population-level variation (Sethna et al., 2020), and could provide some context as to why we found no evidence of strong effects for either individual MHC-II diversity or MHC-II genotype on the TCR repertoire diversity of sticklebacks in our dataset. It is important to consider, however, that neither the experimental design nor the dataset of our study was explicitly designed for investigating MHC-related effects. Even so, the fact that the explanatory power of our statistical models improved upon incorporation of these two factors suggests that our study should not be seen as a disputation of previous results. We rather hope that it will serve as motivation for future analyses to further explore the link between individual TCR and MHC diversity in the stickleback, perhaps by taking into account MHC-I allelic diversity, which studies suggest play a greater role in TCR diversity (Migalska et al., 2019; Krishna et al., 2020).

2.5.3 Inter-individual TCR repertoire overlap is driven by abundant public clonotypes and is infection status- and parasite strain-dependent

The stickleback TCR repertoire consists of clonotypes of various sharing levels among individuals, as is the case in virtually all species studied so far (Robins et al., 2010; Madi et al., 2014; Covacu et al., 2016; Madi et al., 2017; Giorgetti et al., 2021). Interestingly, even though the two stickleback TRB genomic loci do not differ considerably in their number of V, D, and J genes, the TCRβ1 repertoire contained proportionally 5 times more public clonotypes than the TCRβ2 repertoire. Non-exhaustive sequencing preferentially captures highly abundant public clonotypes and thus exaggerates their overall presence relative to lower frequency clones

(Zvyagin et al., 2014). Therefore, given the equal sequencing depth granted to the two repertoires, more TCRβ1 public clonotypes could be the result of shallower sequencing of the – comparatively larger- TCRβ1 repertoire.

We have found that highly shared clonotypes in both repertoires are heavily represented among pairwise overlapping clonotypes – a finding that has been linked to their high relative intra-repertoire abundance (Giorgetti et al., 2021), assumed to be due to peripheral selection and expansion in response to antigens from persistent pathogens (Zvyagin et al., 2014). We would therefore expect that MHC-II haplotype sharing would increase the pairwise repertoire overlap among the TCR repertoires of our control sticklebacks. Yet, no such increase was observed, in agreement with past studies in humans sharing MHC-I alleles (Robins et al., 2010). Conversely, human monozygotic twins, who are genetically identical, have been shown to exhibit more pronounced overlap compared to unrelated individuals (Zvyagin et al., 2014). Here, sticklebacks within a single family shared significantly more clonotypes than with unrelated fish. The above strongly suggest that thymic selection – via different MHC alleles - plays a minor role in determining TCR sharing across individuals (Li et al., 2012). This serves as further indication that the composition of the TCR repertoire is largely defined by the pre-selection step of VDJ recombination, variance in which can lead to quantifiable differences in aspects such as gene usage and overlap among unrelated individuals.

Another notable result in this experiment was the significantly higher TCR repertoire overlap among individuals infected with the virulent Norwegian (NO) strain of *S. solidus*. This overlap was mostly driven by public, non-infection-specific clonotypes. Local sticklebacks experimentally infected with the NO strain have been shown to exhibit altered gene expression at 8-9 weeks post exposure; namely, down-regulation of both MHC-II and TCRβ1 in the head kidney (Piecyk et al., 2019a, 2021), and spleen (Piecyk et al., 2019b). Crucially, no such effect was observed with the less virulent GOT strain (Piecyk et al., 2019b). In full agreement with the above findings, we obtained smaller TCRβ1 and TCRβ2 repertoire samples from NO-infected, but not GOT-infected individuals. The above differences can be viewed as a result of different eco-evolutionary backgrounds, *i.e.*, a parasite strain's higher virulence is a trait evolved under an escalated arms race with its local host population (Piecyk et al., 2019a).

In our experimental system, the use of allopatric *S. solidus* strains precludes local adaptation effects on the parasite's growth (Kalbe et al., 2016). As such, within the course of the experiment, all NO parasites were able to surpass the body mass threshold associated with sexual maturity (Tierney and Crompton, 1992; Hammerschmidt and Kurtz, 2007), as opposed to a small minority of the GOT parasites. Significantly, this stage presumably marks the onset of *S. solidus* immune modulation of its host (Scharsack et al., 2007b; Barber and Scharsack, 2010), a hypothesis backed by the experimental finding of increased co-infection rates in sticklebacks only after *S. solidus* had reached this size (Piecyk et al., 2019a). A series of studies have offered further support to the notion that *S. solidus* can effectively manipulate the stickleback's immune responses (Scharsack et al., 2004; Hammerschmidt and Kurtz, 2005; Barber and Scharsack, 2010). Notably, a positive correlation between the expression levels of the immunosuppressive FoxP3a gene and the *S. solidus* parasite index has been described in wild stickleback populations (Robertson et al., 2016).

A well-studied mechanism through which helminths achieve host immune manipulation is the secretion and excretion of bioactive molecules (Hewitson et al., 2009). It is likely that this also applies to *S. solidus*, as extracellular vesicles have been detected via electron microscopy in plerocercoids in the stickleback's body cavity (Mazanec et al., 2023). Notably, a proteomics study of the excretome of infective-stage *S. solidus* has identified potential immunoregulatory proteins, capable of inhibiting host antigen processing and MHC presentation pathways

(Berger et al., 2021). One of the identified manipulation factors has been proven to substantially inhibit the presentation of selected antigenic epitopes to T cells during nematode infection (Manoury et al., 2001). Recent helminth infection studies in mice provide additional evidence of their T cell-related immunomodulatory effects, which include induction of regulatory T cells (Hartmann et al., 2019), and suppression of by-stander T cell activation after antigenic challenge (Even et al., 2024). Furthermore, apoptosis of lymphocytes in peripheral blood has been reported in humans with persistent helminth infection (Kalinkovich et al., 1998). With regards to *S. solidus*, there is *in vitro* evidence that excretory products from mature parasites induce the simultaneous activation and killing of lymphocyte subsets (Scharsack et al., 2013). Our finding, therefore, is in alignment with the immunomodulatory capacities of *S. solidus*, as exhibited via a contraction of the stickleback's private TCR repertoire in individuals harboring sexually mature parasites- the life stage characterized by the highest secretion of excretory-secretory products (Mazanec et al., 2023). This is further attested to by a strong correlation between parasite index and homeostatic space occupied by private TCR clones.

2.5.4 Differentially expanded clonotypes are private, MHC-restricted, with diverse gene usage and broad antigen specificities

As complex multicellular parasites, helminths express a large number of proteins, peptides of which can potentially be recognized by different T-cell, if presented by antigen presenting cells (Homan and Bremel, 2018). Therefore, and in stark contrast to what can occur during infection with antigenically simple pathogens, such as viruses, where a single immunodominant peptide can trigger public T cell responses (as reviewed by (Venturi et al., 2008b), a helminth infection induces the expansion of clonally diverse and private T cell responses, even among genetically similar mice (Brown et al., 2021). In our study, we identified a group of clonotypes which were differentially expanded (DE) between *S. solidus*-infected and non-infected sticklebacks. The overwhelming majority of these DE clonotypes are private, and could, presumably, include *S. solidus*-specific TCRs.

Most T cells can only recognize antigenic peptides processed by antigen-presenting cells and bound in the groove of MHC molecules- provided there exist the right MHC allele. Due to this MHC-restriction, public clonotypes are expected to emerge as a response to infection only in a subset of the population that shares a common MHC allele (discussed in detail in (Van Rhijn and Moody, 2015). Here, we find that about half of non-private DE clonotypes from both TCRβ1 and TCRβ2 repertoires are indeed present in individuals sharing an MHC-II haplotype. We theorize that the remaining DE clonotypes have expanded in response to MHC-I presented antigens and, as such, will be shared by a set of sticklebacks with a common MHC-I haplotype. This is in accordance with quantitative PCR data suggesting the presence of both $CD8^+$ than $CD4^+$ T cells in the stickleback spleen (discussed in **Chapter 3**). Moreover, the significantly fewer DE clonotypes detected in NO-infected sticklebacks further indicates *S. solidus*'s alleged immunomodulatory properties.

A T cell receptor's specificity to a given peptide-MHC complex is instructed by its three complement determining regions (CDRs), two of which (CDR1 and CDR2) are germline encoded by the corresponding V gene (Krogsgaard and Davis, 2005). TCRs with identical CDRs most probably recognize the same antigen-MHC complex in different individuals. Despite not having information regarding the TCRα, TCRβ clonotypes by themselves can be instructive, as binding of the same p-MHC complex usually comes with shared sequence motifs (Thomas et al., 2014; Klinger et al., 2015; Glanville et al., 2017). We thus examined the potential antigenic specificity of DE and private expanded clonotypes by constructing sequence similar-

ity networks of the hypervariable CDR3, the principal determinant of specificity (Davis and Bjorkman, 1988). DE and other private abundant clonotypes exhibited low connectivity networks, in accordance with previously described private TCR networks of humans and mice (Madi et al., 2017). Such scarce networks signify broad, unfocused antigenic binding scopes, seemingly in support of the notion that expanded clonotypes from different individuals target distinct epitopes. This is not surprising per se, given the sticklebacks' variable genetic and MHC background, and months-long exposure to the non-sterile aquarium environment. The detection of a single DE clonotype shared by individuals infected with GOT and NO *S. solidus* raises the possibility of a common parasitic antigen, and warrants further investigation. In the context of individual-specific repertoires, however, infection failed to disrupt the networks, as would be expected after an immune-triggering challenge (Madi et al., 2017).

2.5.5 Public clonotypes exhibit invariant T cell-like properties and could be activated during virulent helminth infection

A non-negligible number of TCRβ1 and TCRβ2 clonotypes were shared widely across sticklebacks in our dataset, and exhibited high abundance. These public clonotypes, unlike the privately expanded ones, were not confined to individuals sharing an MHC-II haplotype. Lack of MHC-I genotyping data prevents us from ruling out the possibility that a fraction of shared clonotypes has clonally expanded in many MHC-I matched sticklebacks in response to a common peptide. It is, however, unlikely that a majority of the unrelated individuals in our dataset possesses a common MHC-I allele, even with the frequent allele sharing between MHC-I haplotypes (as discussed in **Chapter 1**). No public TCRs operating within an MHC-restricted framework are expected to be found in the majority of individuals from an out-bred population (Van Rhijn and Moody, 2015). This raises the question: how are the observed, seemingly MHC-unrestricted, pubic clonotypes generated in the stickleback and why are they so abundant?

Upon closer examination, we detected a set of distinctive properties of public clonotypes: fewer non-templated nucleotide (N) insertions, and -for TCRβ1 clonotypes- shorter CDR3 length. Both are features commonly observed amongst abundant clonotypes in human TCR repertoires (Venturi et al., 2008b; Robins et al., 2010; Shugay et al., 2013; Zvyagin et al., 2014). In accordance with the above features, public clonotypes also had higher rates of nucleotide convergence, *i.e.*, the process of generation of an identical clonotype through distinct recombination events (Venturi et al., 2008b). This was in agreement with a previous study in zebrafish (Covacu et al., 2016). In silico testing using stickleback-trained recombination models established the higher generation probabilities of public clonotypes. However, further analysis of artificial TCR repertoires revealed that recombination-related biases do not suffice as a sole explanation for the near ubiquitous presence and high frequency of public clonotypes. Instead, public clonotypes must also be selected for, either in the thymus or in the periphery (Elhanati et al., 2018), their presumed biological role computationally supported by higher levels of sequence-specific selection pressures (Sethna et al., 2020).

An additional key observation regarding TCRβ1 and TCRβ2 public clonotypes is their restricted V and J gene usage, skewed towards V1.17/J1.7, and V2.1/J2.8, respectively. Such biased VJ usage profiles are strongly evocative of a subset of T cells in mice and humans called 'unconventional' T cells, which express a semi-invariant $\alpha\beta$ T-cell receptor (Godfrey et al., 2015). Unconventional $\alpha\beta$ T cells, such as invariant Natural Killer T (iNKT) cells, are selected in the thymus in an MHC-independent manner on monomorphic MHC-I-like receptors (Bezbradica et al., 2005; Lepore et al., 2014), which explains their universal sharing among individuals.

The heavy Vβ gene bias of iNKT TCRs is imposed due to their recognizing their receptors exclusively through the germline-encoded CDR2β region (Mallevaey et al., 2009). Non-classical MHC-I genes are also present in all ectotherms (Flajnik, 2018), *e.g.*, invariant TCR-expressing unconventional T cells have been detected in the amphibian *Xenopus* (Edholm et al., 2013). Teleosts also possess an ancient lineage of non-classical MHC-I genes likely to interact NKT-like cells (Dijkstra et al., 2018). In the stickleback, one such non-classical MHC-I allele has already been identified (Grimholt et al., 2015).

A large portion of unconventional T cells in humans are $CD4^{-}CD8^{-}$, and make up for a small share of total T cells in peripheral blood (Godfrey et al., 2015). Most of the relevant literature in humans and mice has focused on TCRs from sorted $CD4^{+}$ and $CD8^{+}$ T cell subsets (Emerson et al., 2017; Pogorelyy et al., 2017; Egorov et al., 2018; Trofimov et al., 2022); consequently, unconventional T cells likely make up for a small portion of the generated repertoires (Pogorelyy et al., 2017; DeWitt et al., 2018). On the other hand, teleost single-cell transcriptomics paint a different picture. In zebrafish, around half of spleen $\alpha\beta$ T-cells are conventional T cells, the rest comprise heterogeneous populations of NKT-like cells (Hu et al., 2023). Likewise, in salmon, over 60% of spleen T-cells are 'unconventional' $CD4^{-}CD8^{-}$, with a proportion resembling innate-like T cells (Sun et al., 2024). Assuming that the stickleback spleen immune cell composition resembles that of other teleosts, we expect that bulk RNA sequencing of unsorted cells results in the inclusion of a substantial subset of unconventional T cells. Further support comes from a TCR repertoire study in minifish, in which a large portion of TCRs comprises invariant T cells (Giorgetti et al., 2021). Therefore, we predict that at least part of the public TCRs we observe in this study represent a subset of unconventional T cells, not yet fully characterized in the stickleback.

Further investigation of the architecture of TCR repertoire similarity networks provides substantial support to this hypothesis. As opposed to the broad antigen specificities of private clonotypes, many public clonotypes self-organized in tight clusters, displaying a limited antigen recognition breadth. Similar public clonotype-formed network structures have already been described in fish and humans, and attributed to invariant TCRs from unconventional T cells (Giorgetti et al., 2021; DeWitt et al., 2018) . Why do public clonotypes cluster together? The strict Vβ restriction of invariant TCRs ensures that each peripheral iNKT cell recognizes a similar spectrum of antigens through a set of hypervariable CDR3β sequences, which interact with the CD1d receptor (Mallevaey et al., 2009). These Vβ-sharing public clonotypes would then cluster together on account of their similar sequences. Yet another indication of the MHC-independent nature of these cluster-forming public clonotypes is that they substantially differ in hydrophobicity and polarity compared to private clonotypes- distinct CDR3β physicochemical properties of surviving TCRs being a sign of MHC-based thymic selection (Lu et al., 2019).

What role could such unconventional T cells play during a stickleback immune response? If our hypothesis regarding the *S. solidus* NO strain immunomodulation is correct, then the higher relative proportion of the -presumably- MHC-independent public clonotypes could be explained by the parasite-induced suppression of MHC-restricted T cells. Mammalian iNKT cells recognize CD1d-bound glycolipids, and can serve an overall tolerogenic role during infection (Wu and Van Kaer, 2011). Helminth cell walls contain a variety of glycolipids and lipoproteins (Yang et al., 2016), and iNKT cells have been shown to secrete immunoregulatory cytokines when exposed to nematode-derived antigens (Aravindhan and Anand, 2017). Furthermore, there is evidence that iNKT cells promote accelerated tissue repair (Linehan et al., 2018), which would be very useful during wound-inducing infection with a fast-growing helminth. It is, therefore, also possible that virulent *S. solidus* infection could trigger the activation of

NKT-like cells through CD1d-presented parasite-derived lipid ligands. Such an expansion of public TCRs would not contradict the overall TCRβ downregulation, because their increased abundance would be masked by other 'canonical' T cells, and so would not be reflected in the total TCRβ gene expression levels (Piecyk et al., 2019a).

Lastly, some of the public clonotypes detected in this study do not exhibit the tell-tale properties of iNKT cells, such as skewed VJ gene usage. This is also the case in human TCR repertoires, in which only a small fraction of public clonotypes is attributable to invariant T cells (Pogorelyy et al., 2017); leading to the hypothesis that the majority of public TCRs are produced prenatally and form a common 'core' repertoire; while prominent at birth, this common core slowly decays with age (Britanova et al., 2016; Pogorelyy et al., 2017). As opposed to clonal expansion through MHC, the high abundance of such 'core' public clonotypes could be the combinatory result of recombination biases (Britanova et al., 2014), and peripheral homeostatic proliferation, the mechanism that maintains naive T cell numbers in the periphery (Takada and Jameson, 2009), but decreases with age (Ferrando-Martínez et al., 2011; Moro-García et al., 2013). During fetal development, these 'core' public TCRs could rapidly proliferate to fill up the empty homeostatic space and form the initial protective repertoire (Pogorelyy et al., 2017). In accordance with this theory, we postulate that, besides invariant T cells, stickleback public TCRs also represent a 'core' repertoire, with a putative role of protecting larvae hatching in pathogen-rich water. It is not inconceivable that a fraction of these public 'core' TCRs exhibit degenerate specificities, including cross-reaction between different MHC alleles (Huseby et al., 2005), given that teleost T cells resemble properties of mammalian innate-like lymphocytes (Scapigliati et al., 2018). Such a theory has been offered as an explanation for the public $\alpha\beta$ TCR clonotypes in the zebrafish, which expanded in response to both self and foreign antigens (Covacu et al., 2016). Similarly, public TCRs in mice exhibit cross-reactive self-antigen specificities (Madi et al., 2014), and shown to arrange in tight clusters in TCR similarity networks (Madi et al., 2017).

2.5.6 Study caveats

In this study, we have discussed virulent *S. solidus* infection in the context of immunomodulation of the stickleback's T cell responses. Our inter-individual repertoire analysis was based on sample size normalization for elimination of sampling biases (Izraelson et al., 2018). We cannot rule out, however, a scenario where an unequal amount of input cDNA molecules due to uneven spleen T cell concentrations, combined with unsaturated sequencing depths, contributes to the higher observed frequency of public clonotypes in NO-infected sticklebacks (Greiff et al., 2015; Zvyagin et al., 2014). Nevertheless, the marginally different spleen size between GOT- and NO-infected individuals grants us confidence that our results are accurate. Future analysis of repertoires obtained from pre-sorted T cells could clarify the extent to which the fewer private responses in NO-infected individuals is the result of helminth infection and not sampling biases inherent to bulk sequencing. Moreover, such targeted T cell sequencing together with MHC-I genotyping would help us decipher whether observed changes in public clonotype frequencies among treatment groups are due to suppression of the private compartment, expansion of the public one, or both. Such an approach would additionally increase the likelihood of detecting infection status-specific differences in the repertoire diversity by eliminating the noise of irrelevant cell populations.

2.6 Conclusion

This study serves as a concerted effort to systematically investigate the TCR repertoire of an important species in eco-evolutionary research, the three-spined stickleback. By fully utilizing the stickleback's natural genetic variation and employing a high-fidelity NGS protocol, we were able to demonstrate, for the first time, that non-MHC genetic effects account for a substantial portion of inter-individual variance in TCR repertoire diversity. We further showed that infection with a virulent parasite can disrupt the relative balance between private and public TCR clonotypes; the latter being more abundant in the subset stemming from the TRB1 genomic locus. It remains to be seen whether these public TCR sequences, which share features with those of evolutionarily distant mammalian species, represent an uncharacterized population of MHC-independent, invariant T cells in the stickleback. We further theorize that the remaining, VJ-unrestricted, public TCRs are part of a 'core' repertoire. The recurrent detection of public TCRs across vertebrate taxa suggest that they constitute an evolutionary conserved feature of adaptive immunity, more prominent in diversity-restricted TCR repertoires of small vertebrates, and offers exciting implications regarding their role during the early stages of the evolution of adaptive immunity.

Chapter 3

Exposure to natural parasitic fauna leads to TCR repertoire activation in the three-spined stickleback

"A living thing is not only structure, but structure in motion. As static, it reveals the superlative combination of compounds of matter; as a moving event, it presents the most intricate time-pattern in nature. Life is exquisitely a time-thing, like music."

– Ernest Everett Just

3.1 Abstract

Parasites can impose great fitness costs on hosts, whose immune responses are vital for controlling ongoing infections. Natural settings, characterized by extensive genetic diversity of parasites and variation in the abiotic environment, can dictate immune system functions. This important natural context is often overlooked in laboratory-based research and can affect conclusions drawn from single infection studies. Here, we investigate the dynamics of the T cell receptor (TCR) repertoire of the three-spined stickleback under semi-natural conditions, aiming to better capture its immunological status in the wild. During late spring and summer, lab-bred naive fish in enclosures were exposed to the local parasite fauna in the lake from which their parents were caught, which is characterized by a high diversity of parasite taxa. Fish were recovered after four and eight weeks of lake exposure, screened for macro-parasites, and their spleen TCRβ repertoires high-throughput sequenced. After eight weeks of exposure, a robust stimulation of the stickleback adaptive immune system was evident through a substantial increase of clonally expanded TCRβ1 clonotypes and simultaneous decrease of repertoire diversity compared to lab-kept controls. This coincided with high average macro-parasite burden, as various co-infecting taxa were detected in each individual. Interestingly, significant sex-based differences were observed in parasite burden among lake-exposed fish, as well as in repertoire diversity of exposed and control fish. In further indication of T cell activation, the TCR repertoire of exposed fish exhibited perturbed similarity network architecture due to lower clonotype connectivity levels, in spite of the persistence of highly shared clonotypes. Finally, comparison of TCR repertoires among control sticklebacks of common genetic background differing in age revealed strikingly similar changes to those described in mammals.

3.2 Introduction

Animal studies on immunization and single infection, like the one described in **Chapter 2**, are extremely valuable for deciphering the dynamics of an immune response to a particular trigger, such as an antigen or an entire pathogen. In these experiments, the systematic control of environmental parameters such as diet, temperature, population density and contact with other species, is an indispensable tool for researchers to 'shield' the already intricate immune reactions from external factors that can affect their duration, intensity, and -ultimately-, outcome. Towards this end, the use of animals that are germ-free, specific-pathogen free (SPF), or harbor a known fauna or flora (gnotobiotic) ensures that the presence of third-party micro- and macro-organisms does not interfere with experimental results.

Unsurprisingly, such highly regulated experimental conditions hardly represent reality. Natural situations are much less predictable than those in a laboratory setting – especially concerning the number of pathogens an organism faces. On a daily basis, a vertebrate encounters a plethora of foreign, and potentially pathogenic, agents; and at any given time point, many vertebrates host more than one kind of parasite (Petney and Andrews, 1998; Cox, 2001; Lafferty, 2010). Given that pathogens constitute a major selective force on their hosts (Fumagalli et al., 2011; Fumagalli and Sironi, 2014), we can reason that the vertebrate immune system has evolved to deal with multiple pathogens at once. The host's immune responses are therefore 'choreographed' by means of timing, spatial compartmentalization, finely-tuned molecular feedback loops, and elaborate mechanisms of cellular regulation, precisely in order to orchestrate a multifaceted defensive effort able to handle simultaneous insults (Germain, 2001; Esterházy et al., 2019). It has even been hypothesized that, such is the extent to which the immune system has grown to anticipate encounters with an array of micro- and macro-organisms, that the absence of adequate exposure during a crucial developmental period can have detrimental effects later in life, autoimmunity being one of them (Rook et al., 2003; Stiemsma et al., 2015).

Immune responses, including those driven by T cells, can vary tremendously depending on whether the host is simultaneously infected with more than one (or, in the case of parasites that grow or reproduce considerably within the host, even the same) species (Lafferty and Kuris, 2009). Once inside the host, a complex array of interactions, both between the host and the pathogens, as well as pathogen-to-pathogen, can unfold. Such multi-parasite within-host interactions have only recently attracted major scientific interest (Lafferty and Kuris, 2009), but -already- the evidence is compelling: they can directly affect the course of infection of each one (Telfer et al., 2010; Alizon et al., 2013). Perhaps there is no better example than cross-immunity, the phenomenon involving the generation of a network of innate immune and regulatory cells, as well as cytokines, in response to a primary infection which confers protection against a succeeding one with a second pathogen (Efstratiou et al., 2020). Conversely, co-infection can also increase host susceptibility, *e.g.*, through the immunosuppressing effect exerted by some helminths, which allows the establishment of secondary infections through lowering of the host's resistance (Graham, 2008).

Regarding the T cell dynamics in the context of co-infection, there are examples of helminth-induced suppression of human T cell responses against bacteria (Resende Co et al., 2007; Li et al., 2015), but also upregulation of regulatory T cells (Tregs) against viruses (Montes et al., 2009). Perhaps the best-known example is the synergistic effect between HIV and *Mycobacterium tuberculosis*; infection with the former can lead to depletion of $CD4^+$ T cells and more Tregs, which in turn increases the risk of reactivation of latent tuberculosis and susceptibility to new infection (Pawlowski et al., 2012). The above examples underline the importance of considering the context in which T cell immune responses are studied in any given experi-

mental or observational study system.

In an effort to deepen our understanding of its T cell-specific immune responses, we sought to emulate the three-spined stickleback's living conditions in the wild. We did so by conducting a field experiment under semi-natural conditions, during which the fish were exposed to their natal lake environment. This is an excellent setup to answer questions regarding the dynamics of the T cell receptor (TCR) repertoire under circumstances of naturally occurring mixed infections: sticklebacks are hosts to a wide range of parasites (Kalbe et al., 2002; MacColl, 2009), while the lake Großer Plöner See, to which they were released, is home to a high parasite diversity (Kalbe et al., 2002; Eizaguirre et al., 2012a). Crucially, the stickleback's individual MHC diversity has been heavily implicated in influencing parasite burden (Wegner et al., 2003a,b, 2008); correspondingly, there's indirect evidence of T cell-associated responses being induced during infection (Scharsack and Kalbe, 2014; Haase et al., 2016; Robertson et al., 2016; Piecyk et al., 2019a). Importantly, as a result of habitat-specific adaptation, lake stickleback ecotypes have been shown to demonstrate a robust immune response to sympatric parasite species (Kalbe and Kurtz, 2006), while evidence of local adaptation of MHC genotypes further implicates an important role for antigen-specific immunity (Eizaguirre et al., 2012a). By subjecting the sticklebacks to the naturally heterogeneous abiotic and biotic factors of lake exposure, all of which can affect the immune function (Lazzaro and Little, 2008), we ensure that any observations regarding the T cell repertoire will accurately reflect the common immunological status of sticklebacks in the wild.

Through this experiment, we set out to test the following hypotheses formulated based on our findings from **Chapter 2**: firstly, that lake exposure will result in T cell activation and expansion, and thus change in the clonal distribution of TCR clonotypes; accordingly, we anticipate a decrease of the TCR repertoire diversity of exposed sticklebacks. Secondly, we expect that exposed sticklebacks will also exhibit a contraction of the public TCR compartment in favor of more private, pathogen-specific responses. Finally, and as a consequence of our second hypothesis, we predict that the TCR similarity networks of exposed individuals will be less connected compared to those of control fish.

3.3 Material and Methods

3.3.1 Fish breeding and maintenance in the aquaria

Fish used in this experiment were the lab-reared offspring of wild-caught sticklebacks from the Großer Plöner See lake collected during April-May 2019, and were bred in the aquaria of the MPI for Evolutionary Biology during the summer of 2019. The experimental fish were raised under standardized summer conditions (18°C, 16/8 hours light/dark regime) in the lab in one 90 L aquarium per sibship (around 80 individuals). To prevent contact with eukaryotic parasites, the aquarium was supplied with constant ultra-filtered freshwater from a nearby lake, and the fish were fed on an ad libitum basis with frozen insect larvae. Following MHC-II genotyping of the parental generation (described in **Chapter 2**), all 252 fish from 10 selected sibships (*i.e.*, families F1-F10) were prepared for the enclosure experiment. Before the start of the experiment, conditions were set to spring (12°C, 12/12 hours light/dark regime) in order for the fish to acclimate to the lake water temperature outdoor light conditions at the time of exposure. Fish body weight and standard-length measurements were taken, while a dorsal spine from each fish was clipped for subsequent DNA extraction. The experiment was approved by the animal ethics committee of the local ministry (Ministerium für Energiewende, Landwirtschaft, Umwelt, Natur und Digitalisierung) of the state of Schleswig Holstein (approval

no. V 244 - 68663/2020 (58-7/20)).

3.3.2 Lake exposure and experimental timeline

Four enclosure cages, measuring 1x0.5x0.5 m with a 5 mm mesh size, were used to expose fish to the natural parasitic fauna of the lake Großer Plöner See. A single cage was stocked with 16 sticklebacks from each one of families F1 and F2 (32 in total) and kept in the lake for an intermediate period of approximately four weeks. Twelve more sticklebacks (six from each one of families F1 and F2) were kept in a 90L aquarium tank at the institute to serve as intermediate control. The remaining three cages were used for a period of approximately eight weeks; these received a total of 147 sticklebacks from 10 families (families F1-F10). In order to control for cage effects, each cage hosted 48–50 fish, corresponding to a standardized 4-5 fish from each family. Lastly, a total of 61 fish from the 10 families (approximately 6 fish/family) were kept in a 90L aquarium at the institute to serve as final control. An overview of the experimental design and timeline is provided in Figure 3.1A.

On May 11th 2021, when fish were between 20 and 21 months old, the cages were moved into the Großer Plöner See at around 1 m beneath surface level, within a natural habitat for sticklebacks. Water temperature was approximately 12℃(Figure 3.1B). Regular removal of algae from the surface of cages was performed so as to allow for free flow of invertebrate intermediate hosts and free-living infective parasite stages. Water temperature was also frequently monitored so as to ensure fish were not exposed to extreme temperatures, as heatwave events can cause mass die-offs (Wegner et al., 2008). The temperature of the aquaria room were both intermediate and final control individuals were kept was changed three times over the course of the experiment to follow that of the lake (Figure 3.1B).

Over the course of the experiment, two fish from the four-week-exposure group as well as eight fish from the eight-week-exposure group died in the outdoor cages, while seven fish of the final control group died in the aquarium (Table C.1). The remaining 30 living fish from the four-week-exposure group were brought to the lab on June 5th and dissected two days later. Fish of the intermediate control group (n=12) were dissected on June 10th. Surviving fish from each one of the three eight-week-exposure cages (n=139) were brought to the lab in three batches on separate consecutive days (July 3th-5th), and dissected in the same order two days later (July 5th-7th). Only one intact corpse of a dead fish was found, due to rapid decomposition. Accurate determination of parasite load in the dead fish was impossible, and so only the surviving fish were dissected. Fish of the final control group (n=54 surviving) were dissected on July 8th.

3.3.3 Dissections and macro-parasite screening

For dissection in the laboratory, sticklebacks were sacrificed with an overdose of the anesthetic MS-222, weighed and their standard length was measured. All surviving four-week-exposed fish (n=30) were dissected and screened for external and internal macro-parasites. For the eight-week-exposure group, 110 out of the 139 fish were randomly chosen for dissection, due to limited dissection capacity. We followed an established protocol targeting multiple stickleback tissues and facilitating the detection local macro-parasite species (Kalbe et al., 2002; Lenz et al., 2009b). Briefly, individuals were first screened externally for macro-parasites with a microscope. Subsequently, all inner organs, including eyes and gills were screened microscopically and parasites were counted and classified at the lowest possible taxonomic level. The spleens of all fish, including those from the control groups, were stored in RNAlater at -20℃

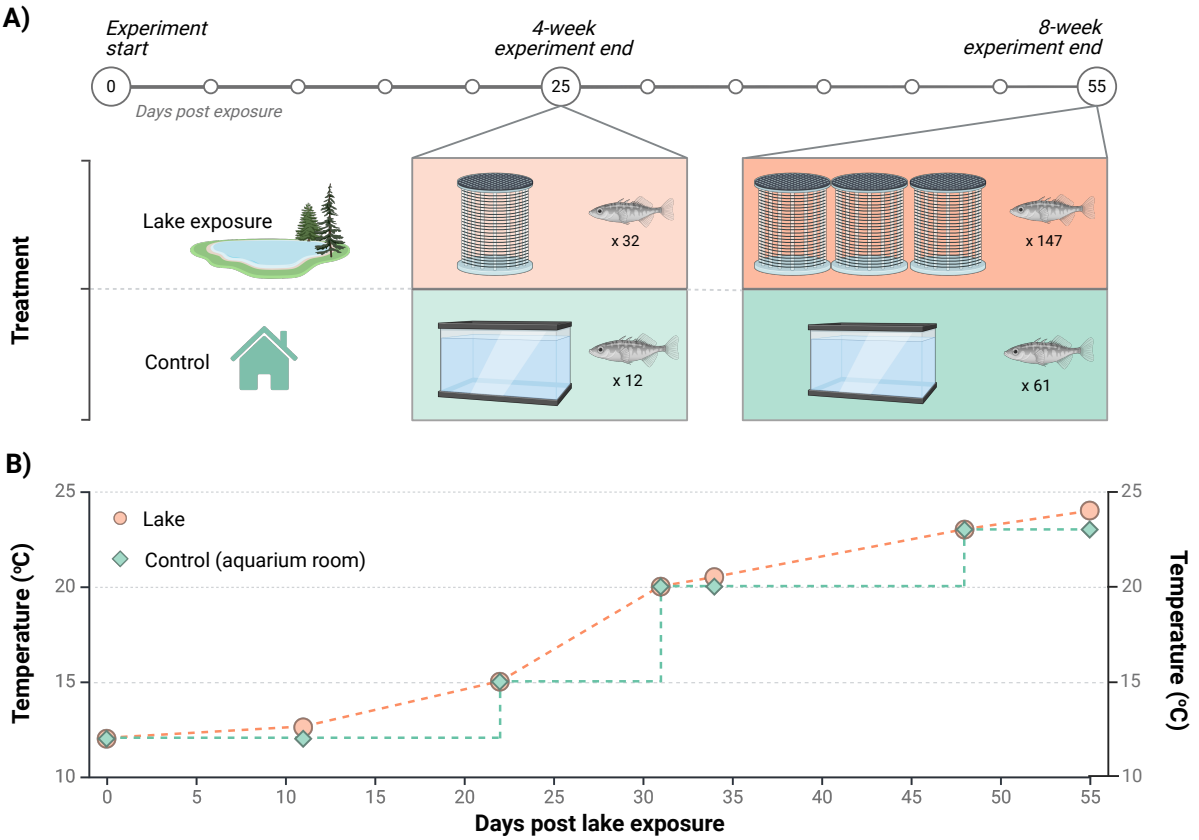


Figure 3.1: A) Diagram of the experimental design and timeline. Fish numbers per family are shown in Table C.1. **B)** Temperature progression of the Großer Plöner See lake water and the aquaria room throughout the experimental period. Created with BioRender.com.

for further use. The tail fin of all dissected fish was clipped for subsequent DNA extraction for molecular sex, MHC-II, and microsatellite typing.

Total parasite load of lake-exposed fish was used as a measure of cumulative intensity of infection by calculating the square root of the sums of all standardized log-transformed count values from all parasite taxa detected within each individual (Wegner et al., 2008). This procedure produces values which follow a normally distribution (Shapiro-Wilk W-test: W = 0.98, P = 0.08), and ensures equal contribution of each taxon to the sum, while also accounting for their abundance. Total parasite load was in high correlation with the sum of different taxa infecting each eight-week-exposed individual (Spearman's R = 0.88, n = 110, P < 0.0001). Stickleback body condition was assessed through the condition factor (CF), calculated by the formula 100 × fish weight/fish lengthb with fish population-specific exponent b (Frischknecht, 1993). Exponent b was computed from all fish prior to exposure and equaled 1.043.

3.3.4 Sex, MHC-II, and microsatellite typing

Two rounds of DNA extractions were performed, one from dorsal spines of the fish collected prior to lake exposure and one from tail fin tissue taken post-experiment. The DNeasy Blood and Tissue kit (Qiagen) was used per the manufacturer's instructions. Identification of fish sex was largely feasible for the lake-exposed fish after recapturing, due to the red coloration developed by males. Further confirmation was provided through molecular sex typing, as described by (Griffiths et al., 2000) and adapted for high-throughput typing via capillary

electrophoresis. Briefly, PCR primers from (Griffiths et al., 2000) were fluorescently labelled and used to amplify an ldh sex-specific marker. 1 µl PCR product was mixed in 10 µl HiDi before a 3-minute denaturation step (94°C) followed by snap-cooling on ice for 5 min. An Applied Biosystem 3730 DNA Analyzer was used for the detection of the fluorescent signal emitted by the DNA fragments. Runs were subsequently analyzed with GeneMarker v.3.0.1 (Softgenetics). Two chromatogram peaks (268bp and 300bp) were present in males, while the presence of a single peak (300bp) represented females. Results from the molecular sex typing were at almost 100% agreement with the visual typing during dissections. All fish were thus successfully sex typed.

In this study, we concentrated on MHC class IIB (MHC-II) genes, as they are critical for activation of the immune system against most extracellular parasites (Murphy and Weaver, 2016). MHC-II genotyping was achieved following the NGS protocol previously developed and described in detail in **Chapter 1** (Efstratiou et al., 2024). Lastly, in order to identify each individual fish's parenthood and match pre- and post-exposure body measurements, we performed microsatellite analysis. Briefly, we assessed the allelic variation at 11 microsatellite loci combined in three different PCR multiplex protocols: the first PCR contained primers for four microsatellites (GAC 1125, GAC 4170, GAC 5196, and GAC 7033), all developed by (Largiadèr et al., 1999). The second PCR targeted another four microsatellites (Stn 18, Stn 32, Stn 75, and Stn 84), and the third PCR amplified the final three loci (Stn 36, Stn 170, Stn 174), first described by (Peichel et al., 2001). This number of microsatellites allowed for high parenthood resolution across the 10 families used in our experiment. We followed the protocol previously described for these multiplex PCR reactions (Kalbe et al., 2009). GeneMarker v.3.0.1 was used for allele calling, and each individual was identified manually. Unambiguous identification was thus achieved for all but 16 individuals, which were subsequently excluded from comparisons of pre- and post-experiment body measurements. In combination with MHC-II genotyping data, family was successfully determined in all but one individual, which was excluded from downstream analyses that involved family as a factor.

3.3.5 TCRβ sequencing and bioinformatics pipeline for repertoire extraction

A total of 192 out of the 236 fish that were retrieved during the experiment were selected for TCRβ sequencing; selection aimed at providing a balanced dataset with regards to number of fish per family, treatment group, and cage (Table C.2). The NGS library preparation and sequencing protocol was the same as described in **Chapter 2**, as were the bioinformatics steps for TCRβ1 and TCRβ2 individual repertoire extraction. A moderate threshold of six reads per UMI group was chosen for both TCRβ1 and TCRβ2 repertoires. Repertoires were filtered so as to retain functional nt sequences, and exclude artifacts consisting of gene segments from both genomic loci. For unbiased inter-individual comparative analyses, all TCRβ1 and TCRβ2 functional repertoires were down-sized to 942 and 199 UMIs, respectively.

3.3.6 Similarity network analysis

Similarity network analysis was carried out for the TCRβ1 and TCRβ2 repertoires separately and as described in detail in **Chapter 2**. For the TCRβ1 repertoire, we focused on the top 250 clonotypes of size-standardized samples. Three samples that fell below this threshold were excluded from the analysis. For the TCRβ2 repertoire, the top 50 clonotypes were used for similarity network construction, and two samples with fewer clonotypes were eliminated.

3.3.7 Repertoire size standardization for age comparisons

We sought to compare the size of the TCRβ1 and TCRβ2 repertoires from the 22-23-month-old fish from the present experiment with those from selected nine-month-old fish of another experiment (see **Chapter 2**), focusing on individuals from three families that were used in both experiments (Table C.8). The size of each repertoire is the total number of functional TCRβ1 and TCRβ2 nt sequences from one individual, which corresponds to the number of starting TCRβ cDNA molecules (TRBMs) present in the initial amount of spleen RNA used. We assume that each TRBM roughly corresponds to a single T cell, and therefore the total number of TRBMs is equal to the amount of T cells in the sample (Britanova et al., 2014, 2016). In our wet lab protocol, we use the same amount of spleen total RNA from each individual per reverse transcription reaction. However, the amount of total spleen RNA varied considerably across control individuals – both between as well as within experiments (9-month-old: mean 1.72 ± 0.75 μg, 22-23-month-old 2.90 ± 2.20 μg). This resulted in a wide range of total spleen RNA proportions being used from each individual across our datasets. To ensure reliable repertoire size comparison across individuals, we proceeded to normalize the TCR repertoire size according to each individual's proportion of spleen RNA usage. We thus calculated the extrapolated number of TRBMs which corresponds to the size of the TCR repertoire we would have theoretically retrieved following the same protocol, provided the entirety of the spleen RNA from each individual had been used.

3.3.8 qPCR for quantification of T cell-specific gene expression

Our TCR sequencing protocol produced substantially smaller T cell repertoire sample sizes in the present experiment compared to the *Schistocephalus solidus* experiment of **Chapter 2**. We sought to externally validate this result -and thus exclude any protocol-related issues- by employing a quantitative real-time PCR (qPCR) approach. In order to obtain an approximation of the relative number T cells present in each sample we chose three T cell-specific genes: CD4.1, CD8α and CD8β, using newly designed stickleback-specific primers. As reference gene we used ubiquitin, previously shown to have the most stable expression among stickleback head kidney samples (Piecyk et al., 2019a). In total, 19 samples from an equal number of individuals were analyzed over two independent runs: 8 samples of control and exposed but uninfected 9-month-old individuals from the *S. solidus* experiment, and 11 samples from 22- to 23-month-old control individuals from the current experiment. All individuals belonged to families F2 and F7. In brief, 1μg of total spleen RNA from each individual was used per reverse transcription reaction with the High-Capacity cDNA reverse transcription kit (ThermoFisher), according to the manufacturer's protocol. 2μl of 1:10 diluted cDNA product were mixed with 0.5μl primer mix (10μM), 2.5μl water, and 5μl of SYBR Green, used as the DNA binding dye. Relative gene expression was measured in duplicates for each sample with the QuantStudio 1 Real-Time-PCR-System (ThermoFisher). Ubc expression levels were used for normalization. Relative expression values were calculated as the average of both duplicates using the $\Delta\Delta$Ct method and exported as log10 transformed CNRQ (calibrated normalized relative quantities).

3.3.9 Statistical analyses and data visualization

All statistical analyses were performed with R (Team, 2021). Cage was found to have a significant effect on parasite load, and was thus incorporated as random effect in all downstream statistical models. To account for any sex-related biases with regards to body condition, parasite load, and T cell responses of individuals, we incorporated sex as a main effect in all

our GLMM models. The effect of individual MHC-II diversity of lake-exposed fish in parasite infection was separately evaluated by controlling for factors sex and cage, as well as non-MHC-related family effects: total parasite load was expressed as residuals of a GLM with total parasite load as the dependent variable and family, sex, and cage and as fixed effects. Based on previous findings of an optimal intermediate MHC-II diversity (Wegner et al., 2003b,a; Kurtz et al., 2004; Wegner et al., 2008), we proceeded to examine the MHC-II effect on the residuals by fitting a quadratic polynomial as the simplest function having a minimum. All statistical tests performed are stated in the text and/or figure legends.

The R package 'ggplot2' (Wickham, 2016) was used for generating figures. Colors for plots and figures were chosen from the ColorBrewer palette (Neuwirth, 2022). Paired boxplots were created with the package 'ggpubr' (Kassambara, 2023). Heatmaps were created using the package ComplexHeatmap (Gu et al., 2016), and package ggmap was used for drawing the map of the experimental site (Kahle and Wickham, 2013). TCR similarity networks were visualized with Cytoscape (Shannon et al., 2003).

3.4 Results

3.4.1 Macro-parasite infection rates post lake exposure

We exposed three-spined sticklebacks from 10 lab-bred families (Table C.1) to their lake of origin for a period of either four or eight weeks; additional sticklebacks from each family were kept at comparable densities in tanks in a temperature-matched aquarium room as controls (Figures 3.1A,3.2A). Lake water temperature was approximately 16℃ at the time of retrieval of the four-week-exposed fish (Figure3.1B). The majority (n=28/30) of the sticklebacks that were exposed to the lake for a four-week period did not harbor any macro-parasites. One of the remaining two four-week-exposed fish was infected with *Diplostomum* sp., and the other with Cyathocotyle sp. (Figure 3.2B, 3.2C). Lake water temperature eight weeks post exposure was approximately 24℃at the time of retrieval of the fully-exposed fish (Figure 3.1B). We found a total of 12 macro-parasites in and/or on the fish exposed in the lake for eight weeks (Figure 3.2B): nematodes *Contracaecum* sp., *Camallanus* sp., and *Raphidascaris* sp., trematodes *Diplostomum* sp., *Apatemon* sp, *Cyathocotyle* sp., *Gyrodactylus* sp., and *Echinochasmus* sp., cestodes *Proteocephalus* sp. and *Schistocephalus solidus*, crustacean *Argulus foliaceus* and mussel larvae (Glochidia). These taxa infect different host tissues, such as inner organs (eyes), the digestive tract, the body cavity, gills and skin (Figure 3.2B).

All long-exposed fish that were screened for parasites (n=110) harbored at least one parasite taxon, with an average of three infecting each individual (Figure 3.2C). With regards to helminths, each individual was infected with an average of 2.8 taxa. Consequently, the individual total parasite load was significantly higher in individuals exposed for eight as opposed to four weeks (Figure 3.2D). This was further confirmed by comparing individuals from families F1 and F2, the only ones represented in both short- and long-term exposure groups ($z = 5.408$, $P = 6.39e\text{-}08$, from Tukey's post-hoc tests following GLMM with family, sex, and duration as fixed effects, and cage as random effect). The individual parasite load among eight-week-exposed fish differed significantly between families, cages, and sexes, but not individual MHC-II diversities (Table C.3, Figure C.1A). The number of parasite taxa detected in each eight-week-exposed individual was affected by sex and cage, but not family or individual MHC-II diversity (Table C.3, Figure C.1B). Female fish harbored more parasite taxa ($z=1.98$, $P=0.048$, Tukey's post-hoc test) than males.

Interestingly, lake exposure was associated with better body condition compared to the

aquaria-kept control groups; this difference was starker in the eight-week-exposed fish (Figure 3.2E). We tested whether individual parasite load affected the change in body condition of long-exposed fish. GLMM analysis showed that change in body condition during the eight-week exposure period was affected by family, but not by parasite load, MHC-II individual diversity, or sex (Table C.4). Additionally, there was no statistically significant correlation between initial body condition and parasite load (Spearman's R=0.17, P= 0.08).

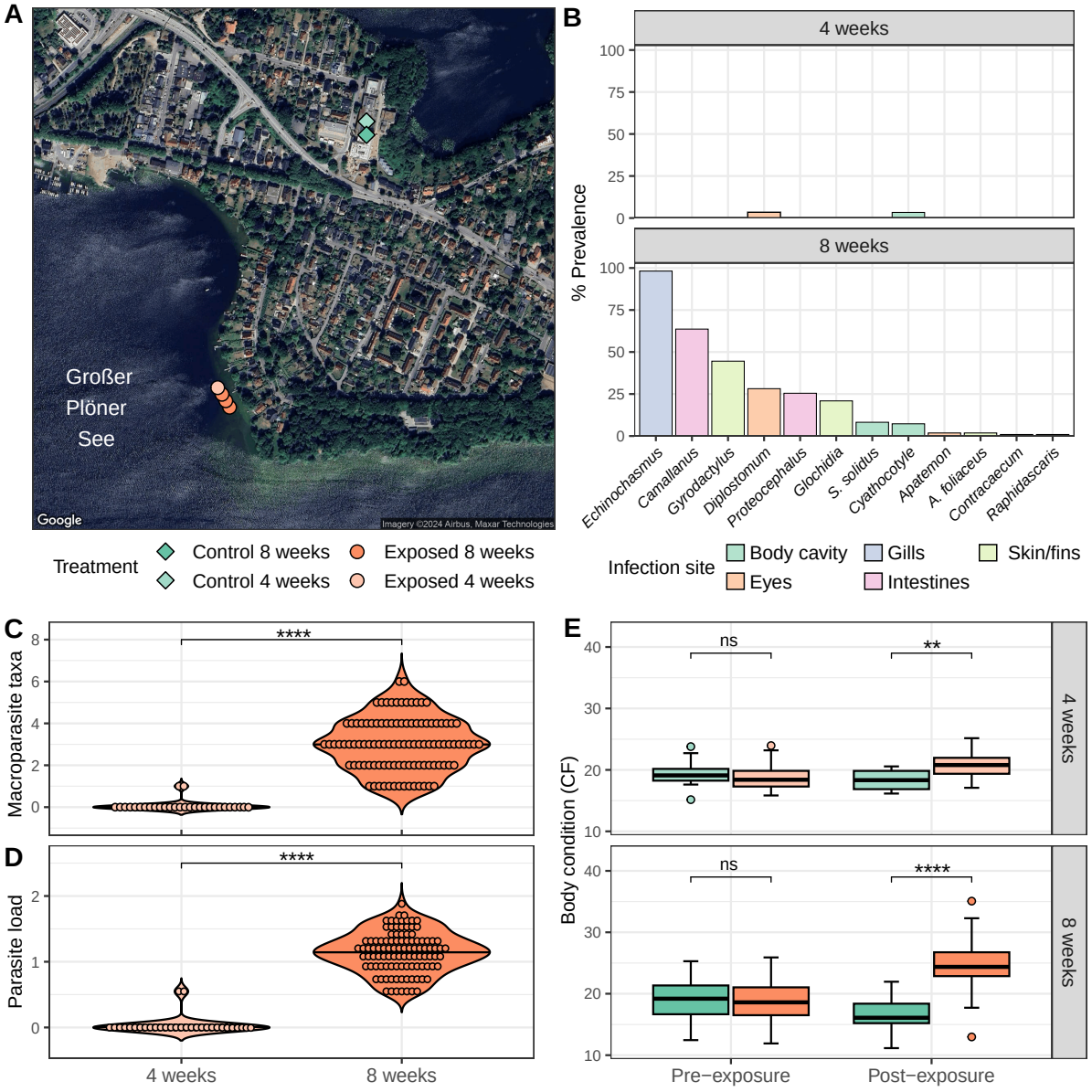


Figure 3.2: A) Map showing the geographic location of the enclosure cages placed in the Großer Plöner See, and the two aquaria tanks at the Max-Planck Institute for Evolutionary Biology used for the control individuals. **B)** Prevalence, *i.e.*, percentage of infected host individuals, of the 12 macro-parasite taxa detected in the lake-exposed individuals by duration of exposure. Bars are color-coded according to the site of infection of the corresponding taxa, as shown in the legend. **C)** Number of macro-parasite taxa detected per individual grouped by duration of exposure. Each dot corresponds to a single exposed individual. Asterisk signifies P value significance level for factor duration as calculated from GLMs with family, sex, duration, and cage as fixed effects (**** $P < 10^{-4}$). **D)** Total parasite load of individuals by experiment duration of exposure. Each dot corresponds to a single exposed individual. Asterisk signifies P value significance level for factor duration as calculated from GLMs with family, sex, duration, and cage as fixed effects (**** $P < 10^{-4}$). **E)** Comparison of condition factor (CF) between control and lake-exposed sticklebacks pre- and post-exposure, grouped by duration of exposure. Asterisks signify P value significance level for factor treatment as calculated from separate GLMs for each duration group, with family, sex, treatment and, when applicable, cage, as fixed effects (ns non-significant P>0.05, **** $P < 10^{-4}$).

3.4.2 TCR repertoire diversity and clonality

Next, we examined the diversity and clonality of size-standardized TCRβ1 and TCRβ2 repertoires. We fit GLMMs with all relevant biological (family, sex, MHC-II diversity) and experimental (treatment, duration and their interaction) variables. Cage was used as random effect (Table C.5). Sex was a significant factor in both repertoires, with males showing higher average diversity compared to females (TCRβ1: z= 2.44, P=0.015, TCRβ2: z= 3.79, P= 0.0002, from Tukey's post-hoc tests). When analyzing the control and exposed groups separately, control individuals showed sex differences only for the TCRβ2 repertoire (GLMM with family, sex, MHC diversity and duration: F=8.06, P= 0.007), while exposed individuals showed sex differences in the total repertoire diversity, *i.e.*, TCRβ1 and TCRβ2 diversity combined (Table C.6).

Interestingly, the interaction of treatment and duration was only important for the TCRβ1 repertoire; the diversity of TCRβ2 repertoire was affected by the duration of the experiment, not by the treatment (Table C.5). Indeed, eight-week-exposed fish, but not four-week-exposed fish, had markedly lower TCRβ1 repertoire diversity compared to control fish (Figure 3.3A, B). Furthermore, family was found to have a significant effect on the diversity of the TCRβ2 but not the TCRβ1 repertoire (Table C.5; C.2A, B). As expected, the lower TCRβ1 repertoire diversity of the eight-week-exposed fish was accompanied by a smaller singleton compartment and more repertoire space occupied by highly expanded clonotypes – a pattern not mirrored in the TCRβ2 repertoire (Figure 3.3C, D). Families only differed significantly in the amount of TCRβ2 repertoire space occupied by singletons (Figure C.2C, D). Overall, family effects alone explained 5% and 8% of the TCRβ1 and TCRβ2 variance in diversity observed across all individuals, respectively. When the cumulative TCRβ diversity is considered, family accounted for 5% of the total variance in diversity.

Because family had a significant effect on TCRβ2 repertoire diversity as well as on individual parasite load, we examined if the two variables were correlated; we calculated the mean TCRβ2 repertoire diversity and mean parasite load of individuals belonging to the same family. Indeed, we found a positive correlation; families with higher average TCRβ2 repertoire diversity after an eight-week exposure period had higher mean parasite loads (Spearman's R=0.66, P=0.044, Figure C.3A). The same significant correlation was also found for the TCRβ1 repertoire (Spearman's R=0.75, P=0.018, Figure C.3B).

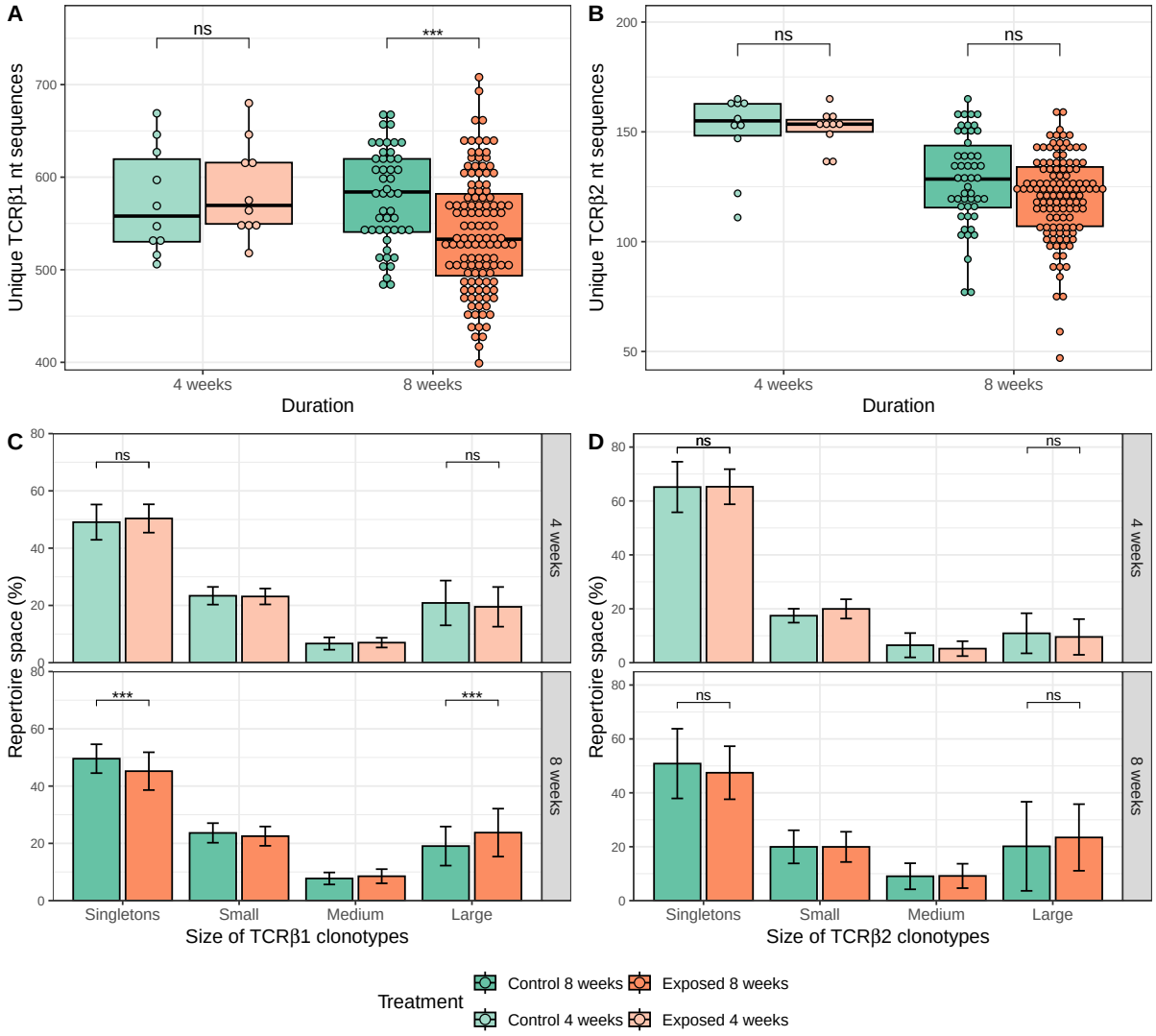


Figure 3.3: A, B) Intra-individual repertoire diversity (number of unique nt sequences) of the TCRβ1 (A) and TCRβ2 (B) repertoires was compared between control and lake-exposed stickleback grouped by exposure duration (four and eight weeks). P values were calculated using GLMMs for each duration group separately, with family, sex, and treatment as fixed effects and cage as random effect. **C, D)** Clonal space homeostasis showing the average percent of clonal space occupied by clones of a given size. Singletons, small, medium and large clones of the TCRβ1 repertoire (C) were defined as corresponding to 1, 2-4,5-7 and >=8 UMIs, respectively. For the TCRβ2 repertoire (D), the same categories correspond to 1, 2-3,4-5 and >=6 UMIs, respectively. Comparisons were made between treatment groups by exposure duration. For singletons and large clonotypes, GLMMs were fitted with family, sex, and treatment as fixed effects and cage as random effect.

3.4.3 Effect of *S. solidus* co-infection on parasite load and TCR repertoire diversity

Out of the 110 fish that were eight-week-exposed and subsequently screened for macro-parasites, nine (8.18%) were infected with *Schistocephalus solidus* (Figure 3.2B); seven individuals had a single *S. solidus* plerocercoid while the remaining two were carrying two plerocercoids in their body cavity. All nine *S. solidus*-infected individuals were simultaneously co-infected with other macro-parasite taxa, ranging in number from one to three. To test whether the average parasite load of *S. solidus*-infected individuals differed from that of non-*S. solidus*-infected individuals, we performed Monte-Carlo sampling. The distribution of mean parasite load of nine randomly sampled individuals from the eight-week-exposure group was calculated over 5,000 iterations. In each iteration, random individuals were matched with the *S. solidus* infected group with regards to family and cage, in order to control for their observed effects on the parasite load (Table C.3). The mean parasite load of *S. solidus*-infected individuals fell well within the 95% confidence interval of the parasite load distribution of non-*S. solidus* infected individuals (Figure C.4). As a second proxy of immune modulation, we also tested the rate of co-infection with *Diplostomum* sp. We observe no statistically significant difference in the proportion of *Diplostomum* sp. co-infection between the two groups (*S. solidus*-infected 11.1%, *S. solidus* uninfected 29.7%, Fisher's exact test: P=0.44).

Similar statistical analysis was employed to explore the diversity of TCRβ1 and TCRβ2 repertoires of *S. solidus*-infected individuals in comparison to non- *S. solidus*-infected ones. Because only seven *S. solidus*-infected individuals were TCR-sequenced, comparisons were made between seven random eight-week exposed fish. For both repertoires, Monte-Carlo sampling was limited to individuals which were sex-matched with the *S. solidus*-infected group, so as to account for its observed effect on repertoire diversity (Table C.7). Results indicate that fish infected with *S. solidus* do not exhibit different diversity levels in either repertoire in comparison to not infected fish (Figure C.5A, B).

3.4.4 Public clonotypes and TCR similarity network architecture

Next, we asked whether exposure to the lake affected the presence and/or abundance of public clonotypes, *i.e.*, clonotypes shared by most than half of the fish in our dataset. Our sequencing protocol produced 19,345 unique TCRβ1 nt sequences, which corresponded to 17,643 clonotypes. Following the logic of the analysis in **Chapter 2**, we defined a clonotype as "private" if it appeared in only one individual, as "relatively private" if it was shared by two to 24 fish, as "relatively public" if shared by 25–96 fish, and as "public" if shared more than half the fish (n>96). The overwhelming majority of clonotypes (17,407 or 98.7%) belong to the private and relatively private compartments, while only 1.3%, or 236 clonotypes, were shared by more than an eighth of the individuals (n>24). Of those, only 8 clonotypes were public (0.05% of total). The TCRβ2 repertoire included a total of 13,651 unique clonotypes; of those, only 5, or 0.04%, belonged to the relatively public compartment, while none were public. Due to the very low count of widely shared clonotypes in both repertoires, we will hereafter consider relatively public and public clonotypes together in our analysis; their presence and abundance will be compared to that of the private and relatively private compartments.

We first examined how lake exposure affected the composition of the stickleback repertoires with regards to public/relatively public clonotypes. Public and relatively public clonotypes occurred in the repertoires of both fish groups. Interestingly, we found that public and relatively public clonotypes from repertoires of individuals exposed to the lake for the full period of eight weeks occupied, on average, more space compared to their control counterparts;

this same was not observed in the four-week exposure group (Figure 3.4A). To examine if the structure of repertoire similarity networks of exposed fish differed from the control, we calculated the number of edges, a measure of network connectivity. We found that TCRβ1 repertoire networks constructed with the top 250 most abundant clonotypes of eight-week-exposed individuals have significantly lower numbers of edges in comparison to the control group (Figure 3.4B). Furthermore, the top 250 most abundant clonotypes of eight-week-exposed individuals had a higher Gini index than those of control fish, indicating a more uneven abundance (Figure 3.4C). The above analysis was extended to the TCRβ2 repertoire; eight-week-exposed individuals did not differ in the proportion repertoire space occupied by (relatively) public clonotypes, neither in the connectivity nor the Gini index of the top 50 clonotypes differed from that of their control counterparts (Figure C.6).

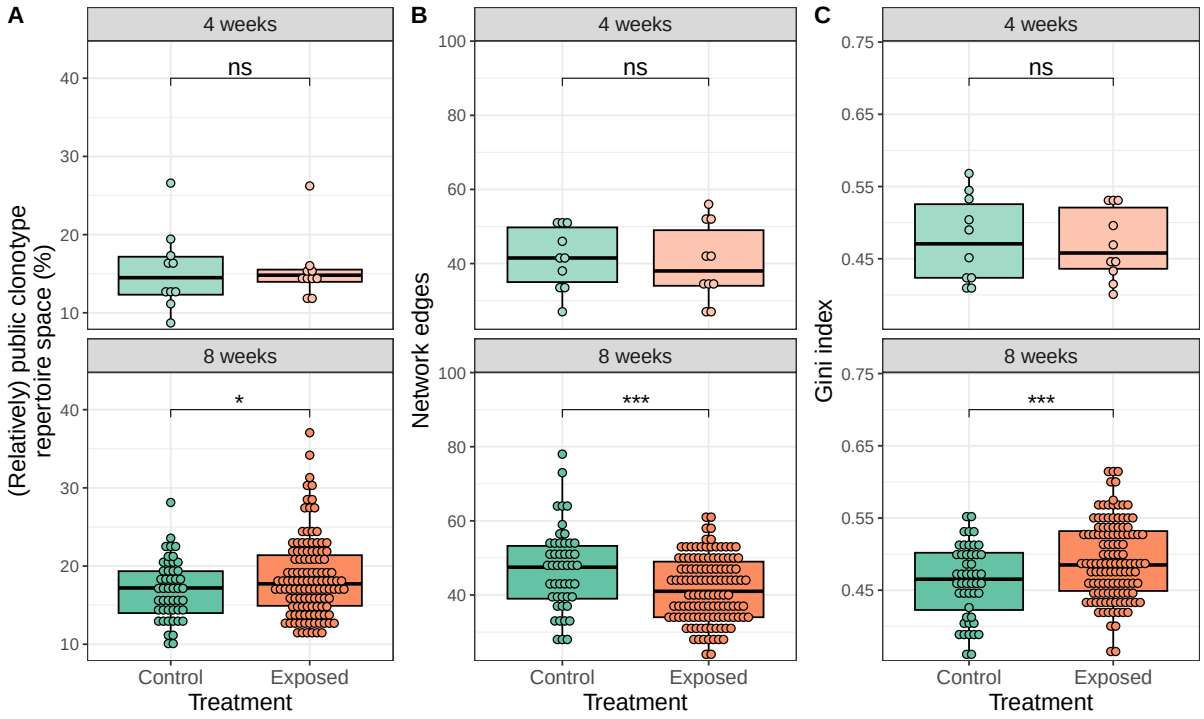


Figure 3.4: TCRβ1 repertoire similarity networks are less connected in eight-week-exposed individuals. A) Comparison of the occupied repertoire space (*i.e.*, cumulative frequency) of public and relatively public (n>24) clonotypes among the entire down-sampled TCRβ1 repertoires of control and lake-exposed sticklebacks, by duration of exposure. **B)** Comparison of the number of edges connecting nodes in each individual network between control and lake-exposed sticklebacks, by duration of exposure. Networks were constructed using the top 250 most abundant clonotypes of each individual. **C)** Comparison of the Gini index (a measure for repertoire evenness) between control and lake-exposed sticklebacks, by duration of exposure. Asterisks in all panels signify P value significance levels for factor treatment, as calculated using GLMMs for each duration group separately, with family, sex, and treatment as fixed effects, and cage as random effect (ns: non-significant $P > 0.05$, * $P < 0.05$, *** $P < 10^{-3}$).

3.4.5 Effect of aging on TCR repertoire size, diversity, and clonality

In this experiment, we observed smaller sample sizes in both TCRβ1 and TCRβ2 repertoires compared to those from the single-infection *Schistocephalus solidus* experiment (**Chapter 2**), despite following the same TCR sequencing protocol. For both experiments we used fish that were lab-bred in the summer of 2019. A clear difference was the age of the sticklebacks at the time of the experiments: those employed in the *S. solidus* experiment were approximately nine months old, whereas individuals used for the Lake exposure experiment were between 22 and 23 months old. Therefore, we decided to conduct a more detailed investigation of the effect of aging on the TCR repertoire of the stickleback. To account for family effects, which we have shown can influence the repertoire diversity, we selected individuals from 3 stickleback families used in both experiments (total of 38 fish, Table C.8). Furthermore, we limited our analysis to individuals belonging to control groups, *i.e.*, fish that were maintained in the lab aquaria throughout their life.

We first confirmed that the selected subset of individuals maintained the representative difference in the size, *i.e.*, total number of functional nt sequences, of both TCRβ1 and TCRβ2 repertoire samples (Figure C.7A); cumulatively, the TCRβ repertoire samples of older sticklebacks were on average 2.5 times smaller compared to those of their younger counterparts (2,808 to 6,999 TRBMs). The same starting amount of RNA was used for all individuals in our protocol; therefore, any difference in repertoire size should correspond to variation in the abundance of TCRβ RNA molecules in the given RNA quantity. We used real-time PCR (qPCR) for quantification of relative expression levels of T-cell-specific genes. Curiously, older fish had a markedly lower (approximately 3.6 times) CD8α relative expression, but CD4.1 and CD8β expression levels were unchanged (Figure C.7B). Nevertheless, as CD8α amounted to two thirds of the within-sample cumulative T-cell-specific gene expression, a significant drop could help explain the smaller size of the repertoire in older sticklebacks.

We thus confirmed that older individuals had a smaller total TCRβ repertoire within a given amount of spleen RNA. In order to account for the difference in the amount of spleen RNA recovered from each individual, we compared the size of the repertoires after normalization for the proportion of spleen RNA used per individual (Materials and Methods). Interestingly, we see a small decline in the normalized size of the TCRβ1 repertoire of old fish (Figure 3.5A). Confirmation of this small reduction in size of the TCRβ1 repertoire in older individuals led us to our next question: does aging also affect the repertoire diversity? To answer this, we standardized all samples to the same size, *i.e.*, 942 TCRβ1 UMIs and 199 TCRβ2 UMIs. Older sticklebacks had a considerably more limited repertoire diversity in both TCRβ1 and TCRβ2 repertoires (Figure 3.5B). Accordingly, the clonal distribution revealed a significantly higher proportion of expanded clonotypes and a smaller proportion of singletons in both repertoires of aged individuals (Figure 3.5C, D).

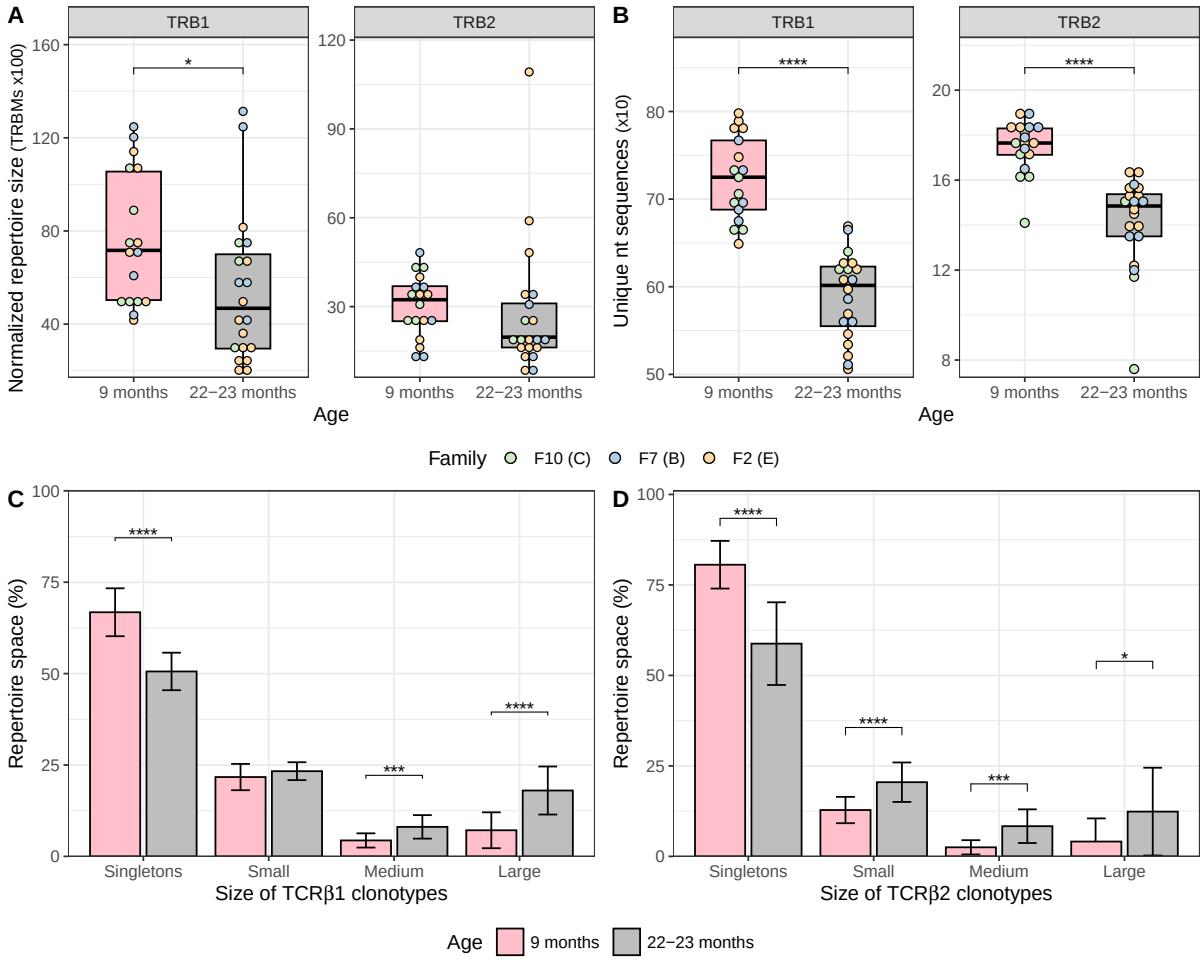


Figure 3.5: Effect of aging on normalized TCR repertoire size, diversity and clonality. A) Comparison of the size of the TCRβ1 (left panel) and TCRβ2 (right panel) repertoires by age group following adjustment for the ratio of spleen RNA used per individual. P values were calculated using GLMs for each repertoire separately, with family, sex, and age as fixed effects. Asterisks signify P value significance levels for factor age (* P < 0.05). **B)** Mean per individual observed diversity (number of unique nt sequences) obtained from down-sampled repertoires was compared between age groups. P values were calculated using GLMs for each repertoire separately, with family, sex, and age as fixed effects. Asterisks signify P value significance levels for factor age (**** P < 10^{-4}). **C, D)** Clonal space homeostasis showing the average percent of clonal space occupied by clones of a given size in the TCRβ1 (C) and TCRβ2 (D) repertoire. For the TCRβ1 repertoire, singletons, small, medium and large clones corresponded to 1, 2-4, 5-7 and >=8 UMIs, respectively. For the TCRβ2 repertoire, singletons, small, medium and large clones corresponded to 1, 2-3, 4-5 and >=6 UMIs, respectively. For each size type, GLMs were fitted; for statistically significant cases, asterisks on top of a category signify P value significance levels for factor age (****P < 10^{-4}, *** P < 10^{-3}, * P < 0.05). Family and sex were also incorporated in all models as fixed effects. Non-significant comparisons are omitted for visual clarity.

3.4.6 Effect of aging on clonotype sharing and TCR repertoire similarity network structure

We proceeded to examine the composition of the down-sampled repertoires with regards to clonotype sharing. The proportion of relatively public and public clonotypes (*i.e.*, clonotypes present in more than a quarter of individuals) in the TCRβ1 repertoires of older fish was almost three times lower compared to younger ones (Figure 3.6C); indeed, we found a total of 253 relatively public and public clonotypes among the repertoires of younger fish, but only 96 such clonotypes among older individuals. In terms of repertoire space, relatively public and public clonotypes occupied 24% of homeostatic space in young sticklebacks on average, but only 12% in older ones ($F_{1,32}$= 59.2, P= 9e-09, from GLM also including sex and family). Notably, even though the V gene usage among private expanded clonotypes was comparable between the two groups, the V gene usage profile of relatively public and public clonotypes in old individuals was markedly different than that of younger fish: usage frequency of gene V1.2, in particular, was elevated, while that of V1.17 was reduced (Figure C.8).

We asked whether there is an overlap between the (relatively) public clonotypes of young and old sticklebacks; we detected 24 such common clonotypes (Table 1), the average intra-repertoire abundance of which did not differ between the two age groups (Figure C.9). By extending our search to private and relatively private clonotypes, we found that 50 (or 52%) of the relatively public and public clonotypes of old individuals were also present in one or more younger individuals. Inversely, 108 (43%) of (relatively) public clonotypes of young individuals were detected in at least one older individual.

To investigate whether the presence of fewer widely shared clonotypes in older individuals affected their repertoire architecture, we used the top 250 most abundant clonotypes from each individual to build sequence similarity networks. As can be observed from the examples of a young and an old individual (Figure 3.6A, B), TCRβ1 repertoires of older fish were less connected, *i.e.*, had significantly reduced number of edges (Figure 3.6D). Furthermore, old individuals had less evenly distributed abundance among the top 250 clonotypes, as evidenced by a significantly higher Gini index (Figure 3.6E). We repeated the above analysis for the TCRβ2 repertoire, and found a statistically significant drop in the percentage of relatively public and public clonotypes in older compared to younger sticklebacks; in fact, there were no such clonotypes in the TCRβ2 repertoires of the 22-23-month-old fish (Figure C.10A). This drop did not translate to any notable change in the connectivity of the repertoire (Figure C.10B); nevertheless, TCRβ2 repertoire evenness, as measured by the Gini index, was reduced in older sticklebacks (Figure C.10C).

Table 1. The 24 relatively public and public clonotypes shared between TCRβ1 repertoires of young (nine-month-old) and old (22-23-month-old) three-spined sticklebacks, and the number of individuals within each group in which they were present.

CDR3β aa sequence	V gene	Young individuals (N=17)	Old individuals (N=20)
CAARDNNEAYF	V1.17	17	16
CAARQGAEAYF	V1.17	17	7
CAARSYAEAIF	V1.5	16	16
CAARDTEAYF	V1.17	15	7
CAVSQGYEAYF	V1.2	14	8
CAARTGYEAYF	V1.17	14	9
CAARDNEAYF	V1.17	13	10
CAARQGNEAYF	V1.17	12	8
CAARQYEAYF	V1.17	12	7
CAAWSYAEAIF	V1.8	11	7
CAARSYGEPAYF	V1.5	11	7
CAASQYEAYF	V1.1	11	10
CAARTAYAEAIF	V1.5	10	7
CAVSLGGYGEPAYF	V1.2	10	6
CAVSQGNYEPAYF	V1.2	9	8
CAVSQGGYGEPAYF	V1.2	9	6
CAARTGNEAYF	V1.17	9	7
CAVSQGYTEAYF	V1.2	8	7
CAASNYEPAYF	V1.17	7	10
CAVSNNYEPAYF	V1.2	7	10
CAVSQGAYAEAIF	V1.2	7	7
CAVSRQYTEAYF	V1.2	7	8
CAARSYGASEAYF	V1.5	6	9
CAVSHSYAEAIF	V1.2	5	7

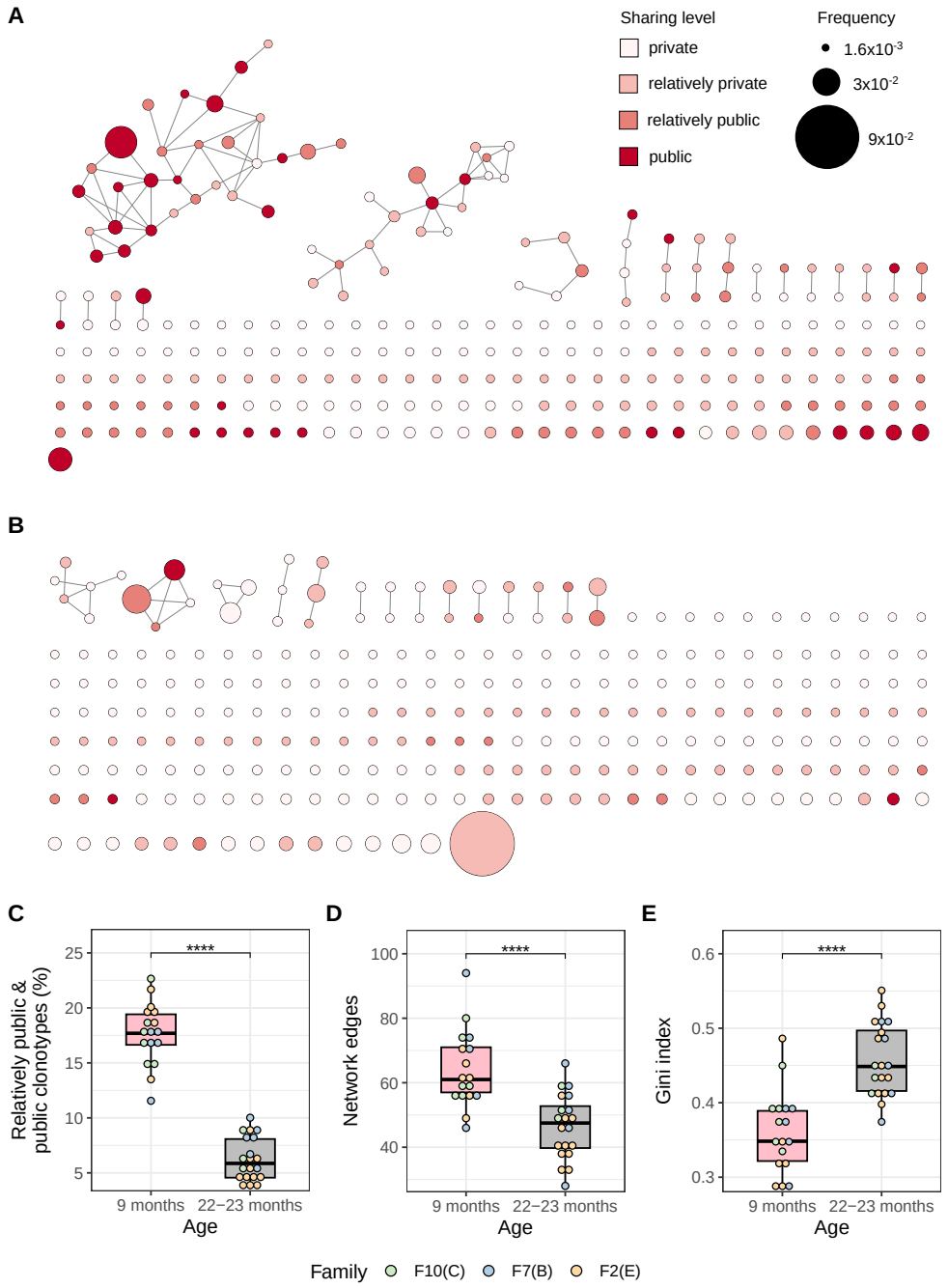


Figure 3.6: Stickleback TCRβ1 repertoire similarity networks become less connected with age. A, B) Network formed by the 250 most frequent CDR3 aa sequences expressed in the splenic TCRβ repertoire from a nine-month-old (A) and a 23-month-old (B) stickleback. Nodes (CDR3 sequences) were connected by edges defined by a Levenshtein distance of 1. Node size reflects the relative frequency of each clonotype among the top 250 (panel A inset). The nodes are colored according to their sharing levels (panel A inset). **C)** Comparison of the proportion of relatively public and public clonotypes among the entire down-sampled TCRβ1 repertoires of young and old sticklebacks. Clonotypes were regarded as relatively public and public if present in more than one quarter of the individuals in each group. **D)** Comparison of the number of edges connecting nodes in each individual network between young and old sticklebacks. (E) Comparison of the Gini index (a measure for repertoire evenness) between young and old sticklebacks. Asterisks in panels C, D, and E signify P value significance levels for factor age, as calculated using GLMs with family, sex, and age as fixed effects (**** $P < 10^{-4}$).

3.5 Discussion

3.5.1 Parasite load is dependent upon duration of exposure, host genetic background and sex

We observed a sharp increase in macro-parasite infection, both in the number of taxa and in parasite load from four- to eight-week-exposed fish. Temperature is an environmental factor that might play a role in this difference. Four-week-exposed fish were retrieved from the lake as water temperature had risen approximately 4℃ since the start of the experiment. Furthermore, for at least the first 11 days of this four-week period, the water temperature was mostly stable at 12℃. Conversely, eight-week-exposed fish underwent a >11℃ increase in temperature, with a large portion of this period spent in temperatures above 20℃. The efficiency of the teleost immune system is highly susceptible to temperature variation, and is compromised following extended exposure to higher temperatures (Scharsack and Franke, 2022). For example, head kidney lymphocyte activation against *S. solidus* is much higher at relatively low (<15℃) temperatures, which also coincide with better control of parasite growth (Franke et al., 2017; Scharsack et al., 2021). Eight-week-exposed sticklebacks could therefore have been more vulnerable to infections, especially towards the end of the exposure period, when temperatures reached severe levels (Wegner et al., 2008). Generally, stickleback infections follow a seasonal pattern, with higher incidence in the summer compared to spring, in accordance with the abundance and level of infection of the invertebrate intermediate parasite hosts (Pennycuick, 1971). Furthermore, pathogen-pathogen interactions might also affect the parasite burden: infection with one parasite can predispose to a secondary infection. For example, *S. solidus* infection is thought to increase the rate of co-infection with *Diplostomum* sp. (Piecyk et al., 2019a). Lastly, it is conceivable that for some parasites, the time needed to become detectable inside the stickleback is prolonged in lower temperatures due to restrained growth, as is the case with *S. solidus* (Scharsack et al., 2021); this could make their early detection challenging, and thus inflate the parasite load of longer-exposed fish.

Individual parasite load differed significantly between families, in agreement with previous stickleback field studies in the lake (Wegner et al., 2008). On the other hand, individual MHC-II diversity did not significantly impact parasite load, contrary to some previous lake exposure results (Wegner et al., 2008), and studies on wild-caught sticklebacks (Wegner et al., 2003b). It should be noted, however, that due to our dataset was unbalanced with regards to individual MHC-II diversity, and thus likely cannot capture the full effect of the latter. Furthermore, the full duration of exposure in the previous field experiment was almost double that of ours (approximately 105 versus 55 days) (Wegner et al., 2008), which is also bound to influence our observations and might explain differences among studies with regard to MHC effects. Family-based differences in parasite resistance could reflect variation in genetic elements pertaining to the function of the innate immune system, which comprises the first line of defense against macro-parasites. The genetic basis of parasite resistance has been demonstrated in invertebrate host–parasite systems, which rely heavily on innate immunity (Carton et al., 2005). Innate immunity has been implicated in the clearance of macro-parasites in the stickleback within the first few days of infection, before the onset of adaptive immune responses (Wedekind and Little, 2004; Wegner et al., 2007; Barber and Scharsack, 2010).

A factor that did have a significant effect on parasite loads was sex: female sticklebacks exposed to the lake for the full experimental duration exhibited higher parasite loads than males, in agreement with previous results (Kalbe et al., 2009). Such differential stickleback parasitism between the sexes has been linked to dietary differences due to higher energy demands of fe-

males (Reimchen and Nosil, 2001; Kalbe et al., 2009). In agreement with this potential dietary link, our results showed a strong cage effect on individual parasite load, as also previously seen (Kalbe et al., 2009). Fish kept in the same cage are likely to consume similar food items, including intermediate hosts of trophically transmitted parasites. Alternatively, this discrepancy could signal lower immunocompetence of female sticklebacks – this hypothesis is further discussed below.

3.5.2 The TCR repertoire of exposed fish exhibits signs of activation

As per our initial hypothesis, the TCRβ repertoire of the lake-exposed fish was significantly changed in terms of diversity and clonality compared to control fish. Specifically, expanded clonotypes of eight-week-exposed fish occupied significantly greater homeostatic space, thereby reducing the overall diversity of the TCR pool. Both are signatures of TCR repertoire mobilization against pathogenic challenges (*e.g.*, Callan et al. (1998); Musvosvi et al. (2023)). One interesting finding is that the above changes were limited to TCRβ1, the diversity of which was not dependent on stickleback family. Conversely, TCRβ2 diversity was genetic background-dependent, but not significantly affected by lake exposure. Notably, the TCRβ2 repertoire was also under stronger family influence in the *S. solidus* experiment (**Chapter 2**). The mechanisms behind this dichotomy is unknown, but might reflect innate differences in the T cell type composition (*e.g.*, cytotoxic, regulatory, or helper T cells, unconventional T cells) of each repertoire. If so, the diverse roles of T cell subsets- which encompass distinct response trajectories following environmental triggers (Liang et al., 2018; Sun et al., 2023)- could help explain the differences in diversity and clonality between the repertoires. It is also possible that the two TRB loci of the stickleback operate under the regulation of distinct genetic or epigenetic mechanisms.

T cell activation, like macro-parasite infection, was time-dependent: there were no meaningful changes in the diversity or the clonality of the TCR repertoire of four-week-exposed fish. Given that the control fish were possibly exposed to microbes (*i.e.*, bacteria and viruses) present in the aquaria freshwater, but not to macro-parasites, the above results possibly hint at a macro-parasite-triggered adaptive immune activation during the full duration of the experiment. This would align with previous findings, linking the resistance of sticklebacks to various macro-parasites with MHC genetics, and, by extension, to adaptive immune responses (Wegner et al., 2003a; Lenz et al., 2009b; Eizaguirre et al., 2011, 2012b,a). One would thus anticipate a relationship between the severity of an individual's parasitic burden and the magnitude of the T cell response. The positive correlation between average per family parasite load and average per family repertoire diversity in both repertoires, could indeed be indicative of protective T cell responses against macro-parasite infection, given that lower repertoire diversity is a common signature of clonal expansion.

Besides contact with the micro- and macro-parasitic fauna, lake exposure also grants access to diverse food sources, such as planktonic invertebrates (copepods and daphnids), insect larvae, and small snails. On the other hand, aquaria-kept control fish were exclusively fed frozen food. The richer diet of the lake fish could have affected the immune system both directly (*i.e.*, more energy to allocate for immune responses), and indirectly through induced changes in the gut microbiome (Bolnick et al., 2014). Recent studies indicate that the composition of symbiotic gut microbial communities can shape the TCR repertoire (Ivanov et al., 2022; Nagashima et al., 2023). Hence, it is conceivable that the dissimilar diets between the two treatment groups contributed to the observed TCR repertoire changes. Our experimental design prohibits the disentanglement of the effects of fish diet and macro-parasite infection

on TCR diversity, especially considering that five out of the twelve macro-parasites found in this study are transmitted trophically to the stickleback (Eizaguirre et al., 2012a). Follow-up experimental studies would be needed to specifically address microbiome influence on the TCR dynamics, and help elucidate to what extent substantial TRC repertoire changes, such as the ones observed here, can be attributed to macro-parasite infection, changes in microbiome composition, or a combination thereof.

3.5.3 *S. solidus* co-infected fish do not exhibit signs of immunomodulation

Specialist parasites are more likely to be locally adapted compared to generalists, *i.e.*, parasites that have a wide range of host species (Lajeunesse and Forbes, 2002). *S. solidus*, in particular, has been proposed to have evolved fast growth and strong immunomodulatory capabilities against a frequently exposed and immunologically adept local stickleback population (Piecyk et al., 2019a). Under such escalated evolutionary 'arms-race' dynamics, highly resistant sticklebacks are expected to successfully impede *S. solidus* growth, and therefore delay the effects of immunomodulation (Kalbe et al., 2016; Piecyk et al., 2019a). On the other hand, the individuals in our experiment originate from a low-resistance local stickleback population, presumably evolved under de-escalated arm-race dynamics: *S. solidus* prevalence is low, and parasite strains from this region exhibit relatively slow growth rate and low virulence (Franke et al., 2014; Piecyk et al., 2019a,b). Indeed, local fish experimentally infected with slow-growth *S. solidus* do not show signs of immune modulation, *i.e.*, no elevated co-infection rates with *Diplostomum* sp. (Piecyk et al., 2019a). We thus did not expect that infection with the local *S. solidus* would impact the co-infection rate with *Diplostomum* sp., specifically, or the mean individual parasite load, generally. Indeed, our results were in line with previous literature. Furthermore, as yet another proxy for immunomodulation, we investigated the TCR repertoire diversity of *S. solidus*-infected fish compared to uninfected ones. No difference was observed in either TCRβ1 or TCRβ2 repertoire; this is in accordance with our results from **Chapter 2**, whereby only infection with a virulent *S. solidus* strain and not a slow-growth one, slightly increased the TCR repertoire diversity.

3.5.4 Sex as important factor shaping TCR repertoire diversity

Unlike our **Chapter 2** experiment, where the lab conditions under which sticklebacks were maintained were not permissive of sexual maturation, in this study, sticklebacks were allowed to mature. Remarkably, stickleback sex was the factor with the strongest overall effect on the TCR diversity. The effect of sex was significant for the diversity of both TCRβ1 and TCRβ2 repertoires of lake-exposed individuals, but was limited to TCRβ2 in control fish. The synergistic effects of environmental temperature and photoperiod (O'Brien et al., 2012), as well as a rich natural diet (**?**Reyes and Baker, 2017), which was accompanied by significantly enhanced body condition (Chellappa et al., 1989), could have propelled the sexual maturation of lake-exposed fish - apparent from the red coloration of males. To the best of our knowledge, this is the first report of sex-related differences in the TCR repertoire diversity of a teleost species. However, this observation is not surprising, as the vertebrate immune system functions in close interaction with the endocrine system (Balm, 1997; Segner et al., 2017), and exhibits sexual dimorphism in its performance across diverse vertebrate taxa, including teleosts (Segner et al., 2017).

What is surprising is that male sticklebacks had generally higher TCR diversity than females, whereas the opposite has been observed in mammals, such as humans (Britanova et al., 2016; Krishna et al., 2020), and small rodents (Migalska et al., 2019). In humans, androgens have been associated with a suppressive effect on the immune system (Trigunaite et al., 2015; Foo et al., 2017). T cells, in particular, show a lower degree of MHC-mediated selection and expansion in males (Schneider-Hohendorf et al., 2018). Unlike mammals, where the overwhelming majority of species offer maternal-only care (Woodroffe and Vincent, 1994), in sticklebacks it is the males that provide the necessary care for the survival of their offspring (Pressley, 1981; Wootton, 1984). Parental care -typically associated with females- has long been linked to better immune function (Rolff, 2002). Remarkably, this observation has been extended to male pipefish and seahorses, both sex-role reversed fish species (Roth et al., 2011; Lin et al., 2016). It is tempting to hypothesize that the similar parental care strategy of sticklebacks could be associated with the higher male TCR repertoire diversity reported here, which was also accompanied by a lower overall parasite load.

3.5.5 Public TCR clonotypes persist following extended lake exposure

Diminished TCRβ1 diversity in lake-exposed sticklebacks was not accompanied, as we predicted, by a smaller public clonotype repertoire space. On the contrary – we observe a slightly larger public TCRβ1 clonotype compartment in eight-week-exposed fish compared to controls. Superficially, this contradicts the principle that T cell activation through immunization and infection results in fewer shared clonotypes (Venturi et al., 2016; Madi et al., 2017), as repertoires diverge from a more similar, naive state (Britanova et al., 2016). There could be a number of reasons behind this contradiction. Firstly, unlike these previous studies, our eight-week-exposed fish faced challenges with multiple pathogens, including helminths. The expansion of public clonotype space seen here resembles the one following virulent *S. solidus* infection described in **Chapter 2**. An explanation for the persisting public clonotypes, therefore, would be a helminth-induced suppression of what would otherwise be prominent private T cell responses to other parasites (Taylor et al., 2012). It is also conceivable that part of the public responses we observe in this experiment are shared only within lake-exposed individuals, and driven by common antigenic stimulation by ubiquitous antigenic triggers (Li et al., 2012). After all, public TCRs do appear in pathogen-specific TCR repertoires (Miles et al., 2011), especially against dominant epitopes (Lim et al., 2000). In support of this, public clonotypes in our experiment are skewed towards usage of the V1.2 gene segment, most commonly used by private expanded clones. Further analysis of our dataset will clarify whether this hypothesis is correct, by calculating what percentage of public clonotypes are exclusive to the lake exposed individuals, and, crucially, whether these public clonotypes are MHC-restricted.

Furthermore, public clonotypes could also include expanded, MHC-unrestricted T cells which target commensal bacteria (Linehan et al., 2018). Diet change of fish in the lake could have elicited shared responses to common gut microbes. This theory is especially fitting, as such commensal-specific responses by unconventional T cells not only promote protection to pathogens, but also accelerate repair of damaged tissue (Linehan et al., 2018), which is a distinctive by-product of helminth infection (Anthony et al., 2007; Chen et al., 2012). The above align well with our hypothesis from **Chapter 2**, according to which part of the public TCRs in the stickleback stem from invariant NK-like T cells. Lastly, we cannot exclude a scenario where an initial first wave of pathogenic insults led to private T cell responses dominating the repertoire space during the first few weeks. Thereafter, public clonotypes return to their

baseline levels. Such recovery of public clonotypes has previously been described in mice following immunization (Madi et al., 2017). In this case, we could have missed any changes in the frequency of public clonotypes due to the late sampling time point. In any case, it is important to acknowledge that elucidating the dynamics of TCR clonotype sharing in a natural setting with simultaneous infections at different stages of progression is a formidable task, and any overly-simplistic interpretations should be approached with caution.

3.5.6 Extended lake exposure leads to perturbed TCR similarity networks

The observed TCRβ1 repertoire mobilization in eight-week-exposed fish was further confirmed in our study by the finding that the similarity networks of corresponding repertoires displayed signs of disruption, *i.e.*, reduced levels of connectivity, and higher abundance inequality amongst top clonotypes. Similar changes in both indices have previously been observed in mouse TCR repertoires following immunization and *in vitro* antigen stimulation (Madi et al., 2017). In both studies, public clonotypes contributed a higher average degree of connectivity than private ones. Nevertheless, here, the reduced network connectivity cannot be attributed to a decrease in public clonotypes as in the study by Madi et al. This is a further indication that at least some public clonotypes present in the repertoires of eight-week-exposed fish were not well-connected within the network, reminiscent of the attributes of isolated clonotypes emerging in response to antigenic stimulation (Madi et al., 2017).

3.5.7 Age diminishes the TCR repertoire diversity and changes its architecture

The stickleback is a relatively short-lived fish; individuals usually live for one to two years, and most do not survive until a second spawning season (Wootton et al., 1978; Wootton, 1984). Fish used in this study, therefore, were already of advanced age (almost two years old) by the end of the experiment. As some of them belong to three families also used in the *S. solidus* experiment described in **Chapter 2**, we took the opportunity to study the effect of age on the TCR repertoire of the stickleback, while controlling for genetic background. Our analysis employed individuals that served in control groups, lacking any contact with macro-parasites for the entirety of the subjects' life.

We observed a small decline in the extrapolated total spleen size of the TCRβ repertoire in older individuals. This stands in agreement with previous studies in humans showing a minor but clear decrease of total T cell counts in blood with aging (Yan et al., 2010; Britanova et al., 2016). We further showed that the TCRβ repertoire decline in older sticklebacks coincides with a decrease in the relative number of $CD8^-$expressing T cells. Interestingly, reports in humans indicate that age-related TCR repertoire attrition is subset-specific and more profound for $CD8^+$ than CD_4^+ T cells (Sun et al., 2022). This age-related contraction of the repertoire has been linked to limited thymic output of new T cells in advanced stages of life, termed thymic involution (Pido-Lopez et al., 2001). A series of human studies demonstrate that thymic atrophy is accompanied by clonal expansions of antigen-driven T cells, leading to a decrease in the number and proportion of repertoire space occupied by naive T cells; consequently, old age is characterized by decreased TCR diversity (Britanova et al., 2014, 2016; Yoshida et al., 2017; Egorov et al., 2018). Our results are therefore in accordance with the above observations, as the TCRβ repertoire diversity of older sticklebacks was significantly constricted compared to younger fish. Again, $CD8^+$ T cells are more affected by this age-related diversity decrease,

suggesting that aging has a more profound effect on cytotoxic as opposed to helper T cell functions (Ahmed et al., 2009; Yoshida et al., 2017; Sun et al., 2022). Future targeted sequencing of sorted T cell subsets of the stickleback will better elucidate these aging dynamics, by also considering its -presumably substantial- $CD4^{-}CD8^{-}$ T cell population, as observed in other teleost species (Sun et al., 2022; Hu et al., 2023).

Alongside the age-induced constriction in TCRβ repertoire diversity of sticklebacks, we also detected a prominent shrinking by 50% of the public clonotype compartment in repertoires of older individuals. Diminishing of public clonotypes with age has been established in humans and mice alike, as these give way to more private, antigen-driven clonotypes (Britanova et al., 2014, 2016; Madi et al., 2017; Pogorelyy et al., 2017). As discussed in **Chapter 2**, public clonotypes in humans are believed to form a common 'core' repertoire; while prominent at birth, this common core slowly decays with age (Britanova et al., 2016; Pogorelyy et al., 2017). Importantly, the frequency of unconventional T cells is also known to decline with advanced age (as reviewed by (Kurioka and Klenerman, 2023). Here, half of public clonotypes observed in repertoires of old individuals were also present in young repertoires. Therefore, our results support the 'core' repertoire theory of prenatally produced public clonotypes slowly fainting throughout an individual's lifetime (Pogorelyy et al., 2017). However, here we also observed that common public clonotypes between young and older fish had the same relative abundance. This suggests that at least a portion of public clonotypes are dependent on selection and clonal expansion, as opposed to the proposed passive maintenance via homeostatic proliferation in the periphery (Britanova et al., 2016). This stands as further evidence towards a portion of the stickleback public clonotypes representing unconventional, MHC-unrestricted T cells serving a biological function (**Chapter 2**).

3.6 Conclusion

This study aimed at exploring the changes pertaining to the TCR repertoire in response to semi-natural settings by employing a state-of-the-art sequencing methodology and the established three-spined stickleback eco-evolutionary model. We show that exposure to a natal lake environment with natural parasites is associated with detectable T cell activation and disruption of clonotype clusters in TCRβ1 repertoire similarity networks- while public clonotypes persevered. To the best of our knowledge, this study is also the first report of age-related changes in the TCR repertoire of a teleost species. Remarkably, the changes are largely analogous to those seen in mammals, including humans. This speaks to the common fundamental principles governing the organization of adaptive immune repertoires across the vertebrate clade throughout an individual's lifetime.

Synthesis

"Nature uses only the longest threads to weave her patterns, so that each small piece of her fabric reveals the organization of the entire tapestry."

– Richard P. Feynman

The main goal of this thesis was to explore fundamental questions regarding the vertebrate adaptive immune system, and more specifically, how the T cell receptor (TCR) repertoire responds to infection. To do so, I used the vertebrate three-spined stickleback (*Gasterosteus aculeatus*) as a model with which I conducted both a laboratory and a field experiment, exploring the response of the repertoire to infection with a single and multiple macro-parasites, respectively. Additionally, by employing a species which captures the existing natural population diversity, I was able to assess the effect of the individual genetic variation of the host on various aspects of the TCR repertoire.

The adaptive immune system of jawed vertebrate relies upon recognition of self from non-self, which is primarily achieved through MHC molecules. In **Chapter 1**, I have shown that the three-spined stickleback has a fully developed and highly diverse set of MHC-I and MHC-II genes showing strong patterns of differential gene expression. This was achieved through the establishment of a reliable and easily adaptable NGS protocol, which allows for the detection and characterization of previously undetected MHC alleles in natural populations of species whose MHC genes are hard to target individually. Furthermore, this protocol is not limited by template, as it can be used with cDNA as starting material. In fact, incorporating cDNA into studies of MHC diversity can provide valuable insight into the levels of expression of each allele – and link to health and disease.

When a vertebrate host is infected with a pathogen, MHC molecules on the surface of antigen-presenting cells present pathogen-derived peptides to T cells, the central mediators of both humoral and cellular adaptive immunity. The T cells exhibit an enormous diversity of T cell receptors, which constitutes the TCR repertoire. In **Chapter 2**, I performed in-depth analysis of the factors that affect the TCR repertoire diversity of the stickleback under controlled experimental conditions, and in response to a helminth infection. An interesting result that stood out was the significant effect of genetic background on an individual's TCRβ1 and TCRβ2 repertoire diversity. In the following two section, I discuss possible underlying mechanisms and its implications for future research.

The TCR repertoire of an unusual genomic configuration

The stickleback is one of the few vertebrate species known to possess two separate TRB loci, TRB1 and TRB2, both constituting independent VDJ recombination units. This configuration, believed to be the product of block duplication (Wölfing, 2012), contrasts the single translocon arrangement containing two distinct C genes described in other teleosts (Zhou

et al., 2003; Meeker et al., 2010). Instead, it resembles a common immunoglobulin locus organization into multiple independent clusters of cartilaginous fish (Smith et al., 2019). This unusual TRB organization found in the stickleback raises multiple questions pertaining to its evolution and mechanisms of allelic exclusion (Malissen et al., 1992). More relevant to the scope of this thesis, however, is the apparent absence of functional TCRs stemming from inter-locus recombination; provided the same number of gene segments, this architecture results in a reduced combinatorial repertoire diversity compared to a single large translocon (Wölfing, 2012). Here, I have shown that both TRB loci contribute considerably to the total TCR repertoire diversity, and despite the stickleback's peculiar TRB arrangement, the total inferred TCRβ spleen diversity is -proportionally- on par with the numbers reported in zebrafish (Covacu et al., 2016), and minifish (Giorgetti et al., 2021). Curiously, the stickleback does not possess a significantly expanded set of V, D, or J genes, and so compares poorly to the combinatorial diversity of the single translocons from these species (Wölfing, 2012). It appears that other factors are compensating for the stickleback's restricted combinatorial capacity, that result in a functional and diverse TCR repertoire.

TCR repertoire diversity: a family affair?

In every individual, the naïve TCR repertoire is generated as T cells initially go through a semi-stochastic VDJ recombination process of acquiring a TCR (Alt et al., 1992), followed by a process of selection for those T cells expressing 'useful' TCRs (Kyewski and Klein, 2006; Klein et al., 2014). Both steps are reliant on genetic elements, from the genes of the TRB locus providing the building blocks of rearrangement, to those encoding the enzymes driving recombination, to the MHC, whose proteins implement self-antigen presentation. This "complex probabilistic tangle of genetically determined biases" (Russell et al., 2022) largely dictates the diversity of the ensuing repertoire, with inter-individual variation being well-documented in all vertebrate species studied so far, including the stickleback, as shown in this thesis. Which are the genetic mechanisms behind this variance?

TCR repertoire diversity and the MHC background

Due to their closely linked function, MHC diversity has long been assumed to be a primary determinant of the TCR repertoire size, and, consequently, its diversity. The prevalent hypothesis, known as the TCR depletion hypothesis, predicts that high individual MHC diversity results the decreased T cell survival due to higher chances of TCR self-reactivity during negative selection, which in turn leads to drastically smaller repertoire (Nowak et al., 1992). The existence of a suggested trade-off between individual MHC diversity and TCR repertoire has been supported by theoretical models (Woelfing et al., 2008). Conversely, it has been proposed that a high number of MHC alleles could lead to survival of more T cells, by means of boosting positive selection (Borghans et al., 2003). Evidence for either side of this debate has proven challenging to find. Recently, and thanks to the advancement of sequencing technologies, human TCR repertoire diversity was shown to be positively correlated with MHC-I diversity (Krishna et al., 2020). A study in bank voles, on the other hand, reported that increased MHC-I diversity restricts the TCR repertoire, while MHC-II diversity was deemed unimpactful (Migalska et al., 2019). Yet, in mice, MHC-II heterozygosity apparently limits the repertoire size (Brown et al., 2024). These wildly contradictory results speak to how difficult it is to 'isolate' MHC effects from all other relevant genetic and environmental factors at play.

Disentangling the role of MHC genotypes from the genetic background is especially difficult (Rauch et al., 2006). In **Chapters 2** and **3**, I have showed that under standardized experimental conditions, individual MHC-II diversity does not significantly impact the stickleback's TCR repertoire diversity. However, neither the experiments nor the datasets were specifically designed to examine this particular association; a key missing part being the sticklebacks' MHC-I genotypes, shown in **Chapter 1** to result in extended allele number variation between individuals, and previously associated with stronger influence on the TCR repertoire than MHC-II (Migalska et al., 2019). Future dedicated experiments should employ balanced datasets of sticklebacks of identical MHC genotypes belonging to different families; ideally, sticklebacks would be of younger age, so as to eliminate the effects of accumulating antigenic exposure that might "cover" the MHC effect by skewing the repertoire diversity. With the above in mind, the results presented within this thesis do not allow us to exclude MHC diversity as an important factor; they do, however, underline that the connection between individual MHC and TCR repertoire diversities is not as straightforward as once thought.

TCR repertoire diversity and the non-MHC genetic background

An effect that has not received nearly as much attention from researchers is the one exerted by the genetic background as a whole, not just the MHC, on the TCR repertoire. The extent to which individual genetic variation is associated with their resulting TCR repertoire diversity remain largely unexplored (Russell et al., 2022). One of the key findings of this thesis is the substantial role that the non-MHC genetic background plays in the TCR repertoire diversity, as confirmed by two separate experiments on adult sticklebacks (**Chapters 2** and **3**). In principle, the impact of the various genetic mechanisms that govern the diversity of the "naïve", *i.e.*, antigen-inexperienced, repertoire is bound to faint over time, as the repertoire is progressively shaped by succeeding exposures to a variety of pathogens, and the combinatorial effects of all other known and unknown environmental aspects, including diet-induced divergence in microbiome communities. This has been demonstrated in humans, where the TCR repertoires of young individuals form a 'core' repertoire from which they diverge with age (Britanova et al., 2016). It is, therefore, not surprising that the variance in cumulative repertoire diversity attributable to family effects decreased 6-fold (30% to 5%) between 9-month-old (**Chapter 2**) and 22-23-month-old (**Chapter 3**) sticklebacks. Besides MHC, what could be the sources of the observed variation in TCR repertoire diversity between stickleback families?

The role of inter-individual variation in VDJ recombination

Computational models show that VDJ recombination, rather than thymic selection, drives most of the population-level variation in the TCR repertoire (Sethna et al., 2020). VDJ recombination generates an extremely diverse, unrefined pool of T cells, the overwhelming majority of which bear non-viable TCRs (Bradley and Thomas, 2019). This primary TCR repertoire diversity is the amalgamation of two independent sources: combinatorial diversity, generated through the choice of V, D, and J genes, and junctional diversity, created via random nucleotide insertions and deletions. Crucially, inter-individual variation in VDJ recombination-related genes can affect both combinatorial and junctional diversities, resulting in differences in the functional TCR repertoire diversity (Russell et al., 2022).

Variability in V, D, and J gene usage between individuals is substantial (Sethna et al., 2020), and can directly affect the combinatorial diversity of the TCR repertoire (Russell et al., 2022). For example, one can easily imagine how severe bias towards specific V and J genes can significantly undermine the generated diversity. Strikingly, much of the V-J combinatorial diversity

in humans is genetically determined (Pham et al., 2012), a fact supported by the observation that TCR repertoires of genetically identical humans exhibit highly similar gene usage profiles (Zvyagin et al., 2014). Accordingly, here I have shown that sticklebacks from different families differ significantly in V and J usage frequencies. We know that variation at the MHC biases germline gene usage (Sharon et al., 2016; Gao et al., 2019). This family effect, however, was even more pronounced in non-functional, unselected TCR repertoires, which is strongly indicative of the role of genetic factors beyond the MHC in the observed gene usage variation – as already evident in humans (Melenhorst et al., 2008). What are the drivers of such family-based variation in V and J gene usage? Recent studies point towards variation in the germline genes themselves - both structural, *i.e.*, difference in gene copy numbers, and in the form of SNPs (Gragert, 2022; Omer et al., 2022; Russell et al., 2022; Corcoran et al., 2023; Corcoran and Karlsson Hedestam, 2024). Furthermore, variation in cis-elements in the TRB locus, such as recombination signal sequences (RSS), enhancers, and promoters can also differ between individuals (Melenhorst et al., 2008; Miles et al., 2011). The picture gets even more complicated when we consider the trans-regulatory elements of VDJ recombination. All of the above can lead to recombination biases, and, in turn, impact the combinatorial diversity of the TCR repertoire.

Another potential explanation for the observed genetic background effect in TCR repertoire diversity of sticklebacks is variation in genes guiding the junctional repertoire diversity. Mutations in the recombination-activating genes (RAG) 1 and 2, the proteins initiating VDJ recombination (Schatz and Swanson, 2011), can result in loss of TCR repertoire diversity and cause severe clinical presentations (Lee et al., 2016; Notarangelo et al., 2016). Variation at the complex genetic regulatory network controlling RAG1/2 could also indirectly impact the TCR repertoire (Y. Ji et al., 2010). Furthermore, it is likely that inter-individual variation in the number of non-templated nucleotides inserted at the junctions by terminal deoxyribonucleotidyl transferase (TdT) also plays a role (Sethna et al., 2020). Tdt is a crucial determinant of TCR repertoire diversity (Cabaniols et al., 2001), and its targeted disruption leads to a sharp drop in diversity (Gilfillan et al., 1993). Importantly, there are several naturally occurring Tdt SNPs among human populations, the varying efficiency of which can impact repertoire diversity (Troshchynsky et al., 2015; Russell et al., 2022). Genetic variants of the enzyme Artemis, the gene thought responsible for junctional nucleotide deletion, have also been identified (Russell et al., 2022).

Lastly, we cannot rule out the possibility that epigenetic, rather than genetic, effects, account for parts of the observed family effects on TCR repertoire diversity described in this thesis. There is accruing evidence of the role of epigenetic factors in the regulation of VDJ recombination, by means of altering the chromatin accessibility at the TRB loci (Jaeger et al., 2013; Ji et al., 2019; Aburajab et al., 2023). Considering the fact that each immature lymphocyte in the stickleback has double the available TRB loci from which a functional TCR can be molded, we expect a strong role for epigenetic mechanisms in allelic exclusion (Bossen et al., 2012). In any case, the significant correlation of intra-individual TCRβ1 and TCRβ2 repertoire diversity in the stickleback further attests to genetic and/ or epigenetic factors exerting a 'central' control over the diversity of the TCR repertoire.

Another interesting result that warrants further investigation is the evidently stronger effect of family background on the TCRβ2 rather than the TCRβ1 repertoire, as observed both in younger and older sticklebacks. Could this difference indicate divergence in specialization in terms of TCR functions between the two translocons? The stickleback provides an ideal system to further investigate the possibility of differential genetic background effects on the two separate TRB loci. Future experiments employing juvenile sticklebacks (*i.e.*, younger than

2 months old), whose TCR repertoire is mostly naïve, coupled with targeted T cell sequencing is likely to reveal an even more pronounced role for genetic background than detected in this thesis, and would therefore make for an even better study system to further elucidate the sources of this inter-individual variation in diversity.

TRB germline variation as overlooked factor in past studies

As discussed above, inter-individual variation at the TRB locus is a potential source of TCR repertoire variability; nevertheless, there is a sizable gap in our knowledge regarding the true extent of germline gene polymorphisms in the population. The development of the TCR gene nomenclature by the ImMunoGeneTics (IMGT) group marked an organized effort to create a comprehensive database of TRB germline genes (Lefranc et al., 1999). However, efforts to identify new allelic variants have long been stagnant – as opposed to the better characterized BCR locus (Omer et al., 2022). Short-read NGS efforts to map the TRB locus face a plethora of technical challenges due to the repetitive and complex nature of the genomic region (Watson et al., 2017). Alternative de novo approaches, through germline gene inference of expressed TCR repertoires, have been rare, as most TCR repertoire sequencing studies are heavily concentrated on the CDR3 (Glanville et al., 2017).

The issue of limited data on existing TRB variation certainly extends to the stickleback. Annotation of its two TRB loci was based on in silico analysis of the Ensembl reference genome of a stickleback from a distant Alaskan population (Jones et al., 2012), combined with screening and sequencing of local stickleback-derived libraries (Wölfing, 2012). In this thesis, utilization of a recently published genome from a local stickleback (Thorburn et al., 2023) for mapping of TCRβ sequences led to the detection of new and seemingly functional V and J genes (**Chapter 2**). Dedicated genomic efforts are needed to clarify whether these genes are ubiquitous in the local stickleback population, or whether they represent CNV between individuals at the population level. Considerable CNV of germline gene segments has been documented between chicken strains (Dunon et al., 1994), while instances of TRBV gene deletions has also recently been reported in human TRB loci (Corcoran et al., 2023). Even more extended deletions at the TRB, resulting in reduced TCR repertoire diversity and thus ability of to recognize antigens, have been described in wild mice (Woodland et al., 1990; Nanda et al., 1991). Even if there is no comparable structural variation within the local stickleback population, we anticipate widespread small polymorphisms, such as SNPs, leading to extensive allelic variation of TRB genes. The original effort at annotating the stickleback TRB loci, despite being constrained by its non-high-throughput capacity, still resulted in the identification of SNPs and small insertions/deletions in the local stickleback population in comparison to the Alaskan genome.

Absence of information regarding germline gene variation in the stickleback restricted our analysis to gene-level identification of V, D and J segments for each TCRβ sequence. The high sequence similarity between certain V gene pairs (Wölfing, 2012) meant that a given TCR's composition could not always be unambiguously assigned to a specific gene. Given the commonplace variation at TCR and BCR loci in other species (Corcoran and Karlsson Hedestam, 2024), it is almost certain that future research into genomic variation of the stickleback TRB will reveal allelic variation at germline genes, which could be associated with observed biases in gene usage patterns (Elhanati et al., 2018). We postulate that the family-based VJ usage differences could be the result of siblings sharing the same germline TRB haplotypes, *i.e.*, sets of specific alleles at each expressed TRB gene carried by multiple individuals, as such distinct TRB haplotypes are associated with different gene usage patterns in humans (Omer et al., 2022). In humans, polymorphic alleles vary considerably in frequency between populations, and individuals have been found to be heterozygous in a large proportion of TRBV genes (Cor-

coran et al., 2023). In sticklebacks, the compactness of the two distant TRB loci (Wölfing, 2012), could theoretically allow for up to four different TRB haplotypes contributing to the diversity of the expressed TCR repertoire within an individual. Increased heterozygosity supported by a diverse pool of TRB haplotypes could be a way for sticklebacks to compensate for limited numbers of V, D and J germlines genes compared to other teleosts possessing a single locus (Covacu et al., 2016; Giorgetti et al., 2021).

Why we need to account for it – path to personalized medicine and evolutionary understanding

In addition to the practical hurdles in characterizing germline TRB variation, systematic research into germline variability has been hindered by the notion that the "shuffling" of V, D, and J genes during somatic recombination probably generates much the same kind of repertoire, regardless of the "starting deck", *i.e.*, the germline genes available to an individual (Omer et al., 2022). This is best reflected by the thoroughness with which MHC genes –the disease "messenger" rather that the effector - have been investigated as disease susceptibility genes (Dendrou et al., 2018); the TCR is tightly associated with the MHC in recognizing antigenic peptides, and yet the germline genes providing the building blocks for its formation are rarely given much attention. Recently, a series of works highlighting the direct impact of germline gene variation on the human TCR repertoire has challenged this idea (Corcoran and Karlsson Hedestam, 2024). For example, certain non-functional allelic variants can cause loss of expression of corresponding genes (Subrahmanyan et al., 2001; Corcoran et al., 2023). Such non-functional gene variants can potentially lead to loss of unconventional T cell subsets expressing invariant TCRs encompassing the particular gene (Corcoran et al., 2023). Furthermore, allelic variation in specific V gene variants has been linked to susceptibility to certain bacterial enterotoxins (Rödström et al., 2015). Another V gene polymorphism has been associated with protection against severe adverse effects of immune checkpoint inhibitor treatment among cancer patients (Stephen et al., 2023). Genetic predisposition to a life threatening condition of cutaneous hypersensitivity - previously only weakly linked to HLA-, has been associated with a specific combination of V and J alleles (Pan et al., 2019).

From the above, it seems that the scarcity of known TCR gene - disease associations is less likely due to the irrelevance of individual TCR genetic variation, but rather a reflection of our lack of knowledge thereof. The advent of novel sequencing tools is bound to rapidly enhance the characterization of both allelic variation and CNV within the TCR (Rodriguez et al., 2022). Detailed cataloguing will enable in-depth understanding of the functional role of this variation, including differential responses to pathogenic epitopes, peptides, molecules, or cellular ligands within populations, due to the presence or absence of specific alleles. In the future, targeted therapeutic approaches based on high-throughput genotyping of both TCR and MHC genes individualized could be designed to maximize the patient's response to medical interventions, such as vaccination (Corcoran and Karlsson Hedestam, 2024). Besides potential clinical applications in humans, the documentation of allelic and structural variation in the TCR gene loci will also be crucial for better understanding the evolutionary forces shaping the highly dynamic TCR regions (Su and Nei, 2001). One would expect that the preservation of a diverse TRB allele pool over evolutionary time would require strong selective forces, such as balancing selection (Gragert, 2022), as is the case with the MHC (Trowsdale, 2011). So far, however, evidence points toward neutral evolutionary processes, rather than selection favoring a more diversified TCR repertoire (Subrahmanyan et al., 2001). The question regarding the evolutionary history behind what is bound to be extensive TRB polymorphism remains.

Beyond (epi)genetics: the TCR repertoire is shaped by infection and aging

Besides the aforementioned (epi)genetically determined biases shaping the naïve TCR repertoire, accumulation of pathogenic exposures over the course of a lifetime further molds an individual's pool of T cells, as these undergo antigen-driven clonal expansions and deletions in the periphery (Woodsworth et al., 2013). Infection typically leads to decreased overall repertoire diversity as naïve T cells, which contribute the most to TCR diversity (Britanova et al., 2016), proportionally concede ground to few antigen-triggered, abundant T cell clones (Hou et al., 2016). This is the case following exposure to various pathogens, *e.g.*, HIV, human herpes virus, and *Mycobacterium tuberculosis* (Baum et al., 2012; Cyktor et al., 2013; Zhu et al., 2013). Moreover, immune activation usually perturbs the architecture of TCR similarity networks, which in naïve individuals is dominated by clusters formed by highly abundant public TCRs (Madi et al., 2017). Within this thesis, a picture of TCR repertoire activation was described in **Chapter 3**, where lake-exposed and heavily parasitized sticklebacks exhibited both decreased overall TCR repertoire diversity and associated clonal expansion, as well as less connected TCR similarity networks. Conversely, in **Chapter 2**, I showed that *Schistocephalus solidus* infection does not significantly alter the TCR repertoire diversity, nor does it perturb the similarity network architecture of infected individuals.

Helminths as master immunomodulators

Infection with most pathogens, such as viruses and bacteria, is a race against time for the host: its immune system has a limited window to produce an effective and targeted response against the invading organism. Failure to do so in a timely manner can prove detrimental, since uncontrolled pathogen replication often bears a progressively greater toll on host fitness. Infection with helminths, on the other hand, differs from that with pathogenic microorganisms: firstly, most parasitic worms don't reproduce inside the host (Begon, 2009), and, secondly, unlike viruses and bacteria, a common strategy of helminths is the establishment of stable, chronic infections (van Riet et al., 2007). Helminth infections are remarkably common amongst humans, which affect approximately a billion people worldwide (Montresor et al., 2020). At the same time, compared to other infectious diseases, infection-induced mortality is relatively rare (Lustigman et al., 2012). These statistics speak to the exceptional ability of helminths to both avoid elimination and minimize severe pathology in the host; this is achieved via immunomodulation, a phenomenon that is conserved across a wide range of vertebrate hosts (Harnett, 2014).

Immunomodulation can be exerted via a number of different mechanisms, as reviewed by Maizels and Yazdanbakhsh (2003): some helminths can inhibit the production of "alarmin" cytokines by epithelial cells, normally induced by signals of cell stress or death, which act as initiators of innate immune responses. Furthermore, helminths can interfere with the host's ability to properly recognize and react to helminth-derived molecules, as mediated by germline-encoded innate immune receptors. In addition to suppression of targeted responses mediated by effector T and B cells, helminths can also induce regulatory T cells, which further contribute to immunosuppression. T cell hypo-responsiveness is especially pronounced during the chronic phase of infection with some helminths, which is marked by T cell "anergy", a state where T cells have lost their capacity to respond to challenge with parasite-derived antigens through clonal expansion (McSorley and Maizels, 2012). Regarding *S. solidus*, there is evidence to suggest that it can manipulate innate immune responses of the stickleback through evasion

and suppression mechanisms (Scharsack et al., 2004; Hammerschmidt and Kurtz, 2005; Barber and Scharsack, 2010); it is also very likely that it dampens the host's adaptive immune responses via down-regulation of MHC-II gene expression, as well as TCRβ1 expression (Piecyk et al., 2019a,b, 2021). The state of parasite-induced immune suppression is believed to be maintained by means of effector molecules (Hewitson et al., 2009). Indeed, accumulating evidence points towards *S. solidus* immune manipulation via excretory/secretory products (Scharsack et al., 2004, 2013; Franke et al., 2014; Berger et al., 2021; Mazanec et al., 2023). Some or all of the above mechanisms can be part of the explanation as to why infection of sticklebacks with *S. solidus* did not seem to elicit detectable T cell-based immune responses **Chapter 2**.

Some results presented in **Chapter 2** indicate that, as opposed to the slower-growing strain, the fast-growing Norwegian strain of *S. solidus* shows signs of suppression of the stickleback's private T cell responses. We hypothesize that *S. solidus* from the Norwegian population might be driven by evolutionary forces towards higher immune suppression phenotypes in response to its local stickleback population being able to mount effective immune responses that can significantly limit its growth (Piecyk et al., 2019a). The Norwegian *S. solidus* strain and its sympatric sticklebacks have likely co-evolved under evolutionary arms race dynamics selecting for increased resistance and virulence (Kalbe et al., 2016). The context of their local biotic and abiotic environments can be instrumental; high *S. solidus* prevalence - and thus high infection rates (Weber et al., 2017) -in the Norwegian lake of origin strongly suggest that local sticklebacks are under high selection pressure to evolve immunocompetence, unlike sticklebacks used in this thesis (Piecyk et al., 2019b). Furthermore, increased virulence in the form of fast growth would also be expected under competition among *S. solidus* parasites in the case of multiple infections (Schulenburg et al., 2009), which is indeed a phenomenon commonly observed in the Norwegian system. Interestingly, the slow-growth *S. solidus* strain used in this thesis also originates from a high prevalence population (Piecyk et al., 2019b), which is indicative of other factors influencing the parasite's growth strategy. One such factor could be the diversity of the local parasitic fauna; as opposed to the natal German lake of sticklebacks used in this thesis, parasite diversity in the Norwegian lake is low (Piecyk et al., 2019b). Abundance of other parasitic species could preclude sticklebacks from evolving costly *S. solidus*-specific immune responses, as they could benefit more from generalized, non-directed defenses against a wider scope of pathogens (Piecyk et al., 2019b). Lastly, the presence of predators besides the stickleback-eating birds that serve as final hosts of *S. solidus*, such as bigger fish, could influence parasite virulence (Jakobsen and Johnsen, 1987). Ecosystems characterized by low non-host predator abundance, where transmission rate to piscivorous birds is high, would favor the evolution of virulent *S. solidus* strains (Bérénos et al., 2009; Kalbe et al., 2016). Future field studies of the habitats of the two *S. solidus* strains used in this thesis would illuminate whether the two populations indeed differ in their prevalence of species predating on sticklebacks, and the overall prevalence of parasitic fauna.

Aging as the accumulation of a lifetime of immune responses

The TCR repertoire represents a dynamic entity: it continuously changes in response to the internal and external environment of the individual (DeWitt et al., 2018); as such, its study through sampling and NGS sequencing represents a snapshot in time, each TCR "caught" at a specific abundance, with those represented at particularly high frequencies possibly participating in ongoing immune responses. This thesis has been centered around the response of the TCR repertoire to external stimuli in the form of parasitic challenges. By expanding our investigation to include snapshots of repertoires of sticklebacks at two different life stages, we were

able to report -for the first time- noteworthy findings pertaining to the response of the TCR repertoire of a teleost to the physiological aging process. Generally speaking, aging induces similar changes to the TCR repertoire of humans as does infection, namely reduced diversity, increased clonal expansion, and altered repertoire architecture (Qi et al., 2014; Britanova et al., 2014, 2016; Yoshida et al., 2017; Egorov et al., 2018; Minervina et al., 2021). Infection-induced changes are short-lived, with the repertoire recovering after resolution of the infection (Madi et al., 2017), while age-induced changes are fairly gradual and permanent (Britanova et al., 2014). This makes intuitive sense, as we can think of aging as the accumulation of pathogenic exposures over the course of an individual's life. Indeed, in **Chapter 3** we observed that the TCR repertoire of sticklebacks of advanced age, near or past their expected lifespan in the wild, displayed all the familiar signs of immune activation also observed in lake-exposed individuals.

There is a striking resemblance between the aging-related effects on the TCR repertoire of the three-spined stickleback and those described in humans: beyond the above-mentioned changes shared with immune activation, aging has been shown to slightly reduce the total number of T cells in the periphery (Yan et al., 2010; Britanova et al., 2016), reduce public TCR abundance (Pogorelyy et al., 2017), and lead to disproportionate decline in $CD8^+$ T cells (Sun et al., 2022). As such, the immune system of an elderly human is characterized by weakened adaptive responses, especially against common viral infections such as influenza; it thus resembles that of a newborn (Simon et al., 2015). In humans, two independent but complementary mechanisms support the maintenance of a steady population of T cells: homeostatic proliferation of naïve T cells in the periphery (Takada and Jameson, 2009), and generation of new T cells in the thymus (Murray et al., 2003). While both have been implicated in age-related changes in the TCR repertoire, there is evidence that homeostatic turnover of naïve T cells is crucial in adults (Egorov et al., 2018), as thymic output sharply declines with age (Goronzy and Weyand, 2005; Naylor et al., 2005; Palmer, 2013). Interestingly, the opposite is true in mice, where thymic output is key for sustaining peripheral naive T cells (den Braber et al., 2012). Are both of these mechanisms present in the stickleback? Future studies will elucidate fundamental questions regarding the adaptive immune system and aging, and perhaps uncover further parallels between higher and lower vertebrates. It has been hypothesized that the evolution of an individual's immune system, whose pathogen-fighting aptitude peaks in early adulthood, reflects the critical role of young adults in the procreation and survival of the species (Simon et al., 2015). An interesting research topic would be to examine whether T cell senescence in the stickleback sets in past the age of sexual maturity.

Public TCRs - an ancient link between innate and adaptive immunity?

An exciting finding within this thesis is that the TCR repertoire of the stickleback contains a non-negligible number of clonotypes that are highly shared across individuals. These clonotypes were independently observed in younger (9-month-old) and older (nearly 2-year-old) sticklebacks in two separate experiments (**Chapters 2** and **3**). The detection of common public TCRs between the two age groups from the two experiments confers further support for their persisting existence throughout an individual's lifetime.

Public TCRs are widespread across vertebrate taxa and likely represent innate-like T cells

Public clonotypes are not unique to the three-spined stickleback. They were first described in mammals, initially in the context of viral epitope-specific responses among MHC-matched individuals (Venturi et al., 2006, 2008b), and their detection was subsequently extended to total TCR repertoires of healthy individuals (Robins et al., 2010; Venturi et al., 2011; Li et al., 2012; Shugay et al., 2013). The existence of public TCRs initially puzzled researchers, given the practically infinite sequences that can be generated through VDJ recombination (Davis and Bjorkman, 1988). It soon became clear, however, that the recombination machinery is heavily biased, and therefore bound to repeatedly produce certain sequences via convergent recombination (Venturi et al., 2008a; Quigley et al., 2010; Li et al., 2012; Elhanati et al., 2018). Part of these public clonotypes are now known to represent unconventional, *i.e.*, MHC-unrestricted, T cells, such as invariant natural killer T cells (NKT cells) and mucosal associated invariant T cells (MAIT cells) (Godfrey et al., 2015). These "innate-like T cells", present in multiple mammalian species (Greenaway et al., 2013), utilize invariant TCR alpha and beta chains (Van Rhijn and Moody, 2015), which have been computationally confirmed to be produced efficiently through convergent recombination, ensuring their presence across individuals (Greenaway et al., 2013). In this thesis, the distinct properties of a fraction of public clonotypes of the stickleback, namely their high generation probability, fewer N insertions, high nucleotide convergence, restricted VJ usage, and CDR3 sequence clustering in similarity networks (**Chapter 2**), perfectly fit the profile of invariant T cells. The presence of a non-classical MHC-I gene in the stickleback genome further attests to these public clonotypes stemming from non-classical MHC-I restricted T cells (Grimholt et al., 2015). I further speculated that the remaining, non VJ restricted public clonotypes represent conventional, *i.e.*, MHC-restricted, T cells whose abundance is explained by homeostatic proliferation in the periphery (Britanova et al., 2016), as described in humans (Pogorelyy et al., 2017).

Significantly, public TCRs have also been detected in non-mammalian species, including amphibians and teleosts (Edholm et al., 2013; Madi et al., 2014; Covacu et al., 2016; Madi et al., 2017; Giorgetti et al., 2021). Although direct evidence that these public TCRs represent unconventional or "innate-like" T cells is still missing, genomic data point towards this direction. Indeed, non-classical MHC-I genes are present in all taxa of jawed vertebrates (Edholm et al., 2014), from cartilaginous fish (Almeida et al., 2020) and other ectotherms (Flajnik, 2018), to birds (Robert and Edholm, 2014), to mammals (Legoux et al., 2017). Genomic studies have also identified an ancient lineage of non-classical MHC-I genes likely to interact NKT-like cells in teleosts (Dijkstra et al., 2018). The case of *Xenopus laevis* is noteworthy, as it is the only non-mammalian species in which unconventional T cells have been functionally characterized and linked to semi-invariant TCRs (Edholm et al., 2013; Robert and Edholm, 2014; Edholm et al., 2015). Thus, public TCRs appear to be a fundamental feature of the TCR repertoire in all vertebrate species. Their universal presence from teleosts to mammals, groups that diverged more than 350 million years ago, strongly suggests an evolutionarily conserved role for unconventional T cells in vertebrates (Edholm et al., 2014).

Public TCRs are overrepresented in lower vertebrates and potentially compensate for limited TCR repertoires diversity

Despite their omnipresence, public TCRs are not equally abundant in the TCR repertoire of all vertebrate species. Widely shared TCRs are a dominant portion of the individual TCR

repertoire of zebrafish (Covacu et al., 2016), minifish (Giorgetti et al., 2021), the tadpole (Edholm et al., 2013), and as shown here, the stickleback. For example, non-private clonotypes made up 40% of the total TCRβ1 amino acid sequences detected in zebrafish (Covacu et al., 2016), and 20% of the total TCRβ1 clonotypes in sticklebacks (**Chapter 2**). In mice, humans, and other mammals, public clonotypes constitute a much smaller fraction of the total TCR repertoire (Miles et al., 2011). Higher TCR sharing levels between mice than between humans have been attributed to the overall lower number of added nucleotides in murine TCRs (Zvyagin et al., 2014). Fewer nucleotide insertions during VDJ recombination is closely associated with the generation of public clonotypes, as it increases the chances of convergent recombination (Venturi et al., 2011). Fittingly, the average number of nucleotide insertions in the TCR repertoire of naive sticklebacks is lower than both humans and mice (Wölfing, 2012). As already discussed in **Chapter 2**, the stickleback TCR repertoire, similar to other teleosts, is restricted in its diversity compared to those of mammals, even when accounting for body size differences.

This observation has led to the theory that species with limited TCR repertoire diversity compensate by employing a set of public cross-reactive TCRs whose specificities is wide enough to target epitopes shared by multiple pathogens (Castro et al., 2017; Covacu et al., 2016). In accordance with this hypothesis, public TCRs actively respond to immunization with both self- and non-self peptides in zebrafish (Covacu et al., 2016), are critical against viral infection in tadpoles (Edholm et al., 2013; Robert and Edholm, 2014; Edholm et al., 2015), expand after parasitic infection in channel catfish (Findly et al., 2017), while their persistence in the repertoires of lake-exposed and heavily parasitized sticklebacks also points towards active participation in immune responses (**Chapter 3**). It therefore appears that invariant T cell populations are critical in lower vertebrates, whose TCR repertoire is optimized to ensure protection with a relatively small number of lymphocytes (Giorgetti et al., 2021). A standing hypothesis is that public TCRs provide a mechanism for the rapid generation of protective immunity against key pathogens (Covacu et al., 2016; Robert and Edholm, 2014). In further support of this idea, rearranged TCR genes encoding parasitic infection-responding public TCRs have been detected in the genome of channel catfish (Findly et al., 2018). Equivalently, B cells expressing germline-encoded public BCRs protect against infection by common pathogens (Castro et al., 2021). This apparent prominent protective role of public TCRs in lower vertebrates suggest that they are less reliant on T cell responses conferred by highly diversified antigen receptors. Indeed, genetically engineered zebrafish incapable of VDJ recombination still mount a specific protection after bacterial re-exposure (Hohn and Petrie-Hanson, 2012).

On the other hand, higher vertebrates capable of forming diverse TCR repertoires appear to be less reliant on public TCRs in their adult stage; their protective role is believed to be important during early ontogeny (Robert and Edholm, 2014). Indeed, human and murine newborns are characterized by sub-optimal classical MHC-I function, but highly expressed nonclassical MHC-I genes (Robert and Edholm, 2014). Thus, survival during early life stages, prior to the development of a mature diversified TCR repertoire, might be dependent upon a pre-existing pool of invariant T cells, able rapidly initiate innate-like immune responses. There is a plethora of studies in humans that supports this hypothesis; Pogorelyy et al. (2017) suggest that a considerable fraction of public clonotypes in humans are created prenatally, a developmental stage that coincides with reduced TdT activity (Benedict et al., 2000). These public TCRs, which likely represent innate-like T cells (Simon et al., 2015), are highly functional and fast-responding to antigens (Davenport et al., 2020). They form a "core" TCR repertoire, which slowly gives way to antigen-experienced, diverse T cells throughout an individual's life (Britanova et al., 2016; Pogorelyy et al., 2017). In adults, persisting populations of iNKT and

MAIT cells have specialized functions at the early stage of infection, especially in regulating subsequent adaptive immune responses (Wu and Van Kaer, 2011; Toubal et al., 2019).

The emerging picture of a binary TCR repertoire architecture

Thus, the picture that emerges from studies on human public TCRs and age-related changes in T cell composition is that of two co-existing TCR repertoire "layers"; a first layer, consisting of prenatal public TCRs of low sequence diversity that persist with age, and a second layer, formed by private, TdT-dependent TCRs that are formed during infancy and increase the TCR repertoire diversity. This theory, supported by an observational study using TCR repertoire data (Trofimov et al., 2022), agrees remarkably well with previous theoretical predictions (Vrisekoop et al., 2014). Interestingly, the two layers seem to have distinct functions: prenatal public TCRs are likely polyreactive to self and non-self, and shown to be associated with known pathogens or autoantigens, while TDT-dependent TCRs can target peptides in a more specific manner (Trofimov et al., 2022). It is hard not to notice that the characteristics of the first TCR repertoire "layer" closely resemble those we find in the prominent public TCRs of sticklebacks (**Chapter 2**): short sequence length, restricted VJ gene usage, and high generation probabilities. Accordingly, the longer, private TdT-dependent sequences observed in the second "layer" of human TCR repertoire are reminiscent of the clonotypes in the stickleback's restricted private TCR compartment. It has been proposed that a layered immune system could be the product of evolutionary processes, during which following descendants obtain predominance during development, giving rise to cell populations responsible for progressively more complex immune activities (Scapigliati et al., 2018), in accordance with Haeckel's famous proposition that "ontogeny recapitulating phylogeny". I argue that this could be the case here; the "core" or first "layer" of the TCR repertoire of mammals is generated early during ontogenesis, at a time of low TdT activity, and is reminiscent of the protective, public TCRs from invariant T cells prominent throughout the life of lower vertebrates (Edholm et al., 2013). If the evident prominence of public TCRs in the repertoires of lower vertebrates does indeed represent an early step during the evolution of adaptive immunity, as already hypothesized (Flajnik and Kasahara, 2010; Boehm et al., 2012), then how did this shift from a public TCR-heavy to a more diversified repertoire take place? And under what evolutionary context was the expression of public TCRs advantageous to ancestral vertebrate species?

Proposed evolutionary mechanism and implications

A groundbreaking study by Giorgetti et al. (2023) offers an elegant answer. The authors show that microhomologies at the ends of V and J genes in the vertebrate genome can guide RAG-mediated recombination, by determining which germline genes will be used for the composition of a rearranged TCRα chain. By looking into various species across the vertebrate clade, they find that the presence and extent of these microhomologies correlates inversely with the level of junctional diversity of the ensuing rearranged TCRα sequences. Therefore, they were able to find an association between the high degree of microhomology-directed recombination with the high degree of public TCRα sequences in the repertoire of the zebrafish, and three other teleost species. Strikingly, due to the dominant role of TRA locus microhomologies, amongst all vertebrates, teleosts exhibit the highest level of predictability in the TCRα generated sequences, and thus have a species-specific public TCRα repertoire. Conversely, the TCRα repertoire of mammalian species exhibits lower predictability levels (and thus, lower levels of public TCRs), due to loss of microhomology signatures. They further

note that in zebrafish, lower levels of TdT, activity of which interferes with microhomology-guided recombination, coincides with higher degrees of public TCRs. Low TdT activity is also a characteristic of fetal life stages of higher vertebrates, during which TCR repertoire generation is more reliant on microhomology-based recombination.

The above findings perfectly align with the prominent public TCR compartment in the three-spined stickleback described within this thesis, as well as the general consensus of lower TRC repertoire diversity found in lower vertebrates, as already discussed. They also provide a much-needed explanation as to the way vertebrates might have gradually moved towards higher TCR repertoire diversity over the course of evolution. Regarding the evolutionary context that favored the presence of public TCRs in ancestral vertebrates, Giorgetti et al. (2021) hypothesize that by constraining TCR sequence diversity, microhomology-guided VDJ recombination might have been crucial in minimizing the potential of autoimmunity during the early stages of vertebrate evolution. As such, this mechanism could have aided in incorporating the intrinsically risky process of somatic diversification into the immune system of jawed vertebrates, an innovation that gave rise to the intricate adaptive immune system we see today.

Whether the above findings can be expanded to include the TRB locus and corresponding TCRβ repertoire of vertebrates is a topic of upcoming research. What is certain is that the adaptive immune system has been successful enough that is represented in the vast majority of extant vertebrate species. The gamble of adaptive immunity has clearly paid off, despite the risks involved. Future studies need to address exciting questions such as: what were the evolutionary forces that drove higher vertebrates towards elevated levels TCR repertoire diversity? Could it be that a more diversified repertoire offers an advantage in the arms race with pathogens by protecting against the emergence of antigen-loss variants (Liston et al., 2021)? And, even more broadly, what primordial need did the emergence of adaptive immunity serve in the vertebrate ancestor in the first place? Is it possible that the need to differentiate between symbiotic and pathogenic microorganisms was a contributor (Lee and Mazmanian, 2010)? Research in the field of evolutionary immunology is bound to provide us with valuable insights in the near future.

Concluding remarks

The past decade has been a time of fast-paced progress in the field of adaptive immune receptor (AIRR) repertoire research, fueled by NGS technologies and single cell transcriptomics. This pace is surely going to increase, with the advent of machine learning approaches and large language models trained on AIRR data for TCR-epitope interaction prediction and disease diagnosis (Liu et al., 2019; Peng et al., 2023); with promising applications in medicine, *e.g.*, in targeted cancer immunotherapy (Joshi et al., 2022; Yang et al., 2023). A case is to be made that basic research will continue to offer - as it so frequently does - novel insights, and guide towards future research directions. In the words of Max Planck, "insight must precede application".

In this thesis, I have used NGS sequencing to conduct an in-depth study of key parts of the immune system in an eco-evolutionary model species, the three-spined stickleback. I have thus provided the first experimental evidence that the diversity of the T cell receptor repertoire - arguably its most decisive feature- is dependent upon the individual's genetic variation. While some of the sources of variance among individuals have already been addressed, *e.g.*, enzymes involved in the somatic recombination process, others have been overlooked. Specifically, future research should address the largely unknown TRB germline variation at the population level. Encouragingly, this field has been granted more attention recently (Luo et al., 2019; Omer et al., 2022). Novel targeted long-read NGS methodologies (Rodriguez et al., 2022) will assist in fully exploring genetic diversity within the TCR gene region, which in turn is likely to reveal novel disease associations, and pave the road to personalized medicine.

Also shown in this thesis is the naturally high MHC-I expressed diversity of the stickleback. As such, the stickleback constitutes a prime model to further disentangle the effects of genetic background and individual MHC diversity, an effort in which the highly reliable NGS genotyping protocol developed here will be of value. This has been an open question for years in the community, with efforts to address it having thus far yielded inconsistent results, hindered by the use of datasets where the factor of individual genetic variation is masked. Experimental work with carefully designed datasets in genetically variable model species will decisively aid this effort.

Lastly, I have provided indirect evidence of public, MHC-unrestricted, germline-encoded TCRs in the stickleback, as already described in mammals, amphibians, and other teleosts. I hypothesize that these TCRs represent a population of "innate-like" T cells, an evolutionarily conserved T cell population, and whose seemingly ubiquitous presence of across wildly divergent vertebrate species attests to their functional importance. The results of this thesis strongly support the notion that a large part of the TCR repertoire of the stickleback forms a "core" of public TCR specificities, in accordance with a proposed hypothesis (Giorgetti et al., 2023) which, if proven right, will mark a huge leap in our understanding of the evolution of the adaptive immune system in jawed vertebrates. The phenotypic make-up of the T cell subsets forming this core in the stickleback, as well as the extent to which this core is involved in the protection against pathogens, is an exciting topic for future research.

Bibliography

R. Aburajab, M. Pospiech, and H. Alachkar. Profiling the epigenetic landscape of the antigen receptor repertoire: the missing epi-immunogenomics data. *Nature methods*, 20(4):477, Apr. 2023. doi: 10.1038/s41592-022-01723-9. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11058354/.

M. Ahmed, K. G. Lanzer, E. J. Yager, P. S. Adams, L. L. Johnson, and M. A. Blackman. Clonal Expansions and Loss of Receptor Diversity in the Naive CD8 T Cell Repertoire of Aged Mice1. *The Journal of Immunology*, 182(2):784–792, Jan. 2009. ISSN 0022-1767. doi: 10.4049/jimmunol.182.2.784. URL https://doi.org/10.4049/jimmunol.182.2.784.

S. Alizon, J. C. de Roode, and Y. Michalakis. Multiple infections and the evolution of virulence. *Ecology Letters*, 16(4):556–567, 2013. ISSN 1461-0248. doi: 10.1111/ele.12076. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ele.12076. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ele.12076.

T. Almeida, P. J. Esteves, M. F. Flajnik, Y. Ohta, and A. Veríssimo. An ancient, MHC-linked, nonclassical class I lineage in cartilaginous fish. *Journal of immunology (Baltimore, Md. : 1950)*, 204(4):892, Jan. 2020. doi: 10.4049/jimmunol.1901025. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7002201/.

F. W. Alt, E. M. Oltz, F. Young, J. Gorman, G. Taccioli, and J. Chen. VDJ recombination. *Immunology Today*, 13(8):306–314, Jan. 1992. ISSN 0167-5699. doi: 10.1016/0167-5699(92)90043-7. URL https://www.cell.com/immunology/abstract/0167-5699(92)90043-7. Publisher: Elsevier.

M. H. Andersen, D. Schrama, P. thor Straten, and J. C. Becker. Cytotoxic T Cells. *Journal of Investigative Dermatology*, 126(1):32–41, Jan. 2006. ISSN 0022-202X. doi: 10.1038/sj.jid.5700001. URL https://www.sciencedirect.com/science/article/pii/S0022202X1532621X.

R. M. Anthony, L. I. Rutitzky, J. F. Urban, M. J. Stadecker, and W. C. Gause. Protective immune mechanisms in helminth infection. *Nature Reviews Immunology*, 7(12):975–987, Dec. 2007. ISSN 1474-1741. doi: 10.1038/nri2199. URL https://www.nature.com/articles/nri2199. Publisher: Nature Publishing Group.

V. Aravindhan and G. Anand. Cell Type-Specific Immunomodulation Induced by Helminthes: Effect on Metainflammation, Insulin Resistance and Type-2 Diabetes. *The American Journal of Tropical Medicine and Hygiene*, 97(6):1650–1661, Dec. 2017. ISSN 0002-9637. doi: 10.4269/ajtmh.17-0236. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5805044/.

B. Arezi and H. H. Hogrefe. Escherichia coli DNA polymerase III subunit increases Moloney murine leukemia virus reverse transcriptase fidelity and accuracy of RT–PCR

procedures. *Analytical Biochemistry*, 360(1):84–91, Jan. 2007. ISSN 00032697. doi: 10.1016/j.ab.2006.10.009. URL https://linkinghub.elsevier.com/retrieve/pii/S0003269706007287.

J. Argüello, A.-M. Little, E. Bohan, J. Goldman, S. Marsh, and J. Madrigal. High resolution HLA class I typing by reference strand mediated conformation analysis (RSCA). *Tissue Antigens*, 52(1):57–66, 1998. ISSN 1399-0039. doi: 10.1111/j.1399-0039.1998.tb03024.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1399-0039.1998.tb03024.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1399-0039.1998.tb03024.x.

C. Arme and R. W. Owen. Infections of the three-spined stickleback, Gasterosteus aculeatus L., with the plerocercoid larvae of Schistocephalus solidus (Müller, 1776), with special reference to pathological effects. *Parasitology*, 57(2):301–314, May 1967. ISSN 0031-1820. doi: 10.1017/s0031182000072103.

T. P. Arstila, A. Casrouge, V. Baron, J. Even, J. Kanellopoulos, and P. Kourilsky. A Direct Estimate of the Human T Cell Receptor Diversity. *Science*, 286(5441):958–961, Oct. 1999. doi: 10.1126/science.286.5441.958. URL https://www.science.org/doi/abs/10.1126/science.286.5441.958. Publisher: American Association for the Advancement of Science.

K. M. Ashby and K. A. Hogquist. A guide to thymic selection of T cells. *Nature Reviews Immunology*, pages 1–15, July 2023. ISSN 1474-1741. doi: 10.1038/s41577-023-00911-8. URL https://www.nature.com/articles/s41577-023-00911-8. Publisher: Nature Publishing Group.

W. Babik. Methods for MHC genotyping in non-model vertebrates. *Molecular Ecology Resources*, 10(2):237–251, 2010. ISSN 1755-0998. doi: 10.1111/j.1755-0998.2009.02788.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1755-0998.2009.02788.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1755-0998.2009.02788.x.

W. Babik, P. Taberlet, M. J. Ejsmond, and J. Radwan. New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. *Molecular Ecology Resources*, 9(3):713–719, 2009. ISSN 1755-0998. doi: 10.1111/j.1755-0998.2009.02622.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1755-0998.2009.02622.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1755-0998.2009.02622.x.

A. Ballesteros-Tato, T. D. Randall, F. E. Lund, R. Spolski, W. J. Leonard, and B. León. T Follicular Helper Cell Plasticity Shapes Pathogenic T Helper 2 Cell-Mediated Immunity to Inhaled House Dust Mite. *Immunity*, 44(2):259–273, Feb. 2016. ISSN 1097-4180. doi: 10.1016/j.immuni.2015.11.017.

P. H. M. Balm. Immune-endocrine interactions. In *Fish Stress and Health in Aquaculture*, pages 195–221. Cambridge University Press, Cambridge, 1997.

I. Barber and J. P. Scharsack. The three-spined stickleback-Schistocephalus solidus system: an experimental model for investigating host-parasite interactions in fish. *Parasitology*, 137(3):411–424, Mar. 2010. ISSN 1469-8161, 0031-1820. doi: 10.1017/S0031182009991466. URL https://www.cambridge.org/core/journals/parasitology/article/threespined-sticklebackschistocephalus-solidus-system-an-experimental-model-for-investigating-hostparasite-interactions-in-fish/E461379C0ECCEB22C84C7C47773CA9C5. Publisher: Cambridge University Press.

P. Barennes, V. Quiniou, M. Shugay, E. S. Egorov, A. N. Davydov, D. M. Chudakov, I. Uddin, M. Ismail, T. Oakes, B. Chain, A. Eugster, K. Kashofer, P. P. Rainer, S. Darko, A. Ransier, D. C. Douek, D. Klatzmann, and E. Mariotti-Ferrandiz. Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. *Nature Biotechnology*, 39(2):236–245, Feb. 2021. ISSN 1546-1696. doi: 10.1038/s41587-020-0656-3. URL https://www.nature.com/articles/s41587-020-0656-3. Publisher: Nature Publishing Group.

L. B. Barreiro and L. Quintana-Murci. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nature Reviews Genetics*, 11(1):17–30, Jan. 2010. ISSN 1471-0064. doi: 10.1038/nrg2698. URL https://www.nature.com/articles/nrg2698. Number: 1 Publisher: Nature Publishing Group.

D. Bates, M. Maechler, B. Bolker [aut, cre, S. Walker, R. H. B. Christensen, H. Singmann, B. Dai, F. Scheipl, G. Grothendieck, P. Green, J. Fox, A. Bauer, P. N. K. s. c. o. simulate.formula), E. Tanaka, and M. Jagan. lme4: Linear Mixed-Effects Models using 'Eigen' and S4, July 2024. URL https://cran.r-project.org/web/packages/lme4/index.html.

P. D. Baum, J. J. Young, D. Schmidt, Q. Zhang, R. Hoh, M. Busch, J. Martin, S. Deeks, and J. M. McCune. Blood T-cell receptor diversity decreases during the course of HIV infection, but the potential for a diverse repertoire persists. *Blood*, 119(15):3469–3477, Apr. 2012. ISSN 0006-4971. doi: 10.1182/blood-2011-11-395384. URL https://doi.org/10.1182/blood-2011-11-395384.

M. Begon. Ecological Epidemiology. In S. A. Levin, S. R. Carpenter, H. C. J. Godfray, A. P. Kinzig, M. Loreau, J. B. Losos, B. Walker, and D. S. Wilcove, editors, *The Princeton Guide to Ecology*, pages 220–226. Princeton University Press, July 2009. ISBN 978-1-4008-3302-3. doi: 10.1515/9781400833023.220. URL https://www.degruyter.com/document/doi/10.1515/9781400833023.220/pdf?licenseType=restricted.

E. b. M. A. Bell and S. A. Foster, editors. *The Evolutionary Biology of the Threespine Stickleback*. Oxford University Press, Oxford, New York, May 1994. ISBN 978-0-19-857728-7.

C. L. Benedict, S. Gilfillan, T. H. Thai, and J. F. Kearney. Terminal deoxynucleotidyl transferase and repertoire development. *Immunological Reviews*, 175:150–157, June 2000. ISSN 0105-2896.

J. Benichou, R. Ben-Hamo, Y. Louzoun, and S. Efroni. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology*, 135(3):183–191, 2012. ISSN 1365-2567. doi: 10.1111/j.1365-2567.2011.03527.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2567.2011.03527.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2567.2011.03527.x.

C. S. Berger, J. Laroche, H. Maaroufi, H. Martin, K.-M. Moon, C. R. Landry, L. J. Foster, and N. Aubin-Horth. The parasite Schistocephalus solidus secretes proteins with putative host manipulation functions. *Parasites & Vectors*, 14(1):436, Aug. 2021. ISSN 1756-3305. doi: 10.1186/s13071-021-04933-w. URL https://doi.org/10.1186/s13071-021-04933-w.

J. S. Bezbradica, T. Hill, A. K. Stanic, L. Van Kaer, and S. Joyce. Commitment toward the natural T (iNKT) cell lineage occurs at the CD4+8+ stage of thymic ontogeny. *Proceedings of the National Academy of Sciences*, 102(14):5114–5119, Apr. 2005. doi: 10.1073/pnas.0408449102. URL https://www.pnas.org/doi/abs/10.1073/pnas.0408449102. Publisher: Proceedings of the National Academy of Sciences.

A. Biedrzycka, A. Sebastian, M. Migalska, H. Westerdahl, and J. Radwan. Testing genotyping strategies for ultra-deep sequencing of a co-amplifying gene family: MHC class I in a passerine bird. *Molecular Ecology Resources*, 17(4):642–655, 2017. ISSN 1755-0998. doi: 10.1111/1755-0998.12612. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12612. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.12612.

M. E. Birnbaum, J. L. Mendoza, D. K. Sethi, S. Dong, J. Glanville, J. Dobbins, E. Ozkan, M. M. Davis, K. W. Wucherpfennig, and K. C. Garcia. Deconstructing the peptide-MHC specificity of T cell recognition. *Cell*, 157(5):1073–1087, May 2014. ISSN 1097-4172. doi: 10.1016/j.cell.2014.03.047.

P. J. Bjorkman, M. A. Saper, B. Samraoui, W. S. Bennett, J. L. Strominger, and D. C. Wiley. The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. *Nature*, 329(6139):512–518, Oct. 1987. ISSN 1476-4687. doi: 10.1038/329512a0. URL https://www.nature.com/articles/329512a0. Number: 6139 Publisher: Nature Publishing Group.

H. Bjørgen and E. O. Koppang. Anatomy of teleost fish immune structures and organs. *Immunogenetics*, 73(1):53–63, Feb. 2021. ISSN 1432-1211. doi: 10.1007/s00251-020-01196-0. URL https://doi.org/10.1007/s00251-020-01196-0.

T. Boehm, N. Iwanami, and I. Hess. Evolution of the immune system in the lower vertebrates. *Annual Review of Genomics and Human Genetics*, 13:127–149, 2012. ISSN 1545-293X. doi: 10.1146/annurev-genom-090711-163747.

K. Bohmann, V. Elbrecht, C. Carøe, I. Bista, F. Leese, M. Bunce, D. W. Yu, M. Seymour, A. J. Dumbrell, and S. Creer. Strategies for sample labelling and library preparation in DNA metabarcoding studies. *Molecular Ecology Resources*, 22(4):1231–1246, 2022. ISSN 1755-0998. doi: 10.1111/1755-0998.13512. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13512. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13512.

T. Bolger and P. L. Connolly. The selection of suitable indices for the measurement and analysis of fish condition. *Journal of Fish Biology*, 34(2):171–182, 1989. ISSN 1095-8649. doi: 10.1111/j.1095-8649.1989.tb03300.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1095-8649.1989.tb03300.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1095-8649.1989.tb03300.x.

D. I. Bolnick, L. K. Snowberg, P. E. Hirsch, C. L. Lauber, R. Knight, J. G. Caporaso, and R. Svanbäck. Individuals' diet diversity influences gut microbial diversity in two freshwater fish (threespine stickleback and Eurasian perch). *Ecology Letters*, 17(8):979–987, 2014. ISSN 1461-0248. doi: 10.1111/ele.12301. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ele.12301. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ele.12301.

D. A. Bolotin, S. Poslavsky, I. Mitrophanov, M. Shugay, I. Z. Mamedov, E. V. Putintseva, and D. M. Chudakov. MiXCR: software for comprehensive adaptive immunity profiling. *Nature Methods*, 12(5):380–381, May 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3364. URL https://www.nature.com/articles/nmeth.3364. Number: 5 Publisher: Nature Publishing Group.

J. A. M. Borghans, A. J. Noest, and R. J. De Boer. Thymic selection does not limit the individual MHC diversity. *European Journal of Immunology*, 33(12): 3353–3358, 2003. ISSN 1521-4141. doi: 10.1002/eji.200324365. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/eji.200324365. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/eji.200324365.

C. Bossen, R. Mansson, and C. Murre. Chromatin topology and the regulation of antigen receptor assembly. *Annual Review of Immunology*, 30:337–356, 2012. ISSN 1545-3278. doi: 10.1146/annurev-immunol-020711-075003.

P. Boudinot, S. Boubekeur, and A. Benmansour. Rhabdovirus Infection Induces Public and Private T Cell Responses in Teleost Fish1. *The Journal of Immunology*, 167(11):6202–6209, Dec. 2001. ISSN 0022-1767. doi: 10.4049/jimmunol.167.11.6202. URL https://doi.org/10.4049/jimmunol.167.11.6202.

P. Boudinot, D. Bernard, S. Boubekeur, M.-I. Thoulouze, M. Bremont, and A. Benmansour. The glycoprotein of a fish rhabdovirus profiles the virus-specific T-cell repertoire in rainbow trout. *Journal of General Virology*, 85(10):3099–3108, 2004. ISSN 1465-2099. doi: 10.1099/vir.0.80135-0. URL https://www.microbiologyresearch.org/content/journal/jgv/10.1099/vir.0.80135-0. Publisher: Microbiology Society,.

P. Bousso, A. Casrouge, J. D. Altman, M. Haury, J. Kanellopoulos, J.-P. Abastado, and P. Kourilsky. Individual Variations in the Murine T Cell Response to a Specific Peptide Reflect Variability in Naive Repertoires. *Immunity*, 9(2):169–178, Aug. 1998. ISSN 1074-7613. doi: 10.1016/S1074-7613(00)80599-3. URL https://www.cell.com/immunity/abstract/S1074-7613(00)80599-3. Publisher: Elsevier.

P. Bradley and P. G. Thomas. Using T Cell Receptor Repertoires to Understand the Principles of Adaptive Immune Recognition. *Annual Review of Immunology*, 37(1):547–570, 2019. doi: 10.1146/annurev-immunol-042718-041757. URL https://doi.org/10.1146/annurev-immunol-042718-041757. _eprint: https://doi.org/10.1146/annurev-immunol-042718-041757.

B. L. Brady, N. C. Steinel, and C. H. Bassing. Antigen Receptor Allelic Exclusion: An Update and Reappraisal. *Journal of immunology (Baltimore, Md. : 1950)*, 185(7):3801, Oct. 2010. doi: 10.4049/jimmunol.1001158. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3008371/.

O. V. Britanova, E. V. Putintseva, M. Shugay, E. M. Merzlyak, M. A. Turchaninova, D. B. Staroverov, D. A. Bolotin, S. Lukyanov, E. A. Bogdanova, I. Z. Mamedov, Y. B. Lebedev, and D. M. Chudakov. Age-Related Decrease in TCR Repertoire Diversity Measured with Deep and Normalized Sequence Profiling. *The Journal of Immunology*, 192(6):2689–2698, Mar. 2014. ISSN 0022-1767, 1550-6606. doi: 10.4049/jimmunol.1302064. URL https://journals.aai.org/jimmunol/article/192/6/2689/1552/Age-Related-Decrease-in-TCR-Repertoire-Diversity.

O. V. Britanova, M. Shugay, E. M. Merzlyak, D. B. Staroverov, E. V. Putintseva, M. A. Turchaninova, I. Z. Mamedov, M. V. Pogorelyy, D. A. Bolotin, M. Izraelson, A. N. Davydov, E. S. Egorov, S. A. Kasatskaya, D. V. Rebrikov, S. Lukyanov, and D. M. Chudakov. Dynamics of Individual T Cell Repertoires: From Cord Blood to Centenarians. *The Journal of Immunology*, 196(12):5005–5013, June 2016. ISSN 0022-1767, 1550-6606. doi: 10.

4049/jimmunol.1600005. URL https://journals.aai.org/jimmunol/article/196/12/5005/105159/Dynamics-of-Individual-T-Cell-Repertoires-From.

A. J. Brown, J. White, L. Shaw, J. Gross, A. Slabodkin, E. Kushner, V. Greiff, J. Matsuda, L. Gapin, J. Scott-Browne, J. Kappler, and P. Marrack. MHC heterozygosity limits T cell receptor variability in CD4 T cells. *Science Immunology*, 9(97):eado5295, July 2024. doi: 10.1126/sciimmunol.ado5295. URL https://www.science.org/doi/full/10.1126/sciimmunol.ado5295. Publisher: American Association for the Advancement of Science.

I. K. Brown, N. Dyjack, M. M. Miller, H. Krovi, C. Rios, R. Woolaver, L. Harmacek, T.-H. Tu, B. P. O'Connor, T. Danhorn, B. Vestal, L. Gapin, C. Pinilla, M. A. Seibold, J. Scott-Browne, R. G. Santos, and R. L. Reinhardt. Single cell analysis of host response to helminth infection reveals the clonal breadth, heterogeneity, and tissue-specific programming of the responding CD4+ T cell repertoire. *PLOS Pathogens*, 17(6):e1009602, June 2021. ISSN 1553-7374. doi: 10.1371/journal.ppat.1009602. URL https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1009602. Publisher: Public Library of Science.

J. H. Brown, T. S. Jardetzky, J. C. Gorga, L. J. Stern, R. G. Urban, J. L. Strominger, and D. C. Wiley. Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature*, 364(6432):33–39, July 1993. ISSN 0028-0836. doi: 10.1038/364033a0.

T. Bråten. Host specificity in Schistocephalus solidus. *Parasitology*, 56(4):657–664, Nov. 1966. ISSN 1469-8161, 0031-1820. doi: 10.1017/S0031182000071687. URL https://www.cambridge.org/core/journals/parasitology/article/abs/host-specificity-in-schistocephalus-solidus/2F92F84495402BBDB3B717BF5E8AA4CA.

R. Burri, M. Promerová, J. Goebel, and L. Fumagalli. PCR-based isolation of multigene families: lessons from the avian MHC class IIB. *Molecular Ecology Resources*, 14(4):778–788, 2014. ISSN 1755-0998. doi: 10.1111/1755-0998.12234. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12234. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.12234.

C. D. Bustamante, A. Fledel-Alon, S. Williamson, R. Nielsen, M. Todd Hubisz, S. Glanowski, D. M. Tanenbaum, T. J. White, J. J. Sninsky, R. D. Hernandez, D. Civello, M. D. Adams, M. Cargill, and A. G. Clark. Natural selection on protein-coding genes in the human genome. *Nature*, 437(7062):1153–1157, Oct. 2005. ISSN 1476-4687. doi: 10.1038/nature04240. URL https://www.nature.com/articles/nature04240. Number: 7062 Publisher: Nature Publishing Group.

C. Bérénos, P. Schmid-hempel, and K. Mathias Wegner. Evolution of host resistance and trade-offs between virulence and transmission potential in an obligately killing parasite. *Journal of Evolutionary Biology*, 22(10):2049–2056, Oct. 2009. ISSN 1010-061X. doi: 10.1111/j.1420-9101.2009.01821.x. URL https://doi.org/10.1111/j.1420-9101.2009.01821.x.

J.-P. Cabaniols, N. Fazilleau, A. Casrouge, P. Kourilsky, and J. M. Kanellopoulos. Most / T Cell Receptor Diversity Is Due to Terminal Deoxynucleotidyl Transferase. *Journal of Experimental Medicine*, 194(9):1385–1390, Nov. 2001. ISSN 0022-1007. doi: 10.1084/jem.194.9.1385. URL https://doi.org/10.1084/jem.194.9.1385.

M. F. Callan, L. Tan, N. Annels, G. S. Ogg, J. D. Wilson, C. A. O'Callaghan, N. Steven, A. J. McMichael, and A. B. Rickinson. Direct visualization of antigen-specific CD8+ T cells during the primary immune response to Epstein-Barr virus In vivo. *The Journal of Experimental Medicine*, 187(9):1395–1402, May 1998. ISSN 0022-1007. doi: 10.1084/jem.187.9.1395.

Y. Carton, A. J. Nappi, and M. Poirie. Genetics of anti-parasite resistance in invertebrates. *Developmental & Comparative Immunology*, 29(1):9–32, Jan. 2005. ISSN 0145-305X. doi: 10.1016/j.dci.2004.05.004. URL https://www.sciencedirect.com/science/article/pii/S0145305X04000801.

C. Carøe and K. Bohmann. Tagsteady: A metabarcoding library preparation protocol to avoid false assignment of sequences to samples. *Molecular Ecology Resources*, 20(6):1620–1631, 2020. ISSN 1755-0998. doi: 10.1111/1755-0998.13227. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13227. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13227.

A. Casrouge, E. Beaudoing, S. Dalle, C. Pannetier, J. Kanellopoulos, and P. Kourilsky. Size estimate of the alpha beta TCR repertoire of naive mouse splenocytes. *Journal of Immunology (Baltimore, Md.: 1950)*, 164(11):5782–5787, June 2000. ISSN 0022-1767. doi: 10.4049/jimmunol.164.11.5782.

R. Castro, D. Bernard, M. P. Lefranc, A. Six, A. Benmansour, and P. Boudinot. T cell diversity and TcR repertoires in teleost fish. *Fish & Shellfish Immunology*, 31(5):644–654, Nov. 2011. ISSN 1050-4648. doi: 10.1016/j.fsi.2010.08.016. URL https://www.sciencedirect.com/science/article/pii/S105046481000269X.

R. Castro, S. Navelsaker, A. Krasnov, L. Du Pasquier, and P. Boudinot. Describing the diversity of Ag specific receptors in vertebrates: Contribution of repertoire deep sequencing. *Developmental & Comparative Immunology*, 75:28–37, Oct. 2017. ISSN 0145-305X. doi: 10.1016/j.dci.2017.02.018. URL https://www.sciencedirect.com/science/article/pii/S0145305X17301015.

R. Castro, S. Navelsaker, B. Collet, L. Jouneau, P. Bochet, E. Quillet, Evensen, J. O. Sunyer, S. Fillatreau, P. Bruhns, T. Rose, F. Huetz, and P. Boudinot. Cutting Edge: Neutralizing Public Antibody Responses Are an Ancient Form of Defense Conserved in Fish and Mammals. *The Journal of Immunology*, 207(2):371–375, July 2021. ISSN 0022-1767. doi: 10.4049/jimmunol.2100149. URL https://doi.org/10.4049/jimmunol.2100149.

D. M. Catron, A. A. Itano, K. A. Pape, D. L. Mueller, and M. K. Jenkins. Visualizing the First 50 Hr of the Primary Immune Response to a Soluble Antigen. *Immunity*, 21(3):341–347, Sept. 2004. ISSN 1074-7613. doi: 10.1016/j.immuni.2004.08.007. URL https://www.cell.com/immunity/abstract/S1074-7613(04)00238-9. Publisher: Elsevier.

A. Chao. Nonparametric Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics*, 11(4):265–270, 1984. ISSN 0303-6898. URL https://www.jstor.org/stable/4615964. Publisher: [Board of the Foundation of the Scandinavian Journal of Statistics, Wiley].

S. Chellappa, F. A. Huntingford, R. H. C. Strang, and R. Y. Thomson. Annual variation in energy reserves in male three-spined stickleback, Gasterosteus aculeatm L. (Pisces, Gasterosteidae). *Journal of Fish Biology*, 35(2):275–286,

1989. ISSN 1095-8649. doi: 10.1111/j.1095-8649.1989.tb02976.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1095-8649.1989.tb02976.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1095-8649.1989.tb02976.x.

S. Chellappa, F. A. Huntingford, R. H. C. Strang, and R. Y. Thomson. Condition factor and hepatosomatic index as estimates of energy status in male three-spined stickleback. *Journal of Fish Biology*, 47(5):775–787, 1995. ISSN 1095-8649. doi: 10.1111/j.1095-8649.1995.tb06002.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1095-8649.1995.tb06002.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1095-8649.1995.tb06002.x.

F. Chen, Z. Liu, W. Wu, C. Rozo, S. Bowdridge, A. Millman, N. Van Rooijen, J. F. Urban, T. A. Wynn, and W. C. Gause. An essential role for TH2-type responses in limiting acute tissue damage during experimental helminth infection. *Nature Medicine*, 18(2):260–266, Feb. 2012. ISSN 1546-170X. doi: 10.1038/nm.2628. URL https://www.nature.com/articles/nm.2628. Publisher: Nature Publishing Group.

A. M. Collins, Y. Wang, K. M. Roskin, C. P. Marquis, and K. J. L. Jackson. The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1676):20140236, Sept. 2015. doi: 10.1098/rstb.2014.0236. URL https://royalsocietypublishing.org/doi/full/10.1098/rstb.2014.0236. Publisher: Royal Society.

M. Corcoran, M. Chernyshev, M. Mandolesi, S. Narang, M. Kaduk, K. Ye, C. Sundling, A. Färnert, T. Kreslavsky, C. Bernhardsson, M. Larena, M. Jakobsson, and G. B. Karlsson Hedestam. Archaic humans have contributed to large-scale variation in modern human T cell receptor genes. *Immunity*, 56(3):635–652.e6, Mar. 2023. ISSN 1074-7613. doi: 10.1016/j.immuni.2023.01.026. URL https://www.sciencedirect.com/science/article/pii/S1074761323000389.

M. M. Corcoran and G. B. Karlsson Hedestam. Adaptive immune receptor germline gene variation. *Current Opinion in Immunology*, 87:102429, Apr. 2024. ISSN 0952-7915. doi: 10.1016/j.coi.2024.102429. URL https://www.sciencedirect.com/science/article/pii/S0952791524000190.

A. I. Costa, D. Koning, K. Ladell, J. E. McLaren, B. P. X. Grady, I. M. M. Schellens, P. van Ham, M. Nijhuis, J. A. M. Borghans, C. Keşmir, D. A. Price, and D. van Baarle. Complex T-Cell Receptor Repertoire Dynamics Underlie the CD8+ T-Cell Response to HIV-1. *Journal of Virology*, 89(1):110–119, Dec. 2014. doi: 10.1128/jvi.01765-14. URL https://journals.asm.org/doi/full/10.1128/jvi.01765-14. Publisher: American Society for Microbiology.

R. Covacu, H. Philip, M. Jaronen, J. Almeida, J. E. Kenison, S. Darko, C.-C. Chao, G. Yaari, Y. Louzoun, L. Carmel, D. C. Douek, S. Efroni, and F. J. Quintana. System-wide Analysis of the T Cell Response. *Cell Reports*, 14(11):2733–2744, Mar. 2016. ISSN 2211-1247. doi: 10.1016/j.celrep.2016.02.056. URL https://www.sciencedirect.com/science/article/pii/S2211124716301772.

F. E. Cox. Concomitant infections, parasites and immune responses. *Parasitology*, 122 Suppl: S23–38, 2001. ISSN 0031-1820. doi: 10.1017/s003118200001698x.

P. Cruaud, J.-Y. Rasplus, L. J. Rodriguez, and A. Cruaud. High-throughput sequencing of multiple amplicons for barcoding and integrative taxonomy. *Scientific Reports*, 7(1):41948, Feb. 2017. ISSN 2045-2322. doi: 10.1038/srep41948. URL https://www.nature.com/articles/srep41948. Number: 1 Publisher: Nature Publishing Group.

J. C. Cyktor, B. Carruthers, G. L. Beamer, and J. Turner. Clonal Expansions of CD8+ T Cells with IL-10 Secreting Capacity Occur during Chronic Mycobacterium tuberculosis Infection. *PLOS ONE*, 8(3):e58612, Mar. 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0058612. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0058612. Publisher: Public Library of Science.

M. P. Davenport, N. L. Smith, and B. D. Rudd. Building a T cell compartment: how immune cell development shapes function. *Nature Reviews Immunology*, 20(8):499–506, Aug. 2020. ISSN 1474-1741. doi: 10.1038/s41577-020-0332-3. URL https://www.nature.com/articles/s41577-020-0332-3. Publisher: Nature Publishing Group.

M. M. Davis and P. J. Bjorkman. T-cell antigen receptor genes and T-cell recognition. *Nature*, 334(6181):395–402, Aug. 1988. ISSN 1476-4687. doi: 10.1038/334395a0. URL https://www.nature.com/articles/334395a0. Publisher: Nature Publishing Group.

M. M. Davis and S. D. Boyd. Recent progress in the analysis of T cell and B cell receptor repertoires. *Current Opinion in Immunology*, 59:109–114, Aug. 2019. ISSN 0952-7915. doi: 10.1016/j.coi.2019.05.012. URL https://www.sciencedirect.com/science/article/pii/S0952791519300482.

I. den Braber, T. Mugwagwa, N. Vrisekoop, L. Westera, R. Mögling, A. B. de Boer, N. Willems, E. H. R. Schrijver, G. Spierenburg, K. Gaiser, E. Mul, S. A. Otto, A. F. C. Ruiter, M. T. Ackermans, F. Miedema, J. A. M. Borghans, R. J. de Boer, and K. Tesselaar. Maintenance of peripheral naive T cells is sustained by thymus output in mice but not humans. *Immunity*, 36(2):288–297, Feb. 2012. ISSN 1097-4180. doi: 10.1016/j.immuni.2012.02.006.

C. A. Dendrou, J. Petersen, J. Rossjohn, and L. Fugger. HLA variation and disease. *Nature Reviews. Immunology*, 18(5):325–339, May 2018. ISSN 1474-1741. doi: 10.1038/nri.2017.143.

W. S. DeWitt, III, A. Smith, G. Schoch, J. A. Hansen, F. A. Matsen, IV, and P. Bradley. Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *eLife*, 7:e38358, Aug. 2018. ISSN 2050-084X. doi: 10.7554/eLife.38358. URL https://doi.org/10.7554/eLife.38358. Publisher: eLife Sciences Publications, Ltd.

J. M. Dijkstra, T. Yamaguchi, and U. Grimholt. Conservation of sequence motifs suggests that the nonclassical MHC class I lineages CD1/PROCR and UT were established before the emergence of tetrapod species. *Immunogenetics*, 70(7):459–476, July 2018. ISSN 1432-1211. doi: 10.1007/s00251-017-1050-2. URL https://doi.org/10.1007/s00251-017-1050-2.

A. Dobson, K. D. Lafferty, A. M. Kuris, R. F. Hechinger, and W. Jetz. Homage to Linnaeus: How many parasites? How many hosts? *Proceedings of the National Academy of Sciences*, 105(supplement_1):11482–11489, Aug. 2008. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0803232105. URL https://pnas.org/doi/full/10.1073/pnas.0803232105.

A. Drews and H. Westerdahl. Not all birds have a single dominantly expressed MHC-I gene: Transcription suggests that siskins have many highly expressed MHC-I genes. *Scientific*

reports, 9(1):19506, Dec. 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-55800-9. URL https://europepmc.org/articles/PMC6925233.

A. Drews, M. Strandh, L. Råberg, and H. Westerdahl. Expression and phylogenetic analyses reveal paralogous lineages of putatively classical and non-classical MHC-I genes in three sparrow species (Passer). *BMC Evolutionary Biology*, 17(1):152, June 2017. ISSN 1471-2148. doi: 10.1186/s12862-017-0970-7. URL https://doi.org/10.1186/s12862-017-0970-7.

L. Du Pasquier. Antibody diversity in lower vertebrates—why is it so restricted? *Nature*, 296 (5855):311–313, Mar. 1982. ISSN 1476-4687. doi: 10.1038/296311a0. URL https://www.nature.com/articles/296311a0. Publisher: Nature Publishing Group.

D. Dunon, J. Schwager, J. Dangy, M. Cooper, and B. Imhof. T cell migration during development: homing is not related to TCR V beta 1 repertoire selection. *The EMBO Journal*, 13(4): 808–815, Feb. 1994. ISSN 0261-4189. doi: 10.1002/j.1460-2075.1994.tb06323.x. URL https://www.embopress.org/doi/abs/10.1002/j.1460-2075.1994.tb06323.x. Num Pages: 815 Publisher: John Wiley & Sons, Ltd.

T. Dupic, Q. Marcou, A. M. Walczak, and T. Mora. Genesis of the T-cell receptor. *PLoS computational biology*, 15(3):e1006874, Mar. 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006874.

E.-S. Edholm, L.-M. Albertorio Saez, A. L. Gill, S. R. Gill, L. Grayfer, N. Haynes, J. R. Myers, and J. Robert. Nonclassical MHC class I-dependent invariant T cells are evolutionarily conserved and prominent from early development in amphibians. *Proceedings of the National Academy of Sciences*, 110(35):14342–14347, Aug. 2013. doi: 10.1073/pnas.1309840110. URL https://www.pnas.org/doi/full/10.1073/pnas.1309840110. Publisher: Proceedings of the National Academy of Sciences.

E.-S. Edholm, L. Grayfer, and J. Robert. Evolution of nonclassical MHC-dependent invariant T cells. *Cellular and Molecular Life Sciences: CMLS*, 71(24):4763, Aug. 2014. doi: 10.1007/s00018-014-1701-5. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4234686/.

E.-S. Edholm, L. Grayfer, F. De Jesús Andino, and J. Robert. Nonclassical MHC-Restricted Invariant V6 T Cells Are Critical for Efficient Early Innate Antiviral Immunity in the Amphibian Xenopus laevis. *The Journal of Immunology*, 195(2):576–586, July 2015. ISSN 0022-1767. doi: 10.4049/jimmunol.1500458. URL https://doi.org/10.4049/jimmunol.1500458.

A. Efstratiou, E. M. S. Galon, G. Wang, K. Umeda, D. Kondoh, M. A. Terkawi, A. Kume, M. Liu, A. E. Ringo, H. Guo, Y. Gao, S.-H. Lee, J. Li, P. F. A. Moumouni, Y. Nishikawa, H. Suzuki, I. Igarashi, and X. Xuan. Babesia microti Confers Macrophage-Based Cross-Protective Immunity Against Murine Malaria. *Frontiers in Cellular and Infection Microbiology*, 10, Apr. 2020. ISSN 2235-2988. doi: 10.3389/fcimb.2020.00193. URL https://www.frontiersin.org/journals/cellular-and-infection-microbiology/articles/10.3389/fcimb.2020.00193/full. Publisher: Frontiers.

A. Efstratiou, A. Gaigher, S. Künzel, A. Teles, and T. L. Lenz. Template-specific optimization of NGS genotyping pipelines reveals allele-specific variation in MHC gene expression. *Molecular Ecology Resources*, 24(4):e13935, 2024. ISSN 1755-0998. doi: 10.1111/1755-0998.13935. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13935. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13935.

E. S. Egorov, E. M. Merzlyak, A. A. Shelenkov, O. V. Britanova, G. V. Sharonov, D. B. Staroverov, D. A. Bolotin, A. N. Davydov, E. Barsova, Y. B. Lebedev, M. Shugay, and D. M. Chudakov. Quantitative Profiling of Immune Repertoires for Minor Lymphocyte Counts Using Unique Molecular Identifiers. *The Journal of Immunology*, 194(12):6155–6163, June 2015. ISSN 0022-1767. doi: 10.4049/jimmunol.1500215. URL https://doi.org/10.4049/jimmunol.1500215.

E. S. Egorov, S. A. Kasatskaya, V. N. Zubov, M. Izraelson, T. O. Nakonechnaya, D. B. Staroverov, A. Angius, F. Cucca, I. Z. Mamedov, E. Rosati, A. Franke, M. Shugay, M. V. Pogorelyy, D. M. Chudakov, and O. V. Britanova. The Changing Landscape of Naive T Cell Receptor Repertoire With Human Aging. *Frontiers in Immunology*, 9, 2018. ISSN 1664-3224. URL https://www.frontiersin.org/articles/10.3389/fimmu.2018.01618.

R. Eisenhofer, J. J. Minich, C. Marotz, A. Cooper, R. Knight, and L. S. Weyrich. Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends in Microbiology*, 27(2):105–117, Feb. 2019. ISSN 0966-842X. doi: 10.1016/j.tim.2018.11.003. URL https://www.sciencedirect.com/science/article/pii/S0966842X18302531.

C. Eizaguirre, T. L. Lenz, R. D. Sommerfeld, C. Harrod, M. Kalbe, and M. Milinski. Parasite diversity, patterns of MHC II variation and olfactory based mate choice in diverging three-spined stickleback ecotypes. *Evolutionary Ecology*, 25(3):605–622, May 2011. ISSN 1573-8477. doi: 10.1007/s10682-010-9424-z. URL https://doi.org/10.1007/s10682-010-9424-z.

C. Eizaguirre, T. L. Lenz, M. Kalbe, and M. Milinski. Divergent selection on locally adapted major histocompatibility complex immune genes experimentally proven in the field. *Ecology Letters*, 15(7):723–731, July 2012a. ISSN 1461-0248. doi: 10.1111/j.1461-0248.2012.01791.x.

C. Eizaguirre, T. L. Lenz, M. Kalbe, and M. Milinski. Rapid and adaptive evolution of MHC genes under parasite selection in experimental vertebrate populations. *Nature Communications*, 3(1):621, Jan. 2012b. ISSN 2041-1723. doi: 10.1038/ncomms1632. URL https://www.nature.com/articles/ncomms1632. Number: 1 Publisher: Nature Publishing Group.

Y. Elhanati, Z. Sethna, C. G. Callan, T. Mora, and A. M. Walczak. Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunological Reviews*, 284(1):167–179, July 2018. ISSN 0105-2896. doi: 10.1111/imr.12665. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6033145/.

R. O. Emerson, W. S. DeWitt, M. Vignali, J. Gravley, J. K. Hu, E. J. Osborne, C. Desmarais, M. Klinger, C. S. Carlson, J. A. Hansen, M. Rieder, and H. S. Robins. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nature Genetics*, 49(5):659–665, May 2017. ISSN 1546-1718. doi: 10.1038/ng.3822. URL https://www.nature.com/articles/ng.3822. Publisher: Nature Publishing Group.

P. Esling, F. Lejzerowicz, and J. Pawlowski. Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Research*, 43(5):2513–2524, Mar. 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv107. URL https://doi.org/10.1093/nar/gkv107.

D. Esterházy, M. C. C. Canesso, L. Mesin, P. A. Muller, T. B. R. de Castro, A. Lockhart, M. ElJalby, A. M. C. Faria, and D. Mucida. Compartmentalized gut lymph node drainage dictates adaptive immune responses. *Nature*, 569(7754):126–130, May 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1125-3. URL https://www.nature.com/articles/s41586-019-1125-3. Publisher: Nature Publishing Group.

Z. Even, A. P. Meli, A. Tyagi, A. Vidyarthi, N. Briggs, D. A. d. Kouchkovsky, Y. Kong, Y. Wang, D. A. Waizman, T. A. Rice, B. D. Kumar, X. Wang, N. W. Palm, J. Craft, M. K. Basu, S. Ghosh, and C. V. Rothlin. The amalgam of naive CD4+ T cell transcriptional states is reconfigured by helminth infection to dampen the amplitude of the immune response. *Immunity*, 57(8):1893–1907.e6, Aug. 2024. ISSN 1074-7613. doi: 10.1016/j.immuni.2024.07.006. URL https://www.cell.com/immunity/abstract/S1074-7613(24)00355-8. Publisher: Elsevier.

D. W. Fadrosh, B. Ma, P. Gajer, N. Sengamalay, S. Ott, R. M. Brotman, and J. Ravel. An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome*, 2(1):6, Feb. 2014. ISSN 2049-2618. doi: 10.1186/2049-2618-2-6. URL https://doi.org/10.1186/2049-2618-2-6.

B. Fang, J. Merilä, F. Ribeiro, C. M. Alexandre, and P. Momigliano. Worldwide phylogeny of three-spined sticklebacks. *Molecular Phylogenetics and Evolution*, 127:613–625, Oct. 2018. ISSN 1055-7903. doi: 10.1016/j.ympev.2018.06.008. URL https://www.sciencedirect.com/science/article/pii/S1055790317308527.

S. Ferrando-Martínez, E. Ruiz-Mateos, A. Hernández, E. Gutiérrez, M. d. M. Rodríguez-Méndez, A. Ordoñez, and M. Leal. Age-related deregulation of naive T cell homeostasis in elderly humans. *AGE*, 33(2):197–207, June 2011. ISSN 1574-4647. doi: 10.1007/s11357-010-9170-8. URL https://doi.org/10.1007/s11357-010-9170-8.

P. G. D. Feulner, F. J. J. Chain, M. Panchal, Y. Huang, C. Eizaguirre, M. Kalbe, T. L. Lenz, I. E. Samonte, M. Stoll, E. Bornberg-Bauer, T. B. H. Reusch, and M. Milinski. Genomics of Divergence along a Continuum of Parapatric Population Differentiation. *PLOS Genetics*, 11(2):e1004966, Feb. 2015. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004966. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004966. Publisher: Public Library of Science.

R. C. Findly, F. D. Niagro, and H. W. Dickerson. The expressed TCR CDR3 repertoire is dominated by conserved DNA sequences in channel catfish. *Developmental & Comparative Immunology*, 68:26–33, Mar. 2017. ISSN 0145-305X. doi: 10.1016/j.dci.2016.11.010. URL https://www.sciencedirect.com/science/article/pii/S0145305X16304037.

R. C. Findly, F. D. Niagro, R. P. Sweeney, A. C. Camus, and H. W. Dickerson. Rearranged T Cell Receptor Sequences in the Germline Genome of Channel Catfish Are Preferentially Expressed in Response to Infection. *Frontiers in Immunology*, 9, 2018. ISSN 1664-3224. URL https://www.frontiersin.org/articles/10.3389/fimmu.2018.02117.

G. J. FitzGerald and R. J. Wootton. Behavioural Ecology of Sticklebacks. In T. J. Pitcher, editor, *The Behaviour of Teleost Fishes*, pages 409–432. Springer US, Boston, MA, 1986. ISBN 978-1-4684-8261-4. doi: 10.1007/978-1-4684-8261-4_16. URL https://doi.org/10.1007/978-1-4684-8261-4_16.

M. F. Flajnik. A cold-blooded view of adaptive immunity. *Nature Reviews Immunology*, 18(7): 438–453, July 2018. ISSN 1474-1741. doi: 10.1038/s41577-018-0003-9. URL https://www.nature.com/articles/s41577-018-0003-9. Publisher: Nature Publishing Group.

M. F. Flajnik and M. Kasahara. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nature Reviews Genetics*, 11(1):47–59, Jan. 2010. ISSN 1471-0064. doi: 10.1038/nrg2703. URL https://www.nature.com/articles/nrg2703. Publisher: Nature Publishing Group.

Y. Z. Foo, S. Nakagawa, G. Rhodes, and L. W. Simmons. The effects of sex hormones on immune function: a meta-analysis. *Biological Reviews*, 92(1):551–571, 2017. ISSN 1469-185X. doi: 10.1111/brv.12243. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/brv.12243. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/brv.12243.

J. Fox, S. Weisberg, B. Price, D. Adler, D. Bates, G. Baud-Bovy, B. Bolker, S. Ellison, D. Firth, M. Friendly, G. Gorjanc, S. Graves, R. Heiberger, P. Krivitsky, R. Laboissiere, M. Maechler, G. Monette, D. Murdoch, H. Nilsson, D. Ogle, B. Ripley, T. Short, W. Venables, S. Walker, D. Winsemius, A. Zeileis, and R-Core. car: Companion to Applied Regression, Sept. 2024. URL https://cran.r-project.org/web/packages/car/.

F. Franke, A. K. Rahn, J. Dittmar, N. Erin, J. K. Rieger, D. Haase, I. E. Samonte-Padilla, J. Lange, P. J. Jakobsen, M. Hermida, C. Fernández, J. Kurtz, T. C. M. Bakker, T. B. H. Reusch, M. Kalbe, and J. P. Scharsack. In vitro leukocyte response of three-spined sticklebacks (Gasterosteus aculeatus) to helminth parasite antigens. *Fish & Shellfish Immunology*, 36(1):130–140, Jan. 2014. ISSN 1050-4648. doi: 10.1016/j.fsi.2013.10.019. URL https://www.sciencedirect.com/science/article/pii/S1050464813008164.

F. Franke, S. A. O. Armitage, M. A. M. Kutzer, J. Kurtz, and J. P. Scharsack. Environmental temperature variation influences fitness trade-offs and tolerance in a fish-tapeworm association. *Parasites & Vectors*, 10(1):252, June 2017. ISSN 1756-3305. doi: 10.1186/s13071-017-2192-7. URL https://doi.org/10.1186/s13071-017-2192-7.

M. Frischknecht. The breeding colouration of male three-spined sticklebacks (Gasterosteus aculeatus) as an indicator of energy investment in vigour. *Evolutionary Ecology*, 7(5):439–450, Sept. 1993. ISSN 1573-8477. doi: 10.1007/BF01237640. URL https://doi.org/10.1007/BF01237640.

M. Fumagalli and M. Sironi. Human genome variability, natural selection and infectious diseases. *Current Opinion in Immunology*, 30:9–16, Oct. 2014. ISSN 1879-0372. doi: 10.1016/j.coi.2014.05.001.

M. Fumagalli, M. Sironi, U. Pozzoli, A. Ferrer-Admetlla, L. Pattini, and R. Nielsen. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS genetics*, 7(11):e1002355, Nov. 2011. ISSN 1553-7404. doi: 10.1371/journal.pgen.1002355.

A. Gaigher, R. Burri, W. H. Gharib, P. Taberlet, A. Roulin, and L. Fumagalli. Family-assisted inference of the genetic architecture of major histocompatibility complex variation. *Molecular Ecology Resources*, 16(6):1353–1364, 2016. ISSN 1755-0998. doi: 10.1111/1755-0998.12537. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12537. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.12537.

A. Gaigher, A. Roulin, W. H. Gharib, P. Taberlet, R. Burri, and L. Fumagalli. Lack of evidence for selection favouring MHC haplotypes that combine high functional diversity. *Heredity*, 120(5):396–406, May 2018. ISSN 1365-2540. doi: 10.1038/s41437-017-0047-9. URL https://www.nature.com/articles/s41437-017-0047-9. Number: 5 Publisher: Nature Publishing Group.

A. Gaigher, A. Rota, F. Neves, A. Muñoz-Mérida, J. Blasco-Aróstegui, T. Almeida, and A. Veríssimo. Extensive MHC class II diversity across multiple loci in the small-spotted catshark (Scyliorhinus canicula). *Scientific Reports*, 13(1):3837, Mar. 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-30876-6. URL https://www.nature.com/articles/s41598-023-30876-6. Number: 1 Publisher: Nature Publishing Group.

M. Galan, E. Guivier, G. Caraux, N. Charbonnel, and J.-F. Cosson. A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. *BMC Genomics*, 11(1):296, May 2010. ISSN 1471-2164. doi: 10.1186/1471-2164-11-296. URL https://doi.org/10.1186/1471-2164-11-296.

K. Gao, L. Chen, Y. Zhang, Y. Zhao, Z. Wan, J. Wu, L. Lin, Y. Kuang, J. Lu, X. Zhang, L. Tian, X. Liu, and X. Qiu. Germline-Encoded TCR-MHC Contacts Promote TCR V Gene Bias in Umbilical Cord Blood T Cell Repertoire. *Frontiers in Immunology*, 10, Aug. 2019. ISSN 1664-3224. doi: 10.3389/fimmu.2019.02064. URL https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2019.02064/full. Publisher: Frontiers.

R. N. Germain. The Art of the Probable: System Control in the Adaptive Immune System. *Science*, 293(5528):240–245, July 2001. doi: 10.1126/science.1062946. URL https://www.science.org/doi/abs/10.1126/science.1062946. Publisher: American Association for the Advancement of Science.

S. Gilfillan, A. Dierich, M. Lemeur, C. Benoist, and D. Mathis. Mice Lacking TdT: Mature Animals with an Immature Lymphocyte Repertoire. *Science*, 261(5125):1175–1178, Aug. 1993. doi: 10.1126/science.8356452. URL https://www.science.org/doi/abs/10.1126/science.8356452. Publisher: American Association for the Advancement of Science.

M. A. F. Gillingham, B. K. Montero, K. Wihelm, K. Grudzus, S. Sommer, and P. S. C. Santos. A novel workflow to improve genotyping of multigene families in wildlife species: An experimental set-up with a known model system. *Molecular Ecology Resources*, 21(3):982–998, 2021. ISSN 1755-0998. doi: 10.1111/1755-0998.13290. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13290. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13290.

O. B. Giorgetti, P. Shingate, C. P. O'Meara, V. Ravi, N. E. Pillai, B.-H. Tay, A. Prasad, N. Iwanami, H. H. Tan, M. Schorpp, B. Venkatesh, and T. Boehm. Antigen receptor repertoires of one of the smallest known vertebrates. *Science Advances*, 7(1):eabd8180, Jan. 2021. doi: 10.1126/sciadv.abd8180. URL https://www.science.org/doi/full/10.1126/sciadv.abd8180. Publisher: American Association for the Advancement of Science.

O. B. Giorgetti, C. P. O'Meara, M. Schorpp, and T. Boehm. Origin and evolutionary malleability of T cell receptor diversity. *Nature*, 619(7968):193–200, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06218-x. URL https://www.nature.com/articles/s41586-023-06218-x. Publisher: Nature Publishing Group.

J. Glanville, H. Huang, A. Nau, O. Hatton, L. E. Wagar, F. Rubelt, X. Ji, A. Han, S. M. Krams, C. Pettus, N. Haas, C. S. L. Arlehamn, A. Sette, S. D. Boyd, T. J. Scriba, O. M. Martinez, and M. M. Davis. Identifying specificity groups in the T cell receptor repertoire. *Nature*, 547 (7661):94–98, July 2017. ISSN 1476-4687. doi: 10.1038/nature22976. URL https://www.nature.com/articles/nature22976. Publisher: Nature Publishing Group.

D. I. Godfrey, A. P. Uldrich, J. McCluskey, J. Rossjohn, and D. B. Moody. The burgeoning family of unconventional T cells. *Nature Immunology*, 16(11):1114–1123, Nov. 2015. ISSN 1529-2916. doi: 10.1038/ni.3298. URL https://www.nature.com/articles/ni.3298. Publisher: Nature Publishing Group.

A. W. Goldrath and M. J. Bevan. Selecting and maintaining a diverse T-cell repertoire. *Nature*, 402(6759):255–262, Nov. 1999. ISSN 0028-0836. doi: 10.1038/46218.

J. J. Goronzy and C. M. Weyand. T cell development and receptor diversity during aging. *Current Opinion in Immunology*, 17(5):468–475, Oct. 2005. ISSN 0952-7915. doi: 10.1016/j.coi.2005.07.020.

L. Gragert. A template for multitudes: Germline immune polymorphism of the T cell receptor loci. *Cell Genomics*, 2(12):100231, Dec. 2022. ISSN 2666979X. doi: 10.1016/j.xgen.2022.100231. URL https://linkinghub.elsevier.com/retrieve/pii/S2666979X22001896.

A. L. Graham. Ecological rules governing helminth-microparasite coinfection. *Proceedings of the National Academy of Sciences of the United States of America*, 105(2):566–570, Jan. 2008. ISSN 1091-6490. doi: 10.1073/pnas.0707221105.

I. GraphPad Software. GraphPad Prism, 2012. URL http://www.graphpad.com/scientific-software/prism/.

H. Y. Greenaway, B. Ng, D. A. Price, D. C. Douek, M. P. Davenport, and V. Venturi. NKT and MAIT invariant TCR sequences can be produced efficiently by VJ gene recombination. *Immunobiology*, 218(2):213–224, Feb. 2013. ISSN 0171-2985. doi: 10.1016/j.imbio.2012.04.003. URL https://www.sciencedirect.com/science/article/pii/S0171298512000861.

V. Greiff, E. Miho, U. Menzel, and S. T. Reddy. Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires. *Trends in Immunology*, 36(11):738–749, Nov. 2015. ISSN 14714906. doi: 10.1016/j.it.2015.09.006. URL https://linkinghub.elsevier.com/retrieve/pii/S1471490615002239.

R. Griffiths, K. L. Orr, A. Adam, and I. Barber. DNA sex identification in the three-spined stickleback. *Journal of Fish Biology*, 57(5):1331–1334, 2000. ISSN 1095-8649. doi: 10.1111/j.1095-8649.2000.tb00490.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1095-8649.2000.tb00490.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1095-8649.2000.tb00490.x.

U. Grimholt, F. Drabløs, S. M. Jørgensen, B. Høyheim, and R. J. M. Stet. The major histocompatibility class I locus in Atlantic salmon (Salmo salar L.): polymorphism, linkage analysis and protein modelling. *Immunogenetics*, 54(8):570–581, Nov. 2002. ISSN 0093-7711. doi: 10.1007/s00251-002-0499-8.

U. Grimholt, K. Tsukamoto, T. Azuma, J. Leong, B. F. Koop, and J. M. Dijkstra. A comprehensive analysis of teleost MHC class I sequences. *BMC Evolutionary Biology*, 15(1):32, Mar. 2015. ISSN 1471-2148. doi: 10.1186/s12862-015-0309-1. URL https://doi.org/10.1186/s12862-015-0309-1.

Z. Gu, R. Eils, and M. Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849, Sept. 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw313. URL https://doi.org/10.1093/bioinformatics/btw313.

D. Haase, J. K. Rieger, A. Witten, M. Stoll, E. Bornberg-Bauer, M. Kalbe, A. Schmidt-Drewello, J. P. Scharsack, and T. B. H. Reusch. Comparative transcriptomics of stickleback immune gene responses upon infection by two helminth parasites, Diplostomum pseudospathaceum and Schistocephalus solidus. *Zoology*, 119(4):307–313, Aug. 2016. ISSN 0944-2006. doi: 10.1016/j.zool.2016.05.005. URL https://www.sciencedirect.com/science/article/pii/S0944200616300307.

K. Hammerschmidt and J. Kurtz. Surface carbohydrate composition of a tapeworm in its consecutive intermediate hosts: Individual variation and fitness consequences. *International Journal for Parasitology*, 35(14):1499–1507, Dec. 2005. ISSN 0020-7519. doi: 10.1016/j.ijpara.2005.08.011. URL https://www.sciencedirect.com/science/article/pii/S0020751905002924.

K. Hammerschmidt and J. Kurtz. *Schistocephalus solidus*: Establishment of tapeworms in sticklebacks – fast food or fast lane? *Experimental Parasitology*, 116(2):142–149, June 2007. ISSN 0014-4894. doi: 10.1016/j.exppara.2006.12.013. URL https://www.sciencedirect.com/science/article/pii/S0014489406003298.

L. Handunnetthi, S. V. Ramagopalan, G. C. Ebers, and J. C. Knight. Regulation of major histocompatibility complex class II gene expression, genetic variation and disease. *Genes & Immunity*, 11(2):99–112, Mar. 2010. ISSN 1476-5470. doi: 10.1038/gene.2009.83. URL https://www.nature.com/articles/gene200983. Number: 2 Publisher: Nature Publishing Group.

J. D. Hansen and S. La Patra. Induction of the rainbow trout MHC class I pathway during acute IHNV infection. *Immunogenetics*, 54(9):654–661, Dec. 2002. ISSN 1432-1211. doi: 10.1007/s00251-002-0509-x. URL https://doi.org/10.1007/s00251-002-0509-x.

W. Harnett. Secretory products of helminth parasites as immunomodulators. *Molecular and Biochemical Parasitology*, 195(2):130–136, July 2014. ISSN 1872-9428. doi: 10.1016/j.molbiopara.2014.03.007.

W. Hartmann, M.-L. Brunn, N. Stetter, N. Gagliani, F. Muscate, S. Stanelle-Bertram, G. Gabriel, and M. Breloer. Helminth Infections Suppress the Efficacy of Vaccination against Seasonal Influenza. *Cell Reports*, 29(8):2243–2256.e4, Nov. 2019. ISSN 22111247. doi: 10.1016/j.celrep.2019.10.051. URL https://linkinghub.elsevier.com/retrieve/pii/S2211124719313658.

J. Hataye, J. J. Moon, A. Khoruts, C. Reilly, and M. K. Jenkins. Naïve and Memory CD4+ T Cell Survival Controlled by Clonal Abundance. *Science*, 312(5770):114–116, Apr. 2006. doi: 10.1126/science.1124228. URL https://www.science.org/doi/abs/10.1126/science.1124228. Publisher: American Association for the Advancement of Science.

S. M. Hedrick, D. I. Cohen, E. A. Nielsen, and M. M. Davis. Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. *Nature*, 308(5955):149–153, Mar. 1984. ISSN 0028-0836. doi: 10.1038/308149a0.

J. P. Hewitson, J. R. Grainger, and R. M. Maizels. Helminth immunoregulation: The role of parasite secreted proteins in modulating host immunity. *Molecular and Biochemical Parasitology*, 167(1-9):1–11, Sept. 2009. ISSN 0166-6851. doi: 10.1016/j.molbiopara.2009.04.008. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2706953/.

L. W. Hillier, W. Miller, E. Birney, W. Warren, R. C. Hardison, C. P. Ponting, P. Bork, D. W. Burt, M. A. M. Groenen, M. E. Delany, J. B. Dodgson, A. T. Chinwalla, P. F. Cliften, S. W. Clifton, K. D. Delehaunty, C. Fronick, R. S. Fulton, T. A. Graves, C. Kremitzki, D. Layman, V. Magrini, J. D. McPherson, T. L. Miner, P. Minx, W. E. Nash, M. N. Nhan, J. O. Nelson, L. G. Oddy, C. S. Pohl, J. Randall-Maher, S. M. Smith, J. W. Wallis, S.-P. Yang, M. N. Romanov, C. M. Rondelli, B. Paton, J. Smith, D. Morrice, L. Daniels, H. G. Tempest, L. Robertson, J. S. Masabanda, D. K. Griffin, A. Vignal, V. Fillon, L. Jacobbson, S. Kerje, L. Andersson, R. P. M. Crooijmans, J. Aerts, J. J. van der Poel, H. Ellegren, R. B. Caldwell, S. J. Hubbard, D. V. Grafham, A. M. Kierzek, S. R. McLaren, I. M. Overton, H. Arakawa, K. J. Beattie, Y. Bezzubov, P. E. Boardman, J. K. Bonfield, M. D. R. Croning, R. M. Davies, M. D. Francis, S. J. Humphray, C. E. Scott, R. G. Taylor, C. Tickle, W. R. A. Brown, J. Rogers, J.-M. Buerstedde, S. A. Wilson, L. Stubbs, I. Ovcharenko, L. Gordon, S. Lucas, M. M. Miller, H. Inoko, T. Shiina, J. Kaufman, J. Salomonsen, K. Skjoedt, G. K.-S. Wong, J. Wang, B. Liu, J. Wang, J. Yu, H. Yang, M. Nefedov, M. Koriabine, P. J. deJong, L. Goodstadt, C. Webber, N. J. Dickens, I. Letunic, M. Suyama, D. Torrents, C. von Mering, E. M. Zdobnov, K. Makova, A. Nekrutenko, L. Elnitski, P. Eswara, D. C. King, S. Yang, S. Tyekucheva, A. Radakrishnan, R. S. Harris, F. Chiaromonte, J. Taylor, J. He, M. Rijnkels, S. Griffiths-Jones, A. Ureta-Vidal, M. M. Hoffman, J. Severin, S. M. J. Searle, A. S. Law, D. Speed, D. Waddington, Z. Cheng, E. Tuzun, E. Eichler, Z. Bao, P. Flicek, D. D. Shteynberg, M. R. Brent, J. M. Bye, E. J. Huckle, S. Chatterji, C. Dewey, L. Pachter, A. Kouranov, Z. Mourelatos, A. G. Hatzigeorgiou, A. H. Paterson, R. Ivarie, M. Brandstrom, E. Axelsson, N. Backstrom, S. Berlin, M. T. Webster, O. Pourquie, A. Reymond, C. Ucla, S. E. Antonarakis, M. Long, J. J. Emerson, E. Betrán, I. Dupanloup, H. Kaessmann, A. S. Hinrichs, G. Bejerano, T. S. Furey, R. A. Harte, B. Raney, A. Siepel, W. J. Kent, D. Haussler, E. Eyras, R. Castelo, J. F. Abril, S. Castellano, F. Camara, G. Parra, R. Guigo, G. Bourque, G. Tesler, P. A. Pevzner, A. Smit, L. A. Fulton, E. R. Mardis, R. K. Wilson, International Chicken Genome Sequencing Consortium, Overall coordination:, s. a. a. Genome fingerprint map, Mapping:, cDNA sequencing:, Other sequencing and libraries:, Analysis and annotation:, and Project management:. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432(7018):695–716, Dec. 2004. ISSN 1476-4687. doi: 10.1038/nature03154. URL https://www.nature.com/articles/nature03154. Number: 7018 Publisher: Nature Publishing Group.

C. Hohn and L. Petrie-Hanson. Rag1-/- mutant zebrafish demonstrate specific protection following bacterial re-exposure. *PloS One*, 7(9):e44451, 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0044451.

J. Holderness, J. F. Hedges, A. Ramstead, and M. A. Jutila. Comparative biology of T cell function in humans, mice, and domestic animals. *Annual Review of Animal Biosciences*, 1: 99–124, Jan. 2013. ISSN 2165-8110. doi: 10.1146/annurev-animal-031412-103639.

E. J. Homan and R. D. Bremel. A Role for Epitope Networking in Immunomodulation by

Helminths. *Frontiers in Immunology*, 9:1763, 2018. ISSN 1664-3224. doi: 10.3389/fimmu.2018.01763.

B. D. Hondowicz, D. An, J. M. Schenkel, K. S. Kim, H. R. Steach, A. T. Krishnamurty, G. J. Keitany, E. N. Garza, K. A. Fraser, J. J. Moon, W. A. Altemeier, D. Masopust, and M. Pepper. Interleukin-2-Dependent Allergen-Specific Tissue-Resident Memory Cells Drive Asthma. *Immunity*, 44(1):155–166, Jan. 2016. ISSN 1097-4180. doi: 10.1016/j.immuni.2015.11.004.

T. Hothorn, F. Bretz, P. Westfall, R. M. Heiberger, A. Schuetzenmeister, and S. Scheibe. multcomp: Simultaneous Inference in General Parametric Models, July 2024. URL https://cran.r-project.org/web/packages/multcomp/index.html.

X.-L. Hou, L. Wang, Y.-L. Ding, Q. Xie, and H.-Y. Diao. Current status and recent advances of next generation sequencing techniques in immunological repertoire. *Genes & Immunity*, 17(3):153–164, Apr. 2016. ISSN 1476-5470. doi: 10.1038/gene.2016.9. URL https://www.nature.com/articles/gene20169. Publisher: Nature Publishing Group.

N. Hozumi and S. Tonegawa. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proceedings of the National Academy of Sciences of the United States of America*, 73(10):3628, Oct. 1976. doi: 10.1073/pnas.73.10.3628. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC431171/.

C.-B. Hu, J. Wang, Y. Hong, H. Li, D.-D. Fan, A.-F. Lin, L.-X. Xiang, and J.-Z. Shao. Single-cell transcriptome profiling reveals diverse immune cell populations and their responses to viral infection in the spleen of zebrafish. *The FASEB Journal*, 37(6):e22951, 2023. ISSN 1530-6860. doi: 10.1096/fj.202201505RRRR. URL https://onlinelibrary.wiley.com/doi/abs/10.1096/fj.202201505RRRR. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1096/fj.202201505RRRR.

E. S. Huseby, J. White, F. Crawford, T. Vass, D. Becker, C. Pinilla, P. Marrack, and J. W. Kappler. How the T cell repertoire becomes peptide and MHC specific. *Cell*, 122(2):247–260, July 2005. ISSN 0092-8674. doi: 10.1016/j.cell.2005.05.013.

D. H. Huson and D. Bryant. Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution*, 23(2):254–267, Feb. 2006. ISSN 0737-4038. doi: 10.1093/molbev/msj030. URL https://doi.org/10.1093/molbev/msj030.

Illumina. Effects of Index Misassignment on Multiplexing and Downstream Analysis, 2017. URL https://emea.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf.

I. I. Ivanov, T. Tuganbaev, A. N. Skelly, and K. Honda. T Cell Responses to the Microbiota. *Annual Review of Immunology*, 40(Volume 40, 2022):559–587, Apr. 2022. ISSN 0732-0582, 1545-3278. doi: 10.1146/annurev-immunol-101320-011829. URL https://www.annualreviews.org/content/journals/10.1146/annurev-immunol-101320-011829. Publisher: Annual Reviews.

M. Izraelson, T. O. Nakonechnaya, B. Moltedo, E. S. Egorov, S. A. Kasatskaya, E. V. Putintseva, I. Z. Mamedov, D. B. Staroverov, I. I. Shemiakina, M. Y. Zakharova, A. N. Davydov, D. A. Bolotin, M. Shugay, D. M. Chudakov, A. Y. Rudensky, and O. V. Britanova. Comparative analysis of murine T-cell receptor repertoires. *Immunology*, 153(2):133–144, 2018. ISSN 1365-2567. doi: 10.1111/imm.12857.

URL https://onlinelibrary.wiley.com/doi/abs/10.1111/imm.12857. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/imm.12857.

S. Jaeger, B. Fernandez, and P. Ferrier. Epigenetic aspects of lymphocyte antigen receptor gene rearrangement or 'when stochasticity completes randomness'. *Immunology*, 139(2):141–150, 2013. ISSN 1365-2567. doi: 10.1111/imm.12057. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/imm.12057. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/imm.12057.

P. J. Jakobsen and G. H. Johnsen. The influence of predation on horizontal distribution of zooplankton species. *Freshwater Biology*, 17(3):501–507, 1987. ISSN 1365-2427. doi: 10.1111/j.1365-2427.1987.tb01070.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2427.1987.tb01070.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2427.1987.tb01070.x.

P. J. Jakobsen, J. P. Scharsack, K. Hammerschmidt, P. Deines, M. Kalbe, and M. Milinski. *In vitro* transition of *Schistocephalus solidus* (Cestoda) from coracidium to procercoid and from procercoid to plerocercoid. *Experimental Parasitology*, 130(3):267–273, Mar. 2012. ISSN 0014-4894. doi: 10.1016/j.exppara.2011.09.009. URL https://www.sciencedirect.com/science/article/pii/S0014489411002815.

C. A. Janeway and R. Medzhitov. Innate immune recognition. *Annual Review of Immunology*, 20:197–216, 2002. ISSN 0732-0582. doi: 10.1146/annurev.immunol.20.083001.084359.

M. K. Jenkins, H. H. Chu, J. B. McLachlan, and J. J. Moon. On the Composition of the Preimmune Repertoire of T Cells Specific for Peptide–Major Histocompatibility Complex Ligands. *Annual Review of Immunology*, 28(Volume 28, 2010):275–294, Apr. 2010. ISSN 0732-0582, 1545-3278. doi: 10.1146/annurev-immunol-030409-101253. URL https://www.annualreviews.org/content/journals/10.1146/annurev-immunol-030409-101253. Publisher: Annual Reviews.

N. K. Jerne. What Precedes Clonal Selection? In *Ciba Foundation Symposium 5 - Ontogeny of Acquired Immunity*, pages 1–15. John Wiley & Sons, Ltd, 1972. ISBN 978-0-470-71988-6. doi: 10.1002/9780470719886.ch1. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470719886.ch1. Section: 1 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470719886.ch1.

Z. Ji, Y. Sheng, J. Miao, X. Li, H. Zhao, J. Wang, C. Cheng, X. Wang, K. Liu, K. Zhang, L. Xu, J. Yao, L. Shen, J. Hou, W. Zhou, J. Sun, L. Li, W.-Q. Gao, and H. H. Zhu. The histone methyltransferase Setd2 is indispensable for V(D)J recombination. *Nature Communications*, 10(1):3353, July 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-11282-x. URL https://www.nature.com/articles/s41467-019-11282-x. Publisher: Nature Publishing Group.

F. C. Jones, M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli, J. Johnson, R. Swofford, M. Pirun, M. C. Zody, S. White, E. Birney, S. Searle, J. Schmutz, J. Grimwood, M. C. Dickson, R. M. Myers, C. T. Miller, B. R. Summers, A. K. Knecht, S. D. Brady, H. Zhang, A. A. Pollen, T. Howes, C. Amemiya, E. S. Lander, F. Di Palma, K. Lindblad-Toh, and D. M. Kingsley. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, 484(7392):55–61, Apr. 2012. ISSN 0028-0836. doi: 10.1038/nature10944. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322419/.

K. Joshi, M. Milighetti, and B. M. Chain. Application of T cell receptor (TCR) repertoire analysis for the advancement of cancer immunotherapy. *Current Opinion in Immunology*, 74:1–8, Feb. 2022. ISSN 1879-0372. doi: 10.1016/j.coi.2021.07.006.

I. Jäger, C. Eizaguirre, S. W. Griffiths, M. Kalbe, C. K. Krobbach, T. B. H. Reusch, H. Schaschl, and M. Milinski. Individual MHC class I and MHC class IIB diversities are associated with male and female reproductive traits in the three-spined stickleback. *Journal of Evolutionary Biology*, 20(5):2005–2015, 2007. ISSN 1420-9101. doi: 10.1111/j.1420-9101.2007.01366.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1420-9101.2007.01366.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1420-9101.2007.01366.x.

D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. *The R Journal*, 5(1):144–161, 2013. ISSN 2073-4859. URL https://journal.r-project.org/archive/2013/RJ-2013-014/index.html.

M. Kalbe and J. Kurtz. Local differences in immunocompetence reflect resistance of sticklebacks against the eye fluke Diplostomum pseudospathaceum. *Parasitology*, 132 (1):105–116, Jan. 2006. ISSN 1469-8161, 0031-1820. doi: 10.1017/S0031182005008681. URL https://www.cambridge.org/core/journals/parasitology/article/abs/local-differences-in-immunocompetence-reflect-resistance-of-sticklebacks-against-the-eye-fluke-diplostomum-pseudospathaceum/C31E45759B61836CE9D09C47736B388C.

M. Kalbe, K. M. Wegner, and T. B. H. Reusch. Dispersion patterns of parasites in 0+ year three-spined sticklebacks: a cross population comparison. *Journal of Fish Biology*, 60(6): 1529–1542, 2002. ISSN 1095-8649. doi: 10.1111/j.1095-8649.2002.tb02445.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1095-8649.2002.tb02445.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1095-8649.2002.tb02445.x.

M. Kalbe, C. Eizaguirre, I. Dankert, T. B. H. Reusch, R. D. Sommerfeld, K. M. Wegner, and M. Milinski. Lifetime reproductive success is maximized with optimal major histocompatibility complex diversity. *Proceedings. Biological Sciences*, 276(1658):925–934, Mar. 2009. ISSN 0962-8452. doi: 10.1098/rspb.2008.1466.

M. Kalbe, C. Eizaguirre, J. P. Scharsack, and P. J. Jakobsen. Reciprocal cross infection of sticklebacks with the diphyllobothriidean cestode Schistocephalus solidus reveals consistent population differences in parasite growth and host resistance. *Parasites & Vectors*, 9(1):130, Mar. 2016. ISSN 1756-3305. doi: 10.1186/s13071-016-1419-3. URL https://doi.org/10.1186/s13071-016-1419-3.

A. Kalinkovich, Z. Weisman, Z. Greenberg, J. Nahmias, S. Eitan, M. Stein, and Z. B. R. Decreased CD_4 and increased CD8 counts with T cell activation is associated with chronic helminth infection. *Clinical and Experimental Immunology*, 114(3):414, Dec. 1998. doi: 10.1046/j.1365-2249.1998.00736.x. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC1905129/.

T. Kanagawa. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of Bioscience and Bioengineering*, 96(4):317–323, Jan. 2003. ISSN 1389-1723. doi: 10.1016/S1389-1723(03)90130-7. URL https://www.sciencedirect.com/science/article/pii/S1389172303901307.

A. Kassambara. ggpubr: 'ggplot2' Based Publication Ready Plots, Feb. 2023. URL https://cran.r-project.org/web/packages/ggpubr/index.html.

S. H. E. Kaufmann. Immunology's foundation: the 100-year anniversary of the Nobel Prize to Paul Ehrlich and Elie Metchnikoff. *Nature Immunology*, 9(7):705–712, July 2008. ISSN 1529-2916. doi: 10.1038/ni0708-705.

S. H. E. Kaufmann. Immunology's Coming of Age. *Frontiers in Immunology*, 10:684, Apr. 2019. doi: 10.3389/fimmu.2019.00684. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6456699/.

J. Kelley, L. Walter, and J. Trowsdale. Comparative genomics of major histocompatibility complexes. *Immunogenetics*, 56(10):683–695, Jan. 2005. ISSN 1432-1211. doi: 10.1007/s00251-004-0717-7. URL https://doi.org/10.1007/s00251-004-0717-7.

M. Kircher, P. Heyn, and J. Kelso. Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genomics*, 12(1):382, July 2011. ISSN 1471-2164. doi: 10.1186/1471-2164-12-382. URL https://doi.org/10.1186/1471-2164-12-382.

J. Klein, R. E. Bontrop, R. L. Dawkins, H. A. Erlich, U. B. Gyllensten, E. R. Heise, P. P. Jones, P. Parham, E. K. Wakeland, and D. I. Watkins. Nomenclature for the major histocompatibility complexes of different species: a proposal. In B. G. Solheim, S. Ferrone, and E. Möller, editors, *The HLA System in Clinical Transplantation: Basic Concepts and Importance*, pages 407–411. Springer, Berlin, Heidelberg, 1993. ISBN 978-3-642-77506-2. doi: 10.1007/978-3-642-77506-2_32. URL https://doi.org/10.1007/978-3-642-77506-2_32.

L. Klein, B. Kyewski, P. M. Allen, and K. A. Hogquist. Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nature Reviews Immunology*, 14(6):377–391, June 2014. ISSN 1474-1741. doi: 10.1038/nri3667. URL https://www.nature.com/articles/nri3667. Publisher: Nature Publishing Group.

M. Klinger, F. Pepin, J. Wilkins, T. Asbury, T. Wittkop, J. Zheng, M. Moorhead, and M. Faham. Multiplex Identification of Antigen-Specific T Cell Receptors Using a Combination of Immune Assays and Immune Receptor Sequencing. *PloS One*, 10(10):e0141561, 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0141561.

C. Kosiol, T. Vinař, R. R. d. Fonseca, M. J. Hubisz, C. D. Bustamante, R. Nielsen, and A. Siepel. Patterns of Positive Selection in Six Mammalian Genomes. *PLOS Genetics*, 4(8):e1000144, Aug. 2008. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000144. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000144. Publisher: Public Library of Science.

C. Krishna, D. Chowell, M. Gönen, Y. Elhanati, and T. A. Chan. Genetic and environmental determinants of human TCR repertoire diversity. *Immunity & Ageing*, 17(1):26, Sept. 2020. ISSN 1742-4933. doi: 10.1186/s12979-020-00195-9. URL https://doi.org/10.1186/s12979-020-00195-9.

M. Krogsgaard and M. M. Davis. How T cells 'see' antigen. *Nature Immunology*, 6(3):239–245, Mar. 2005. ISSN 1529-2916. doi: 10.1038/ni1173. URL https://www.nature.com/articles/ni1173. Publisher: Nature Publishing Group.

A. Kurioka and P. Klenerman. Aging unconventionally: T cells, iNKT cells, and MAIT cells in aging. *Seminars in Immunology*, 69:101816, Sept. 2023. ISSN 1044-5323. doi: 10.1016/j.smim.2023.101816. URL https://www.sciencedirect.com/science/article/pii/S1044532323001070.

J. Kurtz, M. Kalbe, P. B. Aeschlimann, M. A. Häberli, K. M. Wegner, T. B. H. Reusch, and M. Milinski. Major histocompatibility complex diversity influences parasite resistance and innate immunity in sticklebacks. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(1535):197–204, Jan. 2004. doi: 10.1098/rspb.2003.2567. URL https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2003.2567. Publisher: Royal Society.

B. Kyewski and L. Klein. A central role for central tolerance. *Annual Review of Immunology*, 24(Volume 24, 2006):571–606, Apr. 2006. ISSN 0732-0582, 1545-3278. doi: 10.1146/annurev.immunol.23.021704.115601. URL https://www.annualreviews.org/content/journals/10.1146/annurev.immunol.23.021704.115601. Publisher: Annual Reviews.

S. Lachance and G. J. FitzGerald. Parental care tactics of three-spined sticklebacks living in a harsh environment. *Behavioral Ecology*, 3(4):360–366, Dec. 1992. ISSN 1045-2249. doi: 10.1093/beheco/3.4.360. URL https://doi.org/10.1093/beheco/3.4.360.

K. D. Lafferty. Interacting Parasites. *Science*, 330(6001):187–188, Oct. 2010. doi: 10.1126/science.1196915. URL https://www.science.org/doi/abs/10.1126/science.1196915. Publisher: American Association for the Advancement of Science.

K. D. Lafferty and A. M. Kuris. Parasitic castration: the evolution and ecology of body snatchers. *Trends in Parasitology*, 25(12):564–572, Dec. 2009. ISSN 1471-5007. doi: 10.1016/j.pt.2009.09.003.

M. J. Lajeunesse and M. R. Forbes. Host range and local parasite adaptation. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 269(1492):703–710, Apr. 2002. doi: 10.1098/rspb.2001.1943. URL https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2001.1943. Publisher: Royal Society.

T. H. Lam, M. Shen, M. Z. Tay, and E. C. Ren. Unique Allelic eQTL Clusters in Human MHC Haplotypes. *G3 Genes|Genomes|Genetics*, 7(8):2595–2604, Aug. 2017. ISSN 2160-1836. doi: 10.1534/g3.117.043828. URL https://doi.org/10.1534/g3.117.043828.

R. E. Langman and M. Cohn. The E-T (Elephant-Tadpole) paradox necessitates the concept of a unit of b-cell function: The protecton. *Molecular Immunology*, 24(7):675–697, July 1987. ISSN 0161-5890. doi: 10.1016/0161-5890(87)90050-2. URL https://www.sciencedirect.com/science/article/pii/0161589087900502.

C. R. Largiadèr, V. Fries, B. Kobler, and T. C. Bakker. Isolation and characterization of microsatellite loci from the three-spined stickleback (Gasterosteus aculeatus L.). *Molecular Ecology*, 8(2):342–344, Feb. 1999. ISSN 0962-1083.

B. P. Lazzaro and T. J. Little. Immunity in a variable world. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1513):15–26, Oct. 2008. doi: 10.1098/rstb.2008.0141. URL https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2008.0141. Publisher: Royal Society.

Y. K. Lee and S. K. Mazmanian. Has the Microbiota Played a Critical Role in the Evolution of the Adaptive Immune System? *Science*, 330(6012):1768–1773, Dec. 2010. doi: 10.1126/science.1195568. URL https://www.science.org/doi/abs/10.1126/science.1195568. Publisher: American Association for the Advancement of Science.

Y. N. Lee, F. Frugoni, K. Dobbs, I. Tirosh, L. Du, F. A. Ververs, H. Ru, L. Ott de Bruin, M. Adeli, J. H. Bleesing, D. Buchbinder, M. J. Butte, C. Cancrini, K. Chen, S. Choo, R. A. Elfeky, A. Finocchi, R. L. Fuleihan, A. R. Gennery, D. H. El-Ghoneimy, L. A. Henderson, W. Al-Herz, E. Hossny, R. P. Nelson, S.-Y. Pai, N. C. Patel, S. M. Reda, P. Soler-Palacin, R. Somech, P. Palma, H. Wu, S. Giliani, J. E. Walter, and L. D. Notarangelo. Characterization of T and B cell repertoire diversity in patients with RAG deficiency. *Science Immunology*, 1(6):eaah6109, Dec. 2016. ISSN 2470-9468. doi: 10.1126/sciimmunol.aah6109.

A. Leewenhoeck. Microscopical observations about animals in the scurf of the teeth. *Philosophical Transactions (1683-1775)*, 14:568–574, 1684. ISSN 0260-7085. URL https://www.jstor.org/stable/102057. Publisher: The Royal Society.

M.-P. Lefranc, V. Giudicelli, C. Ginestoux, J. Bodmer, W. Müller, R. Bontrop, M. Lemaitre, A. Malik, V. Barbié, and D. Chaume. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Research*, 27(1):209–212, Jan. 1999. ISSN 0305-1048. doi: 10.1093/nar/27.1.209. URL https://doi.org/10.1093/nar/27.1.209.

F. Legoux, M. Salou, and O. Lantz. Unconventional or Preset T Cells: Evolutionarily Conserved Tissue-Resident T Cells Recognizing Nonpeptidic Ligands. *Annual Review of Cell and Developmental Biology*, 33:511–535, Oct. 2017. ISSN 1530-8995. doi: 10.1146/annurev-cellbio-100616-060725.

T. L. Lenz and S. Becker. Simple approach to reduce PCR artefact formation leads to reliable genotyping of MHC and other highly polymorphic loci — Implications for evolutionary analysis. *Gene*, 427(1):117–123, Dec. 2008. ISSN 0378-1119. doi: 10.1016/j.gene.2008.09.013. URL https://www.sciencedirect.com/science/article/pii/S037811190800471X.

T. L. Lenz, C. Eizaguirre, S. Becker, and T. B. Reusch. RSCA genotyping of MHC for high-throughput evolutionary studies in the model organism three-spined stickleback Gasterosteus aculeatus. *BMC Evolutionary Biology*, 9(1):57, Mar. 2009a. ISSN 1471-2148. doi: 10.1186/1471-2148-9-57. URL https://doi.org/10.1186/1471-2148-9-57.

T. L. Lenz, C. Eizaguirre, J. P. Scharsack, M. Kalbe, and M. Milinski. Disentangling the role of MHC-dependent 'good genes' and 'compatible genes' in mate-choice decisions of three-spined sticklebacks Gasterosteus aculeatus under semi-natural conditions. *Journal of Fish Biology*, 75(8):2122–2142, 2009b. ISSN 1095-8649. doi: 10.1111/j.1095-8649.2009.02410.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1095-8649.2009.02410.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1095-8649.2009.02410.x.

M. Lepore, A. Kalinichenko, A. Colone, B. Paleja, A. Singhal, A. Tschumi, B. Lee, M. Poidinger, F. Zolezzi, L. Quagliata, P. Sander, E. Newell, A. Bertoletti, L. Terracciano, G. De Libero, and L. Mori. Parallel T-cell cloning and deep sequencing of human MAIT cells reveal stable oligoclonal TCR repertoire. *Nature Communications*, 5(1):3866, May 2014. ISSN 2041-1723. doi: 10.1038/ncomms4866. URL https://www.nature.com/articles/ncomms4866. Publisher: Nature Publishing Group.

H. Li, C. Ye, G. Ji, and J. Han. Determinants of public T cell responses. *Cell Research*, 22(1): 33–42, Jan. 2012. ISSN 1748-7838. doi: 10.1038/cr.2012.1. URL https://www.nature.com/articles/cr20121. Publisher: Nature Publishing Group.

L.-P. Li, J. C. Lampert, X. Chen, C. Leitao, J. Popović, W. Müller, and T. Blankenstein. Transgenic mice with a diverse human T cell antigen receptor repertoire. *Nature Medicine*, 16(9):1029–1034, Sept. 2010. ISSN 1546-170X. doi: 10.1038/nm.2197. URL https://www.nature.com/articles/nm.2197. Publisher: Nature Publishing Group.

S. Li, M.-P. Lefranc, J. J. Miles, E. Alamyar, V. Giudicelli, P. Duroux, J. D. Freeman, V. D. A. Corbin, J.-P. Scheerlinck, M. A. Frohman, P. U. Cameron, M. Plebanski, B. Loveland, S. R. Burrows, A. T. Papenfuss, and E. J. Gowans. IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nature Communications*, 4(1):2333, Sept. 2013. ISSN 2041-1723. doi: 10.1038/ncomms3333. URL https://www.nature.com/articles/ncomms3333. Publisher: Nature Publishing Group.

X.-X. Li, J.-X. Chen, L.-X. Wang, J. Sun, S.-H. Chen, J.-H. Chen, X.-Y. Zhang, and X.-N. Zhou. Profiling B and T cell immune responses to co-infection of Mycobacterium tuberculosis and hookworm in humans. *Infectious Diseases of Poverty*, 4(1):20, May 2015. ISSN 2049-9957. doi: 10.1186/s40249-015-0046-0. URL https://doi.org/10.1186/s40249-015-0046-0.

Y. Liang, X. Chen, J. Chen, and F. Liu. Differentiation and Function of T Cell Subsets in Infectious Diseases. *Journal of Immunology Research*, 2018:3439025, Sept. 2018. ISSN 2314-8861. doi: 10.1155/2018/3439025. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6174770/.

J. Lighten, C. van Oosterhout, and P. Bentzen. Critical review of NGS analyses for de novo genotyping multigene families. *Molecular Ecology*, 23(16):3957–3972, 2014. ISSN 1365-294X. doi: 10.1111/mec.12843. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.12843. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.12843.

A. Lim, L. Trautmann, M.-A. Peyrat, C. Couedel, F. Davodeau, F. Romagné, P. Kourilsky, and M. Bonneville. Frequent Contribution of T Cell Clonotypes with Public TCR Features to the Chronic Response Against a Dominant EBV-Derived Epitope: Application to Direct Detection of Their Molecular Imprint on the Human Peripheral T Cell Repertoire1. *The Journal of Immunology*, 165(4):2001–2011, Aug. 2000. ISSN 0022-1767. doi: 10.4049/jimmunol.165.4.2001. URL https://doi.org/10.4049/jimmunol.165.4.2001.

T. Lin, D. Zhang, X. Liu, and D. Xiao. Parental care improves immunity in the seahorse (*Hippocampus erectus*). *Fish & Shellfish Immunology*, 58:554–562, Nov. 2016. ISSN 1050-4648. doi: 10.1016/j.fsi.2016.09.065. URL https://www.sciencedirect.com/science/article/pii/S1050464816306325.

J. L. Linehan, O. J. Harrison, S.-J. Han, A. L. Byrd, I. Vujkovic-Cvijin, A. V. Villarino, S. K. Sen, J. Shaik, M. Smelkinson, S. Tamoutounour, N. Collins, N. Bouladoux, A. Dzutsev, S. P. Rosshart, J. H. Arbuckle, C.-R. Wang, T. M. Kristie, B. Rehermann, G. Trinchieri, J. M. Brenchley, J. J. O'Shea, and Y. Belkaid. Non-classical Immunity Controls Microbiota Impact on Skin Immunity and Tissue Repair. *Cell*, 172(4):784–796.e18, Feb. 2018. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2017.12.033. URL https://www.cell.com/cell/abstract/S0092-8674(17)31513-1. Publisher: Elsevier.

A. Liston, S. Humblet-Baron, D. Duffy, and A. Goris. Human immune diversity: from evolution to modernity. *Nature Immunology*, 22(12):1479–1489, Dec. 2021. ISSN 1529-2916. doi: 10.1038/s41590-021-01058-1. URL https://www.nature.com/articles/s41590-021-01058-1. Publisher: Nature Publishing Group.

X. Liu, W. Zhang, M. Zhao, L. Fu, L. Liu, J. Wu, S. Luo, L. Wang, Z. Wang, L. Lin, Y. Liu, S. Wang, Y. Yang, L. Luo, J. Jiang, X. Wang, Y. Tan, T. Li, B. Zhu, Y. Zhao, X. Gao, Z. Wan, C. Huang, M. Fang, Q. Li, H. Peng, X. Liao, J. Chen, F. Li, G. Ling, H. Zhao, H. Luo, Z. Xiang, J. Liao, Y. Liu, H. Yin, H. Long, H. Wu, H. Yang, J. Wang, and Q. Lu. T cell receptor repertoires as novel diagnostic markers for systemic lupus erythematosus and rheumatoid arthritis. *Annals of the Rheumatic Diseases*, 78(8):1070–1078, Aug. 2019. ISSN 1468-2060. doi: 10.1136/annrheumdis-2019-215442.

J. Lu, F. Van Laethem, A. Bhattacharya, M. Craveiro, I. Saba, J. Chu, N. C. Love, A. Tikhonova, S. Radaev, X. Sun, A. Ko, T. Arnon, E. Shifrut, N. Friedman, N.-P. Weng, A. Singer, and P. D. Sun. Molecular constraints on CDR3 for thymic selection of MHC-restricted TCRs from a random pre-selection repertoire. *Nature Communications*, 10:1019, Mar. 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08906-7. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6399321/.

S. Luo, J. A. Yu, H. Li, and Y. S. Song. Worldwide genetic variation of the IGHV and TRBV immune receptor gene families in humans. *Life Science Alliance*, 2(2):e201800221, Apr. 2019. ISSN 2575-1077. doi: 10.26508/lsa.201800221.

S. Lustigman, R. K. Prichard, A. Gazzinelli, W. N. Grant, B. A. Boatin, J. S. McCarthy, and M.-G. Basáñez. A Research Agenda for Helminth Diseases of Humans: The Problem of Helminthiases. *PLoS Neglected Tropical Diseases*, 6(4):e1582, Apr. 2012. doi: 10.1371/journal.pntd.0001582. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3335854/.

A. D. C. MacColl. Parasite Burdens Differ between Sympatric Three-Spined Stickleback Species. *Ecography*, 32(1):153–160, 2009. ISSN 0906-7590. URL https://www.jstor.org/stable/30244661. Publisher: [Nordic Society Oikos, Wiley].

A. Madi, E. Shifrut, S. Reich-Zeliger, H. Gal, K. Best, W. Ndifon, B. Chain, I. R. Cohen, and N. Friedman. T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Research*, 24 (10):1603–1612, Oct. 2014. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.170753.113. URL https://genome.cshlp.org/content/24/10/1603. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

A. Madi, A. Poran, E. Shifrut, S. Reich-Zeliger, E. Greenstein, I. Zaretsky, T. Arnon, F. V. Laethem, A. Singer, J. Lu, P. D. Sun, I. R. Cohen, and N. Friedman. T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences. *eLife*, 6:e22057, July 2017. ISSN 2050-084X. doi: 10.7554/eLife.22057. URL https://doi.org/10.7554/eLife.22057. Publisher: eLife Sciences Publications, Ltd.

B. Magnadóttir. Innate immunity of fish (overview). *Fish & Shellfish Immunology*, 20(2):137–151, Feb. 2006. ISSN 1050-4648. doi: 10.1016/j.fsi.2004.09.006.

R. M. Maizels and M. Yazdanbakhsh. Immune Regulation by helminth parasites: cellular and molecular mechanisms. *Nature Reviews Immunology*, 3(9):733–744, Sept. 2003. ISSN 1474-1741. doi: 10.1038/nri1183. URL https://www.nature.com/articles/nri1183. Publisher: Nature Publishing Group.

M. Malissen, J. Trucy, E. Jouvin-Marche, P.-A. Cazenave, R. Scollay, and B. Malissen. Regulation of TCR and gene allelic exclusion during T-cell development. *Immunology Today*, 13(8): 315–322, Jan. 1992. ISSN 0167-5699. doi: 10.1016/0167-5699(92)90044-8. URL https://www.cell.com/immunology/abstract/0167-5699(92)90044-8. Publisher: Elsevier.

T. Mallevaey, J. P. Scott-Browne, J. L. Matsuda, M. H. Young, D. G. Pellicci, O. Patel, M. Thakur, L. Kjer-Nielsen, S. K. Richardson, V. Cerundolo, A. R. Howell, J. McCluskey, D. I. Godfrey, J. Rossjohn, P. Marrack, and L. Gapin. T cell receptor CDR2 beta and CDR3 beta loops collaborate functionally to shape the iNKT cell repertoire. *Immunity*, 31(1):60–71, July 2009. ISSN 1097-4180. doi: 10.1016/j.immuni.2009.05.010.

L. J. Manley, D. Ma, and S. S. Levine. Monitoring Error Rates In Illumina Sequencing. *Journal of Biomolecular Techniques : JBT*, 27(4):125–128, Dec. 2016. ISSN 1524-0215. doi: 10.7171/jbt.16-2704-002. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5026502/.

B. Manoury, W. F. Gregory, R. M. Maizels, and C. Watts. Bm-CPI-2, a cystatin homolog secreted by the filarial parasite Brugia malayi, inhibits class II MHC-restricted antigen processing. *Current biology: CB*, 11(6):447–451, Mar. 2001. ISSN 0960-9822. doi: 10.1016/s0960-9822(01)00118-x.

Q. Marcou, T. Mora, and A. M. Walczak. High-throughput immune repertoire analysis with IGoR. *Nature Communications*, 9(1):561, Feb. 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-02832-w. URL https://www.nature.com/articles/s41467-018-02832-w. Number: 1 Publisher: Nature Publishing Group.

J. S. Marshall, R. Warrington, W. Watson, and H. L. Kim. An introduction to immunology and immunopathology. *Allergy, Asthma & Clinical Immunology*, 14(2):49, Sept. 2018. ISSN 1710-1492. doi: 10.1186/s13223-018-0278-1. URL https://doi.org/10.1186/s13223-018-0278-1.

P. Massicotte, A. South, and K. Hufkens. rnaturalearth: World Map Data from Natural Earth, Dec. 2023. URL https://cran.r-project.org/web/packages/rnaturalearth/.

H. Mazanec, N. Buskova, Z. Gardian, and R. Kuchta. Secretion of extracellular vesicles during ontogeny of the tapeworm Schistocephalus solidus. *Folia Parasitologica*, 70:2023.003, Jan. 2023. ISSN 1803-6465. doi: 10.14411/fp.2023.003.

J. S. McKinnon and H. D. Rundle. Speciation in nature: the threespine stickleback model systems. *Trends in Ecology & Evolution*, 17(10):480–488, Oct. 2002. ISSN 0169-5347. doi: 10.1016/S0169-5347(02)02579-X. URL https://www.sciencedirect.com/science/article/pii/S016953470202579X.

H. J. McSorley and R. M. Maizels. Helminth Infections and Host Immune Regulation. *Clinical Microbiology Reviews*, 25(4):585–608, Oct. 2012. doi: 10.1128/cmr.05040-11. URL https://journals.asm.org/doi/full/10.1128/cmr.05040-11. Publisher: American Society for Microbiology.

N. D. Meeker, A. C. H. Smith, J. K. Frazer, D. F. Bradley, L. A. Rudner, C. Love, and N. S. Trede. Characterization of the zebrafish T cell receptor beta locus. *Immunogenetics*, 62(1):23–29, Jan. 2010. ISSN 1432-1211. doi: 10.1007/s00251-009-0407-6.

J. J. Melenhorst, M. D. H. Lay, D. A. Price, S. D. Adams, J. Zeilah, E. Sosa, N. F. Hensel, D. Follmann, D. C. Douek, M. P. Davenport, and A. J. Barrett. Contribution of TCR- Locus and HLA to the Shape of the Mature Human V Repertoire1. *The Journal of Immunology*, 180(10):6484–6489, May 2008. ISSN 0022-1767. doi: 10.4049/jimmunol.180.10.6484. URL https://doi.org/10.4049/jimmunol.180.10.6484.

V. Mhanna, H. Bashour, K. Lê Quý, P. Barennes, P. Rawat, V. Greiff, and E. Mariotti-Ferrandiz. Adaptive immune receptor repertoire analysis. *Nature Reviews Methods Primers*, 4(1):1–25, Jan. 2024. ISSN 2662-8449. doi: 10.1038/s43586-023-00284-1. URL https://www.nature.com/articles/s43586-023-00284-1. Publisher: Nature Publishing Group.

M. Migalska, A. Sebastian, and J. Radwan. Major histocompatibility complex class I diversity limits the repertoire of T cell receptors. *Proceedings of the National Academy of Sciences*, 116 (11):5021–5026, Mar. 2019. doi: 10.1073/pnas.1807864116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1807864116. Publisher: Proceedings of the National Academy of Sciences.

J. J. Miles, D. C. Douek, and D. A. Price. Bias in the T-cell repertoire: implications for disease pathogenesis and vaccination. *Immunology & Cell Biology*, 89(3):375–387, 2011. ISSN 1440-1711. doi: 10.1038/icb.2010.139. URL https://onlinelibrary.wiley.com/doi/abs/10.1038/icb.2010.139. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1038/icb.2010.139.

A. A. Minervina, E. A. Komech, A. Titov, M. Bensouda Koraichi, E. Rosati, I. Z. Mamedov, A. Franke, G. A. Efimov, D. M. Chudakov, T. Mora, A. M. Walczak, Y. B. Lebedev, and M. V. Pogorelyy. Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T-cell memory formation after mild COVID-19 infection. *eLife*, 10:e63502, Jan. 2021. ISSN 2050-084X. doi: 10.7554/eLife.63502. URL https://elifesciences.org/articles/63502.

M. Montes, C. Sanchez, K. Verdonck, J. E. Lake, E. Gonzalez, G. Lopez, A. Terashima, T. Nolan, D. E. Lewis, E. Gotuzzo, and A. C. W. Jr. Regulatory T Cell Expansion in HTLV-1 and Strongyloidiasis Co-infection Is Associated with Reduced IL-5 Responses to Strongyloides stercoralis Antigen. *PLOS Neglected Tropical Diseases*, 3(6):e456, June 2009. ISSN 1935-2735. doi: 10.1371/journal.pntd.0000456. URL https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0000456. Publisher: Public Library of Science.

A. Montresor, D. Mupfasoni, A. Mikhailov, P. Mwinzi, A. Lucianez, M. Jamsheed, E. Gasimov, S. Warusavithana, A. Yajima, Z. Bisoffi, D. Buonfrate, P. Steinmann, J. Utzinger, B. Levecke, J. Vlaminck, P. Cools, J. Vercruysse, G. Cringoli, L. Rinaldi, B. Blouin, and T. W. Gyorkos. The global progress of soil-transmitted helminthiases control in 2020 and World Health Organization targets for 2030. *PLOS Neglected Tropical Diseases*, 14(8):e0008505, Aug. 2020. ISSN 1935-2735. doi: 10.1371/journal.pntd.0008505. URL https://dx.plos.org/10.1371/journal.pntd.0008505.

J. J. Moon, H. H. Chu, M. Pepper, S. J. McSorley, S. C. Jameson, R. M. Kedl, and M. K. Jenkins. Naive CD4+ T Cell Frequency Varies for Different Epitopes and Predicts Repertoire

Diversity and Response Magnitude. *Immunity*, 27(2):203–213, Aug. 2007. ISSN 1074-7613. doi: 10.1016/j.immuni.2007.07.007. URL https://www.cell.com/immunity/abstract/S1074-7613(07)00366-4. Publisher: Elsevier.

M. A. Moro-García, R. Alonso-Arias, and C. Lopez-Larrea. When Aging Reaches CD4+ T-Cells: Phenotypic and Functional Changes. *Frontiers in Immunology*, 4, May 2013. ISSN 1664-3224. doi: 10.3389/fimmu.2013.00107. URL https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2013.00107/full. Publisher: Frontiers.

K. Murphy and C. Weaver. *Janeway's Immunobiology*. Number 9780815345053. New York: Garland Science, New York, 9th ed. edition, Mar. 2016. ISBN 978-1-315-53324-7. Google-Books-ID: GmPLCwAAQBAJ.

J. M. Murray, G. R. Kaufmann, P. D. Hodgkin, S. R. Lewin, A. D. Kelleher, M. P. Davenport, and J. J. Zaunders. Naive T cells are maintained by thymic output in early ages but by proliferation without phenotypic change after age twenty. *Immunology and Cell Biology*, 81 (6):487–495, Dec. 2003. ISSN 0818-9641. doi: 10.1046/j.1440-1711.2003.01191.x.

M. Musvosvi, H. Huang, C. Wang, Q. Xia, V. Rozot, A. Krishnan, P. Acs, A. Cheruku, G. Obermoser, A. Leslie, S. M. Behar, W. A. Hanekom, N. Bilek, M. Fisher, S. H. E. Kaufmann, G. Walzl, M. Hatherill, M. M. Davis, and T. J. Scriba. T cell receptor repertoires associated with control and disease progression following Mycobacterium tuberculosis infection. *Nature Medicine*, 29(1):258–269, Jan. 2023. ISSN 1546-170X. doi: 10.1038/s41591-022-02110-9. URL https://www.nature.com/articles/s41591-022-02110-9. Publisher: Nature Publishing Group.

K. Nagashima, A. Zhao, K. Atabakhsh, M. Bae, J. E. Blum, A. Weakley, S. Jain, X. Meng, A. G. Cheng, M. Wang, S. Higginbottom, A. Dimas, P. Murugkar, E. S. Sattely, J. J. Moon, E. P. Balskus, and M. A. Fischbach. Mapping the T cell repertoire to a complex gut bacterial community. *Nature*, 621(7977):162–170, Sept. 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06431-8. URL https://www.nature.com/articles/s41586-023-06431-8. Publisher: Nature Publishing Group.

N. K. Nanda, R. Apple, and E. Sercarz. Limitations in plasticity of the T-cell receptor repertoire. *Proceedings of the National Academy of Sciences*, 88(21):9503–9507, Nov. 1991. doi: 10.1073/pnas.88.21.9503. URL https://www.pnas.org/doi/abs/10.1073/pnas.88.21.9503. Publisher: Proceedings of the National Academy of Sciences.

K. Naylor, G. Li, A. N. Vallejo, W.-W. Lee, K. Koetz, E. Bryl, J. Witkowski, J. Fulbright, C. M. Weyand, and J. J. Goronzy. The influence of age on T cell generation and TCR diversity. *Journal of Immunology (Baltimore, Md.: 1950)*, 174(11):7446–7452, June 2005. ISSN 0022-1767. doi: 10.4049/jimmunol.174.11.7446.

E. Neuwirth. RColorBrewer: ColorBrewer Palettes, Apr. 2022. URL https://cran.r-project.org/web/packages/RColorBrewer/index.html.

J. Nikolich-Žugich, M. K. Slifka, and I. Messaoudi. The many important facets of T-cell repertoire diversity. *Nature Reviews Immunology*, 4(2):123–132, Feb. 2004. ISSN 1474-1741. doi: 10.1038/nri1292. URL https://www.nature.com/articles/nri1292. Publisher: Nature Publishing Group.

L. D. Notarangelo, M.-S. Kim, J. E. Walter, and Y. N. Lee. Human RAG mutations: biochemistry and clinical implications. *Nature Reviews. Immunology*, 16(4):234–246, Apr. 2016. ISSN 1474-1741. doi: 10.1038/nri.2016.28.

M. A. Nowak, K. Tarczy-Hornoch, and J. M. Austyn. The optimal number of major histocompatibility complex molecules in an individual. *Proceedings of the National Academy of Sciences*, 89(22):10896–10899, Nov. 1992. doi: 10.1073/pnas.89.22.10896. URL https://www.pnas.org/doi/abs/10.1073/pnas.89.22.10896. Publisher: Proceedings of the National Academy of Sciences.

J. Oksanen, G. L. Simpson, F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, P. Solymos, M. H. H. Stevens, E. Szoecs, H. Wagner, M. Barbour, M. Bedward, B. Bolker, D. Borcard, G. Carvalho, M. Chirico, M. D. Caceres, S. Durand, H. B. A. Evangelista, R. FitzJohn, M. Friendly, B. Furneaux, G. Hannigan, M. O. Hill, L. Lahti, D. McGlinn, M.-H. Ouellette, E. R. Cunha, T. Smith, A. Stier, C. J. F. T. Braak, and J. Weedon. vegan: Community Ecology Package, Aug. 2024. URL https://cran.r-project.org/web/packages/vegan/index.html.

A. Omer, A. Peres, O. L. Rodriguez, C. T. Watson, W. Lees, P. Polak, A. M. Collins, and G. Yaari. T cell receptor beta germline variability is revealed by inference from repertoire data. *Genome Medicine*, 14(1):2, Jan. 2022. ISSN 1756-994X. doi: 10.1186/s13073-021-01008-4. URL https://doi.org/10.1186/s13073-021-01008-4.

T. S. C. Orr, C. A. Hopkins, and G. H. Charles. Host specificity and rejection of Schistocephalus solidus. *Parasitology*, 59(3):683–690, Aug. 1969. ISSN 1469-8161, 0031-1820. doi: 10.1017/S0031182000031206. URL https://www.cambridge.org/core/journals/parasitology/article/abs/host-specificity-and-rejection-of-schistocephalus-solidus/A842B3A78841C6D6C2D683DCA9039FBF.

C. S. O'Brien, R. Bourdo, W. E. Bradshaw, C. M. Holzapfel, and W. A. Cresko. Conservation of the photoperiodic neuroendocrine axis among vertebrates: evidence from the teleost fish, Gasterosteus aculeatus. *General and Comparative Endocrinology*, 178(1):19–27, Aug. 2012. ISSN 0016-6480. doi: 10.1016/j.ygcen.2012.03.010. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3389224/.

D. B. Palmer. The Effect of Age on Thymic Function. *Frontiers in Immunology*, 4, Oct. 2013. ISSN 1664-3224. doi: 10.3389/fimmu.2013.00316. URL https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2013.00316/full. Publisher: Frontiers.

R.-Y. Pan, M.-T. Chu, C.-W. Wang, Y.-S. Lee, F. Lemonnier, A. W. Michels, R. Schutte, D. A. Ostrov, C.-B. Chen, E. J. Phillips, S. A. Mallal, M. Mockenhaupt, T. Bellón, W. Tassaneeyakul, K. D. White, J.-C. Roujeau, W.-H. Chung, and S.-I. Hung. Identification of drug-specific public TCR driving severe cutaneous adverse reactions. *Nature Communications*, 10(1):3569, Aug. 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-11396-2. URL https://www.nature.com/articles/s41467-019-11396-2. Publisher: Nature Publishing Group.

Z. Pancer and M. D. Cooper. The evolution of adaptive immunity. *Annual Review of Immunology*, 24:497–518, 2006. ISSN 0732-0582. doi: 10.1146/annurev.immunol.24.021605.090542.

P. Parham. Innate immunity: The unsung heroes. *Nature*, 423(6935):20–20, May 2003. ISSN 1476-4687. doi: 10.1038/423020a. URL https://www.nature.com/articles/423020a. Publisher: Nature Publishing Group.

I. Patil. Visualizations with statistical details: The 'ggstatsplot' approac. Technical report, 2021. URL https://doi.org/10.31234/osf.io/p7mku. Type: article.

S. A. Pavey, M. Sevellec, W. Adam, E. Normandeau, F. C. Lamaze, P.-A. Gagnaire, M. Filteau, F. O. Hebert, H. Maaroufi, and L. Bernatchez. Nonparallelism in MHCII diversity accompanies nonparallelism in pathogen infection of lake whitefish (Coregonus clupeaformis) species pairs as revealed by next-generation sequencing. *Molecular Ecology*, 22(14):3833–3849, July 2013. ISSN 1365-294X. doi: 10.1111/mec.12358.

A. Pawlowski, M. Jansson, M. Sköld, M. E. Rottenberg, and G. Källenius. Tuberculosis and HIV Co-Infection. *PLOS Pathogens*, 8(2):e1002464, Feb. 2012. ISSN 1553-7374. doi: 10.1371/journal.ppat.1002464. URL https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1002464. Publisher: Public Library of Science.

E. Peatman, J. Terhune, P. Baoprasertkul, P. Xu, S. Nandi, S. Wang, B. Somridhivej, H. Kucuktas, P. Li, R. Dunham, and Z. Liu. Microarray analysis of gene expression in the blue catfish liver reveals early activation of the MHC class I pathway after infection with Edwardsiella ictaluri. *Molecular Immunology*, 45(2):553–566, Jan. 2008. ISSN 0161-5890. doi: 10.1016/j.molimm.2007.05.012. URL https://www.sciencedirect.com/science/article/pii/S0161589007002325.

C. L. Peichel, K. S. Nereng, K. A. Ohgi, B. L. E. Cole, P. F. Colosimo, C. A. Buerkle, D. Schluter, and D. M. Kingsley. The genetic architecture of divergence between threespine stickleback species. *Nature*, 414(6866):901–905, Dec. 2001. ISSN 1476-4687. doi: 10.1038/414901a. URL https://www.nature.com/articles/414901a. Publisher: Nature Publishing Group.

X. Peng, Y. Lei, P. Feng, L. Jia, J. Ma, D. Zhao, and J. Zeng. Characterizing the interaction conformation between T-cell receptors and epitopes with deep learning. *Nature Machine Intelligence*, 5(4):395–407, Apr. 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00634-4. URL https://www.nature.com/articles/s42256-023-00634-4. Publisher: Nature Publishing Group.

L. Pennycuick. Seasonal variations in the parasite infections in a population of three-spined sticklebacks, Gasterosteus aculeatus L. *Parasitology*, 63(3):373–388, Dec. 1971. ISSN 1469-8161, 0031-1820. doi: 10.1017/S0031182000079919. URL https://www.cambridge.org/core/journals/parasitology/article/abs/seasonal-variations-in-the-parasite-infections-in-a-population-of-threespined-sticklebacks-gasterosteus-aculeatus-l/1709C58113BD951CA1A3135A840DDA6E.

J. Perera, L. Meng, F. Meng, and H. Huang. Autoreactive thymic B cells are efficient antigen-presenting cells of cognate self-antigens for T cell negative selection. *Proceedings of the National Academy of Sciences of the United States of America*, 110(42):17011–17016, Oct. 2013. ISSN 1091-6490. doi: 10.1073/pnas.1313001110.

A.-C. Persson, R. J. M. Stet, and L. Pilström. Characterization of MHC class I and 2-microglobulin sequences in Atlantic cod reveals an unusually high number of expressed class I genes. *Immunogenetics*, 50(1):49–59, Oct. 1999. ISSN 1432-1211. doi: 10.1007/s002510050685. URL https://doi.org/10.1007/s002510050685.

E. W. Petersdorf and C. O'hUigin. The MHC in the era of next-generation sequencing: Implications for bridging structure with function. *Human Immunology*, 80(1):67–78, Jan. 2019. ISSN 01988859. doi: 10.1016/j.humimm.2018.10.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0198885918309121.

T. N. Petney and R. H. Andrews. Multiparasite communities in animals and humans: frequency, structure and pathogenic significance. *International Journal for Parasitology*, 28(3): 377–393, Mar. 1998. ISSN 0020-7519. doi: 10.1016/S0020-7519(97)00189-6. URL https://www.sciencedirect.com/science/article/pii/S0020751997001896.

H.-P. Pham, M. Manuel, N. Petit, D. Klatzmann, S. Cohen-Kaminsky, A. Six, and G. Marodon. Half of the T-cell repertoire combinatorial diversity is genetically determined in humans and humanized mice. *European Journal of Immunology*, 42(3):760–770, 2012. ISSN 1521-4141. doi: 10.1002/eji.201141798. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/eji.201141798. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/eji.201141798.

J. Pido-Lopez, N. Imami, and R. Aspinall. Both age and gender affect thymic output: more recent thymic migrants in females than males as they age. *Clinical & Experimental Immunology*, 125(3):409–413, 2001. ISSN 1365-2249. doi: 10.1046/j.1365-2249.2001.01640.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2249.2001.01640.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2249.2001.01640.x.

A. Piecyk, M. Ritter, and M. Kalbe. The right response at the right time: Exploring helminth immune modulation in sticklebacks by experimental coinfection. *Molecular Ecology*, 28(10):2668–2680, 2019a. ISSN 1365-294X. doi: 10.1111/mec.15106. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.15106. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.15106.

A. Piecyk, O. Roth, and M. Kalbe. Specificity of resistance and geographic patterns of virulence in a vertebrate host-parasite system. *BMC Evolutionary Biology*, 19(1):80, Mar. 2019b. ISSN 1471-2148. doi: 10.1186/s12862-019-1406-3. URL https://doi.org/10.1186/s12862-019-1406-3.

A. Piecyk, M. A. Hahn, O. Roth, N. M. Dheilly, D. C. Heins, M. A. Bell, and M. Kalbe. Cross-continental experimental infections reveal distinct defence mechanisms in populations of the three-spined stickleback *Gasterosteus aculeatus*. *Proceedings of the Royal Society B: Biological Sciences*, 288(1959):20211758, Sept. 2021. ISSN 0962-8452, 1471-2954. doi: 10.1098/rspb.2021.1758. URL https://royalsocietypublishing.org/doi/10.1098/rspb.2021.1758.

S. B. Piertney and M. K. Oliver. The evolutionary ecology of the major histocompatibility complex. *Heredity*, 96(1):7–21, Jan. 2006. ISSN 1365-2540. doi: 10.1038/sj.hdy.6800724. URL https://www.nature.com/articles/6800724. Number: 1 Publisher: Nature Publishing Group.

M. V. Pogorelyy, Y. Elhanati, Q. Marcou, A. L. Sycheva, E. A. Komech, V. I. Nazarov, O. V. Britanova, D. M. Chudakov, I. Z. Mamedov, Y. B. Lebedev, T. Mora, and A. M. Walczak. Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires. *PLOS Computational Biology*, 13(7):e1005572, July 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005572. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005572. Publisher: Public Library of Science.

J. C. F. Poole and A. J. Holladay. Thucydides and the Plague of Athens. *The Classical Quarterly*, 29(2):282–300, Dec. 1979. ISSN 1471-6844, 0009-8388. doi: 10.1017/S0009838800035928. URL https://www.cambridge.org/core/journals/classical-quarterly/article/abs/thucydides-and-the-plague-of-athens/95602E6BCFACB32F3832B965091C0CFE.

P. H. Pressley. Parental Effort and the Evolution of Nest-Guarding Tactics in the Threespine Stickleback, Gasterosteus aculeatus L. *Evolution*, 35(2):282–295, 1981. ISSN 0014-3820. doi: 10.2307/2407838. URL https://www.jstor.org/stable/2407838. Publisher: [Society for the Study of Evolution, Wiley].

Q. Qi, Y. Liu, Y. Cheng, J. Glanville, D. Zhang, J.-Y. Lee, R. A. Olshen, C. M. Weyand, S. D. Boyd, and J. J. Goronzy. Diversity and clonal selection in the human T-cell repertoire. *Proceedings of the National Academy of Sciences*, 111(36):13139–13144, Sept. 2014. doi: 10.1073/pnas.1409155111. URL https://www.pnas.org/doi/full/10.1073/pnas.1409155111. Publisher: Proceedings of the National Academy of Sciences.

M. F. Quigley, H. Y. Greenaway, V. Venturi, R. Lindsay, K. M. Quinn, R. A. Seder, D. C. Douek, M. P. Davenport, and D. A. Price. Convergent recombination shapes the clonotypic landscape of the naïve T-cell repertoire. *Proceedings of the National Academy of Sciences*, 107(45):19414–19419, Nov. 2010. doi: 10.1073/pnas.1010586107. URL https://www.pnas.org/doi/10.1073/pnas.1010586107. Publisher: Proceedings of the National Academy of Sciences.

J. Radwan, W. Babik, J. Kaufman, T. L. Lenz, and J. Winternitz. Advances in the Evolutionary Understanding of MHC Polymorphism. *Trends in Genetics*, 36(4):298–311, Apr. 2020. ISSN 0168-9525. doi: 10.1016/j.tig.2020.01.008. URL https://www.sciencedirect.com/science/article/pii/S0168952520300214.

G. Rauch, . M. Kalbe, and T. B. H. Reusch‡. Relative importance of MHC and genetic background for parasite load in a field experiment. *Evolutionary Ecology Research*, 8(2):373–386, 2006. ISSN 1522-0613. URL https://www.evolutionary-ecology.com/abstracts/v08/1930.html. Publisher: Evolutionary Ecology, Ltd.

T. E. Reimchen and P. Nosil. Ecological causes of sex-biased parasitism in threespine stickleback. *Biological Journal of the Linnean Society*, 73(1):51–63, May 2001. ISSN 0024-4066. doi: 10.1111/j.1095-8312.2001.tb01346.x. URL https://doi.org/10.1111/j.1095-8312.2001.tb01346.x.

S. L. Rekdal, J. A. Anmarkrud, A. Johnsen, and J. T. Lifjeld. Genotyping strategy matters when analyzing hypervariable major histocompatibility complex-Experience from a passerine bird. *Ecology and Evolution*, 8(3):1680–1692, 2018. ISSN 2045-7758. doi: 10.1002/ece3.3757. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.3757. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ece3.3757.

T. Resende Co, C. S. Hirsch, Z. Toossi, R. Dietze, and R. Ribeiro-Rodrigues. Intestinal helminth co-infection has a negative impact on both anti-Mycobacterium tuberculosis immunity and clinical response to tuberculosis therapy. *Clinical and Experimental Immunology*, 147(1):45–52, Jan. 2007. ISSN 0009-9104. doi: 10.1111/j.1365-2249.2006.03247.x. URL https://doi.org/10.1111/j.1365-2249.2006.03247.x.

T. B. Reusch and Langefors. Inter- and Intralocus Recombination Drive MHC Class IIB Gene Diversification in a Teleost, the Three-Spined Stickleback Gasterosteus aculeatus. *Journal of Molecular Evolution*, 61(4):531–541, Oct. 2005. ISSN 1432-1432. doi: 10.1007/s00239-004-0340-0. URL https://doi.org/10.1007/s00239-004-0340-0.

T. B. H. Reusch, K. M. Wegner, and M. Kalbe. Rapid genetic divergence in post-glacial populations of threespine stickleback (Gasterosteus aculeatus): the role of habitat type, drainage and geographical proximity. *Molecular Ecology*, 10(10):2435–2445, 2001. ISSN 1365-294X. doi: 10.1046/j.0962-1083.2001.01366.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1046/j.0962-1083.2001.01366.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.0962-1083.2001.01366.x.

T. B. H. Reusch, H. Schaschl, and K. M. Wegner. Recent duplication and inter-locus gene conversion in major histocompatibility class II genes in a teleost, the three-spined stickleback. *Immunogenetics*, 56(6):427–437, Sept. 2004. ISSN 1432-1211. doi: 10.1007/s00251-004-0704-z. URL https://doi.org/10.1007/s00251-004-0704-z.

M. L. Reyes and J. A. Baker. The consequences of diet limitation in juvenile threespine stickleback: growth, lipid storage and the phenomenon of compensatory growth. *Ecology of Freshwater Fish*, 26(2):301–312, 2017. ISSN 1600-0633. doi: 10.1111/eff.12276. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/eff.12276. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/eff.12276.

J. Robert and E.-S. Edholm. A prominent role for invariant T cells in the amphibian Xenopus laevis tadpoles. *Immunogenetics*, 66(9-10):513–523, Oct. 2014. ISSN 1432-1211. doi: 10.1007/s00251-014-0781-6.

S. Robertson, J. E. Bradley, and A. D. C. MacColl. Measuring the immune system of the three-spined stickleback – investigating natural variation by quantifying immune expression in the laboratory and the wild. *Molecular Ecology Resources*, 16(3):701–713, 2016. ISSN 1755-0998. doi: 10.1111/1755-0998.12497. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12497. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.12497.

H. S. Robins, P. V. Campregher, S. K. Srivastava, A. Wacher, C. J. Turtle, O. Kahsai, S. R. Riddell, E. H. Warren, and C. S. Carlson. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood*, 114(19):4099–4107, Nov. 2009. ISSN 1528-0020. doi: 10.1182/blood-2009-04-217604.

H. S. Robins, S. K. Srivastava, P. V. Campregher, C. J. Turtle, J. Andriesen, S. R. Riddell, C. S. Carlson, and E. H. Warren. Overlap and Effective Size of the Human CD8+ T Cell Receptor Repertoire. *Science Translational Medicine*, 2(47):47ra64–47ra64, Sept. 2010. doi: 10.1126/scitranslmed.3001442. URL https://www.science.org/doi/abs/10.1126/scitranslmed.3001442. Publisher: American Association for the Advancement of Science.

J. T. Robinson, H. Thorvaldsdóttir, A. M. Wenger, A. Zehir, and J. P. Mesirov. Variant Review with the Integrative Genomics Viewer. *Cancer Research*, 77(21):e31–e34, Oct. 2017. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-17-0337. URL https://doi.org/10.1158/0008-5472.CAN-17-0337.

O. L. Rodriguez, C. A. Silver, K. Shields, M. L. Smith, and C. T. Watson. Targeted long-read sequencing facilitates phased diploid assembly and genotyping of the human T cell receptor alpha, delta, and beta loci. *Cell Genomics*, 2(12):100228, Dec. 2022. ISSN 2666-979X. doi: 10.1016/j.xgen.2022.100228.

S. Rodriguez-Martinez, J. Klaminder, M. A. Morlock, L. Dalén, and D. Y.-T. Huang. The topological nature of tag jumping in environmental DNA metabarcoding studies. *Molecular Ecology Resources*, n/a(n/a), Nov. 2022. ISSN 1755-0998. doi: 10.1111/1755-0998.13745. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13745. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13745.

J. Rolff. Bateman's principle and immunity. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 269(1493):867–872, Apr. 2002. doi: 10.1098/rspb.2002.1959. URL https://royalsocietypublishing.org/doi/10.1098/rspb.2002.1959. Publisher: Royal Society.

F. Ronquist and J. P. Huelsenbeck. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574, Aug. 2003. ISSN 1367-4803. doi: 10.1093/bioinformatics/btg180. URL https://doi.org/10.1093/bioinformatics/btg180.

G. A. W. Rook, R. Martinelli, and L. R. Brunet. Innate immune responses to mycobacteria and the downregulation of atopic responses. *Current Opinion in Allergy and Clinical Immunology*, 3(5):337–342, Oct. 2003. ISSN 1528-4050. doi: 10.1097/00130832-200310000-00003.

E. Rosati, C. M. Dowds, E. Liaskou, E. K. K. Henriksen, T. H. Karlsen, and A. Franke. Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnology*, 17(1):61, July 2017. ISSN 1472-6750. doi: 10.1186/s12896-017-0379-9. URL https://doi.org/10.1186/s12896-017-0379-9.

D. B. Roth. V(D)J Recombination: Mechanism, Errors, and Fidelity. *Microbiology spectrum*, 2(6): 10.1128/microbiolspec.MDNA3, Dec. 2014. doi: 10.1128/microbiolspec.MDNA3-0041-2014. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5089068/.

O. Roth, J. P. Scharsack, I. Keller, and T. B. H. Reusch. Bateman's principle and immunity in a sex-role reversed pipefish. *Journal of Evolutionary Biology*, 24(7):1410–1420, July 2011. ISSN 1010-061X. doi: 10.1111/j.1420-9101.2011.02273.x. URL https://doi.org/10.1111/j.1420-9101.2011.02273.x.

J. Roved, B. Hansson, M. Stervander, D. Hasselquist, and H. Westerdahl. MHCtools – an R package for MHC high-throughput sequencing data: Genotyping, haplotype and supertype inference, and downstream genetic analyses in non-model organisms. *Molecular Ecology Resources*, 22(7):2775–2792, 2022. ISSN 1755-0998. doi: 10.1111/1755-0998.13645. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13645. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13645.

J. Rozas, A. Ferrer-Mata, J. C. Sánchez-DelBarrio, S. Guirao-Rico, P. Librado, S. E. Ramos-Onsins, and A. Sánchez-Gracia. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Molecular Biology and Evolution*, 34(12):3299–3302, Dec. 2017. ISSN 0737-4038. doi: 10.1093/molbev/msx248. URL https://doi.org/10.1093/molbev/msx248.

M. L. Russell, A. Souquette, D. M. Levine, S. A. Schattgen, E. K. Allen, G. Kuan, N. Simon, A. Balmaseda, A. Gordon, P. G. Thomas, F. A. Matsen, IV, and P. Bradley. Combining genotypes and T cell receptor distributions to infer genetic loci determining V(D)J recombination probabilities. *eLife*, 11:e73475, Mar. 2022. ISSN 2050-084X. doi: 10.7554/eLife.73475. URL https://doi.org/10.7554/eLife.73475. Publisher: eLife Sciences Publications, Ltd.

K. E. J. Rödström, P. Regenthal, and K. Lindkvist-Petersson. Structure of Staphylococcal Enterotoxin E in Complex with TCR Defines the Role of TCR Loop Positioning in Superantigen Recognition. *PLOS ONE*, 10(7):e0131988, July 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0131988. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0131988. Publisher: Public Library of Science.

A. Sato, F. Figueroa, C. O'hUigin, N. Steck, and J. Klein. Cloning of major histocompatibility complex (MHC) genes from threespine stickleback. *Molecular marine biology and biotechnology*, 7:221–31, Oct. 1998.

G. Scapigliati, A. M. Fausto, and S. Picchietti. Fish Lymphocytes: An Evolutionary Equivalent of Mammalian Innate-Like Lymphocytes? *Frontiers in Immunology*, 9:971, 2018. ISSN 1664-3224. doi: 10.3389/fimmu.2018.00971.

J. P. Scharsack and F. Franke. Temperature effects on teleost immunity in the light of climate change. *Journal of Fish Biology*, 101(4):780–796, 2022. ISSN 1095-8649. doi: 10.1111/jfb.15163. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jfb.15163. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jfb.15163.

J. P. Scharsack and M. Kalbe. Differences in susceptibility and immune responses of three-spined sticklebacks (Gasterosteus aculeatus) from lake and river ecotypes to sequential infections with the eye fluke Diplostomum pseudospathaceum. *Parasites & Vectors*, 7:109, Mar. 2014. ISSN 1756-3305. doi: 10.1186/1756-3305-7-109.

J. P. Scharsack, M. Kalbe, R. Derner, J. Kurtz, and M. Milinski. Modulation of granulocyte responses in three-spined sticklebacks Gasterosteus aculeatus infected with the tapeworm Schistocephalus solidus. *Diseases of Aquatic Organisms*, 59(2):141–150, May 2004. ISSN 0177-5103. doi: 10.3354/dao059141.

J. P. Scharsack, M. Kalbe, C. Harrod, and G. Rauch. Habitat-specific adaptation of immune responses of stickleback (Gasterosteus aculeatus) lake and river ecotypes. *Proceedings of the Royal Society B: Biological Sciences*, 274(1617):1523, Apr. 2007a. doi: 10.1098/rspb.2007.0210. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC2176159/.

J. P. Scharsack, K. Koch, and K. Hammerschmidt. Who is in control of the stickleback immune system: interactions between Schistocephalus solidus and its specific vertebrate host. *Proceedings of the Royal Society B: Biological Sciences*, 274(1629):3151–3158, Oct. 2007b. doi: 10.1098/rspb.2007.1148. URL https://royalsocietypublishing.org/doi/full/10.1098/rspb.2007.1148. Publisher: Royal Society.

J. P. Scharsack, A. Gossens, F. Franke, and J. Kurtz. Excretory products of the cestode, Schistocephalus solidus, modulate in vitro responses of leukocytes from its specific host, the three-spined stickleback (Gasterosteus aculeatus). *Fish & Shellfish Immunology*, 35 (6):1779–1787, Dec. 2013. ISSN 10504648. doi: 10.1016/j.fsi.2013.08.029. URL https://linkinghub.elsevier.com/retrieve/pii/S1050464813007432.

J. P. Scharsack, B. Wieczorek, A. Schmidt-Drewello, J. Büscher, F. Franke, A. Moore, A. Branca, A. Witten, M. Stoll, E. Bornberg-Bauer, S. Wicke, and J. Kurtz. Climate change facilitates a parasite's host exploitation via temperature-mediated immunometabolic processes. *Global Change Biology*, 27(1):94–107, Jan. 2021. ISSN 1365-2486. doi: 10.1111/gcb.15402.

H. Schaschl and K. M. Wegner. Polymorphism and signature of selection in the MHC class I genes of the three-spined stickleback Gasterosteus aculeatus. *Journal of Fish Biology*, 69 (sb):177–188, 2006. ISSN 1095-8649. doi: 10.1111/j.1095-8649.2006.01223.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1095-8649.2006.01223.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1095-8649.2006.01223.x.

H. Schaschl and K. M. Wegner. Contrasting mode of evolution between the MHC class I genomic region and class II region in the three-spined stickleback (Gasterosteus aculeatus L.; Gasterosteidae: Teleostei). *Immunogenetics*, 59(4):295–304, Apr. 2007. ISSN 1432-1211. doi: 10.1007/s00251-007-0192-z. URL https://doi.org/10.1007/s00251-007-0192-z.

D. G. Schatz and P. C. Swanson. V(D)J recombination: mechanisms of initiation. *Annual Review of Genetics*, 45:167–202, 2011. ISSN 1545-2948. doi: 10.1146/annurev-genet-110410-132552.

M. Schirmer, U. Z. Ijaz, R. D'Amore, N. Hall, W. T. Sloan, and C. Quince. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, 43(6):e37, Mar. 2015. ISSN 0305-1048. doi: 10.1093/nar/gku1341. URL https://doi.org/10.1093/nar/gku1341.

M. Schirmer, R. D'Amore, U. Z. Ijaz, N. Hall, and C. Quince. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*, 17(1):125, Mar. 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-0976-y. URL https://doi.org/10.1186/s12859-016-0976-y.

T. Schneider-Hohendorf, D. Görlich, P. Savola, T. Kelkka, S. Mustjoki, C. C. Gross, G. C. Owens, L. Klotz, K. Dornmair, H. Wiendl, and N. Schwab. Sex bias in MHC I-associated shaping of the adaptive immune system. *Proceedings of the National Academy of Sciences*, 115(9):2168–2173, Feb. 2018. doi: 10.1073/pnas.1716146115. URL https://www.pnas.org/doi/abs/10.1073/pnas.1716146115. Publisher: Proceedings of the National Academy of Sciences.

I. B. Schnell, K. Bohmann, and M. T. P. Gilbert. Tag jumps illuminated – reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*, 15(6):1289–1303, 2015. ISSN 1755-0998. doi: 10.1111/1755-0998.12402. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12402. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.12402.

H. Schulenburg, J. Kurtz, Y. Moret, and M. T. Siva-Jothy. Introduction. Ecological immunology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1513):3–14, Jan. 2009. ISSN 0962-8436. doi: 10.1098/rstb.2008.0249. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2666701/.

A. Sebastian, M. Herdegen, M. Migalska, and J. Radwan. Amplisas: a web server for multilocus genotyping using next-generation amplicon sequencing data. *Molecular Ecology Resources*, 16(2):498–510, 2016. ISSN 1755-0998. doi: 10.1111/1755-0998.12453. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12453. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.12453.

H. Segner, B. M. L. Verburg-van Kemenade, and M. Chadzinska. The immunomodulatory role of the hypothalamus-pituitary-gonad axis: Proximate mechanism for reproduction-immune trade offs? *Developmental & Comparative Immunology*, 66:43–60, Jan. 2017. ISSN 0145-305X. doi: 10.1016/j.dci.2016.07.004. URL https://www.sciencedirect.com/science/article/pii/S0145305X1630221X.

R. Sender, Y. Weiss, Y. Navon, I. Milo, N. Azulay, L. Keren, S. Fuchs, D. Ben-Zvi, E. Noor, and R. Milo. The total mass, number, and distribution of immune cells in the human body. *Proceedings of the National Academy of Sciences*, 120(44):e2308511120, Oct. 2023. doi: 10.1073/pnas.2308511120. URL https://www.pnas.org/doi/10.1073/pnas.2308511120. Publisher: Proceedings of the National Academy of Sciences.

Z. Sethna, Y. Elhanati, C. G. Callan, A. M. Walczak, and T. Mora. OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. *Bioinformatics*, 35(17):2974–2981, Sept. 2019. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btz035. URL https://academic.oup.com/bioinformatics/article/35/17/2974/5292315.

Z. Sethna, G. Isacchini, T. Dupic, T. Mora, A. M. Walczak, and Y. Elhanati. Population variability in the generation and selection of T-cell repertoires. *PLOS Computational Biology*, 16(12):e1008394, Dec. 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008394. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008394. Publisher: Public Library of Science.

P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504, Nov. 2003. ISSN 1088-9051. doi: 10.1101/gr.1239303. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC403769/.

E. Sharon, L. V. Sibener, A. Battle, H. B. Fraser, K. C. Garcia, and J. K. Pritchard. Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nature Genetics*, 48(9):995–1002, Sept. 2016. ISSN 1546-1718. doi: 10.1038/ng.3625. URL https://www.nature.com/articles/ng.3625. Publisher: Nature Publishing Group.

Y. Shi, A. Strasser, D. R. Green, E. Latz, A. Mantovani, and G. Melino. Legacy of the discovery of the T-cell receptor: 40 years of shaping basic immunology and translational work to develop novel therapies. *Cellular & Molecular Immunology*, 21(7):790–797, July 2024. ISSN 2042-0226. doi: 10.1038/s41423-024-01168-4. URL https://www.nature.com/articles/s41423-024-01168-4. Publisher: Nature Publishing Group.

M. Shugay, D. A. Bolotin, E. V. Putintseva, M. V. Pogorelyy, I. Z. Mamedov, and D. M. Chudakov. Huge Overlap of Individual TCR Beta Repertoires. *Frontiers in Immunology*, 4, Dec. 2013. ISSN 1664-3224. doi: 10.3389/fimmu.2013.00466. URL https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2013.00466/full. Publisher: Frontiers.

M. Shugay, D. V. Bagaev, M. A. Turchaninova, D. A. Bolotin, O. V. Britanova, E. V. Putintseva, M. V. Pogorelyy, V. I. Nazarov, I. V. Zvyagin, V. I. Kirgizova, K. I. Kirgizov, E. V. Skorobogatova, and D. M. Chudakov. VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. *PLOS Computational Biology*, 11(11):e1004503, Nov. 2015. ISSN 1553-7358.

doi: 10.1371/journal.pcbi.1004503. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004503. Publisher: Public Library of Science.

A. K. Simon, G. A. Hollander, and A. McMichael. Evolution of the immune system in humans from infancy to old age. *Proceedings of the Royal Society B: Biological Sciences*, 282(1821):20143085, Dec. 2015. doi: 10.1098/rspb.2014.3085. URL https://royalsocietypublishing.org/doi/10.1098/rspb.2014.3085. Publisher: Royal Society.

A. Six, E. Mariotti-Ferrandiz, W. Chaara, S. Magadan, H.-P. Pham, M.-P. Lefranc, T. Mora, V. Thomas-Vaslin, A. M. Walczak, and P. Boudinot. The Past, Present, and Future of Immune Repertoire Biology – The Rise of Next-Generation Repertoire Analysis. *Frontiers in Immunology*, 4, Nov. 2013. ISSN 1664-3224. doi: 10.3389/fimmu.2013.00413. URL https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2013.00413/full. Publisher: Frontiers.

N. C. Smith, M. L. Rise, and S. L. Christian. A Comparison of the Innate and Adaptive Immune Systems in Cartilaginous Fish, Ray-Finned Fish, and Lobe-Finned Fish. *Frontiers in Immunology*, 10, 2019. ISSN 1664-3224. URL https://www.frontiersin.org/articles/10.3389/fimmu.2019.02292.

J. D. Smyth. Studies on tapeworm physiology, the cultivation of Schistocephalus solidus in vitro. *The Journal of Experimental Biology*, 23(1):47–70, Oct. 1946. ISSN 0022-0949. doi: 10.1242/jeb.23.1.47.

J. D. Smyth. Studies on tapeworm physiology. VII. Fertilization of *Schistocephalus solidus in vitro*. *Experimental Parasitology*, 3(1):64–71, Jan. 1954. ISSN 0014-4894. doi: 10.1016/0014-4894(54)90019-3. URL https://www.sciencedirect.com/science/article/pii/0014489454900193.

S. Sommer, A. Courtiol, and C. J. Mazzoni. MHC genotyping of non-model organisms using next-generation sequencing: a new methodology to deal with artefacts and allelic dropout. *BMC Genomics*, 14(1):542, Aug. 2013. ISSN 1471-2164. doi: 10.1186/1471-2164-14-542. URL https://doi.org/10.1186/1471-2164-14-542.

L. G. Spurgin and D. S. Richardson. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proceedings. Biological Sciences*, 277(1684):979–988, Apr. 2010. ISSN 1471-2954. doi: 10.1098/rspb.2009.2084.

B. Stephen, J. Hajjar, S. Sarda, D. Y. Duose, J. M. Conroy, C. Morrison, A. Alshawa, M. Xu, A. Zarifa, S. P. Patel, Y. Yuan, E. Kwiatkowski, L. Wang, J. Rodon Ahnert, S. Fu, F. Meric-Bernstam, G. M. Lowman, T. Looney, and A. Naing. T-cell receptor beta variable gene polymorphism predicts immune-related adverse events during checkpoint blockade immunotherapy. *Journal for Immunotherapy of Cancer*, 11(8):e007236, Aug. 2023. ISSN 2051-1426. doi: 10.1136/jitc-2023-007236.

L. T. Stiemsma, L. A. Reynolds, S. E. Turvey, and B. B. Finlay. The hygiene hypothesis: current perspectives and future therapies. *Immunotargets and Therapy*, 4:143–157, July 2015. ISSN 2253-1556. doi: 10.2147/ITT.S61528. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4918254/.

N. Stoler and A. Nekrutenko. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics and Bioinformatics*, 3(1):lqab019, Mar. 2021. ISSN 2631-9268. doi: 10.1093/nargab/lqab019. URL https://doi.org/10.1093/nargab/lqab019.

W. E. Stutz and D. I. Bolnick. Stepwise threshold clustering: a new method for genotyping MHC loci using next-generation sequencing technology. *PloS One*, 9(7):e100587, 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0100587.

C. Su and M. Nei. Evolutionary Dynamics of the T-Cell Receptor VB Gene Family as Inferred from the Human and Mouse Genomic Sequences. *Molecular Biology and Evolution*, 18(4): 503–513, Apr. 2001. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a003829. URL https://doi.org/10.1093/oxfordjournals.molbev.a003829.

L. Subrahmanyan, M. A. Eberle, A. G. Clark, L. Kruglyak, and D. A. Nickerson. Sequence Variation and Linkage Disequilibrium in the Human T-Cell Receptor (TCRB) Locus. *American Journal of Human Genetics*, 69(2):381, June 2001. doi: 10.1086/321297. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC1235310/.

J. Sun, R. Ruiz Daniels, A. Balic, A. M. S. Andresen, H. Bjørgen, R. Dobie, N. C. Henderson, E. O. Koppang, S. A. M. Martin, J. H. Fosse, R. S. Taylor, and D. J. Macqueen. Cell atlas of the Atlantic salmon spleen reveals immune cell heterogeneity and cell-specific responses to bacterial infection. *Fish & Shellfish Immunology*, 145:109358, Feb. 2024. ISSN 1050-4648. doi: 10.1016/j.fsi.2024.109358. URL https://www.sciencedirect.com/science/article/pii/S1050464824000020.

L. Sun, Y. Su, A. Jiao, X. Wang, and B. Zhang. T cells in health and disease. *Signal Transduction and Targeted Therapy*, 8(1):1–50, June 2023. ISSN 2059-3635. doi: 10.1038/s41392-023-01471-y. URL https://www.nature.com/articles/s41392-023-01471-y. Publisher: Nature Publishing Group.

X. Sun, T. Nguyen, A. Achour, A. Ko, J. Cifello, C. Ling, J. Sharma, T. Hiroi, Y. Zhang, C. W. Chia, W. W. III, W. W. Wu, L. Zukley, J.-N. Phue, K. G. Becker, R.-F. Shen, L. Ferrucci, and N.-p. Weng. Longitudinal analysis reveals age-related changes in the T cell receptor repertoire of human T cell subsets, Sept. 2022. URL https://www.jci.org/articles/view/158122/pdf. Publisher: American Society for Clinical Investigation.

J. O. Sunyer. Fishing for mammalian paradigms in the teleost immune system. *Nature Immunology*, 14(4):320–326, Apr. 2013. ISSN 1529-2916. doi: 10.1038/ni.2549. URL https://www.nature.com/articles/ni.2549. Publisher: Nature Publishing Group.

K. Takada and S. C. Jameson. Naive T cell homeostasis: from awareness of space to a sense of place. *Nature Reviews Immunology*, 9(12):823–832, Dec. 2009. ISSN 1474-1741. doi: 10.1038/nri2657. URL https://www.nature.com/articles/nri2657. Publisher: Nature Publishing Group.

M. D. Taylor, N. v. d. Werf, and R. M. Maizels. T cells in helminth infection: the regulators and the regulated. *Trends in Immunology*, 33(4):181–189, Apr. 2012. ISSN 1471-4906, 1471-4981. doi: 10.1016/j.it.2012.01.001. URL https://www.cell.com/trends/immunology/abstract/S1471-4906(12)00002-6. Publisher: Elsevier.

R. C. Team. R: A Language and Environment for Statistical Computing, 2021. URL https://www.R-project.org/.

S. Telfer, X. Lambin, R. Birtles, P. Beldomenico, S. Burthe, S. Paterson, and M. Begon. Species interactions in a parasite community drive infection risk in a wildlife population. *Science (New York, N.Y.)*, 330(6001):243–246, Oct. 2010. ISSN 1095-9203. doi: 10.1126/science.1190333.

N. Thomas, K. Best, M. Cinelli, S. Reich-Zeliger, H. Gal, E. Shifrut, A. Madi, N. Friedman, J. Shawe-Taylor, and B. Chain. Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. *Bioinformatics*, 30(22):3181–3188, Nov. 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu523. URL https://doi.org/10.1093/bioinformatics/btu523.

D.-M. J. Thorburn, K. Sagonas, M. Binzer-Panchal, F. J. J. Chain, P. G. D. Feulner, E. Bornberg-Bauer, T. B. H. Reusch, I. E. Samonte-Padilla, M. Milinski, T. L. Lenz, and C. Eizaguirre. Origin matters: Using a local reference genome improves measures in population genomics. *Molecular Ecology Resources*, 23(7):1706–1723, 2023. ISSN 1755-0998. doi: 10.1111/1755-0998.13838. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13838. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13838.

J. F. Tierney and D. W. Crompton. Infectivity of plerocercoids of Schistocephalus solidus (Cestoda: Ligulidae) and fecundity of the adults in an experimental definitive host, Gallus gallus. *The Journal of Parasitology*, 78(6):1049–1054, Dec. 1992. ISSN 0022-3395.

A. Toubal, I. Nel, S. Lotersztajn, and A. Lehuen. Mucosal-associated invariant T cells and disease. *Nature Reviews Immunology*, 19(10):643–657, Oct. 2019. ISSN 1474-1741. doi: 10.1038/s41577-019-0191-y. URL https://www.nature.com/articles/s41577-019-0191-y. Publisher: Nature Publishing Group.

A. Trigunaite, J. Dimo, and T. N. Jørgensen. Suppressive effects of androgens on the immune system. *Cellular Immunology*, 294(2):87–94, Apr. 2015. ISSN 0008-8749. doi: 10.1016/j.cellimm.2015.02.004. URL https://www.sciencedirect.com/science/article/pii/S0008874915000313.

A. Trofimov, P. Brouillard, J.-D. Larouche, J. Séguin, J.-P. Laverdure, A. Brasey, G. Ehx, D.-C. Roy, L. Busque, S. Lachance, S. Lemieux, and C. Perreault. Two types of human TCR differentially regulate reactivity to self and non-self antigens. *iScience*, 25(9):104968, Sept. 2022. ISSN 2589-0042. doi: 10.1016/j.isci.2022.104968. URL https://www.sciencedirect.com/science/article/pii/S2589004222012408.

A. Troshchynsky, I. Dzneladze, L. Chen, Y. Sheng, V. Saridakis, and G. E. Wu. Functional analyses of polymorphic variants of human terminal deoxynucleotidyl transferase. *Genes and Immunity*, 16(6):388–398, Sept. 2015. ISSN 1476-5470. doi: 10.1038/gene.2015.19.

J. Trowsdale. The MHC, disease and selection. *Immunology Letters*, 137(1):1–8, June 2011. ISSN 0165-2478. doi: 10.1016/j.imlet.2011.01.002. URL https://www.sciencedirect.com/science/article/pii/S0165247811000071.

S. J. Turner, N. L. L. Gruta, K. Kedzierska, P. G. Thomas, and P. C. Doherty. Functional implications of T cell receptor diversity. *Current opinion in immunology*, 21(3):286, June 2009. doi: 10.1016/j.coi.2009.05.004. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC2706259/.

P. van den Elsen. Expression Regulation of Major Histocompatibility Complex Class I and Class II Encoding Genes. *Frontiers in Immunology*, 2, 2011. ISSN 1664-3224. URL https://www.frontiersin.org/articles/10.3389/fimmu.2011.00048.

T. van der Valk, F. Vezzi, M. Ormestad, L. Dalén, and K. Guschanski. Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. *Molecular Ecology Resources*, 20(5):1171–1181, 2020. ISSN 1755-0998. doi: 10.1111/1755-0998.13009. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13009. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13009.

I. Van Rhijn and D. B. Moody. Donor Unrestricted T Cells: A Shared Human T Cell Response. *The Journal of Immunology*, 195(5):1927–1932, Sept. 2015. ISSN 0022-1767. doi: 10.4049/jimmunol.1500943. URL https://doi.org/10.4049/jimmunol.1500943.

E. van Riet, F. C. Hartgers, and M. Yazdanbakhsh. Chronic helminth infections induce immunomodulation: Consequences and mechanisms. *Immunobiology*, 212(6):475–490, June 2007. ISSN 0171-2985. doi: 10.1016/j.imbio.2007.03.009. URL https://www.sciencedirect.com/science/article/pii/S0171298507000307.

C. Vandiedonck, M. S. Taylor, H. E. Lockstone, K. Plant, J. M. Taylor, C. Durrant, J. Broxholme, B. P. Fairfax, and J. C. Knight. Pervasive haplotypic variation in the spliceo-transcriptome of the human major histocompatibility complex. *Genome Research*, 21(7):1042–1054, July 2011. ISSN 1088-9051. doi: 10.1101/gr.116681.110. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.116681.110.

V. Venturi, K. Kedzierska, D. A. Price, P. C. Doherty, D. C. Douek, S. J. Turner, and M. P. Davenport. Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proceedings of the National Academy of Sciences of the United States of America*, 103(49):18691–18696, Dec. 2006. ISSN 0027-8424. doi: 10.1073/pnas.0608907103. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1693724/.

V. Venturi, H. Y. Chin, D. A. Price, D. C. Douek, and M. P. Davenport. The role of production frequency in the sharing of simian immunodeficiency virus-specific CD8+ TCRs between macaques. *Journal of Immunology (Baltimore, Md.: 1950)*, 181(4):2597–2609, Aug. 2008a. ISSN 1550-6606. doi: 10.4049/jimmunol.181.4.2597.

V. Venturi, D. A. Price, D. C. Douek, and M. P. Davenport. The molecular basis for public T-cell responses? *Nature Reviews Immunology*, 8(3):231–238, Mar. 2008b. ISSN 1474-1741. doi: 10.1038/nri2260. URL https://www.nature.com/articles/nri2260. Number: 3 Publisher: Nature Publishing Group.

V. Venturi, M. F. Quigley, H. Y. Greenaway, P. C. Ng, Z. S. Ende, T. McIntosh, T. E. Asher, J. R. Almeida, S. Levy, D. A. Price, M. P. Davenport, and D. C. Douek. A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. *Journal of Immunology (Baltimore, Md.: 1950)*, 186(7):4285–4294, Apr. 2011. ISSN 1550-6606. doi: 10.4049/jimmunol.1003898.

V. Venturi, K. Nzingha, T. G. Amos, W. C. Charles, I. Dekhtiarenko, L. Cicin-Sain, M. P. Davenport, and B. D. Rudd. The Neonatal CD8+ T Cell Repertoire Rapidly Diversifies during Persistent Viral Infection. *The Journal of Immunology*, 196(4):1604–1616, Feb. 2016. ISSN 0022-1767. doi: 10.4049/jimmunol.1501867. URL https://doi.org/10.4049/jimmunol.1501867.

N. Vrisekoop, J. P. Monteiro, J. N. Mandl, and R. N. Germain. Revisiting Thymic Positive Selection and the Mature T Cell Repertoire for Antigen. *Immunity*, 41(2):181–190, Aug. 2014. ISSN 1074-7613. doi: 10.1016/j.immuni.2014.07.007. URL https://www.cell.com/immunity/abstract/S1074-7613(14)00264-7. Publisher: Elsevier.

H.-J. Wallny, D. Avila, L. G. Hunt, T. J. Powell, P. Riegert, J. Salomonsen, K. Skjødt, O. Vainio, F. Vilbois, M. V. Wiles, and J. Kaufman. Peptide motifs of the single dominantly expressed class I molecule explain the striking MHC-determined response to Rous sarcoma virus in chickens. *Proceedings of the National Academy of Sciences*, 103(5):1434–1439, Jan. 2006. doi: 10.1073/pnas.0507386103. URL https://www.pnas.org/doi/abs/10.1073/pnas.0507386103. Publisher: Proceedings of the National Academy of Sciences.

C. T. Watson, F. A. Matsen, K. J. L. Jackson, A. Bashir, M. L. Smith, J. Glanville, F. Breden, S. H. Kleinstein, A. M. Collins, and C. E. Busse. Comment on "A Database of Human Immune Receptor Alleles Recovered from Population Sequencing Data". *Journal of Immunology (Baltimore, Md.: 1950)*, 198(9):3371–3373, May 2017. ISSN 1550-6606. doi: 10.4049/jimmunol.1700306.

C. T. Watson, J. T. Kos, W. S. Gibson, L. Newman, G. Deikus, C. E. Busse, M. L. Smith, K. J. Jackson, and A. M. Collins. A comparison of immunoglobulin IGHV, IGHD and IGHJ genes in wild-derived and classical inbred mouse strains. *Immunology & Cell Biology*, 97(10):888–901, 2019. ISSN 1440-1711. doi: 10.1111/imcb.12288. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/imcb.12288. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/imcb.12288.

J. N. Weber, M. Kalbe, K. C. Shim, N. I. Erin, N. C. Steinel, L. Ma, and D. I. Bolnick. Resist Globally, Infect Locally: A Transcontinental Test of Adaptation by Stickleback and Their Tapeworm Parasite. *The American Naturalist*, 189(1):43–57, Jan. 2017. ISSN 1537-5323. doi: 10.1086/689597.

C. Wedekind. The infectivity, growth, and virulence of the cestode Schistocephalus solidus in its first intermediate host, the copepod Macrocyclops albidus. *Parasitology*, 115 (Pt 3): 317–324, Sept. 1997. ISSN 0031-1820. doi: 10.1017/s0031182097001406.

C. Wedekind and T. J. Little. The clearance of hidden cestode infection triggered by an independent activation of host defense in a teleost fish. *The Journal of Parasitology*, 90(6): 1329–1331, Dec. 2004. ISSN 0022-3395. doi: 10.1645/GE-225R.

K. M. Wegner, M. Kalbe, J. Kurtz, T. B. H. Reusch, and M. Milinski. Parasite Selection for Immunogenetic Optimality. *Science*, 301(5638):1343–1343, Sept. 2003a. doi: 10.1126/science.1088293. URL https://www.science.org/doi/full/10.1126/science.1088293. Publisher: American Association for the Advancement of Science.

K. M. Wegner, T. B. H. Reusch, and M. Kalbe. Multiple parasites are driving major histocompatibility complex polymorphism in the wild. *Journal of Evolutionary Biology*, 16(2): 224–232, 2003b. ISSN 1420-9101. doi: 10.1046/j.1420-9101.2003.00519.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1420-9101.2003.00519.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1420-9101.2003.00519.x.

K. M. Wegner, M. Kalbe, G. Rauch, J. Kurtz, H. Schaschl, and T. B. H. Reusch. Genetic variation in MHC class II expression and interactions with MHC sequence polymorphism

in three-spined sticklebacks: MHC CLASS IIB EXPRESSION IN STICKLEBACKS. *Molecular Ecology*, 15(4):1153–1164, Mar. 2006. ISSN 09621083, 1365294X. doi: 10.1111/j.1365-294X.2006.02855.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1365-294X.2006.02855.x.

K. M. Wegner, M. Kalbe, and T. B. H. Reusch. Innate versus adaptive immunity in sticklebacks: evidence for trade-offs from a selection experiment. *Evolutionary Ecology*, 21(4):473–483, July 2007. ISSN 1573-8477. doi: 10.1007/s10682-006-9129-5. URL https://doi.org/10.1007/s10682-006-9129-5.

K. M. Wegner, M. Kalbe, M. Milinski, and T. B. Reusch. Mortality selection during the 2003 European heat wave in three-spined sticklebacks: effects of parasites and MHC genotype. *BMC Evolutionary Biology*, 8(1):124, Apr. 2008. ISSN 1471-2148. doi: 10.1186/1471-2148-8-124. URL https://doi.org/10.1186/1471-2148-8-124.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, 2016. ISBN 978-3-319-24277-4. doi: 10.1007/978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

B. Woelfing, A. Traulsen, M. Milinski, and T. Boehm. Does intra-individual major histocompatibility complex diversity keep a golden mean? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1513):117–128, Oct. 2008. doi: 10.1098/rstb.2008.0174. URL https://royalsocietypublishing.org/doi/full/10.1098/rstb.2008.0174. Publisher: Royal Society.

D. L. Woodland, B. L. Kotzin, and E. Palmer. Functional consequences of a T cell receptor D beta 2 and J beta 2 gene segment deletion. *The Journal of Immunology*, 144(1):379–385, Jan. 1990. ISSN 0022-1767. doi: 10.4049/jimmunol.144.1.379. URL https://doi.org/10.4049/jimmunol.144.1.379.

R. Woodroffe and A. Vincent. Mother's little helpers: Patterns of male care in mammals. *Trends in Ecology & Evolution*, 9(8):294–297, Aug. 1994. ISSN 0169-5347. doi: 10.1016/0169-5347(94)90033-7.

D. J. Woodsworth, M. Castellarin, and R. A. Holt. Sequence analysis of T-cell repertoires in health and disease. *Genome Medicine*, 5(10):98, Oct. 2013. ISSN 1756-994X. doi: 10.1186/gm502. URL https://doi.org/10.1186/gm502.

R. J. Wootton. *A Functional Biology of Sticklebacks*. Springer US, Boston, MA, 1984. ISBN 978-1-4615-8515-2 978-1-4615-8513-8. doi: 10.1007/978-1-4615-8513-8. URL http://link.springer.com/10.1007/978-1-4615-8513-8.

R. J. Wootton, G. W. Evans, and L. Mills. Annual cycle in female Three-spined sticklebacks (Gasterosteus aculeatus L.) from an upland and lowland population. *Journal of Fish Biology*, 12(4):331–343, 1978. ISSN 1095-8649. doi: 10.1111/j.1095-8649.1978.tb04178.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1095-8649.1978.tb04178.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1095-8649.1978.tb04178.x.

L. Wu and L. Van Kaer. Natural killer T cells in health and disease. *Frontiers in bioscience (Scholar edition)*, 3:236–251, Jan. 2011. ISSN 1945-0516. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3626278/.

B. Wölfing. *Beyond MHC: The stickleback T-cell repertoire and evolutionary dynamics*. PhD thesis, Christian-Albrechts-Universität zu Kiel, Kiel, 2012.

T. Yamaguchi and J. M. Dijkstra. Major Histocompatibility Complex (MHC) Genes and Disease Resistance in Fish. *Cells*, 8(4):378, Apr. 2019. ISSN 2073-4409. doi: 10.3390/cells8040378. URL https://www.mdpi.com/2073-4409/8/4/378. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

J. Yan, J. M. Greer, R. Hull, J. D. O'Sullivan, R. D. Henderson, S. J. Read, and P. A. McCombe. The effect of ageing on human lymphocyte subsets: comparison of males and females. *Immunity & Ageing*, 7(1):4, Mar. 2010. ISSN 1742-4933. doi: 10.1186/1742-4933-7-4. URL https://doi.org/10.1186/1742-4933-7-4.

Y. Yanagi, Y. Yoshikai, K. Leggett, S. P. Clark, I. Aleksander, and T. W. Mak. A human T cell-specific cDNA clone encodes a protein having extensive homology to immunoglobulin chains. *Nature*, 308(5955):145–149, Mar. 1984. ISSN 1476-4687. doi: 10.1038/308145a0. URL https://www.nature.com/articles/308145a0. Publisher: Nature Publishing Group.

J. Yang, Y. Chen, Y. Jing, M. R. Green, and L. Han. Advancing CAR T cell therapy through the use of multidimensional omics data. *Nature Reviews Clinical Oncology*, 20(4):211–228, Apr. 2023. ISSN 1759-4782. doi: 10.1038/s41571-023-00729-2. URL https://www.nature.com/articles/s41571-023-00729-2. Publisher: Nature Publishing Group.

J.-Q. Yang, Y. Zhou, and R. R. Singh. Effects of Invariant NKT Cells on Parasite Infections and Hygiene Hypothesis. *Journal of Immunology Research*, 2016(1):2395645, 2016. ISSN 2314-7156. doi: 10.1155/2016/2395645. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2016/2395645. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2016/2395645.

Z. Yang. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591, Aug. 2007. ISSN 0737-4038. doi: 10.1093/molbev/msm088. URL https://doi.org/10.1093/molbev/msm088.

R. Yazawa, G. A. Cooper, M. Beetz-Sargent, A. Robb, L. McKinnel, W. S. Davidson, and B. F. Koop. Functional adaptive diversity of the Atlantic salmon T-cell receptor gamma locus. *Molecular Immunology*, 45(8):2150–2157, Apr. 2008. ISSN 0161-5890. doi: 10.1016/j.molimm.2007.12.007. URL https://www.sciencedirect.com/science/article/pii/S0161589007008905.

K. Yoshida, J. B. Cologne, K. Cordova, M. Misumi, M. Yamaoka, S. Kyoizumi, T. Hayashi, H. Robins, and Y. Kusunoki. Aging-related changes in human T-cell repertoire over 20years delineated by deep sequencing of peripheral T-cell receptors. *Experimental Gerontology*, 96:29–37, Oct. 2017. ISSN 1873-6815. doi: 10.1016/j.exger.2017.05.015.

M. Zagalska-Neubauer, W. Babik, M. Stuglik, L. Gustafsson, M. Cichoń, and J. Radwan. 454 sequencing reveals extreme complexity of the class II Major Histocompatibility Complex in the collared flycatcher. *BMC Evolutionary Biology*, 10(1):395, Dec. 2010. ISSN 1471-2148. doi: 10.1186/1471-2148-10-395. URL https://doi.org/10.1186/1471-2148-10-395.

A. Zapata and C. T. Amemiya. Phylogeny of lower vertebrates and their immunological structures. *Current Topics in Microbiology and Immunology*, 248:67–107, 2000. ISSN 0070-217X. doi: 10.1007/978-3-642-59674-2_5.

A. G. Zapata. The fish spleen. *Fish & Shellfish Immunology*, 144:109280, Jan. 2024. ISSN 1050-4648. doi: 10.1016/j.fsi.2023.109280. URL https://www.sciencedirect.com/science/article/pii/S1050464823007660.

J. Zhang, R. Nielsen, and Z. Yang. Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. *Molecular Biology and Evolution*, 22(12):2472–2479, Dec. 2005a. ISSN 0737-4038. doi: 10.1093/molbev/msi237. URL https://doi.org/10.1093/molbev/msi237.

N. Zhang, H. Hartig, I. Dzhagalov, D. Draper, and Y. W. He. The role of apoptosis in the development and function of T lymphocytes. *Cell Research*, 15(10):749–769, Oct. 2005b. ISSN 1748-7838. doi: 10.1038/sj.cr.7290345. URL https://www.nature.com/articles/7290345. Publisher: Nature Publishing Group.

H. Zhou, E. Bengtén, N. W. Miller, L. W. Clem, and M. Wilson. The T Cell Receptor Locus of the Channel Catfish, Ictalurus punctatus, Reveals Unique Features1. *The Journal of Immunology*, 170(5):2573–2581, Mar. 2003. ISSN 0022-1767. doi: 10.4049/jimmunol.170.5.2573. URL https://doi.org/10.4049/jimmunol.170.5.2573.

J. Zhu, T. Peng, C. Johnston, K. Phasouk, A. S. Kask, A. Klock, L. Jin, K. Diem, D. M. Koelle, A. Wald, H. Robins, and L. Corey. Immune surveillance by CD8+ skin-resident T cells in human herpes virus infection. *Nature*, 497(7450):494–497, May 2013. ISSN 1476-4687. doi: 10.1038/nature12110. URL https://www.nature.com/articles/nature12110. Publisher: Nature Publishing Group.

L. Zinger, A. Bonin, I. G. Alsos, M. Bálint, H. Bik, F. Boyer, A. A. Chariton, S. Creer, E. Coissac, B. E. Deagle, M. De Barba, I. A. Dickie, A. J. Dumbrell, G. F. Ficetola, N. Fierer, L. Fumagalli, M. T. P. Gilbert, S. Jarman, A. Jumpponen, H. Kauserud, L. Orlando, J. Pansu, J. Pawlowski, L. Tedersoo, P. F. Thomsen, E. Willerslev, and P. Taberlet. DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. *Molecular Ecology*, 28(8):1857–1862, 2019. ISSN 1365-294X. doi: 10.1111/mec.15060. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.15060. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.15060.

I. V. Zvyagin, M. V. Pogorelyy, M. E. Ivanova, E. A. Komech, M. Shugay, D. A. Bolotin, A. A. Shelenkov, A. A. Kurnosov, D. B. Staroverov, D. M. Chudakov, Y. B. Lebedev, and I. Z. Mamedov. Distinctive properties of identical twins' TCR repertoires revealed by high-throughput sequencing. *Proceedings of the National Academy of Sciences*, 111(16):5980–5985, Apr. 2014. doi: 10.1073/pnas.1319389111. URL https://www.pnas.org/doi/full/10.1073/pnas.1319389111. Publisher: Proceedings of the National Academy of Sciences.

List of manuscripts

Peer-reviewed publications

Chapter 1:

Efstratiou, A., Gaigher, A., Künzel, S., Teles, A. & Lenz, T.L. (2024). Template-specific optimization of NGS genotyping pipelines reveals allele-specific variation in MHC gene expression. Molecular Ecology Resources, doi:10.1111/1755-0998.13935

Author contributions

Thesis title:
Dynamics of the T cell receptor repertoire of a small vertebrate in response to infection

Chapter 1
Tobias L. Lenz and Artemis Efstratiou designed the study. Artemis Efstratiou conducted most of the molecular work and data analysis and wrote the manuscript. Tobias L. Lenz helped with the interpretation of the data and manuscript writing. Arnaud Gaigher and Ana Teles provided feedback on the manuscript. Arnaud Gaigher performed the MHC-I allele sequence diversity and phylogenetic analysis. Ana Teles provided the RNA samples used in our MHC-I dataset and performed RSCA typing for a subset of samples in our MHC-II data set. Sven Künzel performed part of the MHC-I and MHC-II Illumina library preparations and sequencing.

Chapter 2
Tobias L. Lenz, Ana Teles, and Artemis Efstratiou designed the study. Manfred Milinski contributed with conceptual discussions on the study design. Tobias L. Lenz, Nina Wildenhayn, and Ana Teles performed the fish breeding. Nina Wildenhayn coordinated the experimental infections and fish dissections. Artemis Efstratiou and Ana Teles conducted the RNA extractions and molecular protocol for TCR sequencing. Sören Franzenburg and Andre Franke contributed to NGS library design and performed TCR sequencing. Artemis Efstratiou conducted sequencing data analysis and wrote the chapter.

Chapter 3
Tobias L. Lenz, Ana Teles, and Artemis Efstratiou designed the study. Nina Wildenhayn, Artemis Efstratiou, Ana Teles, and Tobias L. Lenz performed the fish breeding. Tobias L. Lenz, Ana Teles, and Artemis Efstratiou prepared the fish for lake exposure, and performed macroparasite screening. Artemis Efstratiou conducted the RNA extractions and molecular protocol for TCR sequencing. Sören Franzenburg and Andre Franke contributed to NGS library design and performed TCR sequencing. Artemis Efstratiou performed data analysis towards genotyping of MHC-II and microsatellites, as well as sex typing, and conducted TCR data analysis and wrote the chapter.

Acknowledgements

First and foremost, I thank my supervisor, Tobias Lenz, for his constant support, guidance and gentle encouragement, without which this work would have never been completed. Being a member of our group has been a privilege, and something I am thankful for even when physical distance has kept its members further apart than we would have liked. To Ana, Arnaud, Onur, Clinton, Joanna, Britta, Alex and all other members throughout the years, thank you for the lovely atmosphere, and for the fond memories of fun outings in nature. I would like to especially thank Malavi, with whom sharing an office was a blast and whose puns would help lighten any gloomy day.

I would like to thank members of my Thesis Advisory Committee, Thorsten Reusch and John Baines, for their helpful advice and encouraging feedback. A big thank you to Olivia Roth for agreeing to be an examiner of this thesis. I also thank Hinrich Schulenburg for agreeing to be a member of the defense committee, and Marc Bramkamp for being the chair of the defense.

Without the many helping hands of the technicians, the work of this thesis could not have come to be. My gratitude goes to Nicole Thomsen, Irene Tomaschewski, and Sybille Liedtke for overall assistance in the laboratory. I also thank Sven Künzel, whose expertise in library preparation and sequencing proved more than valuable. A big thanks to Daniel Martens for his valuable help with maintaining and breeding the sticklebacks; it was good to know that our fish were always cared for by you. Nina Wildenhayn, thank you for your precious contributions with the fish and dissection work, especially during the peak of the pandemic. I owe a special shout-out to Angela Donner and Derk Wachsmuth, for always being extremely helpful and making dealing with bureaucratic and technical issues all the easier.

Plön might be a beautiful place, but it's the people that make it so. To all the friendships I formed here- the ones who still remain and those who have already gone-, without whom life in this small place would have been a dud. 2019 will forever remain in my memory as the most fun I've had – any positive correlation with the volume of alcohol consumed that year is purely coincidental. Andrea, your deeply kind yet fiery spirit and incredible work ethic has been an inspiration to me for years. I love you and value our friendship deeply. To the Kiel gang: I still reminisce upon our earlier days, and trip to Amsterdam, with a smile. To Gustavo, with whom our paths overlapped shortly but left a mark on this place: our zoom meetings during corona lockdowns helped keep me sane and entertained, in equal measure.

To Marianna, whose friendship has kept me going when it felt like an impossible task. Love you dearly. Don't let this PhD fool you - between the two of us, I'm still very much counting on you to be the one to get that yacht.

To my family, who have stood by me throughout this journey, which I set upon as a university freshman more than fifteen years ago: thank you. I owe you more than words could ever express, and can only hope to one day repay you to the fullest.

Last but not least, to my partner, Natasha, who has been there almost since the beginning: I consider myself the luckiest person in the world, to get to grow alongside you. Can't wait to see where the future takes us.

Declaration

I hereby declare that:

i) Apart from my supervisor's guidance, the content and design of this thesis is the product of my own work and only using the sources listed. All quotations from other works have been identified as such and properly acknowledged. My own contributions and the contributions of my co-authors are listed in detail in the corresponding section;

ii) this thesis has not been submitted either partially or wholly as part of a doctoral degree to another examining body, and no other parts have been published or submitted for publication other than indicated in the thesis;

iii) this thesis has been prepared with regard to the Rules of Good Scientific Practice of the German Research Foundation;

iv) an academic degree has never been withdrawn.

Signature (Artemis Efstratiou)	Date and location

Appendix A

Template-specific optimization of NGS genotyping pipelines reveals allele-specific variation in MHC gene expression

A.1 Supplementary Materials and Methods

A.1.1 Traditional approach: RSCA genotyping of MHC-II

Initially, all individuals in our dataset were genotyped for MHC-II with the established RSCA protocol (Lenz, Eizaguirre, Becker, et al., 2009), as were our artificial plasmid combination samples to confirm the presence of the expected alleles. This protocol targets all currently known MHC-II loci of the three-spined stickleback using the primer pair GAIIEx2startF (5'-GTC TTT AAC TCC ACG GAG CTG AAG G-3') and GAIIExon2R_RSCA (5'-ACT CAC CGG ACT TAG TCA G-3'). The product is a 247 bp fragment of exon 2 of the MHC-IIß genes, encoding the b1 domain, which, together with the a1 domain, form the highly polymorphic peptide-binding region of the MHC-II molecule (Brown et al., 1993). The PCR reaction contained either 100 ng of genomic DNA or cDNA samples, or 0.01 ng of the artificial samples, following the PCR protocol as previously described (Lenz, Eizaguirre, Becker, et al., 2009), and using two reference alleles for allele identification (Lenz and Becker (2008). This MHC genotyping protocol has been used extensively in local stickleback populations (Lenz, Eizaguirre, Scharsack, et al., 2009; Eizaguirre et al., 2011, 2012), allowing us to obtain reliable haplotype identity and unambiguously genotype all samples in our dataset, hence ensuring the basis for proper comparison of genotyping pipelines.

A.1.2 NGS approach: Illumina library preparation and sequencing of MHC-II

The first PCR reaction contained either 100 ng of genomic DNA or cDNA samples, or 0.01 ng of plasmid combination samples, 1× GeneAmp PCR Buffer II, 5 mM $MgCl_2$, 50 μM of each dNTP, 0.5 μM of each primer, and 1 unit of AmpliTaq Gold polymerase. We used the following PCR program: 94℃ for 10 minutes, 22 cycles of denaturation at 94℃ for 30 seconds, annealing at 58℃ for 30 seconds and elongation at 72℃ for 60 seconds, finally 72℃ for 5 minutes. An additional 'reconditioning' PCR (Thompson et al., 2002) of 8 cycles was employed for

adequate MHC-II amplification from cDNA samples. Subsequently, 5 μL of PCR product underwent a multiplexing PCR, with primers incorporating Illumina-specific adapters and 8-bp indices. For maximum multiplexing capacity, we employed a dual-indexing saturated design, with each primer used in multiple combinations (Galan et al., 2010). The reaction volume of 25 μl included 1× PCR Phusion Buffer, 200 μM of each dNTP, 0.5 μM of each primer, and 1 unit of Phusion Hot Start II DNA Polymerase (Thermo Fisher). The thermal cycler program was: 98℃ for 30 seconds, 8 cycles of 98℃ for 9 seconds, 50℃ for 30 seconds and 72℃ for 2 minutes, and 72℃ for 10 minutes.

A.1.3 Development of MHC-I-targeting primer pairs

In the three-spined stickleback, MHC class I genes have so far not been as thoroughly explored as MHC class II genes. We therefore started the development of a reliable MHC-I genotyping protocol with careful primer design and evaluation: two species-specific primer pairs were tested on the basis of MHC I allele amplification using a sub-dataset of 32 samples (representing all 10 sibships, chosen at random from our 185-sample dataset). Both primer pairs target exon 2 of the MHC-I gene (encoding the a1 domain of the molecule), which -together with exon 3, encoding the a2 domain- encodes for the highly polymorphic peptide-binding region of the MHC-I molecule (Bjorkman et al., 1987). The first pair comprised two previously published primers, the forward primer 5'- GTC CCA AAC TTC CCA GAG TTT GTG -3' (Sato et al., 1998) and the reverse primer 5'-ACC TCC AGT TTG GTT GAA GCG T-3' (Schaschl & Wegner, 2006); this pair amplifies a 228 bp fragment and has been shown to successfully amplify at least 22 distinct MHC-I variants (Schaschl & Wegner, 2006). The second primer pair (forward primer GAI_Ex2_For 5'-GAG TCC CAA ACT TCC CAG AGT-3', reverse primer GAI_Ex2_Rev 5'-AAA CAA CGC TTC AAC CAA ACT GGA-3'), tested here for the first time, was designed to amplify the same part of exon 2 (227 bp in length) from all known MHC-I loci based on sequence information from both the stickleback reference genome (Jones et al., 2012), and a novel genome sequence of a local three-spined stickleback from the same population as the fish of this study (Thorburn et al., 2023).

A.1.4 NGS approach: Illumina library preparation and sequencing of MHC-I

As was the case in our MHC-II protocol, both MHC-I primer pairs included heterogeneity spacers used in various combinations (Table S5). PCR amplification with the first primer pair was achieved via initial heating at 95° C for 10 min, followed by 22 cycles of denaturation at 95° C for 30 s, annealing at 60° C for 30 s, extension at 72° C for 1 min and ending with a 5 min extension step at 72° C. The PCR reaction (25 μl) contained 2 μl cDNA template, 50 μM dNTPs, 2.5 mM MgCl2, 1 U AmpliTaq Gold polymerase (Thermo Fischer), and 0.5 μM of each primer. Identical PCR reactions and conditions were followed for the second primer set, with the exception of the annealing temperature, which was set at 58° C. We followed a similar protocol for library preparation as described for MHC-II, with adaptors and indices added for multiplexing during a 2nd PCR, followed by a purification step. Libraries were sequenced separately on Illumina MiSeq platform using a MiSeq Reagent Kit v2 Micro. Demultiplexed sample reads were merged and quality filtered for a minimum Phred score of 30 using the AmpliSAT web server, then processed with AmpliSAS at default parameters, with a 25% clustering threshold, to avoid clustering of similar true alleles of varying amplification efficiencies (Biedrzycka et al., 2017), and a low per amplicon frequency filtering threshold of 0.5%, to ac-

count for the higher expected MHC-I allelic diversity. In an effort to further ensure reliability of genotyping results, pipeline-called alleles were regarded as true only if determined upon manual inspection to be neither contaminants nor erroneous sequences originating from 1-2 substitutions from a more abundant parental variant which 'escaped' the clustering step, as shown to occasionally ensue in our MHC-II dataset.

A.2 Supplementary Results

A.2.1 RSCA-based genotyping of MHC-II

Using our RSCA genotyping protocol for MHC-II, we were able to successfully genotype all individuals in our dataset. In total, we detected 31 distinct alleles, ranging from one to seven per individual, with an average of four. An additional very divergent allele, which has been previously described (Reusch & Langefors, 2005; Lenz, Eizaguirre, Scharsack, et al., 2009; Lenz et al., 2013), was detected in every gDNA sample and multiple cDNA samples. This variant (NCBI accession no DQ399896) likely originates from a separate highly conserved and non-classical MHC locus, and has been excluded from further analysis in the present study. Using our allele library with known mobility values, all 31 alleles were identified unambiguously. When two alleles were not distinguishable due to their very similar mobility values when hybridized with the first reference allele, the use of the second reference allele provided clearly distinct peaks and the variants were successfully identified (example shown in Fig. S2A). The 31 identified alleles belonged to 35 different MHC-II genotypes, from various combinations of 15 distinct MHC-II haplotypes (Tables S1 and S2), largely overlapping with the haplotypes detected in previous studies of the local lake population (Lenz, Eizaguirre, Scharsack, et al., 2009; Kaufmann et al., 2017). The haplotypes contained between one and four alleles. When typing our 18 plasmid combination samples, in total encompassing 35 alleles, identification was straightforward in samples containing up to 6 alleles. However, for samples consisting of 8 and 10 alleles, distinguishing between different variants proved a challenging task, as peaks were often indistinguishable. In certain cases, allele identification was difficult even with mobility values from both reference alleles, due to multiple overlapping allele peaks (Fig. S2B). All alleles were named following the nomenclature proposed by Klein et al. (1990) and deposited to NCBI (accession numbers to be provided upon acceptance). Previously published sequences retained their given numerical suffix, with an 'X' as the locus designating letter, since alleles could not be assigned to specific loci (e.g. Gaac-DXB*03, see Tables S1, S3).

A.2.2 Evaluation of novel stickleback-specific MHC-I primers

In order to best capture the stickleback MHC-I diversity, two sets of primers were tested with a subset of cDNA samples from 32 individuals. Of those, 30 were successfully sequenced with both primers and were used for comparison. The newly designed stickleback-specific primer pair (GAI_Ex2_For, GAI_Ex2_Rev) consistently amplified a higher number of MHC-I variants in our comparison dataset (Table S9); 44 vs. 34 allelic variants in total, and 11.8 (SD 3.5) vs. 9.4 (SD 2.9) alleles per individual.

A.3 Supplementary Tables

Haplotypes	Alleles	Previous nomenclature	GenBank accession no (100%ID, %cover)	Samples	
				gDNA	cDNA
II-A	*Gaac*-DXB*15	-	KT885240.1(100%)	14	13
	Gaac-DXB*16	*Gaac*-DXB*16	AY687846.1(97%)		
II-B	*Gaac*-DXB*17	*Gaac*-DBB*01	DQ016399.1(97%)	4	2
	Gaac-DXB*18	*Gaac*-DAB*01	DQ016400.1(97%)		
II-C	*Gaac*-DXB*03	*Gaac*-DCB*03	DQ016417.1(97%)	9	8
	Gaac-DXB*35	*Gaac*-DXB*35	DQ016410.1(97%)		
II-D	*Gaac*-DXB*11	-	KT885246.1(100%)	6	7
	Gaac-DXB*12	*Gaac*-DXB*12	AY687842.1(97%)		
II-E	*Gaac*-DXB*14	-	KT885262.1(100%)	3	2
	***Gaac*-DXB*31**	*Gaac*-DXB*31	GQ277654.1(100%)		
II-F	*Gaac*-DXB*01	*Gaac*-DXB*01	AY687829.1(97%)	8	10
II-G	***Gaac*-DXB*27**	*Gaac*-DXB*27	DQ016402.1(97%)	4	3
	***Gaac*-DXB*28**	-	KT885235.1(100%)		
	Gaac-DXB*30	*Gaac*-DXB*30	DQ016405.1(97%)		
II-H	*Gaac*-DXB*04	-	-	3	4
	Gaac-DXB*05	-	-		
II-L	*Gaac*-DXB*02	-	KT885315.1 (100%)	3	3
	Gaac-DXB*13	*Gaac*-DXB*13	AY687843.1(97%)		
II-O	*Gaac*-DXB*06	-	KT885241.1(100%)	4	4
	Gaac-DXB*07	-	-		
	Gaac-DXB*08	-	KT885259.1(100%)		
II-P	*Gaac*-DXB*09	-	FJ360532.1(100%)	1	0
	Gaac-DXB*10	-	FJ360533.1(100%)		
II-R	*Gaac*-DXB*19	-	JN827380.1(97%)	1	2
	Gaac-DXB*20	-	KT885410.1 (100%)		
II-T	*Gaac*-DXB*21	-	KT885414.1(100%)	1	2
	Gaac-DXB*22	-	KT885359.1(100%)		
II-V	*Gaac*-DXB*23	-	FJ360540.1(100%)	5	6
	Gaac-DXB*24	-	-		
	Gaac-DXB*25	-	-		
II-W	***Gaac*-DXB*27**	*Gaac*-DXB*27	DQ016402.1(97%)	7	9
	***Gaac*-DXB*28**	-	KT885235.1(100%)		
	Gaac-DXB*29	-	-		
	***Gaac*-DXB*31**	*Gaac*-DXB*31	GQ277654.1(100%)		
15	**31**				

Alleles found in more than one haplotype are marked in bold.

Table A.1: Overview of all haplotypes and corresponding alleles present in the three-spined stickleback gDNA and cDNA samples of our MHC-II genotyping dataset.

Genotypes	Number of alleles	Samples		
		gDNA	cDNA	Total
II-A/II-A	2	4	4	8
II-A/II-B	4	1	0	1
II-A/II-C	4	1	1	2
II-A/II-D	4	1	1	2
II-A/II-F	3	1	0	1
II-A/II-O	5	1	0	1
II-A/II-T	4	0	1	1
II-A/II-V	5	2	3	5
II-A/II-W	6	3	3	6
II-B/II-G	5	1	0	1
II-B/II-H	4	1	2	3
II-B/II-L	4	1	0	1
II-C/II-C	2	3	2	5
II-C/II-D	4	1	1	2
II-C/II-E	4	1	1	2
II-C/II-F	3	1	2	3
II-C/II-L	4	1	1	2
II-C/II-P	4	1	0	1
II-D/II-D	2	0	1	1
II-D/II-E	4	1	1	2
II-D/II-F	3	2	2	4
II-D/II-L	4	0	1	1
II-D/II-W	6	1	0	1
II-E/II-W	5	1	0	1
II-F/II-H	3	2	2	4
II-F/II-L	3	1	1	2
II-F/II-O	4	1	1	2
II-F/II-W	5	0	2	2
II-G/II-O	6	2	3	5
II-G/II-V	6	1	0	1
II-R/II-T	4	1	1	2
II-R/II-W	6	0	1	1
II-V/II-V	3	0	1	1
II-V/II-W	7	2	2	4
II-W/II-W	4	0	1	1
35	**2-7**	**40**	**42**	**82**

Table A.2: Overview of the MHC-II genotypes of the three-spined stickleback gDNA and cDNA samples included in the pipeline evaluation dataset.

Sample	Allele no / sample	Alleles
3a	3	*Gaac*-DXB*11, 20, **45**
3b	3	*Gaac*-DXB*17, 21, **32**
3c	3	*Gaac*-DXB*04, 11, **47**
4a	4	*Gaac*-DXB***26**, **46**, **57**, **62**
4b	4	*Gaac*-DXB*07, 20, 31, **48**
4c	4	*Gaac*-DXB*06, 23, **33**, **52**
5a	5	*Gaac*-DXB*08, 18, 22, **53**, **61**
5b	5	*Gaac*-DXB*05, 17, **42**, **43**, **60**
5c	5	*Gaac*-DXB*11, 21, 23, **26**, **59**
6a	6	*Gaac*-DXB*09, 18, **32, 47, 55**, **58**
6b	6	*Gaac*-DXB*10, 11, 17, 20, **52**, **56**
6c	6	*Gaac*-DXB*09, 19, **33**, **45**, **46**, **62**
8a	8	*Gaac*-DXB*06, 21, 22, **43**, **46**, **55**, **57**, **61**
8b	8	*Gaac*-DXB*05, 07, 08, 31, **42**, **48**, **56**, **59**
8c	8	*Gaac*-DXB*19, 21, 23, 31, **53**, **58**, **60**, **62**
10a	10	*Gaac*-DXB*05, 11, 17, 19, 22, **26**, **33**, **45**, **57**, **61**
10b	10	*Gaac*-DXB*04, 09, 20, 31, **32**, **46**, **48**, **56**, **60**, **62**
10c	10	*Gaac*-DXB*06, 07, 10, 11, 18, 21, **43**, **52**, **58**, **59**
18	3-10	Plasmid alleles used in total = 35

Alleles not present in the biological three-spined stickleback samples (gDNA and cDNA) of our MHC-II genotyping dataset are marked in bold (*N*=19).

Table A.3: Plasmid allele composition of artificial combination samples included in our MHC-II genotyping dataset, used to evaluate the performance of bioinformatics pipelines at genotyping MHC-II genes of the three-spined stickleback.

Primer	Nucleotide sequence (5' → 3')
Fw 1	ACACTCTTTCCCTACACGACGCTCTTCCGATCT*GTCTTTAACTCCACGGAGCTGAAGG*
Fw 2	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**T***GTCTTTAACTCCACGGAGCTGAAGG*
Fw 3	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**AC***GTCTTTAACTCCACGGAGCTGAAGG*
Fw 4	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**CAA***GTCTTTAACTCCACGGAGCTGAAGG*
Rv 1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT*ACTCACCGGACTTAGTCAG*
Rv 2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**T***ACTCACCGGACTTAGTCAG*
Rv 3	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**GT***ACTCACCGGACTTAGTCAG*
Rv4	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**CGG***ACTCACCGGACTTAGTCAG*

Table A.4: MHC-II primers used in the first PCR step of our NGS protocol. Underlined nucleotides represent the multiplexing read 1/2 sequencing primers. Heterogeneity spacers are highlighted in bold. MHC-II-specific sequences are shown in italics.

Primer	Nucleotide sequence (5' → 3')
Fw 1	ACACTCTTTCCCTACACGACGCTCTTCCGATCT*GAGTCCCAAACTTCCCAGAGT*
Fw 2	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**T***GAGTCCCAAACTTCCCAGAGT*
Fw 3	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**AC***GAGTCCCAAACTTCCCAGAGT*
Fw 4	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**CAA***GAGTCCCAAACTTCCCAGAGT*
Rv 1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT*TCCAGTTTGGTTGAAGCGTTGTTT*
Rv 2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**A***TCCAGTTTGGTTGAAGCGTTGTTT*
Rv 3	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**GC***TCCAGTTTGGTTGAAGCGTTGTTT*
Rv4	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**CGG***TCCAGTTTGGTTGAAGCGTTGTTT*

Table A.5: MHC-I primers used in the first PCR step of our NGS protocol. Underlined nucleotides represent the multiplexing read 1/2 sequencing primers. Heterogeneity spacers are highlighted in bold. MHC-I-specific sequences are shown in italics.

Allele	Nucleotide sequence
MHCI_X_918500	GTGACTCACTTGCTGAAGTTTTTCCTCACTGGATCCTCTGGAGTCCCAAACTTCCCAGA GTTTGTGGTTGTTGGGCTGCTGGATGAAGTTGAGGTTGTTCACTATGACAGTGACACC TGGAGAGTAGAACCCAGACAGGACTGGATGAGCAGAGTCAGAAAGGATCTGCCTTGG GACTGGTTAGCGCTGACTCAGAACGCTTTGGTCGCCCAGCAGGAATTGAAAGCCTAC ATTGAAATATTAAAACGACGCTTCAACCAAACTGGAGCCGGT
MHCI_X_1123300	GTGACTCACTCTATGAAGTATTTCCTCACTGCGTCCTCTGGAGTCCCAAACTTCCCAG AGTTTGTGGTTGTTGGGCTGCTGGATGAAGTTGAGGTTGTTCACTATGACGGTAACAC AACGAGAGCAGAACCCAGACAGGACTGGATGAGCAGACTCAAAGAGGATGATCCTC AGTACTGGAAGAGTCGGACTGAGCTCGCTATGTACACCCAGCAGGACTTCAAAGGCA ACATTGAAACAGCAAAACAACGCTTCAACCAAACTGGAGCAGGT
MHCI_X_19000	CTCACTCGCTGAAGAATTTCTACACTGGATCCTCTGGAGTCCCAAACTTCCCAGAGTT TGTGGTTGTTGGGCTGCTGGATGAAGTTGAGATTAGTCACTATGACAGTAACACCAGG AGAGAAGAACCCAGACAGGACTGGATGAGCAGAGTCACAGAGGATGATCCTCAGTAC TGGAAGAGTGAGACTGAGATCCTTATGGGCCAACAGCAGGGCTTCAAAGTCAACATT GAAACAGCAAAACAACGCTTCAACCAAACTGGAGCAGGT
MHCI_X_520748:521355:1	GTGACTCACTCGCTGAAGTTTTTCTACACTGCGTCCTCTGGAGTCCCAAACTTCCCAG AGTTTGTGGCTGTTGGGCTGCTGGATGAAGTTGAGGTTGTTCACTATGACAGTGACAC CAGGAGACTAGAACCCAGACAGAACTGGGTGAGCAGACTCAGAGGGGATGATCCTCA GTACTGGAAGAGAAACACTAAGAACTCTATGGACGCCCAGCAGGTCTTCAAAGGCAA CATTGAAACACTAAAACAACGCTTCAACCAAACTGGAGGTT
MHCI_X_1093652:1093957:1	GTGACTCACTCGCTGAAGTATTTCTACACTGGATCCTCTGGAGTCCCAAACTTCCCAG AGTTTGTGGTTGTTGGGCTGCTGGATGAAGTTGAGATTGTTCACTATGACGGTAACAC CAGGAGAACAGAACCCAGACAGGACTGGATGAGCAGACTCAGAGAGGATGATCCTC AGTTCTGGGATATTCAGACTGATATCGCTATGGACGCCCAGCAGGACTTCAAAGGCCA CACTGAAACAGCAAAACAAAGCTTCAACCAAACTGGAGGTTTGTTTATGTTTCACACT
MHCI_X_1098372:1098675:1	GTGACTCACTCGCTGAAGTATTTCTACACTGGATCCTCTGGAGTCCCAAACTTCCCAG AGTTTGTGGTTGTTGGGCTGCTGGATGAAGTTGAGATTGTTCACTATGACGGTAACAC CAGGAGAACAGAACCCAGACAGGACTGGATGATCAAAGTCATAGATGATGATCCTCA GTACTGGAAGAGTCAGACTGATATCGCTATGGACACCCAGCAGGCCTTCAAAGGCTA CATTGAAATAGCAAAACAAAGCTTCAACCAAACTGGAGGTTTGTTTATGTTTCAC
MHCI_X_426565:427172:1T	ACTCACTCGCTGAAGTATTTCCTCACTGCGTCCTCTGGACTCCCAAACTTCCCAGAGTT TGTGGGTGTTGGGCTGCTGGATGAAGTTGAGTTTGTTCACTATGACAGTGACACCAAG AGAGCAGAACCCAGACAGGACTGGATGAGCAGACTCACAGAGGATGATCCTCAGTAC TGGAAGAGGAACACTGAGCTCGCTATGGACACCCAGCAGGTCTACAAACGCCACATT GAAATATTAAACCAAAGCTTCAACCAAACTGGAGGTTTGTTTTTGTTTCACACTT
MHCI_X_802052:802355:1C	ACTCGCTGAAGTATTTCTACACTGGATCCTCTGGACTCCCAAACTTCCCAGAGTTTGT GATTGTTGGGCTGGTGGATGAAGTTGAGGTTGTTCACTATGACAGTGACACCTGGAGA CTAGAACCCAGACAGGATTGGGTGAGCAGAGTCACAGAGGACGATCATTGGTTCTGG AACTGGCAGACTGGGCTCGCTATGAACGCCCAGCGGGAATTTAAAGGCTACATTAAA ACAGCAAAACAACGCTTCAACCAAACTGGAGGTTTGTTTGTTTCACACATTTTGTTTA TATTTGCATCCCTTT

Table A.6: MHC-I alleles retrieved from the Ensembl reference genome of the three-spined stickleback.

	Replicate 1			Replicate 2		
Step	**gDNA**	**cDNA**	**plasmid**	**gDNA**	**cDNA**	**plasmid**
Raw reads	668,008	734,773	331,186	1,426,874	854,562	753,120
Merged†	652,358	719,260	323,564	1,406,705	839,923	741,403
Quality-filtered †	540,745	604,223	273,589	1,240,893	731,205	653,128
N	40	42	18	39	42	18
Mean replicate depth ± SD (reads)	13,518± 11,502	14,386± 11,656	15,199± 12,024	31,817± 23,758	17,409± 16,913	36,284± 29,200
Mean sample depth (both replicates)	22,678± 16,017	15,897± 10,849	25,742± 24,694			

† Merging of Illumina paired-end reads and quality filtering for AmpliSAS was performed with the AmpliMerge and AmpliClean function of the AmpliSAT online tool, respectively. Only 247-253 bp reads with an average Phred quality score of 30 were retained. A similar threshold (q = 30, p = 90) was used for quality filtering with the ACACIA pipeline. Numbers presented here correspond to AmpliSAT outputs.

Table A.7: Processing details of raw sequencing data obtained from our duplicate MHC-II genotyping dataset. N, number of samples successfully sequenced.

Step	**Replicates 1**	**Replicates 2**
Raw reads	1,524,882	3,500,392
Merged†	1,481,433	3,464,692
Quality-filtered †	1,346,748	3,176,338
N	180	184
Mean replicate depth ± SD (reads)	7,482 ± 4,426	17,262 ± 9,963
Mean individual depth from both replicates ± SD (reads)	12,385 ± 5,892 (N=179)	

† Merging of Illumina paired-end reads and quality filtering was performed with the AmpliMerge and AmpliClean function of the AmpliSAT online tool, respectively.

Table A.8: Processing details of raw sequencing data obtained from our duplicate MHC-I genotyping dataset. N, number of samples successfully sequenced.

Method	NGS (Illumina)			Cloning		BAC cloning	Cloning	Published genomes	
Material	cDNA	gDNA		gDNA	cDNA	gDNA	gDNA		
Primers	'New'	'Old' primers targeting exon 2					'Old' [exons 2/3]		
Study / **Alleles**	Present study		Unpublished data, Kasala et al.	Unpublished data, Lenz et al.	Schaschl & Wegner 2006	Schaschl & Wegner 2007	Sato et al., 1998	European gynogen genome §	Ensembl gene ID [ENSGACG000000], % identity, nomenclature¶
Gaac-UX*51	✓	✓	KT885581.1	✓	×			×	01941, 95%, GA11(U)
Gaac-UX*17	✓	✓	×	✓	DQ399889.1			×	01941, 95.1%, GA11(U)
Gaac-UX*10	✓	✓	×	✓	DQ399882.1			×	01288, 93.8%, GA27(U)
Gaac-UX*44	✓	✓	KT885428.1	✓	×		AJ230208.1‡‡	×	01837, 92.7%, GA1(U)
Gaac-UX*27	✓	✓	KT885705.1	✓	DQ399899.1		AJ230216.1	✓	01919, 99.5%, GA5(U)
Gaac-UX*52	✓	×	×	×	×	✓		×	01973, 96.2%, GA13 (U)
Gaac-UX*53	✓	✓	KT885455.1	×	×			×	01941, 94.5%, GA11(U)
Gaac-UX*08	✓	✓	KT885704.1	✓	DQ399880.1			×	01288, 93.7%, GA27(U)
Gaac-UX*18	✓	✓	KT885745.1, KT885699.1	✓	DQ399890.1			×	01979, 99.5%, GA17 (U)
Gaac-UX*19	✓	✓	KT885535.1 ‡	✓	DQ399891.1	✓		×	01837, 91.7%, GA1(U)
Gaac-UX*54	✓	✓	KT885514.1	×	×			×	01837, 93.7%, GA1(U)
***Gaac*-UX*55**	✓	×	×	×	×			×	scaffold_653:9561-9742, 97.8%, GA28(U)
***Gaac*-UX*56**	✓	×	×	×	×			×	01973, 95.1%, GA13 (U)
Gaac-UX*22	✓	✓	KT885630.1 ‡	✓	DQ399894.1			×	01837, 90.6%, GA1(U)
Gaac-UX*11	✓	✓	×	✓	DQ399883.1			×	00151, 95.5%, GA26(U)
***Gaac*-UX*57**	✓	×	×	×	×			×	01973, 95.1%, GA13 (U)
Gaac-UX*46	✓	✓	KT885434.1	✓	×			×	01913, 93.6%, GA4 (U)
***Gaac*-UX*58**	✓	×	×	×	×			✓‡	01978, 92.3%, GA16 (U)
Gaac-UX*24	✓	✓	×	✓	DQ399896.1			×	00151, 95.1%, GA26(U)
Gaac-UX*59	✓	✓	KT885451.1	×	×			×	01935, 91.5%, GA8(U)
Gaac-UX*40	✓	✓	KT885430.1	✓	×			×	01910, 92.7%, GA2(U)
***Gaac*-UX*60**	✓	×	×	×	×			×	00116, 93.3%, GA19(U)
Gaac-UX*61	✓	✓	KT885422.1	×	×			×	01973, 97.6%, GA13 (U)
Gaac-UX*47	✓	✓	KT885424.1	✓	×			✓	01935, 90.9%, GA8(U)
***Gaac*-UX*62**	✓	×	×	×	×			×	01978, 94%, GA16 (U)
***Gaac*-UX*63**	✓	×	×	×	×			×	01978, 94%, GA16 (U)
***Gaac*-UX*64**	✓	×	×	×	×			×	01973, 94.5%, GA13 (U)
Gaac-UX*65	✓	✓	KT885443.1	×	×			×	01941, 92.9%, GA11(U)
Gaac-UX*66	✓	✓	KT885426.1	×	×			✓	01935, 95.6%, GA8(U)
Gaac-UX*67	✓	✓	KT885458.1	×	×			×	01979, 97.8%, GA17 (U)
Gaac-UX*50	✓	✓	KT885583.1	✓	×			×	00122, 94%, GA23(U)
Gaac-UX*21	✓	✓	×	✓	DQ399893.1			×	00122, 92.2%, GA22 (U)
Gaac-UX*13	✓	✓	×	✓	DQ399885.1		AJ230213.1	×	00122, 93.2%, GA22 (U)
Gaac-UX*26	✓	✓	×	✓	DQ399898.1			×	01837, 94.4%, GA1(U)
Gaac-UX*14	✓	✓	KT885474.1	✓	DQ399886.1			✓‡	01910, 93.2%, GA2(U)
Gaac-UX*68	✓	✓	KT885442.1	×	×			×	01973, 91.8%, GA13 (U)
Gaac-UX*15	✓	✓	×	✓	DQ399887.1			×	00141, 93.8%, GA24(U)
Gaac-UX*42	✓	✓	KT885431.1	✓	×		AJ230215.1	✓‡	01919, 94%, GA5(U)
***Gaac*-UX*69**	✓	×	×	×	×			✓	01978, 92.9%, GA16 (U)
Gaac-UX*41	✓	✓	KT885533.1	✓	×			×	01913, 93.7%, GA4 (U)
Gaac-UX*70	✓	×	KT885588.1 ‡	×	×	✓		×	01979, 98.4%, GA17 (U)
Gaac-UX*71	✓	✓	KT885421.1	×	×			×	01973, 01976, 91.5%, GA13/14(U)
Gaac-UX*16	✓	✓	KT885634.1 ‡	✓	DQ399888.1			×	00141, 93.4%, GA24(U)
Gaac-UX*09	✓	✓	×	✓	DQ399881.1			×	01973, 92.2%, GA13(U)
Gaac-UX*20	✓	✓	×	✓	DQ399892.1			×	01837, 96.1%, GA1(U)
Gaac-UX*72†	✓	?	KT885772.1	×	×			×	scaffold_653:9561-9742, 98.4%, GA28(U)
***Gaac*-UX*73†**	✓	?	×	×	×			×	scaffold_653:9561-9742, 98.4%, GA28(U)
***Gaac*-UX*74†**	✓	?	×	×	×			×	00122, 93.2%, GA21 (U)
Total	**45-48**	**35-37**	**26**	**25**	**17/22**	**3/3**	**4**	**7**	**18/22**

Alleles amplified only by the new primer pair are marked in bold. Crosses (†) designate alleles not present in the primer pair comparison dataset (32 samples), but successfully amplified by the new primer pair in the complete dataset (185 samples). Undetermined amplification by the old primer pair is indicated with a question mark. Double daggers indicate single (‡) or double (‡‡) base substitutions. § published by Thorburn et al. 2023, ¶ According to Grimholt et al., 2015.

Table A.9: Comparison of the two MHC-I exon 2 primer pairs tested at amplifying alleles in the three-spined stickleback. Detection status from previous studies (with corresponding NCBI accession numbers) and sequence identity with published three-spined stickleback genomes are shown. N, number of samples successfully sequenced.

Family	Parent 1	Parent 2	Sibship size	Offspring	Number of individuals
1	I-A/I-B	I-A/I-C	23	I-A/I-A	5
				I-A/I-B	7
				I-A/I-C	5
				I-B/I-C	6
2	I-D/I-E	I-D/I-F	19	I-D/I-D	4
				I-D/I-E	7
				I-D/I-F	4
				I-E/I-F	4
3	I-C/I-H	I-C/I-G	21	I-C/I-C	3
				I-C/I-G	10
				I-C/I-H	5
				I-G/I-H	3
4	I-A/I-D	I-C/I-I	24	I-A/I-C	2
				I-C/I-D	8
				I-A/I-I	11
				I-D/I-I	3
5	I-A/I-C	I-F/I-I	20	I-A/I-F	4
				I-A/I-I	2
				I-C/I-F	2
				I-C/I-I	10
				I-A:C/I-I*	2
6	I-C/I-G	I-C/I-G	21	I-C/I-C	6
				I-C/I-G	9
				I-G/I-G	6
7	I-A/I-C	I-C/I-F	15	I-A/I-C	3
				I-C/I-C	3
				I-A/I-F	3
				I-C/I-F	4
				I-A:C/I-F*	2
8	I-D/I-E	I-G/I-J	21	I-D/I-J	5
				I-D/I-G	4
				I-E/I-G	6
				I-E/I-J	5
				I-D:E/I-J*	1
9	I-C/I-C	I-C/I-G	9	I-C/I-C	2
				I-C/I-G	7
10	I-C/I-C	I-C/I-J	6	I-C/I-C	1
				I-C/I-J	5

Recombinant genotypes are marked with an asterisk (*), and recombinant haplotypes shown in bold.

Table A.10: Inferred parental and observed offspring MHC-I genotypes of the 10 families included in our dataset.

A.4 Supplementary Figures

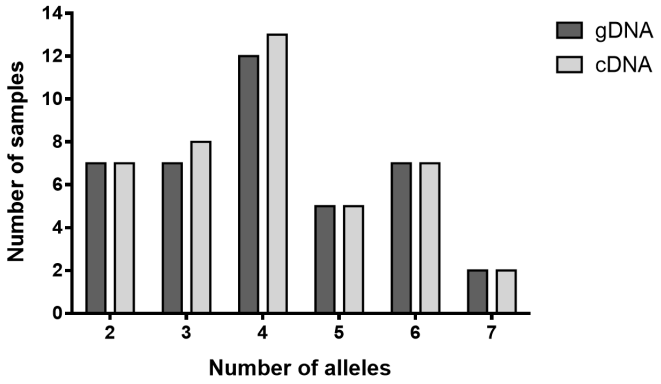


Figure A.1: Distribution of allele numbers in the 40 genomic and 42 cDNA samples comprising our pipeline evaluation dataset.

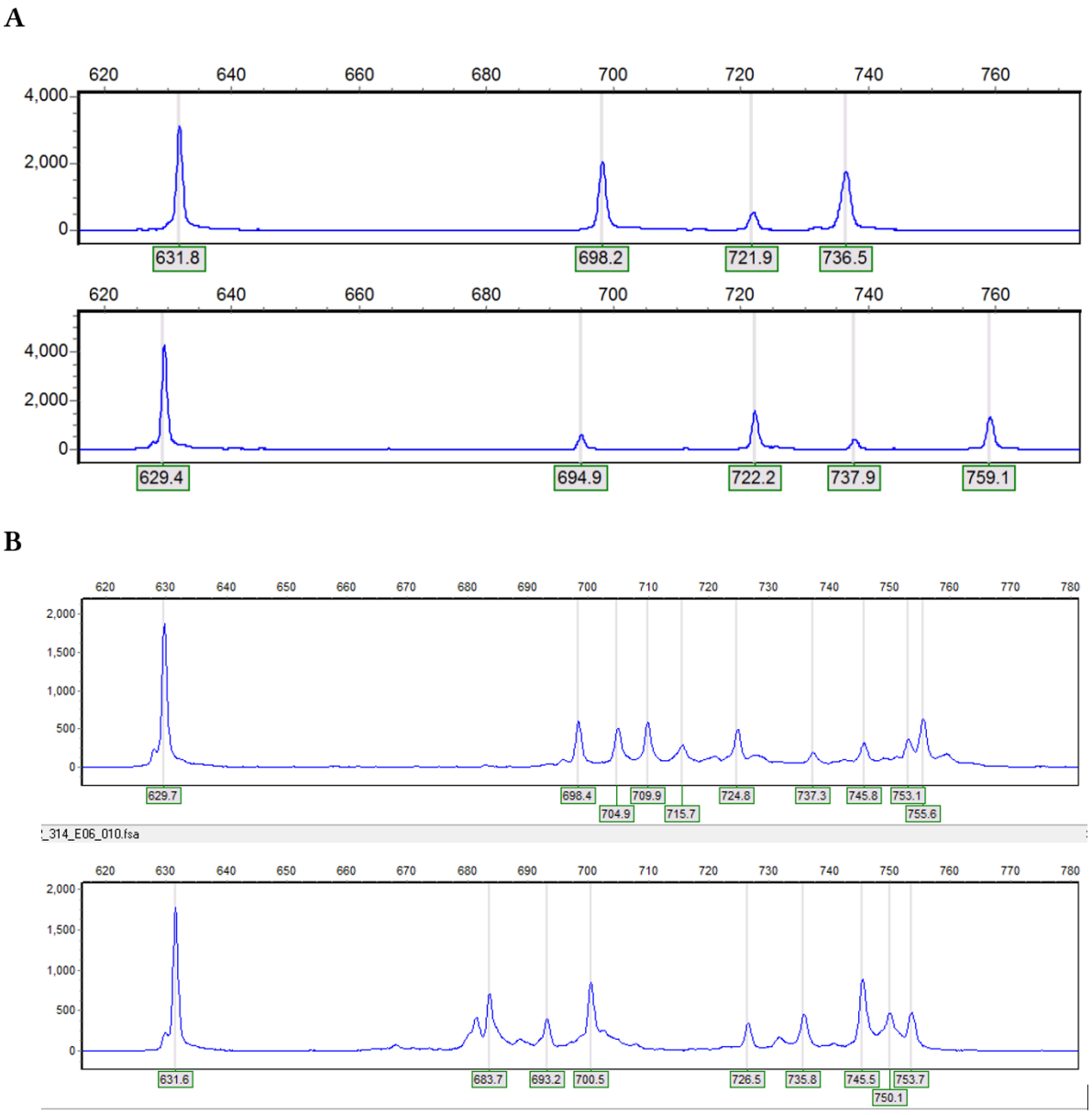


Figure A.2: RSCA chromatograms after hybridization to FLR_3_1_4 (top panel) and FLR_2_2_8 (bottom panel) of A) a genomic DNA sample containing four alleles, and B) a plasmid combination sample containing ten alleles. Each chromatogram shows the homoduplex of the FLR (leftmost peak) and variable number of heteroduplex peaks for the individual alleles.

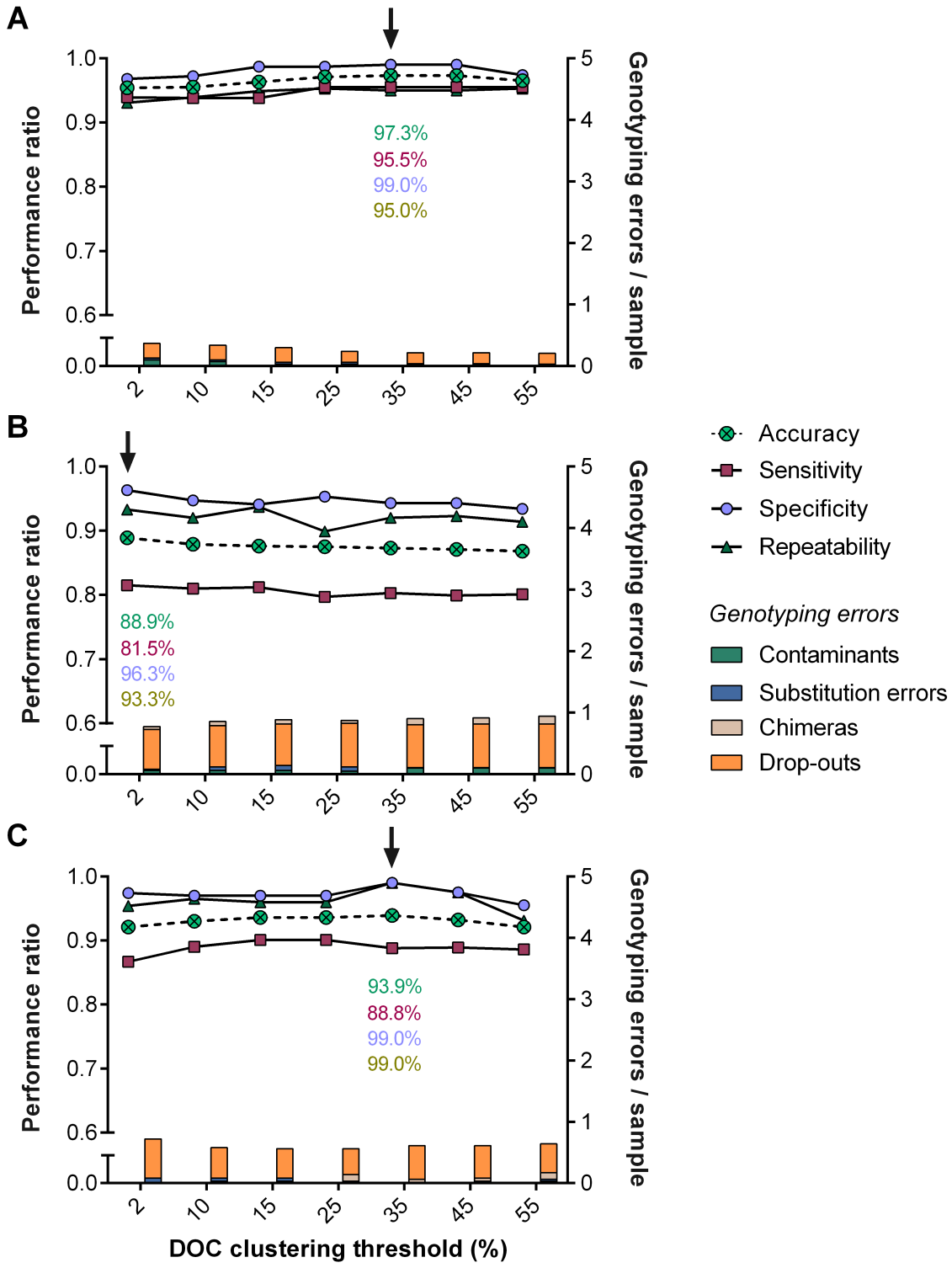


Figure A.3: Genotyping performance of the Degree of Change (DOC) method with the gDNA (A), cDNA (B), and 'artificial' plasmid combination (C) datasets (left y axis), presented alongside corresponding average per sample numbers of genotyping errors (right y axis). Optimal thresholds for each dataset are pinpointed with an arrow.

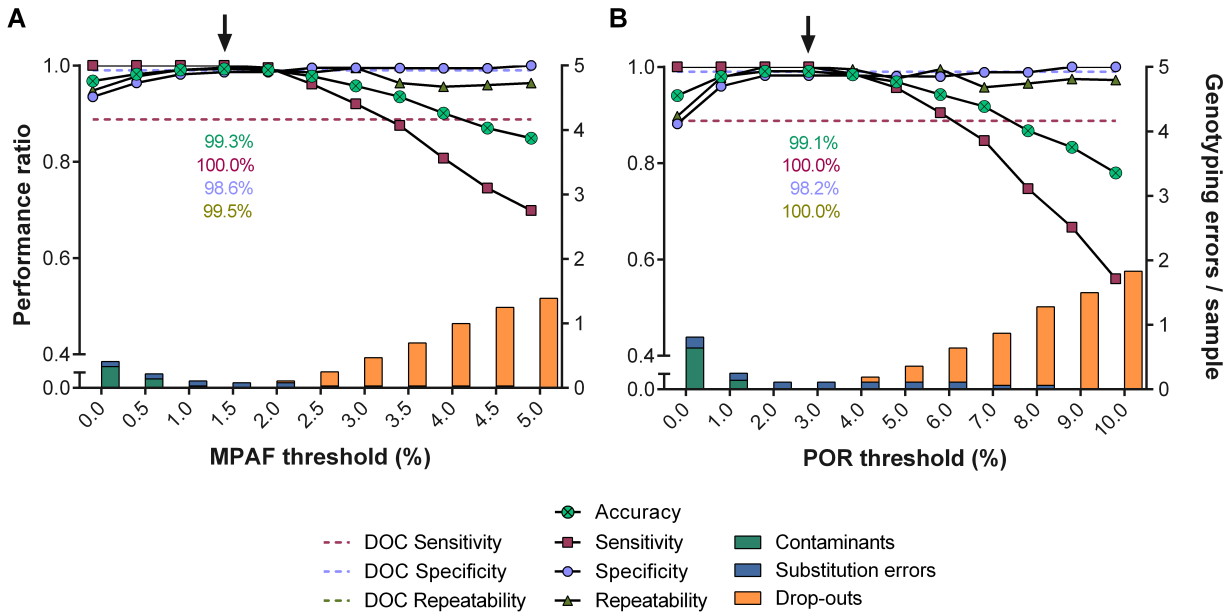


Figure A.4: Genotyping performance of AmpliSAS (A) and ACACIA (B) with the 'artificial' plasmid combination dataset (left y axis), presented alongside corresponding average per sample numbers of genotyping errors (right y axis). As the DOC method doesn't employ an adjustable filtering step, its optimum performance is depicted as horizontal dotted lines. Optimal filtering thresholds are pinpointed with an arrow. POR: percentage of reads filtering threshold of ACACIA. MPAF: Minimum per amplicon frequency filtering threshold of AmpliSAS. Note the truncated y-axis scale. Optimal thresholds for each dataset are pinpointed with an arrow.

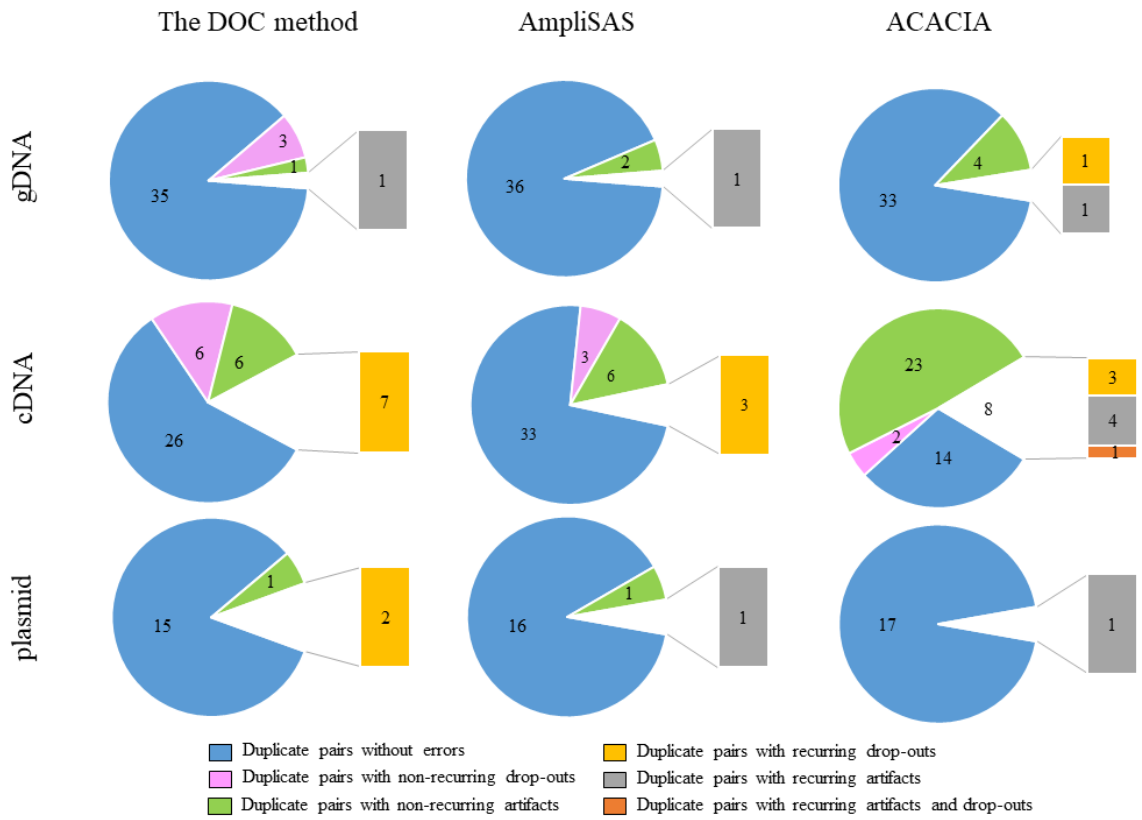


Figure A.5: Occurrence of duplicate pairs with and without recurring genotyping errors across our gDNA, cDNA and plasmid sample subsets for each genotyping pipeline at optimal configuration. The 'missing' part of each pie represents duplicates with recurring errors, which are categorized further based on the type of errors (drop-outs or artifacts) on the side bar chart. Note that the sum of duplicates might differ across pipelines, due to the presence of pairs including more than one error types.

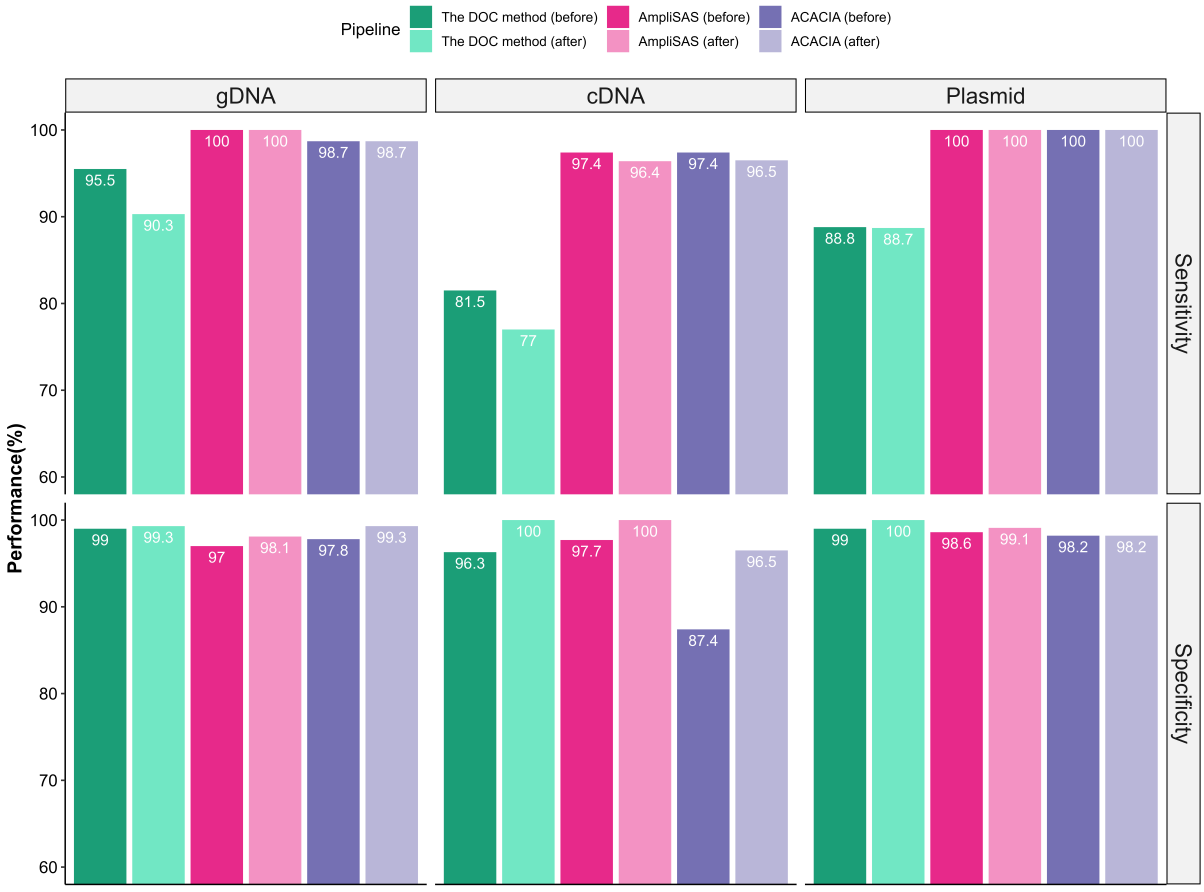


Figure A.6: Genotyping specificity and sensitivity rates of the three pipelines at optimal parameter configurations before and after manual screening of allele calls.

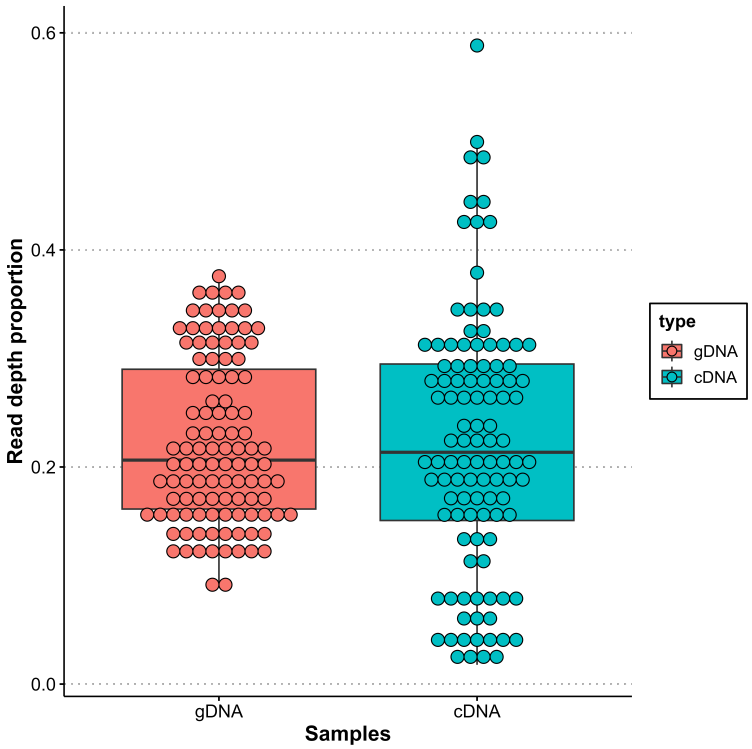


Figure A.7: Comparison of allelic read depth variance between the 24 gDNA and cDNA samples of our "paired" MHC-II dataset. For each allele, read depth proportion was calculated relative to the cumulative read depth of all true alleles within its corresponding gDNA or cDNA sample. Each dot represents the average read depth of a single allele calculated from two replicates. There is a statistically significant difference in the variances between the two sample groups, as confirmed by Levene's test (P = 0.0001236).

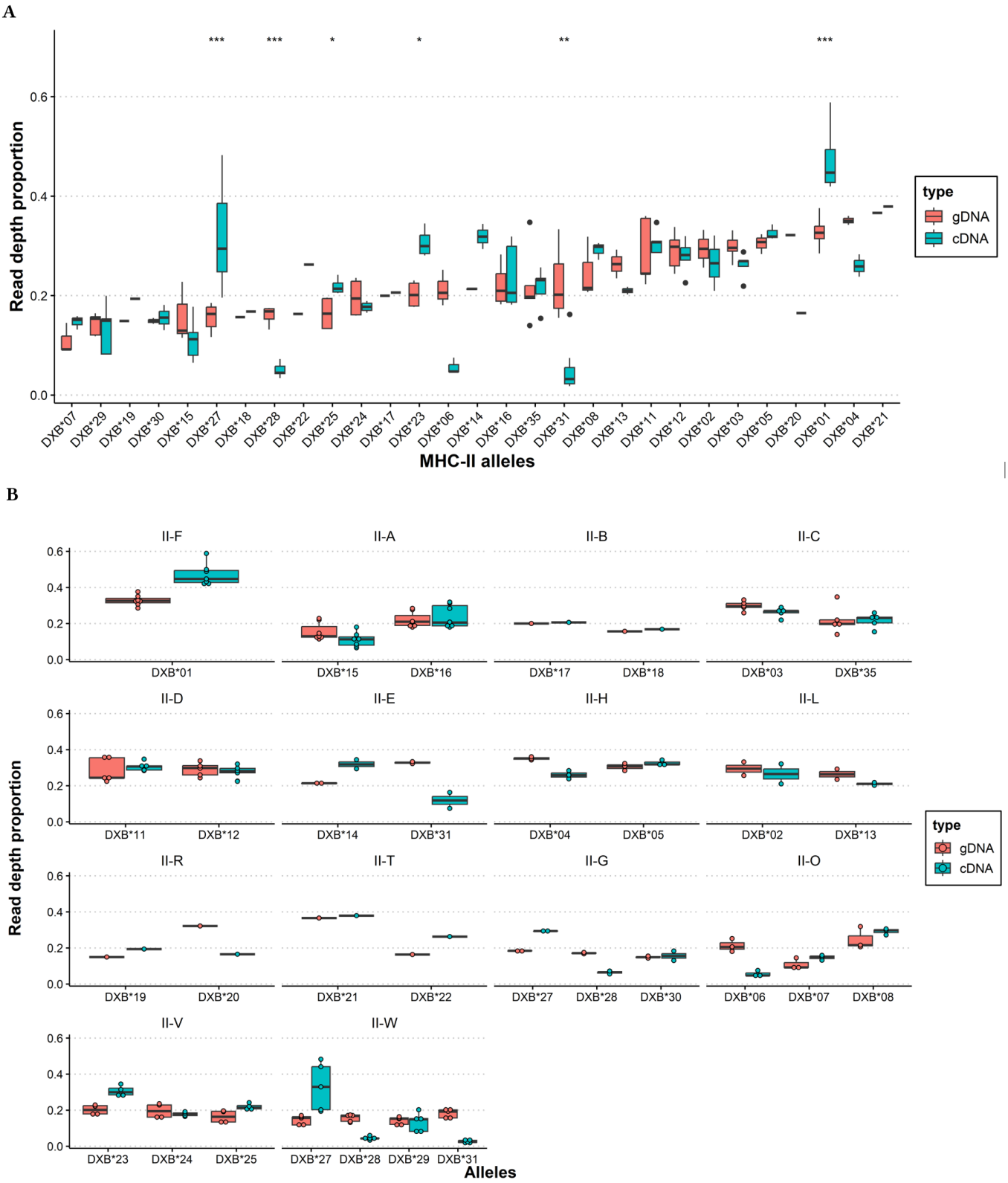
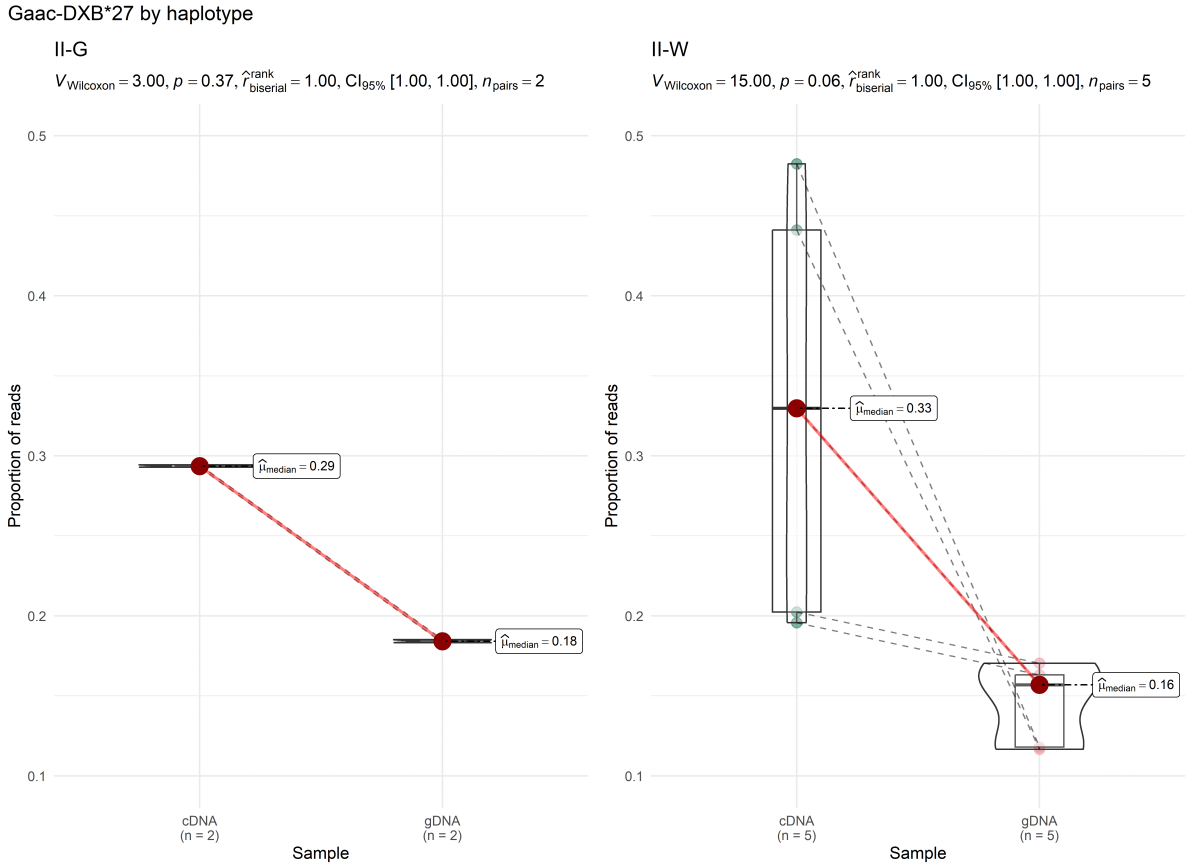


Figure A.8: A) Comparison of sequencing read depth proportions of alleles between gDNA and cDNA samples. Boxplots represent the average read depth of alleles from both replicate samples, calculated as a proportion of the total read depth of all true alleles within each sample. Alleles are arranged along the x axis in order of increasing mean gDNA depth. Asterisks signify statistically significant differences (*P < 0.05, ** P<0.01, *** P<0.001, Wilcoxon signed-rank test). **B) Comparison of read depth proportions of alleles from gDNA and cDNA samples grouped by haplotype.**


Gaac-DXB*27 by haplotype
II-G
$V_{Wilcoxon} = 3.00$, $p = 0.37$, $\hat{r}^{rank}_{biserial} = 1.00$, $CI_{95\%}$ [1.00, 1.00], $n_{pairs} = 2$
II-W
$V_{Wilcoxon} = 15.00$, $p = 0.06$, $\hat{r}^{rank}_{biserial} = 1.00$, $CI_{95\%}$ [1.00, 1.00], $n_{pairs} = 5$
Proportion of reads
0.5
0.4
0.3
0.2
0.1
$\hat{\mu}_{median} = 0.29$
$\hat{\mu}_{median} = 0.18$
$\hat{\mu}_{median} = 0.33$
$\hat{\mu}_{median} = 0.16$
cDNA (n = 2)
gDNA (n = 2)
cDNA (n = 5)
gDNA (n = 5)
Sample

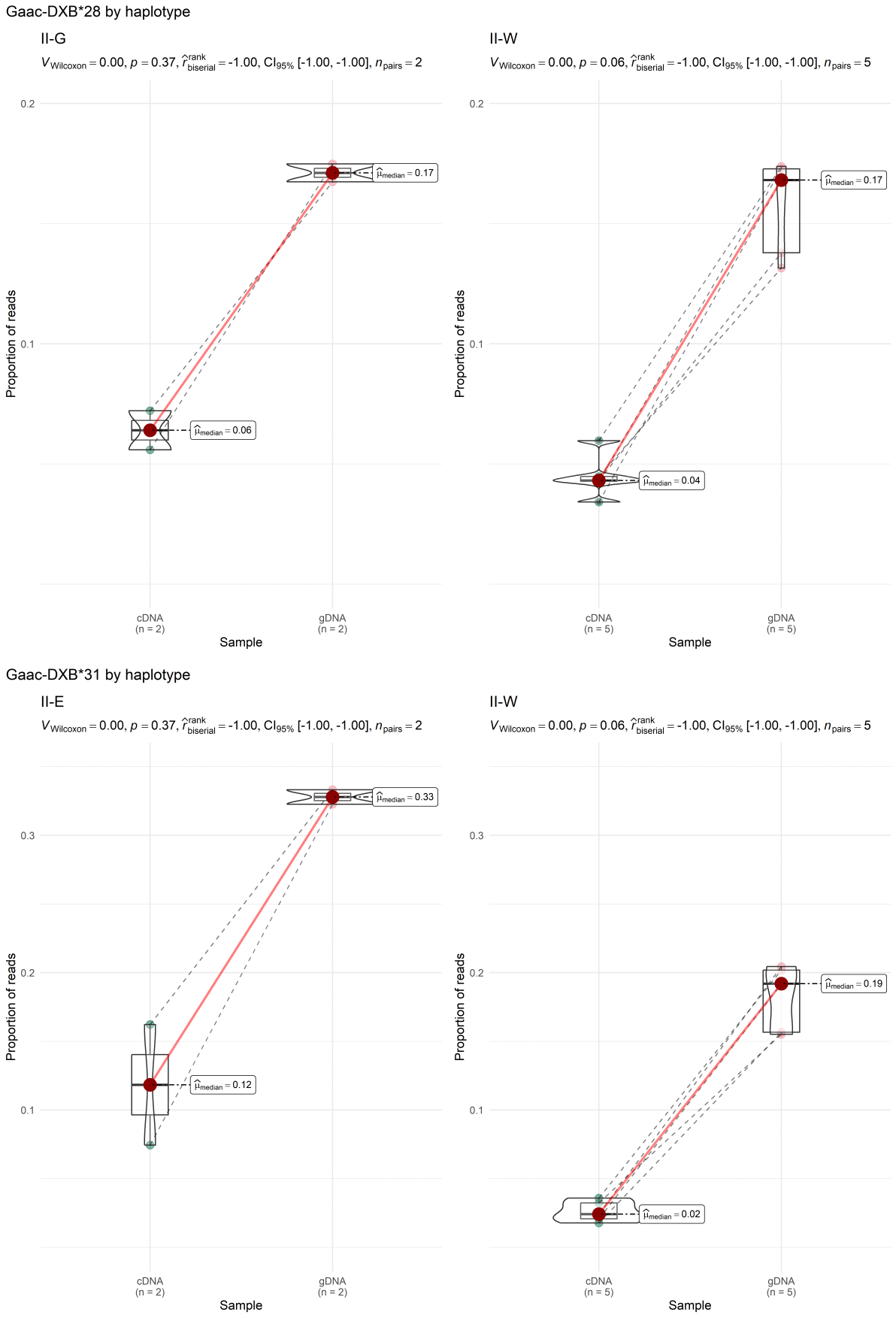


Figure A.9: Comparison of allele sequencing depth between gDNA and cDNA samples for the three alleles (Gaac-DXB*27, Gaac-DXB*28, and Gaac-DXB*31) in our 'paired' MHC-II dataset found in more than one haplotype. Results are grouped by the two haplotypes in which each allele is present. P values from Wilcoxon signed-rank test.

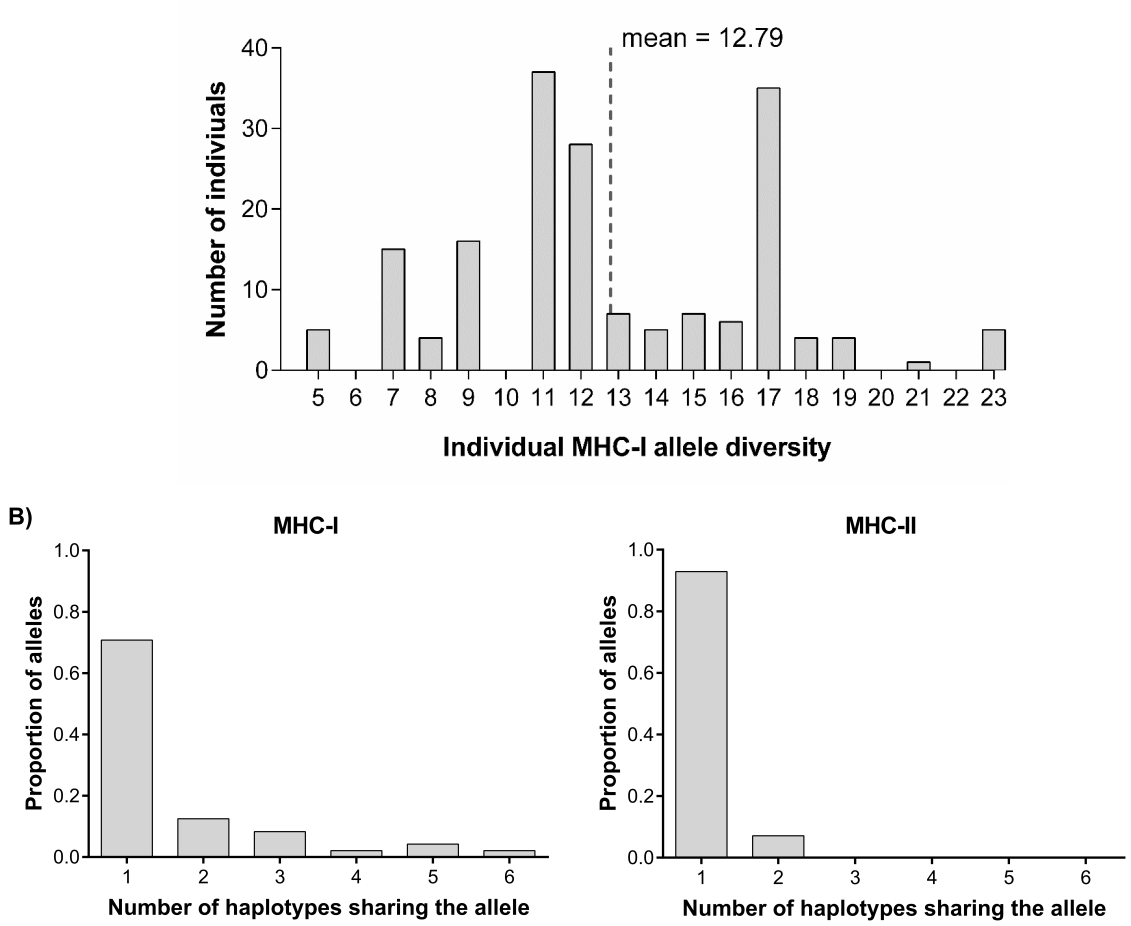


Figure A.10: A) Distribution of individual MHC-I allele diversity across the 179 individuals in our MHC-I dataset. B) Sharing of alleles among all different MHC-I (left) haplotypes. Allele sharing among corresponding MHC-II haplotypes (right) are shown for comparison.

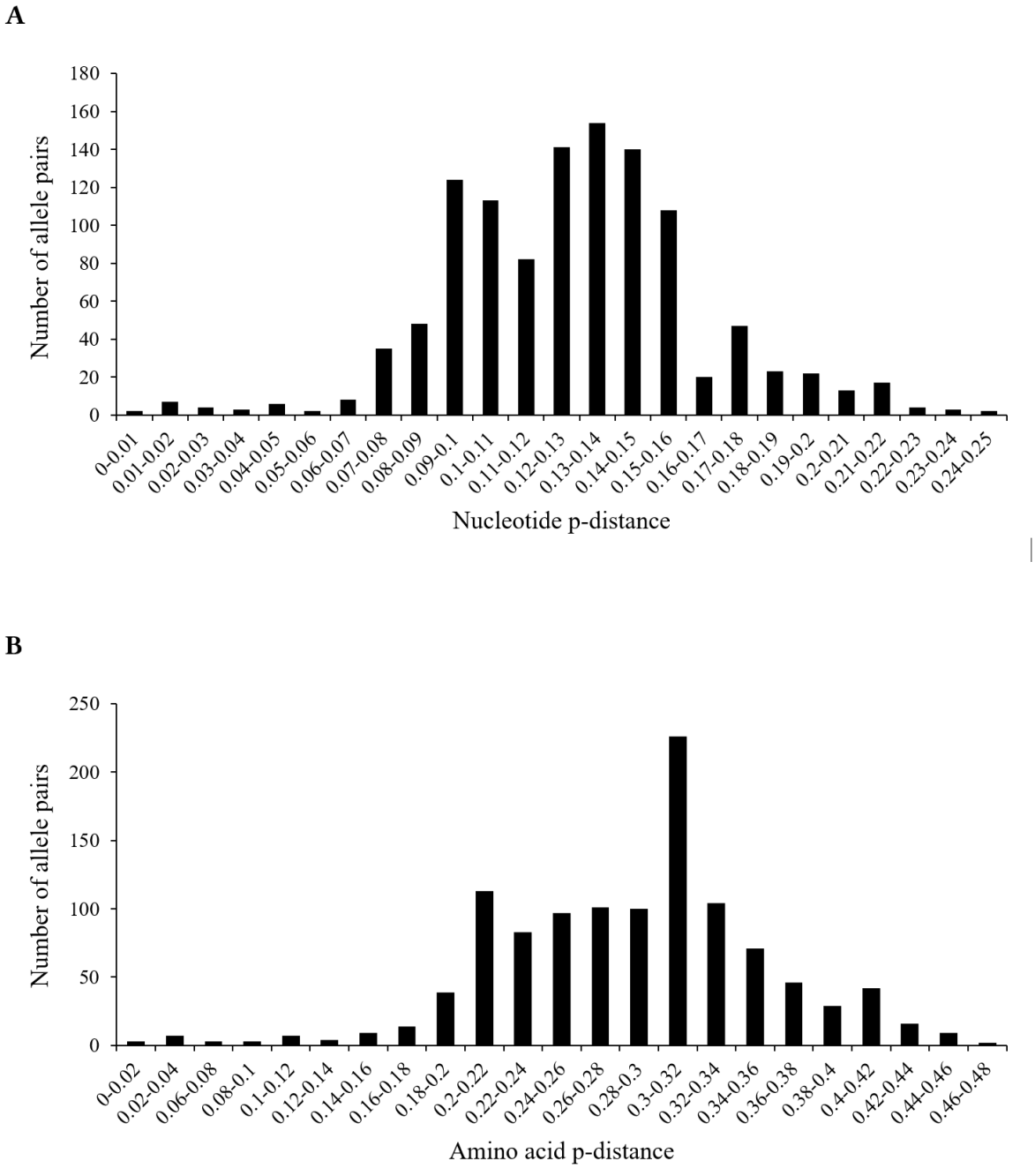


Figure A.11: Histogram of the nucleotide (A) and amino acid (B) p-distance between the 48 MHC-I alleles in the three-spined stickleback.

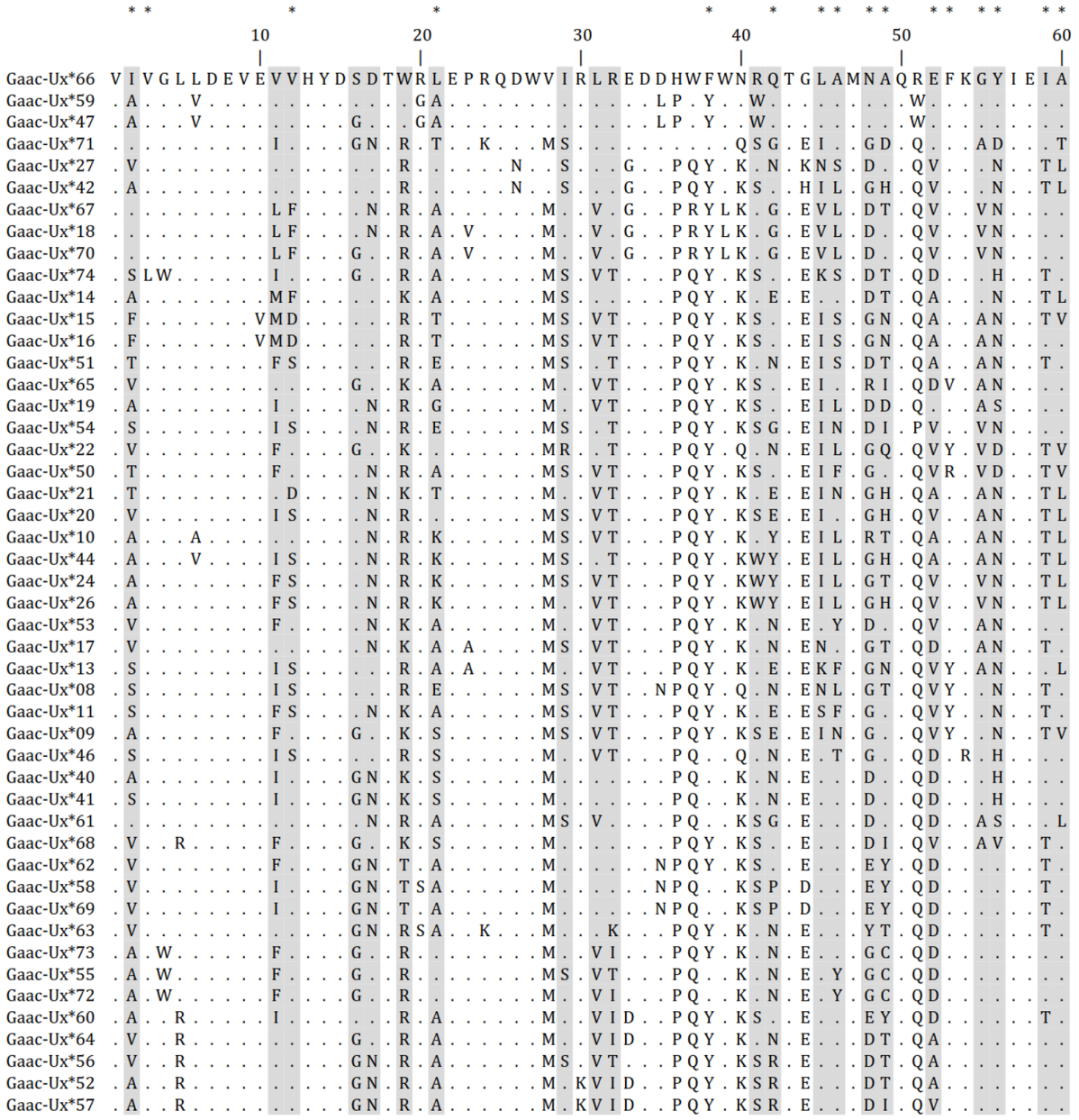


Figure A.12: Alignment of amino acid sequences of the 48 observed MHC-I alleles. Grey sites were identified to evolve under positive selection according to the M8 model (P > 95%). *, residues involved in the peptide binding (based on human HLA structure).

Appendix B

Effect of genetic background and helminth infection on the TCR repertoire of the three-spined stickleback

B.1 Supplementary Materials and Methods

B.1.1 RNA extraction and TCRβ-specific cDNA synthesis with barcoding

Total RNA from the spleen of 114 sticklebacks was extracted using TRIzol reagent according to the manufacturer's protocol. cDNA synthesis was performed by using a 5'RACE protocol with unique molecular identifiers (UMI) coupled with a SmartNNNa 5'-template switch adapter and a reverse primer for the constant regions of the TCRβ chain following (Egorov et al., 2015). 1 μg (or 60% on average) of total RNA from each fish was used per reverse transcription reaction with the SMARTScribe Reverse Transcriptase (Takara Bio) with the stickleback C-segment-specific primer 'Cgen1' (5-TGAAGAAGCTGACGGTGCAGGTG-3') and the SmartNNNa 5'-template switch adapter (5-AAGCAGUGGTAUCAACGCAGAGUNNNNUNNNNUNNNNU-CTT [rGrGrGrG]-3'). The Ns of the template switch adapter represent the UMI. The synthesized cDNA was treated with uracil-DNA glycosylase and purified using the AMPure XP Beads (Beckman Coulter) following the manufacturer's instructions and with a ratio 1 (cDNA product): 1.8 (beads). Finally, purified cDNA was eluted with DEPC-treated water (final volume 40 μl).

B.1.2 Library preparation

Next, we implemented a two-step PCR protocol targeting both TCRβ chain loci (Figure S1). For the first PCR, we used a step-out primer mix that targets the template switch adapter, and a nested primer specific to the C segment of the TCRβ1 or the TCRβ2 gene. Two aliquots of the cDNA of each individual were used as technical replicates. Each reaction included 10 μl of purified cDNA, so that a total of approximately 1/3 of the original total RNA material per fish was subjected to analysis. The cDNA was mixed with 0.02 μM of Step-out1 primer (5- CACTCTATCCGACAAGCAGTGGTATCAACGCAG -3'), 0.2 μM of Step-out2 primer (5-CACTCTATCCGACAAGCAGT -3'), 1× Q5 buffer, 0.2 mM dNTP, 0.2 μM reverse primer 'C1b'

(5- CGTGGTCCGGATAGAATCCGC -3'), 0.2 μM reverse primer 'C2c' (5- CACATTCCTTCG-GCGAAGGTC -3'), 0.02 U/μl Q5 High-Fidelity DNA polymerase (New England Biolabs), and water to 25 μl. The cycler protocol is as follows: 95°C for 1m 30 s followed by 24 cycles of 95°C for 10 s, 60°C for 20 s, and 72°C for 40 s, followed by 4 min final extension at 72°C. Afterwards, the two duplicates were pooled together and purified with AMPure XP Beads (following the manufacturer's instructions (ratio PCR product: beads equal to 1:1). Finally, purified PCR product was eluted with water (final volume 40 μl).

In the second PCR, each target locus (TCRβ1 and TCRβ2) was amplified separately. Illumina sequencing adaptors and heterogeneity spacers were inserted together with the forward ('M1') and reverse ('C1', 'C2') primers as per (Fadrosh et al., 2014). All 2nd PCR primers are listed in Table S4. For each locus, a duplicate reaction was performed for each sample. 3 μl of 1st PCR product were used per 25 μl reaction volume, using 0.4 μM of forward primer 'M1', 0.4 μM of reverse primer 'C1' or 'C2', and the same concentrations of remaining components as described above for the 1st PCR. The cycler protocol was identical to that of the 1st PCR, except for the number of cycles; 16 cycles were employed for C1, while 14 were used for C2. Duplicates for each locus (i.e. C1+C1) were pooled together, purified with AMPure XP beads (ratio 1:1), and eluted with 40 μl of water. Samples were run on an agarose gel and the amount of product (expected band btw 600-700 bp) in each sample was calculated using the software ImageJ. Subsequently, equal amounts of C1 or C2 PCR product from different samples were pooled together. Each of the C1 and C2 pools were size-separated on agarose gels, and the area of the target band was excised. DNA was extracted using the Nucleo Spin MN Gel and PCR clean-up kit (Macherey-Nagel) following the manufacturer's protocol. Lastly, DNA concentration of each library was measured on a NanoDrop™ 3300 Fluorospectrometer (ThermoFisher). Paired-end sequencing was performed in an Illumina NovaSeq 6000 instrument at a read length of 250 bp.

B.1.3 Sample size normalization and diversity analysis

Because of considerable variation in the size (i.e. total UMIs or starting cDNA molecules) of the functional TCRβ1 and TCRβ2 repertoires across individuals (Supplementary Results), a down-sampling step was implemented in order to avoid sampling-related bias. Such bias stems from uneven sampling and sequencing depth, which are known to significantly affect diversity measures (Shugay et al., 2015). Here, we normalized samples to a UMI size, inferred by calculating the 20-th quantile across all TCRβ1 or TCRβ2 samples (Q20), and taking the minimal sample which is above 0.5*Q20, as implemented by MiXCR. Down-sampled repertoires were employed for all downstream analyses with the exception of two TCRβ1 and two TCRβ2 samples that fell below the minimum required size and were excluded from the dataset. Basic comparative post-analyses of the diversity, clonality, V-J gene segment usage, and overlap of individual down-sampled TCR repertoires was carried out with MiXCR version 4.3.2 and VDJtools version 1.2.1 (Shugay et al., 2015). For our diversity analysis we calculated the observed diversity, i.e. number of unique nucleotide sequences in each repertoire. We calculated the mean observed diversity using normalized repertoires, after subsampling each repertoire 10 times. Through the above steps of normalization to a fixed repertoire size and repeated random subsampling, we are confident to yield an interpretable measure that is robust to variation in repertoire size and sequencing depth (Bradley & Thomas, 2019).

B.2 Supplementary Results

B.2.1 Extrapolation of total spleen $\alpha\beta$ T cell repertoire diversity

Hereafter we will use the term TCRβ cDNA molecules (TRBMs) for describing clonotype count units. The size of our TCRβ1 repertoire samples, defined as the sum of TRBMs, ranged from approximately 2,000 to 8,300 TRBMs (mean 4,700 ± 1,200 s.d.), while the TCRβ2 repertoire sample size was less than half the size, between 860-4,200 TRBMs (mean 2000 ± 600 s.d.). We used a nonparametric, incidence-based estimator,Chao1, to deal with the "unseen species" problem in our dataset (Chao, 1984). Chao1 uses abundance data to estimate a lower boundary of TCRβ repertoire diversity within a sample and accounts for missed species due to sampling. We calculated a lower bound total diversity of 7,322±2,733 TCRβ1 and 3,070±991 TCRβ2 unique nt sequences. This amounts to a TCRβ diversity of at least 10^4 TCRβ unique nt sequences in the stickleback spleen. We arrive at a similar result by following a different approach, as described by (Giorgetti et al., 2021): the average number of TCRβ unique nt sequences was 2,264, and given the use of approximately 1/3 of total spleen RNA for sequencing (Materials and Methods), this would point towards a minimum of 7,000 distinct TCRβ nt sequences. However, for a plethora of reasons addressed in the Discussion, this serves as a lower-bound estimate and caution should be taken for its interpretation, as it most likely represents an underestimation of the true spleen TCRβ repertoire diversity of the stickleback.

B.2.2 Rarefaction analysis and size normalization

In order to evaluate the saturation level achieved through our sampling methodology, we performed rarefaction analysis using total TCRβ1 and TCRβ2 repertoire samples. A rarefaction curve shows the expected number of distinct clonotypes, i.e. observed clonotype diversity, discovered as a function of the total number of TRBMs within a sample. For both TCRβ1 and TCRβ2 samples (Figure S4A, B), the decreasing curve slope indicates we have managed to capture a good portion of the existing diversity in our samples, yet the absence of a plateau means we haven't reached complete saturation, i.e., greater sampling and thus higher numbers of TRBMs would have resulted in higher observed diversity. Given the non-saturated rarefaction curves and the considerable variation in repertoire size for both TCRβ1 and TCRβ2 across individuals, as described above, we proceeded to normalize sample sizes by down-sampling TCRβ1 and TCRβ2 repertoires to 2395 and 860 TRBMs, respectively (see Supplementary Materials and Methods).

B.3 Supplementary Tables

Family	Number of offspring	Parental MHC-II genotypes	Offspring MHC-II genotypes	Individual MHC-II diversity level (allele numbers)
A	30	♂CL x ♀CE	CC	Low (2)
			CE, CL, EL	Intermediate (4)
B	30	♂AE x ♀CC	AC, CE	Intermediate (4)
C	30	♂DW x ♀AE	AD, DE	Intermediate (4)
			EW	High (5)
			AW	High (6)
D	30	♂AN x ♀AC	AA	Low (2)
			AC	Intermediate (4)
			AN, CN	High (5)
E	30	♂AC x ♀AA	AA	Low (2)
			AC	Intermediate (4)
F	30	♂AD x ♀CE	AC, AE, CD, DE	Intermediate (4)

Table B.1: Overview of lab-bred stickleback families and their MHC-II genotypes used in this study.

Family	Control	GOT-treated	NO-treated	Total infected
A	10	10 (3)	10 (5)	8
B	10	10 (3)	10 (0)	3
C	10	10 (4)	10 (4)	8
D	10*	10 (4)	10 (4)	8
E	10	10 (2)	10 (3)	5
F	10	10 (2)	10 (8)	10
Sum	60*	60 (18)	60 (24)	42

*2 control fish died before the end of the experiment and were excluded from downstream analyses.

Table B.2: Summary table of the experimental design. Sample sizes and infection outcome are presented by fish family and *S. solidus* strain. In parentheses are the numbers of *Schistocephalus solidus*-infected fish within the GOT-treated and NO-treated groups.

Family \ Treatment	Control		Exposed		Infected		Total
	GOT	NO	GOT	NO	GOT	NO	
A	3	3	3	3	3	5	20
B	3	3	3	3	3	0	15
C	3	3	3	3	4	4	20
D	1	5	3	3	4	4	20
E	3	3	3	3	2	3	17
F	3	3	4	2	2	8	22
Sum	16	20	19	17	18	24	**114**

Table B.3: Numbers of fish used for TCRβ sequencing by infection status, *S. solidus* strain treatment group, and fish family.

Forward Primers	Nucleotide sequence (5' → 3')
M1A	ACACTCTTTCCCTACACGACGCTCTTCCGATCT*AAGCAGTGGTATCAACGCAGAGT*
M1B	ACACTCTTTCCCTACACGACGCTCTTCCGATCT*AAGCAGTGGTATCAACGCAGAGT*
M1C	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**T***AAGCAGTGGTATCAACGCAGAGT*
M1D	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**T***AAGCAGTGGTATCAACGCAGAGT*
M1E	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**GT***AAGCAGTGGTATCAACGCAGAGT*
M1F	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**GT***AAGCAGTGGTATCAACGCAGAGT*
M1G	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**CGA***AAGCAGTGGTATCAACGCAGAG T*
M1H	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**CGA***AAGCAGTGGTATCAACGCAGAG T*
M1I	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**ATGA***AAGCAGTGGTATCAACGCAGA GT*
M1J	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**ATGA***AAGCAGTGGTATCAACGCAGA GT*

Reverse Primers	Nucleotide sequence (5' → 3')
C1A	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT*CGATGGTCTTCCTCGCCTTGTC*
C1B	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT*CGATGGTCTTCCTCGCCTTGTC*
C1C	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**A***CGATGGTCTTCCTCGCCTTGTC*
C1D	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**A***CGATGGTCTTCCTCGCCTTGTC*
C1E	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**TC***CGATGGTCTTCCTCGCCTTGTC*
C1F	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**TC***CGATGGTCTTCCTCGCCTTGTC*
C1G	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**CTA***CGATGGTCTTCCTCGCCTTGTC*
C1H	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**CTA***CGATGGTCTTCCTCGCCTTGTC*
C1I	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**GATA***CGATGGTCTTCCTCGCCTTGT C*
C1J	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**GATA***CGATGGTCTTCCTCGCCTTGT C*
C1K	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**ACTCA***CGATGGTCTTCCTCGCCTTG TC*
C1L	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**ACTCA***CGATGGTCTTCCTCGCCTTG TC*

C2A	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT*GAAGGTCGGAGCACTCGGACG*
C2B	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT*GAAGGTCGGAGCACTCGGACG*
C2C	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**A***GAAGGTCGGAGCACTCGGACG*
C2D	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**A***GAAGGTCGGAGCACTCGGACG*
C2E	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**TC***GAAGGTCGGAGCACTCGGACG*
C2F	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**TC***GAAGGTCGGAGCACTCGGACG*
C2G	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**CTA***GAAGGTCGGAGCACTCGGACG*
C2H	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**CTA***GAAGGTCGGAGCACTCGGACG*
C2I	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**GATA***GAAGGTCGGAGCACTCGGAC G*
C2J	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**GATA***GAAGGTCGGAGCACTCGGAC G*
C2K	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**ACTCA***GAAGGTCGGAGCACTCGG ACG*
C2L	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**ACTCA***GAAGGTCGGAGCACTCGG ACG*

Table B.4: TCRβ primers used in the second PCR step of our NGS protocol. Underlined nucleotides represent the multiplexing read 1/2 sequencing primers. Heterogeneity spacers are highlighted in bold. TCRβ-specific sequences are shown in italics.

TRB1 locus	
Gene	Coding sequence (5' → 3')
V1.1*	GTCTTCCAAACGCCTCCGAGCATCATTAAGGAACTTGGGCAATCTGTTGTCGGCAAGG AGATCAGCTGTTCACACAACATGCCAAATTACGACGTATTCCTCTGGTACAAAAGAGA AAAAGAACGAGGAGATCTAAAGCTCTTGGGTTTCCTGAATGTGGACTACCCAAACCCA GAGGACGACGTAAAAGGCAAAATAGACTTGAAAGGAGACGCCCGTGATATTGCAACC CTCACCATTTCGGATCTCGTACTAAATGACACTGGCGTGTACTTCTGTGCAGCCAGTCA G
V1.2	GATGGCAGTGATGTCACCCAGACTGCGCTGCTGTGGAAAGACAAGGGTGCCGAGGCA ACAATGAACTGCAGCCACACCAAGGGGAGCAGCTACTACCAGATGTACTGGTACAGA CAGCTGCCAGGAGAAACTATGAAACAGATCGTGTACACGTCCTTGGGCGTGAATCATG ATTTTGGAAATTTCAGCAAGGAGAAATTTTCAGCCACCAAGCCCGACGCTGACAGTGG GACGTTCACAGTGAAGAACCTGCAGCCGCAAGATAAGGGCTTGTACTTCTGTGCTGTG AGCCAA
V1.3	AGTGACCAAGTTCACCAGACTCCAGCTGATATTTACAACAAACCCACAGAAGAAGCCA CAATCAACTGTTCACACAGTATTCCCAACTACGATCGAATCTTCTGGTACAAACTGTCC AGGGACAACAAACTGCAGTTTCTGGGATACGTGTTAAAAACCCTCAGAATCCCAGAAG CTGGACTAGGCGTAAAGATGGATGGGAGTGCTGATAAAAACCAAATGTGCCAATTAA GAATTGAGGGCCTGGTGCTGAACAGTAGCGCGGTGTACTTCTGTGCTGCTAGTCTA
V1.4†	TGCCAGGGTGTGACCCAACACCCTGAAATCAACTGGAGTCTTGTTTCCAAATCGGCAG AGATGAACTGCAGCCACAACAAAGATGCCGGTCACAACCCAGATGTACTGGTACAGA CAGCGTCCAGGAGAGACCATGACCCTGGTGGTC
V1.5	AGCGACCAGGTGCACCAGACTCCGTCTGAGATGTTCACCAGACCTGAAGGAACAGCCC ATATCTACTGCTCACACAGCATTAAGGACTACGATAAAATCCTCTGGTACAAACAATC CAGGGATGACGCGATGCTGTTCCTGGGTTACATGTTTGCAAATAATAAAAACCCAGAG GTTGGACTGGGTGTGAATATGACAGGGAGTGCAAATCAACATCAGACCTGCACATTGA CAATGGAGAGAGTGGCCCTGAACAGCAGTGCTGTGTACTTCTGTGCTGCGAGGTCA
V1.6	AGTGACCAAGTTCACCAGACTCCAGCTGATATTTACAACAAACCCACAGAAGAAGCCA CAATCAACTGTTCACACAGTATTCCCAACTACGATCGAATCTTCTGGTACAAACTGTCC AGGGACAACAAACTGCAGTTTCTGGGATACGTGTTAAAAACCCTCAGAATCCCAGAAG CTGGACTAGGCGTAAAGATGGATGGGAGTGCTGATAAAAACCAAATGTGCCAATTAA GAATTGAGGGCCTGGTGCTGAACAGTAGCGCGGTGTACTTCTGTGCTGCTAGTCTA
V1.7	TGCCAGGGTGTGACCCAACACCCTGAAATCAACTGGAGTCTTGTTTCCAAATCGGCAG AGATGAACTGCAGCCACAACAAAGATGCCGGTCACAACCAGATGTACTGGTACAGAC AGCGTCCAGGAGAGACCATGACCCTGGTGGTCTACACAGTTTTTGGCGGACAGCCAGA TTACGGGAAGCTCCCTCAGACCAAATATTCAGCCGTTAAAGAAAGTTCTGAGAGCGGC GCTCTTACTGTGAAGGACTTGCAGCTGGAGGACGCCGGCGTGTATTTTTGTGCTGTCAG CAAA
V1.8	AGCGACCATGTGCACCAGACTCCATCTGAGATGTTCACCAGACCTGAAGGAACAGCCC ATATCTACTGCTCACACAGCATTAAGGACTACAATAAAATCCTCTGGTACAAACAATC CAGGGATGACGCGATGCTGTTCCTGGGTTACATGTTTGCAAATAATAAAAACCCAGAG GTTGGACTGGGTGTGAATATGACAGGGAGTGCAAATCAACATCAGACCTGCACATTGA CAATGGAGAGAGTGGCCCTGAACAGCAGTGCTGTGTACTTCTGTGCTGCGTGGTCA
V1.9†	GCTCGTCTCTCAGTGACCAAGTTCACCAGACTCCAGCTGATATTTACAACAAACCCAC AGAAGAAAACGTGACCGTTCACACAGCATCGACACCCACAACCAGATCCTGGCTTAAT TCAATCCAATTCAATCCATTTTATTTTGTATAGCCCATAAT

V1.10 TGCCAGGGTGTGACGCAACACCCTGAAATCAACTGGAGTCTTGCTTCCAAATCGGCAG
AGATGAACTGCAGCCACAACAAAGATGCCAGTCACACCCAGATGTACTGGTACAGAC
AGCGTCCAGGAGAGACCATGACCCTGGTGGTCTACTCAGTTTTTGGCGGACAGCCAGA
TTACGGGAAGCTCCCTCAGACCAGATATTCAGCCGTTAAAGAAAGTTCTGAGAGCGGC
GCTCTTACTGTGAAGGACTTGCAGCTGGAGGACGCCGGCGTGTATTTTTGTGCTGTCAG
CAAA

V1.11 CTTTTAATCAAACAATCACCAGTGCATGTTTTGATTAAACCTGATCAAGTAAAAGCAC
ATTTTGGCTGTTACCATGGCGACAACAACTACCCCTACATGCTTTGGTACCGGTCGGCA
GCGGGAGGCCGGAGGGACATGGAGCTGATCGGACTGCTCCACTATGAAAACGCAAAC
CTTGAGAAGAACTTTGAGGCCCGTTGCAACATGACGGGCCACTCAAAGGGCCGAGCTG
AGCTACTCATCTCCAACGCGAAGCCAGCAGACAGCGCTGTGTACTTCTGTGCAGCGAG
TCAG

V1.12 AGCGACCAGGTGCACCAGACTCCATCTGAGATGTTCACCAGACCTGAAGGAACAGCCC
ATATCTACTGCTCACACAGCATTAAGGACTACGATAAAATCCTCTGGTACAAACAATC
CAGGGATGACGCGATGCTGTTCCTAGGTTACATGTTTGCAACTAATAAAAACCCAGAG
GTTGGACTGGGTGTTAATATGACAGGGAGTGCAAATCAACATCAGACCTGCACATTGA
CAATGGAGAGAGTGGCCCTGAACAGCAGTGCTGTGTACTTCTGTGCTGCGTGGTCA

V1.13 AGTGACCAAGTTCACCAGACTCCAGCTGATATTTACAACAAACCCACAGAAGAAGCCA
CAATCAACTGTTCACACAGTATTCCCAACTACGATCAAATCCTCTGGTACAAACTATCC
AGGGACAACAAACTGCAGTTTCTGGGATACGTGTTAAAAACCCTCAGAATCCCAGAAG
CTGGACTAGGCGTAAAGATGGATGGGAGTGCAGATAAAGACCAAATGTGCCAATTGA
GAATTGAGGGCCTGGTGCTGAACAGTAGCGCGGTGTACTTTTGTGCCGCTAGTCTA

V1.14 TGCCAGGGTGTCACGCAACACCCTGAAATCAACTGGAGTCTTGCTTCCAAATCGGCAG
AGATGAACTGCAGCCACAACAAAGATGCCAGTCACACCCAGATGTACTGGTACAGAC
AGCGTCCAGGAGAGACCATGACCCTGGTGGTCTACACAGTTTTTGGCGGACAGCCAGA
TTACGGGAAGCTCCCTCAGACCAAATATTCAGCCGTTAAAGAAAGTTCTGAGAGCGGC
GCTCTTACTGTGAAGGACTTGCAGCTGGAGGACGCCGGCGTGTATTTTTGTGCTGTCAG
CAAA

V1.15 CTTTTAATCAAACAATCACCAGTGCATGTTTTGATTAAACCTGAACAAGTAAAAGCAC
ATTTTGGCTGTTACCATGGCGACAACAACTACCCCTACATGCTTTGGTACCGGTCGGCA
GCGGGAGGCCGGAGGGACATGGAGCTGATCGGACTTCTCAGCTATGGAAACGCAAAC
CTTGAGAAGAACTTTGAGGCCCGTTGCAACATGACGGGCCACTCAACGGGCCGGGCTG
AGCTCGTCATCTCCAACGCGAAGCCAGCAGACAGCGCTGTGTACTTCTGTGCAGCGAG
TCAG

V1.16* TCCGTGCTGATCACCCAGTGGCCCCGCTACATCTCCAGTCCTCCCAACGGCTCTGCAGA
AATGACCTGCTACCAGAATGACACCAACTATGACTATCTGTACTGGTACCAACAGCTG
AGTGGAAGGAGCTTTCAGTTAATAGTGTCTACAGTAGGTGGCTTTTCAAGCTTTGAAG
TCGGATTCAAGTCTGGTTTTCAGTCGGTGAAGTTCAGTGAGAAGCAATGGTCTCTGAC
AGTAACCGGCCTTCAGGGGAATCACGCGGCTGTCTATCTGTGTGCTGCCAGTCTA

V1.17 AGCGCTGCCACATTCCAACAATCTCCCCCCCCAGATAGTGGAGGAGAGCACTGAGGTCC
ACATAAATTGCAGCCACGACGACTCTAACCTGGTTGTAATGCTGTGGTATCAGCAGAG
ACGGGACAATGGGTCCTTGGCCCTCATAGGCTACGGGTGGGTAGGCAGCCCAAACTAT
GAGGGTCAGTTTAAAGAACAGTTTGAGCTGACGAGAGCAGACACAATGAGAGGAGCT
CTGGTTGTGCGCACAGCGGCCCCGTCACACTCGGCTGTTTATTTCTGCGCTGCCAGTA

V1.18 AAGAGTGTGAGGTTTGAGCCACAGTCTCCCAGAATAGTCAACGACAAAGCCAGGGTG
GAAATAAAATGCAGTCACAATGACAGCAGTTTTTATGTAATGCTCTGGTACCAGCAAA
CCCAGAGAGGTCTGGTTCATTTCATCGGATACAGCTACGCTGGCAGTGCGCCCGTGTA
TGAGAAGCAGTTTGAAGTTGGCTTCGAGATAACGAGGGAAGGCACCCAAACGGGGGC
TCTGGTTGTTGCCAGGGCGAATCAATCCGACTCGGCTGTGTATTTCTGCGCTGCCAGTA

V1.19 CTTGCGGTTGTAGTCCTTCAGTCTGGAGATCAAACGTCCCACCCGGGTCACGCGGTGA
GCTTCGTGTGCAACATGGGTCAAGGGGTCAGCATGGGCAGCTACACCATGCTCTGGTA

	CCGACAAACCCACCGTGGAGCTCCGCTGGACTTCCTCTCCAAGGAATACGACCAGACC ACGGGGCACTTCCAGTCGTCCATCGACACATCCAGAAACAACTTCTCTCTTCAAATTAC TGAGGTGCTCCTGAATGACAGCAGTACCTACTACTGTGCTGCCAGT
D1.1	GGGACAGGGGG
J1.1	ACACTGCATCTGAGGCTTACTTTGGAAAGGGGACAAAACTGACTGTTTTGG
J1.2	AGCTACAGCGAACCTGCTTACTTTGGCCAAGGAACCAAACTGACTGTTTTGG
J1.3	ACAACAACGAGGCTTATTTTGGAAATGGAACCAAACTGACTGTTTTGG
J1.4	TCAGTACGAAGCCTATTTCGGCCAGGGAACGAAACTGACTGTTCTCG
J1.5	ACTATGGGGCTTCAGAGGCTTACTTTGGTGGAGGAACAAAGTTAACGGTTTTAG
J1.6	TTCCTACGCCGAAGCAATCTTTGGAGCTGGAACTAAACTCACGGTTTTAG
J1.7	AGTATACTGAGGCTTACTTTGGAGCTGGAACTAAACTCACTGTTTTAG
J1.8	AATAACTACGAACCTGCGTACTTTGGCAGCGGCACCAGACTGACGGTGTTGG
TRB2 locus	
V2.1	GGCGGCAGCGACGTCGTCCAGAACCCGATGCTCTGGGAAGTCGAGGGTGTTGATGCAA CAATGAACTGCACACACACCAAGGGGGCTCAGTTCTTCCAGATGTACTGGTACAGACA GCTGCCAGGAGAAACCATGAAGCAGATCGTGTACACGTCCACCGCCAGCGCAGATCA CGATTTTGGAAATTTCAGCAAAGACAAGTTTTCCGCCACCAAGCCGGACGCCGCCAGC GGGACGCTGACGGTGAAGAAGCTGCAGCAGGGAGACAAAGGCTGGTACTTCTGTGCA GTGAGTGAA
V2.2	AGCACTGAGGTCCACCAGGATCCACCTTCACTATTTGGGACATCTCTTGTTCCAGCAAC AATCAGCTGCTCCCATTCCATCAGTTCATACTATACAATACTGTGGTACCAGCAGCCGG TGGGCGACTCCAGCCTGAAACTCATCGGATATGTTCTGTATAACAACCCGACCATTGA GGAAACATTCAAATCGCAATTTAATGTGACAGGAGACGGCTCATCGCGATCTGAGCTG CAGGTGCTGAACCCCACCAGCGGCGCGTACTTCTGTGCGGCTAGTAGG
V2.3	AAGAACAGTGTTGCTCAGAGTCCCGGTGCTCTGATCAAACACCCCGGAGAAGAAGTCC GGATAGATTGTAACCACTCGAACACGGATTTCGATATGATCCAGTGGTACAAACAGTC TGCCGGGGAGAGCGACATGCTCCTGGTTGGCTACGTACGATTCACCTCCACAGTCGTG GAAGCTCCGTTCAAGGACTCGTACAACGTGAGCGGTGACGGGGAAAGTCGCGCTTGTC TCCACATCCCAAAGCCGGGTGTCCCGGAGGACGGGGCCGTGTACTTCTGCGCTGCCAG TAGAG
V2.4	GGTCTAGATGTCCGTCAGTCCGTCTCAGATCTCATCATGAAACCTGGGGACGAAGAGC AGATTTCCTGCAGCCATGACAGGACCGACTACAGGGTGATGCTCTGGTACCAGCAGTC CCCTGGAGACACGGCTATGAAACTCATTGGATATTTGCACTATAAAGCTCCCACAATG GAAGAGTCTTATAAGGATCATTTTATTATCTCTGGAGATTTGGGAGGAACTACAGCAA AGAATGGCTCGCTTGTAGTCAGAGGTGTAAAACAGGAGCACAGTGCAGTTTACTACTG TGCAGCGTGTGAAG
V2.5	GGCGGCAGCGACGTCGTCCAGACTTCGATGCTCTGGGAAGTCGAGGGTGTTGATGCAA CAATGAACTGCACACACACCAAGGGACCTGGGTTCTTCCAGATGTACTGGTACAGACA GCTGCCAGGAGAAACCATGAAGCAGATCGTGTACACGTCCTCCGCCAGCGCGGATCAC GATTTTGGAAATTTCAGCAAAGACAAATTTTTCCGCCACCAAGCTGGACGCCGCCAGC GGGACGCTGACGGTGAAGAAGCTGCAGCAGGGAGACAAAGGCTGGTACTTCTGTGCA GTGAGTGAA
V2.6	AGCACTGAGGTCCACCAGGATCCACCTTCACTATTTGGGACATCTCTTGTTCCAGCAAC AATCAGCTGCTCCCATTCCATCAGTTCATACAATACAATACTGTGGTACCAGCAGCCG GTGGGCGACTCCAGCCTGAAACTCATCGGATATTCTTATTTTAACAACCCGACCATTGA GGACACATTCAAATCGCAATTTAATGTGACAGGACACGGCTCATCGCGATCTGAGCTG CAGGTGCTGAATCCCACCAGCGGCGCGTACTTCTGTGCGGCTAGTAGG
V2.7	AAGAACAGTGTTGCTCAGAGTCCCGGTGCTCTGATCAAACACCCCGGAGAAGAAGTCC GGATAGATTGTAACCACTCGAACACGGATTTCGATATGATCCAGTGGTACAAACAGTC TGCCGGGGAGAGCGACATGCTCCTGGTTGGCTACGTACAATTCACCTCCACAGTCGTG

	GAAGCTCCGTTCAAGGACTCGTACAACGTGAGCGGTAACGGGGGAAGTCACGCTTGTC TCCACATCCCAAAGCCGGGTGTCCCGGAGGACGGGGCCGTGTACTTCTGCGCTGCCAG TAGAG
V2.8	GGTCTAGATGTCCGTCAGTCCGTCTCAGATCTCATCATGAAACCTGGGGACGAAGAGC AGATTTCCTGCAGCCATGACAGGACCGACTACAGGGTGATGCTCTGGTACCAGCAGTC CCCTGGAGACACGGCTATGAAACTCATTGGATATTTGTACTTTAAAGATCCCACAATG GAAGAGTCTTATAAGGATAATTTTAATATCTCTGGCGATTTGGGAGGAAACACAGCAA AGAATGGCTCGCTTGTAGTCAGAGGTGTAAAACAGGAGCACAGTGCAGTTTACTACTG TGCAGCGCGTGAAG
V2.9	GGCGGCAGCGACGTCGTCCAGAACCCGATGCTCTGGGAAGTCGAGGGTGTCGATGCA ACAATGAACTGCACACACACCAAGGGGCCTCTGTTCTTCCAGATGTACTGGTACAGAC AGCTGCCAGGAGAAACCATGAAGCAGATCGTGTACACGTCCACCGCCAGCGCGGATC ACGATTTTGGAAATTTCAGCAAAGACAAATTTTCCGCCACCAAGCCGGACGCCGCCAG CGGGACGCTGACGGTGAAGAAGCTGCAGCAGGGAGACAAAGGCTGGTACTTCTGTGC AGTGAGTGAA
V2.10	AGCACTGAGGTCCACCAGGATCCACCCTCACTGTTTGGGACATCTCTTGTTCCAGCAAC AATCAGCTGCTCCCATTCCATCAGTTCATACTATACAATACTGTGGTACCAGCAGCCGG TGGGCGACTCCAGCCTGAAACTCATCGGATATGTTCTGTATAACAACCCGACCATTGA GGAAACATTCAAATCGCAATTTAATGTGACAGGACACGGCTCATCGCGATCTGAGCTG CAGGTGCTGAACCCCACCAGCGGCGCGTACTTCTGTGCGGCTAGTAGG
V2.11	AAGAACAGTGTTGCTCAGAGTCCCGGTGCTCTGATCAAACACCCCGGAGAAGAAGTCC GGATAGATTGTAACCACTCGAACACGGATTTCAATATGATCCAGTGGTACAAACAGTC TGCCGGGGAGAGCGACATGCTCCTGGTTGGCTACGTACGATTCAACTTCCCGGTCTTG GAAGCTCCGTTCAAGGACTCGTACAACGTGAGCGGTAACGGGGGAAGTCGCGCTTGTC TCCACATCCCAAAGCCGGGTGTCCCGGAGGACGGGGCCGTGTACTTCTGCGCTGCCAG TAGAG
V2.12	GGTCTAGATGTCCGTCAGTCCGTCTCAGATCTCATCATGAAACCTGGGGACGAAGAGC AGATTTCCTGCAGCCATGACAGGACCGACTACAGACAGATGCTCTGGTACCAGCAGTC CCCTGGAGACACGGCTATGAAACTCATTGGATATTTGAACTTTAAAGCTTCCACAATG GAAGAGTCTTATAAGGATCATTTTATTATCTCTGGAGATTTGGGAGGAAACACAGCAA AGAATGGCTCGCTTGTAGTCAGAGGTGTAAAACAGGAGCACAGTGCAGTTTACTACTG TGCAGCGTGTGAAG
V2.13	ATGCTGTGGTACCAGCAGCGCCCCGGAGACAACGCCGTCAAACTCATCGGACACGGAT
*	ACACCGTGTTCAGGAACGACAGCGCGGAGGAGCCGTTCAGAAAGACTTTCAAACTAA CCGGAGACCTGAGTGACAAAACCAAAAGCGGCGTGCTTTCCATAACGAAGCTGGACG TGCGTGAACACACCGCGACGTACTTCTGTGCAGCCAGTCAG
V2.14	GTCAAAAAAGTCTCCCAGAATCCGGCCGATGTGCTTCTCGGGCCAAATAGCGAGGCCA CCCTGAGCCTCCGCCATGAGATCCCGAGCTACGACACCGTCCTCTGGTATCAGCGCCC GGCGGGAGGCTCGGCCCTGAAGCTGATCGGCTACGTGTATTACAGAAGTCCTACGGTG GAGGTAGATTTCAAAGGCCGCTTCAACGTGAGCGGAAATGGAGAGAGTACCGCTCATC TTCACCTGCTCAGCCCGAGACACGCCGAAGACAGCGGCACGTATTTCGGTGCAGCCAG TATTA
D2.1	GGGACAGGGGGGC
J2.1	ACACTGGGACTGAAGCGTACTTTGGAAAGGGAACCAAAGTCTCTGTTTTGG
J2.2	ACCTACACTGATCCCGCTTACTTCGGCCAAGGAACCAAGCTGACGGTTTTAG
J2.3	ACAATGCAAACGAGGCTTACTTCGGACCAGGAACCAAACTGACTGTTTTGG
J2.4	CCAATTACGAGGCGTATTTCGGCCCAGGAACCAAACTGACTGTTCTCG
J2.5	ACACTGGTGGCGGTGAAGCCTTTTTTGGTGGAGGAACAAAGTTAACAGTTTTAG
J2.6	ACTCTGGTGCTTATGAAGCTTTCTTTGGTGGAGGAACAAAGTTAACGGTGTTAG
J2.7	ACTCAGACTGCAGAGGCTTACTTTGGTGGAGGAACAAGGTTAACAGTTTTAG

J2.8	CTCCTTCAACGAGGCTTACTTTGGAGCTGGAACTAAGCTCACCGTTTTAG
J2.9*	AACTACGACCCTGCTTACTTTGGCGGTGGCACCAGACTCACGGTGTTGG

Asterisks (*) mark gene segments added after alignment to local stickleback genome.
Crosses (†) mark pseudogenes.

Table B.5: Nucleotide sequences of V, D, and J coding gene segments of the TRB1 and TRB2 loci of the three-spined stickleback.

Infection rate				
Data subset	Explanatory	Df	Chisq	Pr(>Chisq)
All infected	(Intercept)	1	1.5076	0.2195
	Host family	5	1.8538	0.8690
	Parasite strain	1	0.8194	0.3654
	Family: strain	5	4.6608	0.4587

Table B.6: Family, parasite strain and their interaction effects on infection rates. Results from Type III chi-square tests of generalized linear mixed effect models (GLMMs) using fish sex as random intercept.

Parasite index				
Data subset	Explanatory	Df	F value	Pr(>F)
All infected	**Parasite strain**	**1**	**86.5493**	**0.0003**
	Host family	5	1.1399	0.4756
	MHC allelic diversity	2	1.1561	0.3288
GOT-infected	Host family	5	1.4581	0.2856
	MHC allelic diversity	2	3.0119	0.0947
NO-infected	Host family	5	0.4979	0.7377
	MHC allelic diversity	2	1.6854	0.2156

Table B.7: Effect of host family, individual MHC-II allelic diversity, and parasite strain on parasite indices. Generalized linear mixed model (GLMMs) included host family, individual MHC-II allelic diversity, and parasite strain as fixed effects, as well as fish sex and tank as random terms. Bold numbers indicate significance with an alpha of 0.05.

Host body condition factor (CF)				
Data subset	Explanatory	Df	F value	Pr(>F)
All fish	**Host family**	**5**	**13.8645**	**2.546e-05** ***
	Infection status	2	1.1530	0.3255
	Strain	1	0.0359	0.8520
	Infection status: strain	2	0.7818	0.4641
Hepatosomatic index (HSI)				
Data subset	Explanatory	Df	F value	Pr(>F)
All fish	Host family	5	1.2432	0.3498742
	Infection status	**2**	**20.8524**	**8.135e-06** ****
	Strain	1	0.5053	0.4926599
	Infection status: strain	**2**	**11.3286**	**0.0004095** **
Head kidney index (HKI)				
Data subset	Explanatory	Df	F value	Pr(>F)
All fish	**Host family**	**5**	**6.2851**	**0.0015307** *
	Infection status	**2**	**23.0511**	**3.369e-07** ****
	Strain	**1**	**15.8988**	**0.0009866** **
	Infection status: strain	**2**	**16.6457**	**7.266e-06** ****
Splenosomatic index (SSI)				
Data subset	Explanatory	Df	F value	Pr(>F)
All fish	**Host family**	**5**	**5.0806**	**0.000232** **
	Infection status	**2**	**12.7154**	**7.273e-06** ****
	Strain	**1**	**6.3880**	**0.012423** *
	Infection status: strain	2	1.1351	0.323881

Table B.8: Effects of family, infection status, parasite strain, and the interaction between infection status and strain on host condition and immunological indices. Results from Type III Wald F tests of generalized linear mixed effect models (GLMMs) using fish sex and tank as random intercepts when appropriate.

Hepatosomatic index (HSI)					
Data subset	Contrast (infection status)	Estimate	Std. Error	z value	Pr(>|z|)
All fish	Exposed - Control	0.09121	0.1620	0.563	0.574
	Infected - Control	**-0.99659**	**0.18946**	**-5.260**	**2.16e-07** ****
	Infected - Exposed	**-1.08780**	**0.17893**	**-6.079**	**3.62e-09** ****
GOT treatment	Exposed - Control	0.1056	0.2197	0.481	0.631
	Infected - Control	-0.1901	0.2720	-0.699	0.631
	Infected - Exposed	-0.2957	0.2537	-1.166	0.631
NO treatment	Exposed - Control	0.1919	0.2066	0.929	0.353
	Infected - Control	**-1.6786**	**0.2284**	**-7.349**	**2.99e-13** ****
	Infected - Exposed	**-1.8705**	**0.2215**	**-8.446**	**< 2e-16** ****
Head kidney index (HKI)					
Data subset	Contrast (infection status)	Estimate	Std. Error	z value	Pr(>|z|)
All fish	Exposed - Control	0.006405	0.018404	0.348	0.728
	Infected - Control	**0.142120**	**0.021506**	**6.608**	**5.83e-11** ****
	Infected - Exposed	**0.135715**	**0.020315**	**6.680**	**5.83e-11** ****
GOT treatment	Exposed - Control	-0.005411	0.020102	-0.269	0.788
	Infected - Control	0.011261	0.024892	0.452	0.788
	Infected - Exposed	0.016672	0.023212	0.718	0.788
NO treatment	Exposed - Control	0.01836	0.02533	0.725	0.469
	Infected - Control	**0.24171**	**0.02806**	**8.614**	**< 2e-16** ****
	Infected - Exposed	**0.22335**	**0.02700**	**8.272**	**3.33e-16** ****
Splenosomatic index (SSI)					
Data subset	Contrast (infection status)	Estimate	Std. Error	z value	Pr(>|z|)
All fish	Exposed - Control	0.007067	0.005406	1.307	0.191
	Infected - Control	**0.031466**	**0.006317**	**4.981**	**1.9e-06** ****
	Infected - Exposed	**0.024400**	**0.005967**	**4.089**	**6.5e-05** ***
GOT treatment	Exposed - Control	0.007975	0.006832	1.167	0.2431
	Infected - Control	**0.021265**	**0.008460**	**2.514**	**0.0358** *
	Infected - Exposed	0.013290	0.007889	1.685	0.1381
NO treatment	Exposed - Control	0.005629	0.008137	0.692	0.489059
	Infected - Control	**0.039331**	**0.008998**	**4.371**	**3.71e-05** ***
	Infected - Exposed	**0.033701**	**0.008717**	**3.866**	**0.000166** **

Table B.9: Effect of exposure and infection on host indices according to *S. solidus* strain. Sticklebacks were either sham-exposed (control), *S. solidus* exposed but uninfected (exposed), or *S. solidus* infected (infected). Response variables on which infection status was shown to have a significant effect in Table S8 were used. LMMs included infection status as fixed effect, and fish sex as random term. Tukey's all-pair comparisons or comparisons of manually defined contrasts were performed. Bold numbers indicate significance post FDR correction according to Benjamini and Hochberg with an alpha of 0.05.

Index	**Contrast (strain)**	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
PI	NO - GOT	11.813	0.8851	13.35	**<2e-16**
CF	NO - GOT	0.02483	0.07447	0.333	0.739
HSI	NO - GOT	-0.8653	0.1899	-4.558	**5.17e-06**
HKI	NO - GOT	0.20894	0.04974	4.2	**2.67e-05**
SSI	NO - GOT	0.02539	0.01199	2.118	**0.0342**

Table B.10: Effect of *S. solidus* strain on condition and immunological indices of infected sticklebacks. Response variables were the parasite index (PI), body condition factor (CF), hepatosomatic index (HSI), as well as the splenosomatic index (SSI) and head kidney index (HKI). GLMMs included parasite strain as fixed effect and fish sex and tank as random terms. Post hoc tests are based on Tukey's comparisons. Bold numbers indicate significance with an alpha of 0.05.

Hepatosomatic index (HSI)				
Data subset	Explanatory	Df	F value	Pr(>F)
NO- infected fish	Parasite index	1	0.0819	0.7780
	Family	4	0.6302	0.6475
GOT- infected fish	Parasite index	1	1.1604	0.3044
	Family	5	0.9632	0.4802
Head kidney index (HKI)				
Data subset	Explanatory	Df	F value	Pr(>F)
NO- infected fish	**Parasite index**	**1**	**7.9642**	**0.011376**
	Family	**4**	**5.2052**	**0.006113**
GOT- infected fish	**Parasite index**	**1**	**7.1538**	**0.02161**
	Family	5	2.0169	0.15433
Splenosomatic index (SSI)				
Data subset	Explanatory	Df	F value	Pr(>F)
NO- infected fish	Parasite index	1	0.5274	0.4775
	Family	4	0.5224	0.7206
GOT- infected fish	Parasite index	1	0.0033	0.9552
	Family	5	0.9394	0.4947
Condition factor				
Data subset	Explanatory	Df	F value	Pr(>F)
NO- infected fish	Parasite index	1	1.4527	0.243700
	Family	**4**	**5.1943**	**0.005826**
GOT- infected fish	Parasite index	1	0.4962	0.4958
	Family	5	0.9907	0.4661

Table B.11: Effects of family and parasite index on host body and immunological indices of NO- and GOT-infected fish. Results from Type III Wald F tests of generalized linear mixed effect models (GLMMs) using fish sex as random intercept.

Repertoire	Family	Within-family	Between-families	*p*
TCRβ1	A	**125.8 +-16 (n=15)**	**112.5+-20.6 (n=156)**	**0.0068**
	B	92.7 +-27 (n=10)	100.5 +-29.5 (n=135)	0.422
	C	97.5 +-11.2 (n=15)	100.1 +-20 (n=156)	0.702
	D	100.7+-12.5 (n=3)	103.2+-18.1 (n=87)	0.779
	E	106+-10.9 (n=15)	103.9+-19.3 (n=156)	0.479
	F	110.5+- 21.9 (n=15)	107.8 +-24.6 (n=156)	0.619
	Average	107.2 +-20.3 (n=73)	104.9 +-22.9 (n=423)	0.266
TCRβ2	A	19.9 +-5 (n=15)	18.4 +-6.5 (n=156)	0.194
	B	22.7 +-11 (n=10)	19.9 +-9 (n=135)	0.43
	C	13.7 +-3.6 (n=15)	16.1 +-5.2 (n=156)	0.0738
	D	11+-4.4 (n=3)	13.5+-4.9 (n=87)	0.305
	E	16.1+-4.9 (n=15)	17.1+-5.7 (n=156)	0.555
	F	**24.9+- 9.1 (n=15)**	**19.3+-7.3 (n=156)**	**0.0102**
	Average	18.9+-8 (n=73)	17.7 +-6.9 (n=423)	0.289

Table B.12: Mean pairwise intra- and inter-family repertoire overlap for each family in our dataset. Results are given for the TCRβ1 and TCRβ2 repertoires separately. Overlap was calculated using the top 500 (TCRβ1) or top 250 (TCRβ2) most abundant aa clonotypes from each control sample. Mean+-SD values of overlapping clonotypes shown. Parentheses denote number of pairwise comparisons, p values from Wilcoxon test. Statistically significant differences highlighted in bold.

Repertoire	MHC-II haplotype	Within-haplotype	Between- haplotypes	*p*
TCRβ1	A (n=17)	98.7+-15 (n=136)	102.4+-19.5 (n=345)	0.0596
	C (n=17)	104+-25.4 (n=136)	105.1+-23.9 (n=359)	0.5647
	D (n=5)	121+-26.2 (n=10)	107.8+-21.9 (n=109)	0.0812
	E (n=13)	112.3+-28.9 (n=78)	108.1+-23.5 (n=287)	0.2174
	L (n=2)	145 (n=1)	117.6+-18.8 (n=47)	NA
	W (n=2)	79 (n=1)	92.5+-16.7 (n=43)	NA
	Average	104.8 +-23.7 (n=315)	105.9+-20.3 (n=181)	0.349
TCRβ2	A (n=17)	15.3+-5.4 (n=136)	16.4+-6 (n=315)	0.08793
	C (n=17)	20.8 +- 8.5 (n=136)	19.1+-7.3 (n=325)	0.06421
	D (n=5)	15.4+-5.2 (n=10)	15.7+-5.4 (n=104)	0.829
	E (n=13)	20.8+-8.3 (n=78)	18.9+-7.4 (n=299)	0.05451
	L (n=2)	18 (n=1)	19+-6.4 (n=59)	NA
	W (n=2)	16 (n=1)	16.2+-4.5 (n=48)	NA
	Average	18.4 +-7.7 (n=315)	16.9+-5.7 (n=181)	0.121

Table B.13: Mean pairwise intra- and inter- haplotype repertoire overlap for each MHC-II haplotype in our dataset. Results are given for the TCRβ1 and TCRβ2 repertoires separately. Overlap was calculated using the top 500 (TCRβ1) or top 250 (TCRβ2) most abundant aa clonotypes from each control individual. Mean +-SD values of overlapping clonotypes shown. Parentheses denote number of pairwise comparisons, p values from Wilcoxon test.

Repertoire	Subset	Comparison	diff	*p*
TCRβ1	Whole	Exposed-Control	0.1882377	0.5812143
		Infected-Control	**-0.4533490**	**0.0479836**
		Infected-Exposed	**-0.6415867**	**0.0028807**
	NO-treated	Exposed-Control	0.09049563	0.9265791
		Infected-Control	**-0.96885940**	**0.0007016**
		Infected-Exposed	**-1.05935503**	**0.0002187**
	GOT-treated	Exposed-Control	0.26517945	0.5125328
		Infected-Control	-0.03799091	0.9861413
		Infected-Exposed	-0.30317037	0.4193462
TCRβ2	Whole	Exposed-Control	0.05076352	0.7611971
		Infected-Control	-0.10107335	0.3432619
		Infected-Exposed	-0.15183688	0.0936236
	NO-treated	Exposed-Control	0.03035415	0.9611533
		Infected-Control	-0.22010351	0.1380025
		Infected-Exposed	-0.25045766	0.0797965
	GOT-treated	Exposed-Control	0.063081407	0.7252370
		Infected-Control	-0.007899421	0.9949332
		Infected-Exposed	-0.070980828	0.6664072

Table B.14: Post-hoc Tukey's multiple comparisons of means of distances from centroid (dispersion) across infection status groups based on similarity-matrix of TCRβ1 and TCRβ2 repertoires. Calculated with 95% family-wise confidence level. Analyses were performed with the entire subset, and by strain.

TCRβ1 repertoire				
Data subset	Explanatory	Df	F value	Pr(>F)
GOT	Family	5	3.7758	0.1931
	Infection status	2	0.8802	0.4515
NO	Family	5	1.4346	0.23297
	Infection status	**2**	**3.2786**	**0.04802** *
TCRβ2 repertoire				
Data subset	Explanatory	Df	F value	Pr(>F)
GOT	**Family**	**5**	**4.2444**	**0.003445** **
	Infection status	2	0.8633	0.429498
NO	Family	5	1.8425	0.3805
	Infection status	2	2.6980	0.1746

Table B.15: Effects of family and infection status on TCRβ1 and TCRβ2 clonotype abundance inequality, as measured with the Gini index by treatment group. Results from Type III Wald F tests of generalized linear mixed effect models (GLMMs) using tank as random intercept.

B.4 Supplementary Figures

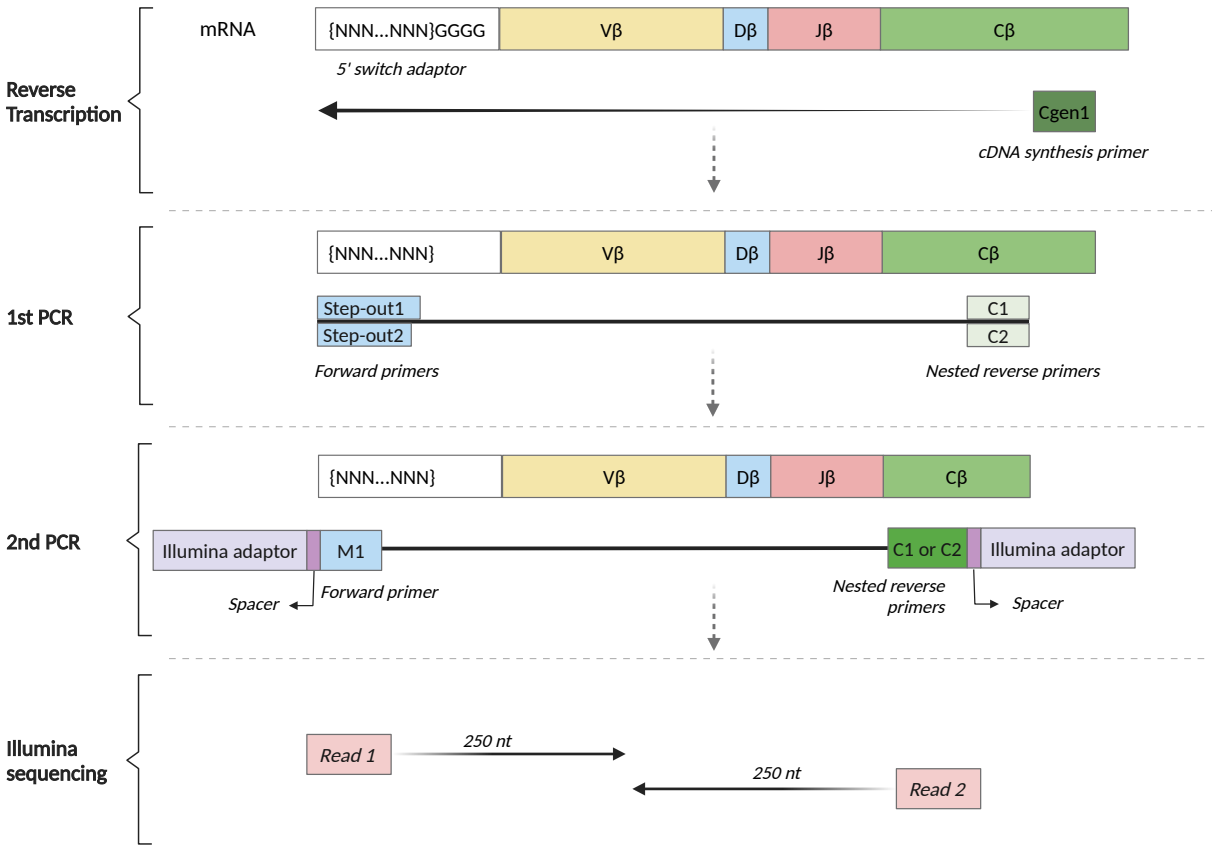


Figure B.1: Schematic overview of the 5' RACE protocol for TCRβ targeting, amplification, and library preparation for Illumina sequencing. Created with Biorender.com.

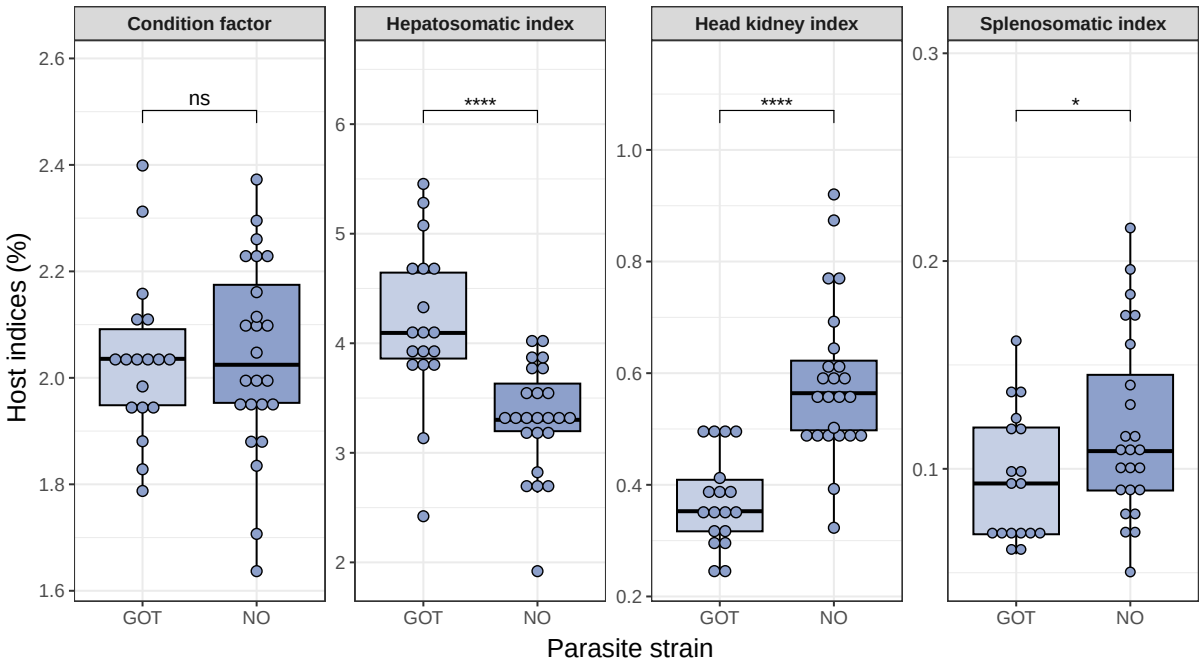


Figure B.2: Comparison of various host indices between sticklebacks experimentally infected with two allopatric strains of the parasite *S. solidus*. Comparison of two body condition indices (condition factor and hepatosomatic index) and two immunological condition indices (head kidney index and splenosomatic index) by parasite strain. P values shown were calculated based on the GLMMs described in Table S10 (**** $p < 10^{-5}$, * $p < 0.05$).

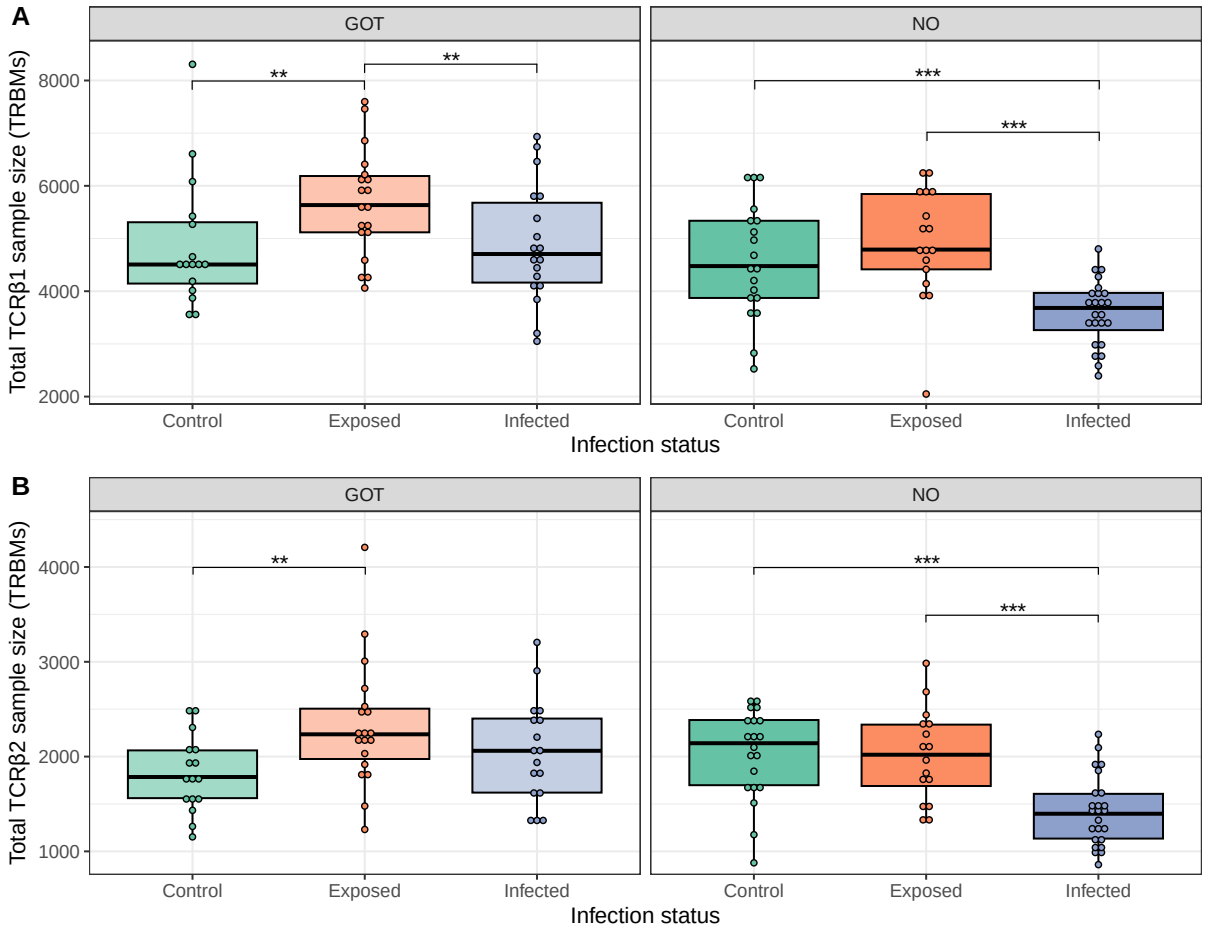


Figure B.3: Total TCRβ repertoire sample sizes from our sequenced three-spined stickleback dataset grouped by infection status and *S. solidus* strain. Shown are the sizes of the **(A)** TCRβ1 and **(B)** TCRβ2 functional repertoires obtained post-bioinformatic extraction and error correction, as described in Materials and Methods. Sample sizes are expressed in total TCRβ cDNA molecules (TRBM). For each repertoire, samples were grouped by strain and GLMMs were fitted with family and infection status as fixed effects, and sex and tank as random effects. When infection status had a significant effect, we performed Tukey's post-hoc tests, with asterisks signifying P value significance levels after Holm correction (***P< 10^{-4}, ** P < 10^{-3}). Non-significant comparisons are omitted for visual clarity.

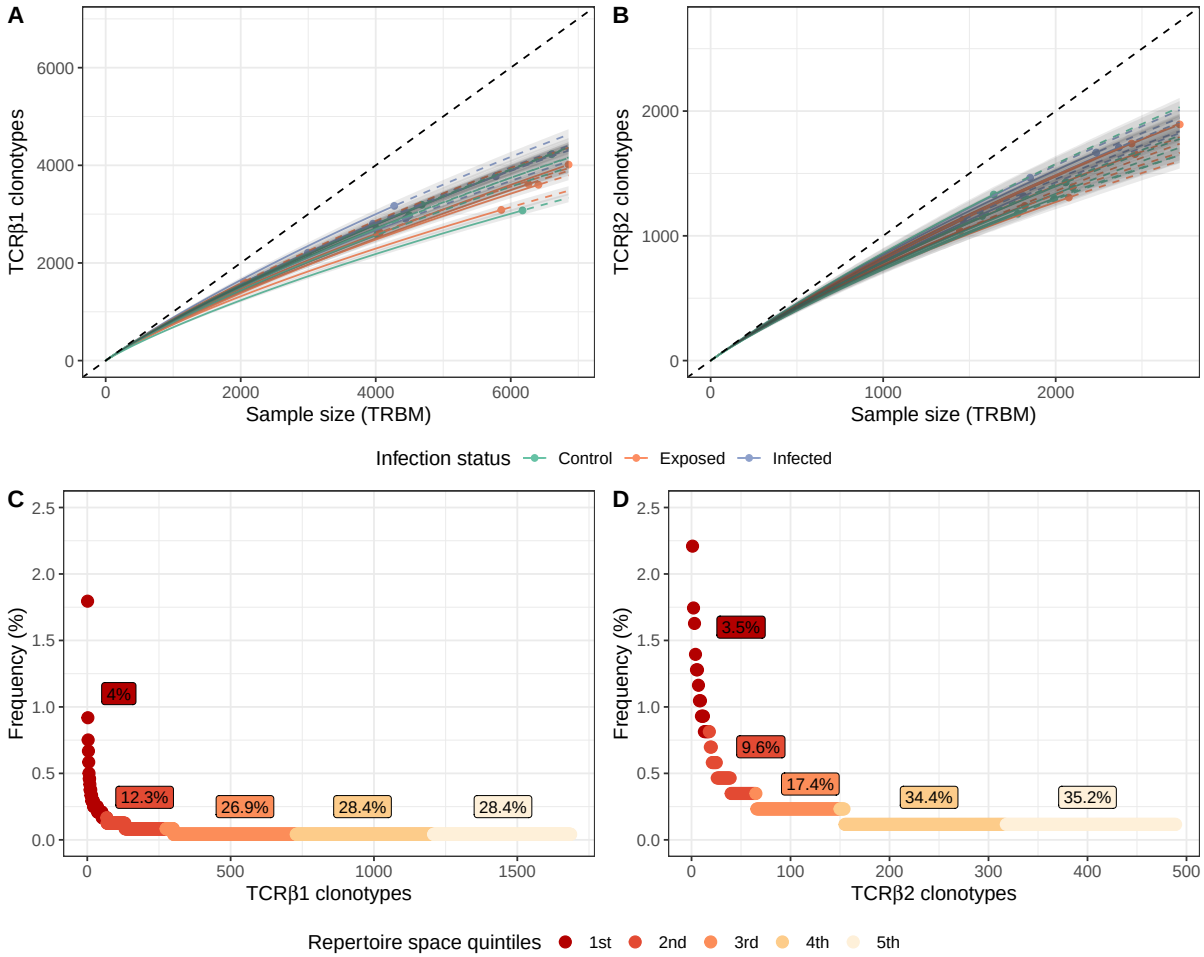


Figure B.4: TCR repertoire rarefaction analysis. Rarefaction plots of the total TCRβ1 **(A)** and TCRβ2 **(B)** functional repertoires from a single family (family E). The plots show the relationship between repertoire size on the x axis (*i.e.*, total cDNA molecules, TRBM) and clonotype diversity (*i.e.*, unique TCRβ nucleotide sequences) on the y axis. Each line represents a single sample / individual (n=16). Solid and dashed lines mark interpolated and extrapolated regions of rarefaction curves respectively, points mark exact sample size and diversity of each sample. Shaded areas mark 95% confidence intervals. Interpolation and extrapolation up to the size of the largest sample are based on multinomial models, as implemented by VDJtools. All samples fall beneath the dashed diagonal (representing a state of complete unsaturation), however our samples did not reach sequencing saturation (*i.e.*, plateauing of curves). C, D) Rank-abundance plots of clonotypes from a control individual's TCRβ1 and TCRβ2 repertoire, showing that a small number of highest ranking clonotypes (based on their clonal frequency) occupy a disproportionate amount of the total repertoire space. Numbers refer to the percentage of total clonotypes occupying each one of the five space quintiles. Lowest frequency clonotypes (*i.e.*, singletons) make up for 3/5 and 2/5 of the TCRβ1 and TCRβ2 repertoire space, respectively.

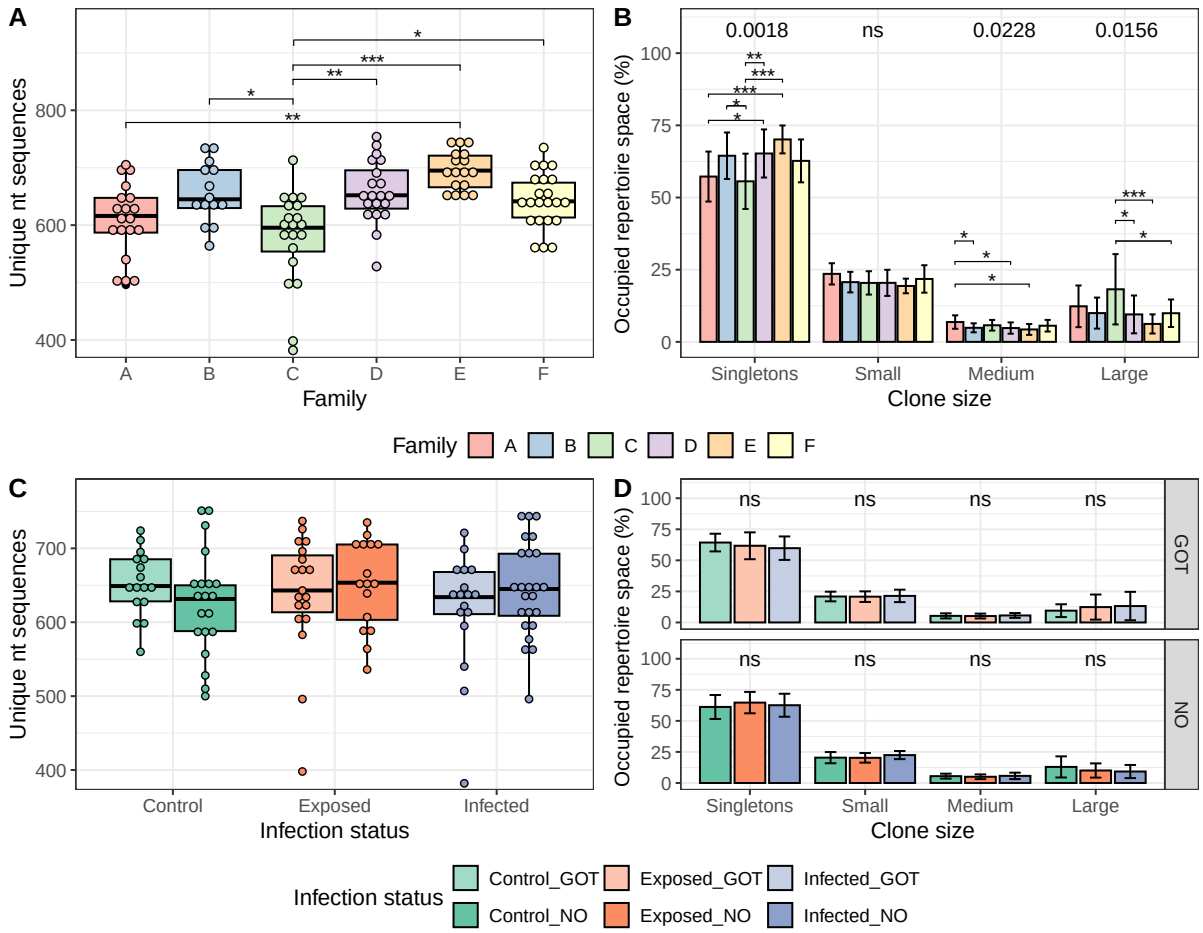


Figure B.5: TCRβ2 repertoire diversity and clonality is family-dependent. Mean values of observed diversity (*i.e.*, unique nt sequences) obtained from down-sampled repertoires of 860 T cells was compared between stickleback families **(A)** and infection status grouped by parasite strain **(C)**. Where GLMM showed a statistically significant effect, Tukey's post-hoc tests were performed with asterisks signifying P value significance levels *** p< 10^{-4}, ** p< 10^{-3}, * p< 0.05). B, D) Clonal space homeostasis showing the average percent of clonal space occupied by clones of a given size. Singletons, small, medium and large clones were defined as constituting <0.15, <0.35, <0.6 or >= 0.6% of analyzed T cells, respectively. In terms of starting cDNA molecules, these frequencies correspond to 1, 2-3, 4-5, and >5 UMIs, respectively. Comparisons were made between families **(B)** and infection status by parasite strain **(D)**. For each size type, GLMMs were fitted (P values on top of each category, ns non-significant); for statistically significant cases, they were followed by Tukey's post-hoc tests, with asterisks signifying P value significance levels after Holm correction (***P< 10^{-4}, ** P < 10^{-3}, * P < 0.05). In all models, tank was used as random effect. Non-significant comparisons are omitted for visual clarity.

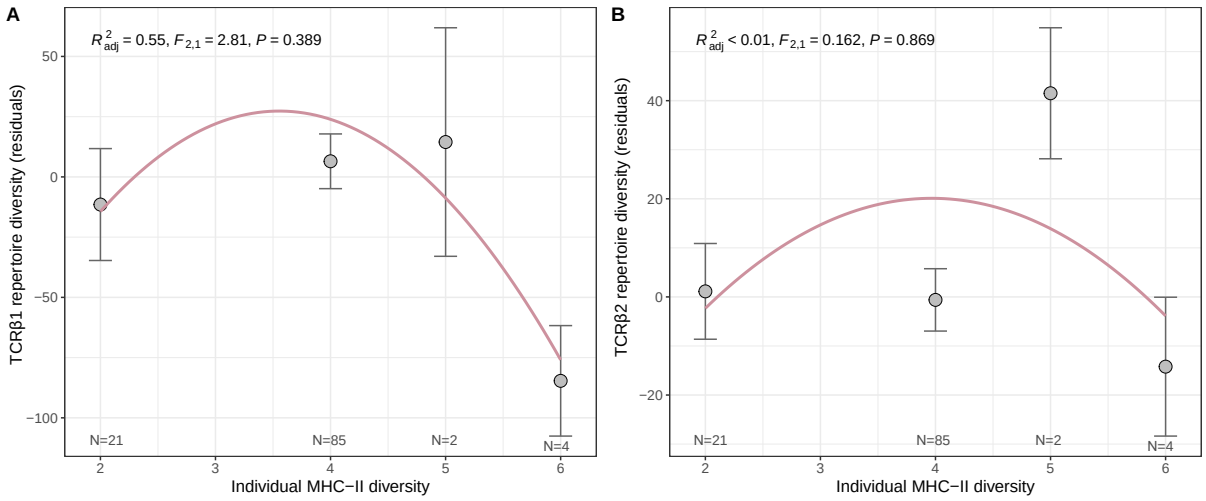


Figure B.6: Effect of individual MHC-II diversity on TCR diversity. Mean (A) TCRβ1 repertoire diversity residuals ± S.E. and (B) TCRβ2 repertoire diversity± S.E of sticklebacks grouped by individual MHC-II diversity. Grey circles indicate the mean value of all individuals possessing the given number of MHC-II alleles. To reduce variation introduced by effects of family and infection status, residuals of a GLMM were used for subsequent analysis, with tank as random effect (Table 1). The displayed quadratic polynomials are fitted to the means. N indicates sample sizes for each diversity level.

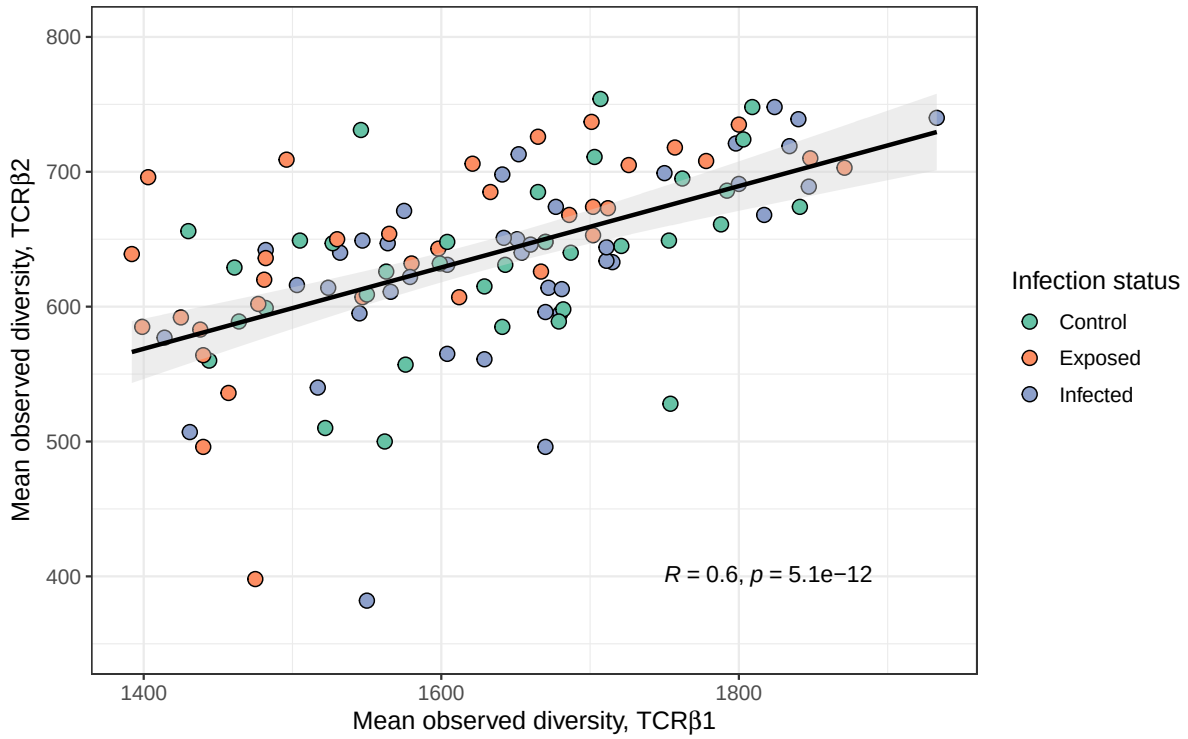


Figure B.7: Correlation between mean intra-individual observed diversity of TCRβ1 and TCRβ2 repertoires. Spearman's R and P value depicted on the graph. 95% confidence intervals are indicated by the grey shaded area.

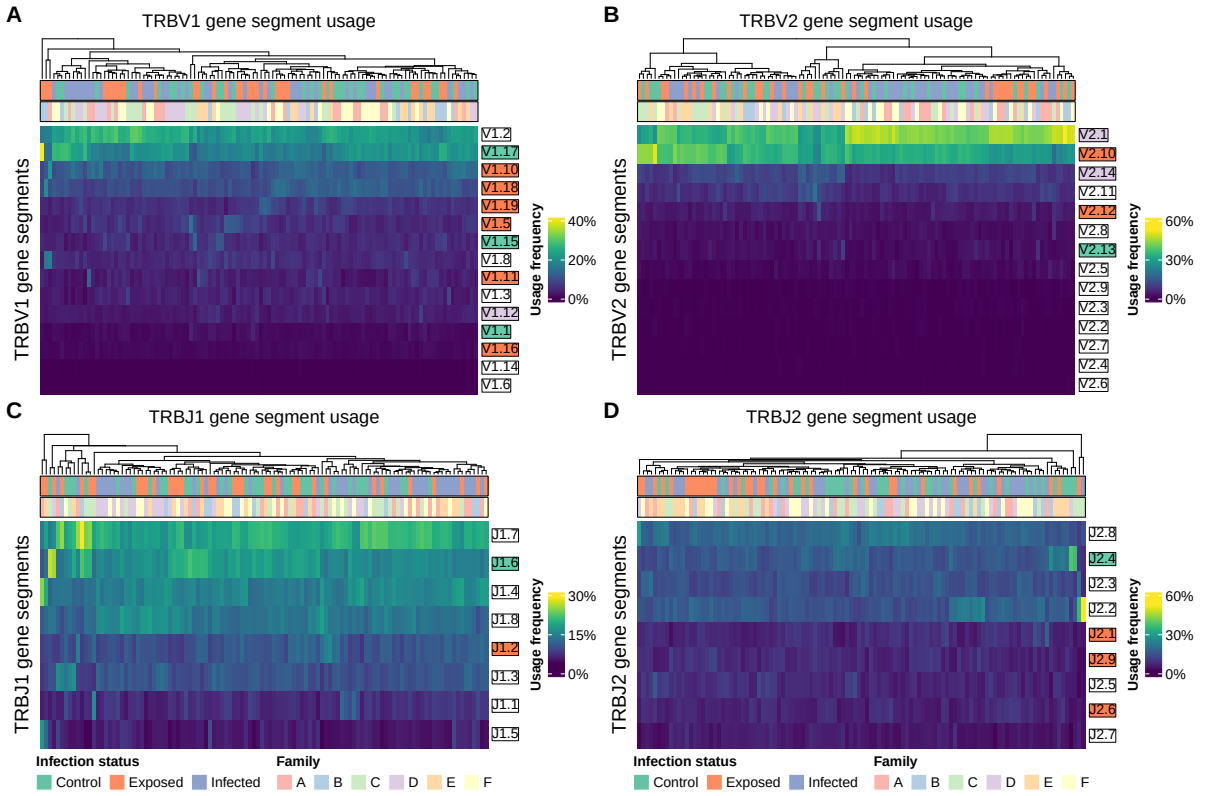


Figure B.8: V and J gene segment usage in the functional TCRβ repertoire. Heatmaps present the frequency of usage of V and J gene segments of the TCRβ1 (A, C) and TCRβ2 (B, D) loci. Weighted usage profiles were used, and hierarchical clustering of samples based on the frequency of usage was performed using Euclidean distance. Annotation bars on top of each heatmap show the family and infectious status of each sample. Names of gene segment highlighted in color are those whose frequency differs significantly between families (orange), infection status (green), or both (purple), as inferred from GLMM analyses with family and infection status as fixed effects, and tank as random effect.

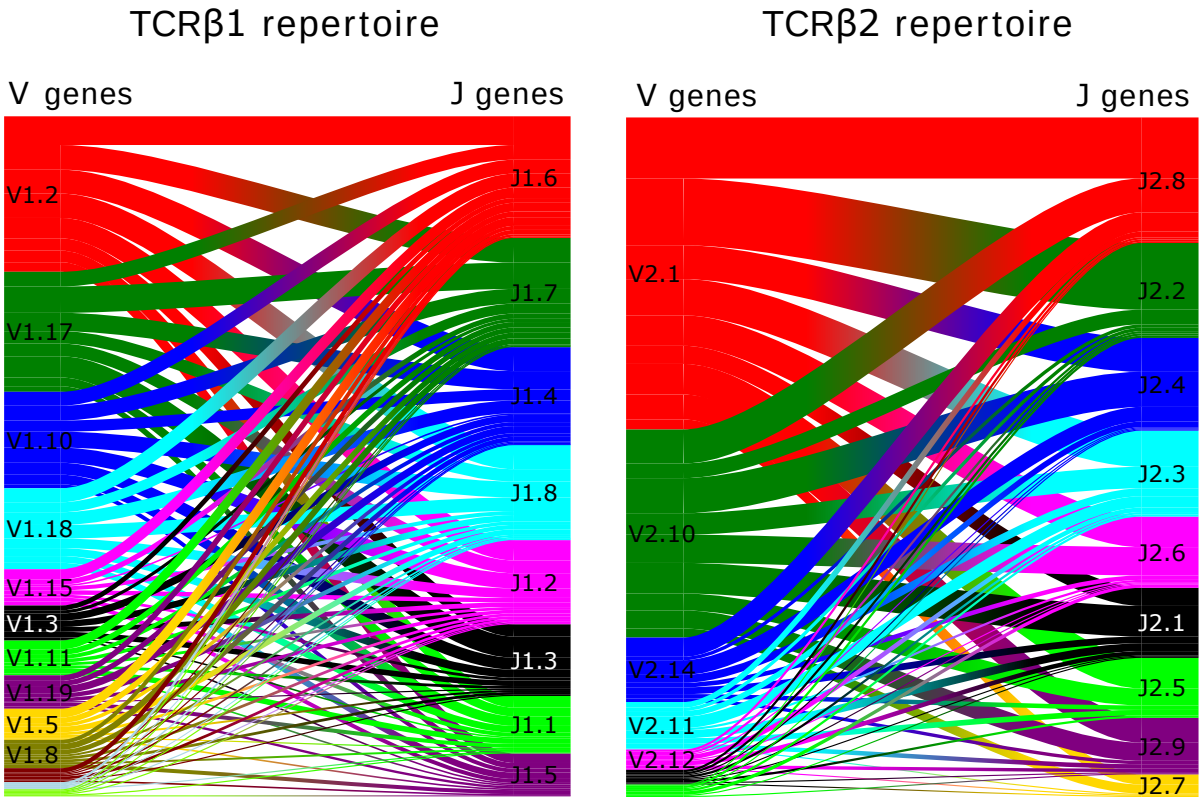


Figure B.9: Sankey diagrams illustrating the pattern of V and J gene pairings of the TCR repertoire. Shown are the structures of the gene-segment pairings of the TCRβ1 (left) and the TCRβ2 (right) repertoires from a representative individual belonging to the control group, family A. In each panel, the two V and J genes are arrayed left to right. Below each gene is a color-stack representing how the clonotypes break down into the different segments for that gene. The curved segments joining neighboring gene-stacks show how the two segment distributions pair up, with their thickness corresponding to the unweighted frequency of clonotypes having those two segments.

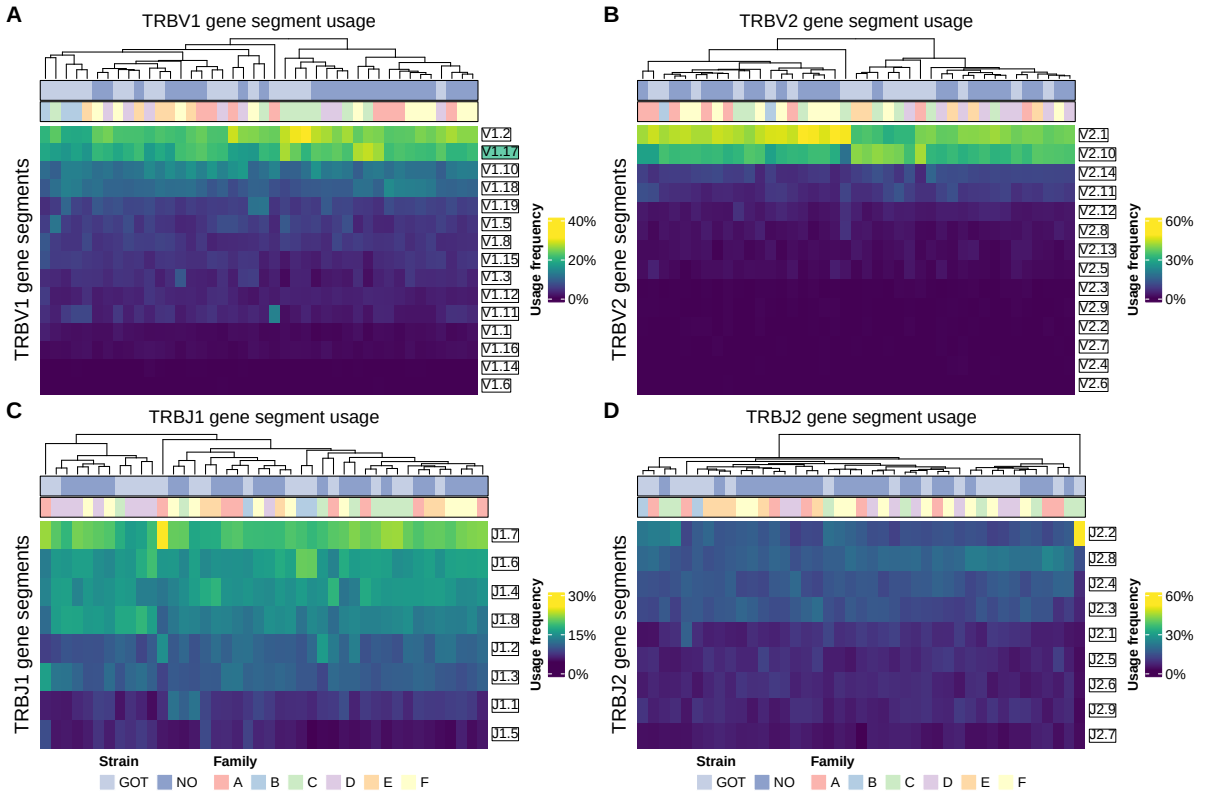


Figure B.10: V and J gene segment usage in the functional TCRβ repertoire of infected individuals. Heatmaps present the frequency of usage of V and J gene segments of the TCRβ1 (A, C) and TCRβ2 (B, D) loci. Weighted usage profiles were used, and hierarchical clustering of samples based on the frequency of usage was performed using Euclidean distance. Annotation bars on top of each heatmap show the family and parasite strain of each sample. Names of gene segment highlighted in green are those whose frequency differs significantly between strains, as inferred from GLMM analyses with family and strain as fixed effects, and tank as random effect.

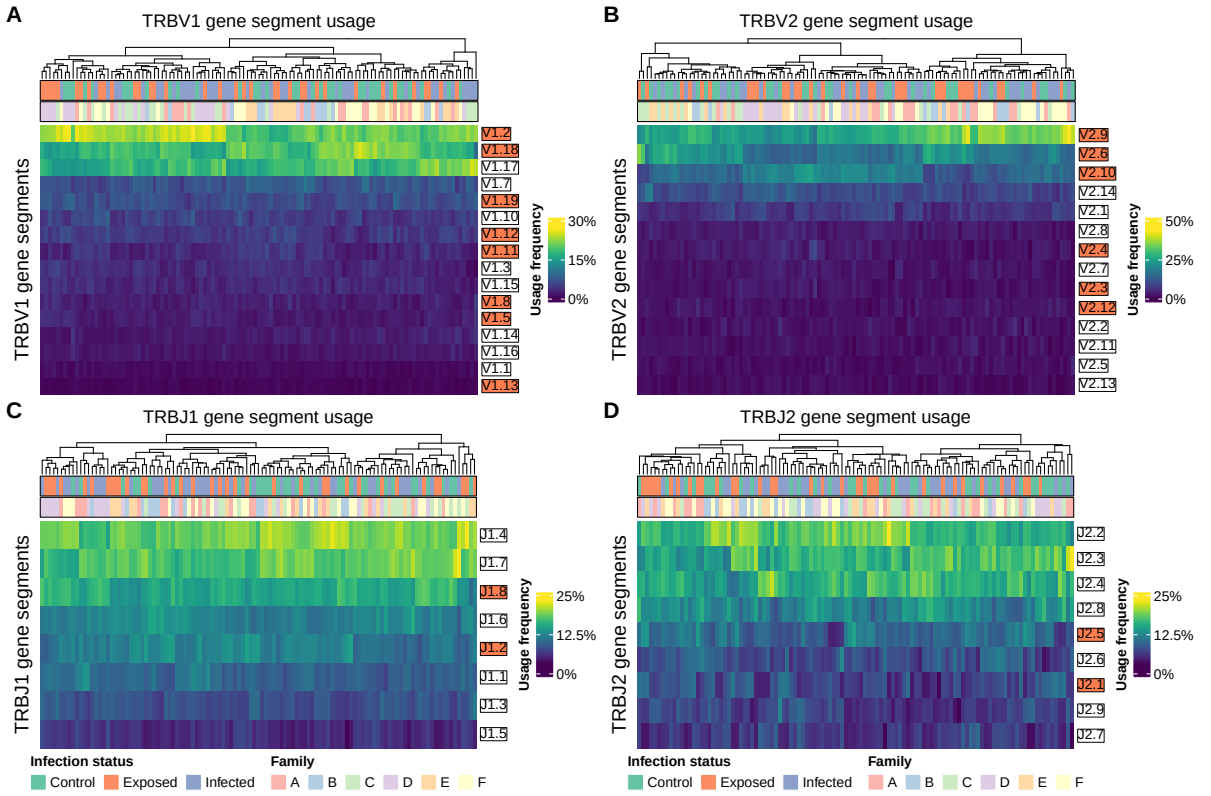


Figure B.11: V and J gene segment usage in the non-functional TCRβ repertoire. Heatmaps present the frequency of usage of V and J gene segments of the TCRβ1 **(A, C)** and TCRβ2 **(B, D)** loci. Unweighted usage profiles were used, and hierarchical clustering of samples based on the frequency of usage was performed using Euclidean distance. Annotation bars on top of each heatmap show the family and infectious status of each sample. Names of gene segment highlighted in orange are those whose frequency differs significantly between families, as inferred from GLMM analyses with family as fixed and tank as random effect.

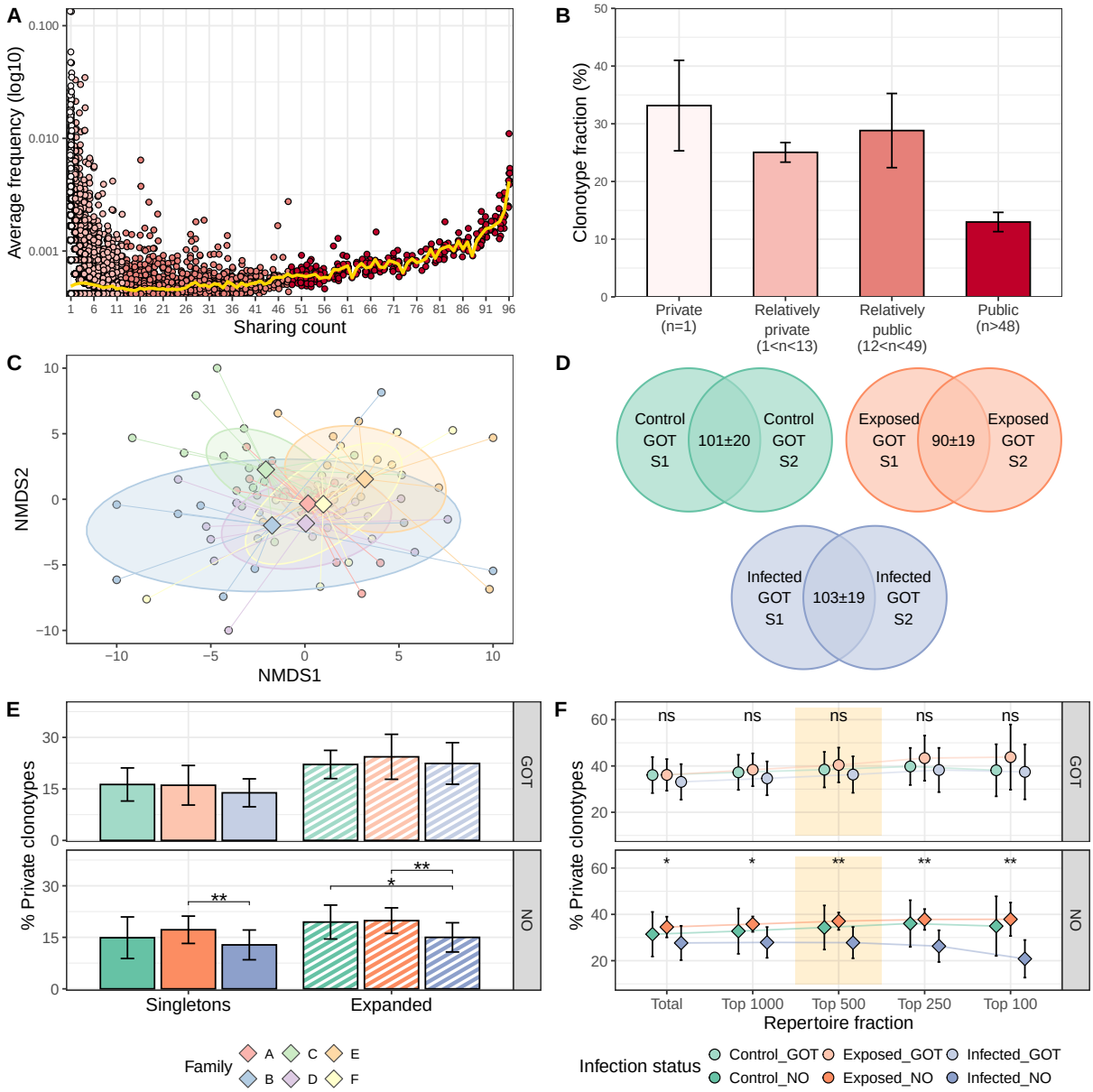


Figure B.12: Infection-non-specific public clonotypes drive the extensive TCRβ1 repertoire overlap between individuals infected with NO *S. solidus*. **(A)** Relationship between clonotype average within-repertoire frequency (abundance) and sharing across different individuals. The golden line connects the mean frequency of clonotypes at each sharing level. **(B)** Bar plot showing average distribution (%) of clonotypes by publicness level within each TCRβ1 repertoire sample in our dataset. Bars denote SD. **(C)** Non-metric multidimensional scaling (NMDS) plots on inter-repertoire Euclidean distances and two dimensions comparing data from different stickleback families. NMDS was based on the number of overlapping clonotypes between repertoires down-sampled to the top 500 most abundant clonotypes. Each dot represents the repertoire of one individual; colors refer to stickleback family. Ellipses represent 95% confidence intervals, and diamonds the centroids of each family. Outlier samples were moved to the plot periphery for better visualization. **(B)** Average pairwise repertoire overlap percentages with control, exposed and infected individuals of the GOT strain treatment. **(D)** Average number of pairwise overlapping clonotypes of each infection status group in the NO treatment. **(E)** Percentage of private clonotypes by infection status and parasite strain for different fractions of the TCRβ1 repertoires. For the NO-infected group, we observe a decline towards the more abundant repertoire ends. P values pertain to the effect of infection status for each repertoire fraction, as obtained from GLMMs with family and infection status as main effects, and tank as random effect. **(F)** Percentage of singletons (clonotypes of clonal size equal to 1) and expanded clonotypes (>1) by infection status and parasite strain, within the top 500 repertoire fraction. P values from Tukey's post hoc tests, following GLMMs with family and infection status as main effects, and tank as random effect.

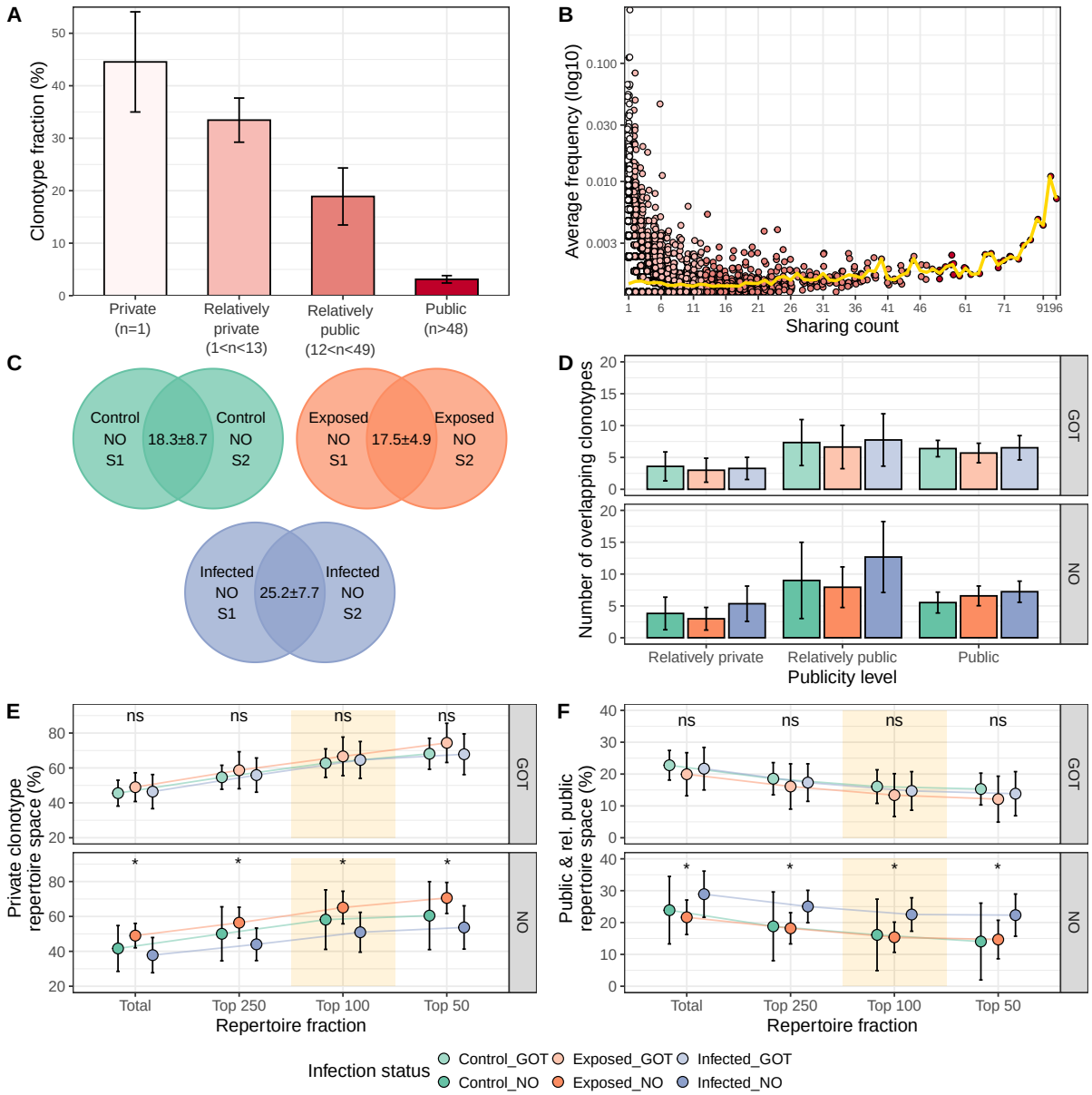


Figure B.13: Infection-non-specific public clonotypes drive the extensive TCRβ2 repertoire overlap between individuals infected with NO S. solidus. (A) Bar plot showing average distribution (%) of clonotypes by publicness level within each TCRβ2 repertoire sample in our dataset. Bars denote SD. **(B)** Relationship between clonotype average within-repertoire frequency (abundance) and sharing across different individuals. The golden line shows the average frequency of clonotypes at each specific sharing level. **(C)** Average pairwise repertoire overlap percentages with control, exposed and infected individuals of the NO treatment. Shown are the average numbers of overlapping clonotypes +- SD between TCRβ2 repertoires down-sampled to the top 250 most abundant clonotypes. **(D)** Number of pairwise overlapping clonotypes by publicness level for each infection status and by strain. **(E)** Analysis of the repertoire space occupied (*i.e.*, cumulative frequency) of private clonotypes by infection status and parasite strain for different fractions of the TCRβ2 repertoire. NO-infected repertoires have fewer private clonotypes on average compared to their exposed and control counterparts. P values for each fraction refer to the statistical significance of the effect of infection status, obtained from fitting GLMM models with family and infection status as main effects, and tank as random effect. **(F)** Similar analysis as described in panel E, for the repertoire space occupied by public and relatively public clonotypes.

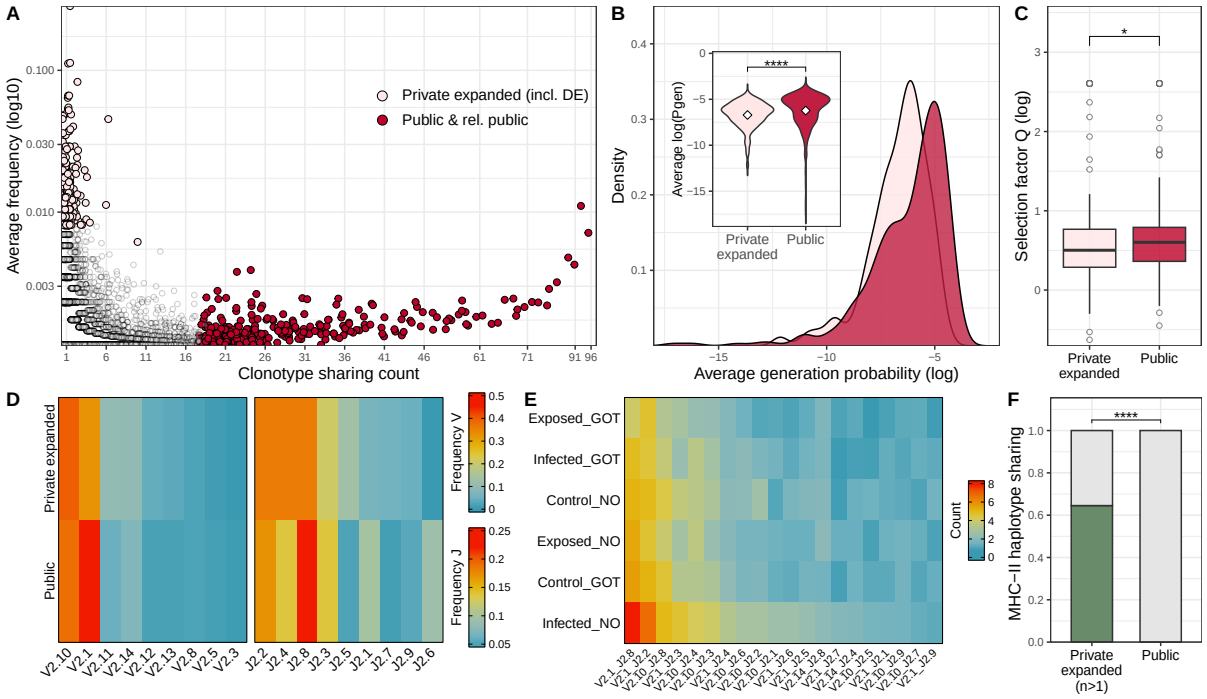


Figure B.14: Public TCRβ2 clonotypes have distinct properties that set them apart from private expanded clonotypes. A) Relationship between clonotype average within-repertoire frequency (abundance) and sharing across different individuals. Each dot represents one clonotype. Both the private/relatively private and public/relative public compartments include expanded clonotypes, *i.e.*, clonotypes of clone size >1. We focus on the 300 most shared clonotypes (in red) and 300 most expanded private/relatively private clonotypes, including 59 differentially expanded (DE) clonotypes across infection status groups (in pink). **B)** Distribution of generation probabilities Pgen for private expanded and public clonotypes. The Pgen value of each clonotype was calculated as an average of all corresponding nt sequences. Inset asterisks signify statistical significance as inferred from Wilcoxon rank-sum test (P= 7.58e-09). **C)** Comparison of average Q values between private expanded and public clonotypes. Q is a proxy of the selection pressure of each clonotype in the thymus and the periphery, as calculated with the software Sonnia. Asterisks signify statistical significance as inferred from Wilcoxon rank-sum test (P= 0.0265). **D)** Heatmaps of frequency of usage of V (left) and J (right) gene segments within the private expanded and the public clonotypes. **E)** Heatmap of the average number of pairwise overlapping clonotypes by VJ gene usage within each of the six treatment groups. Shown are the most frequent VJ gene combinations, *i.e.*, those accounting on average for >2% of all overlapping clonotypes across treatment groups. **F)** Proportion of relatively private expanded (n=31) and public (n=300) clonotypes that are shared exclusively by individuals sharing at least one MHC-II haplotype. Green color denotes sharing. Asterisks signify statistical significance as inferred from Chi-squared test (P= 8.08e-31).

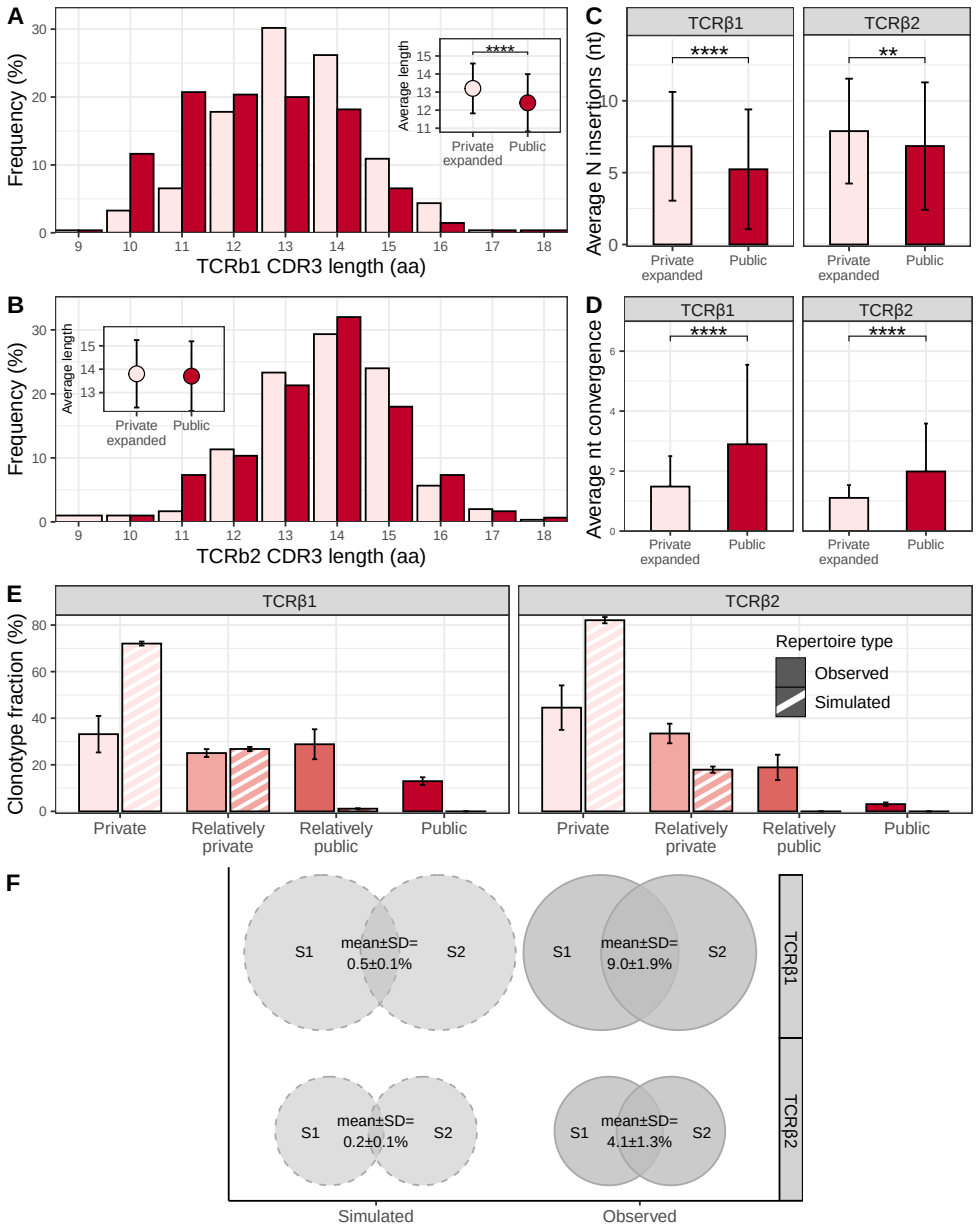


Figure B.15: VDJ recombination biases cannot explain the presence of highly public clonotypes. A, B) Frequency distribution of CDR3 amino acid lengths of private expanded and public TCRβ1 and TCRβ2 clonotypes, respectively. Insets show comparison of average lengths, with asterisks signifying statistical significance inferred from Wilcoxon rank-sum tests (TCRβ1: P= 8.21e-10, TCRβ2: P= 0.326). **C)** Comparison of average number of non-templated nucleotide (N) insertions between private expanded and public clonotypes. Asterisks signify statistical significance inferred from Wilcoxon rank-sum tests (TCRβ1: P= 6.38e-06, TCRβ2 P= 0.00428). **D)** Comparison of average number of unique nucleotide sequences corresponding to each private expanded and public TCRβ1 and TCRβ2 clonotype. Asterisks signify statistical significance inferred from Wilcoxon rank-sum tests (TCRβ1: P= 2.87e-14, TCRβ2: P= 3.08e-25). **E)** Average percentage of clonotypes across publicness levels within TCRβ1 (left panel) and TCRβ2 (right panel) samples in our observed and simulated repertoires. Simulated repertoires were generated with the OLGA software. Bars denote SD. **F)** Pairwise overlap of observed and simulated TCRβ1 and TCRβ2 control repertoires. Shown are the mean percent of overlapping clonotypes +- SD from all pairwise comparisons (n=496). For the simulated dataset, values were computed using the mean and SD from 5,000 iterations (see Materials and Methods). The overlap level of observed repertoires was not reached in any of the 5,000 iterations.

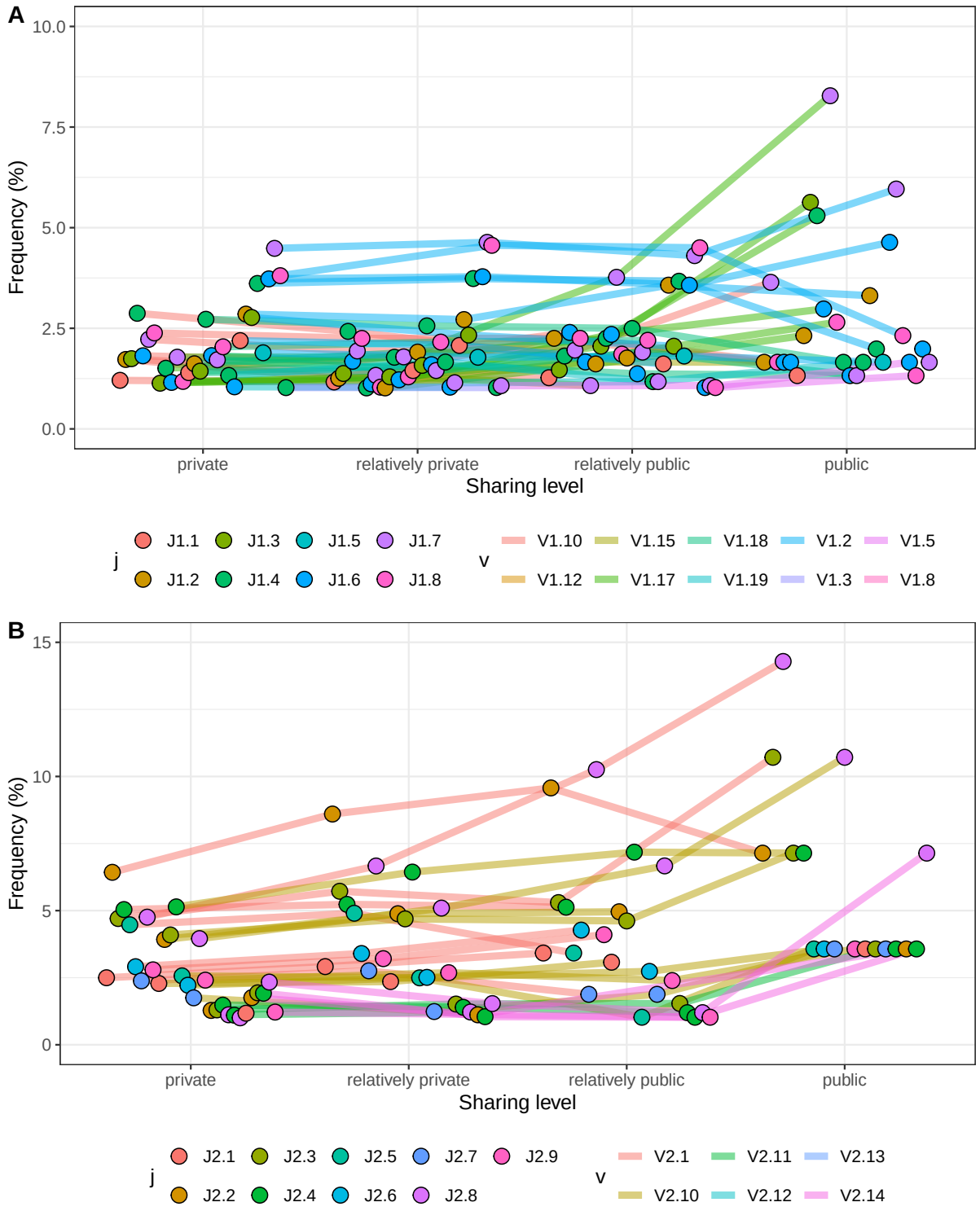


Figure B.16: Clonotypes of certain VJ combinations are more frequent in the public and relatively public compartments. Frequency of clonotypes bearing unique VJ combinations by sharing level in the **(A)** TCRβ1 and **(B)** TCRβ2 repertoires.

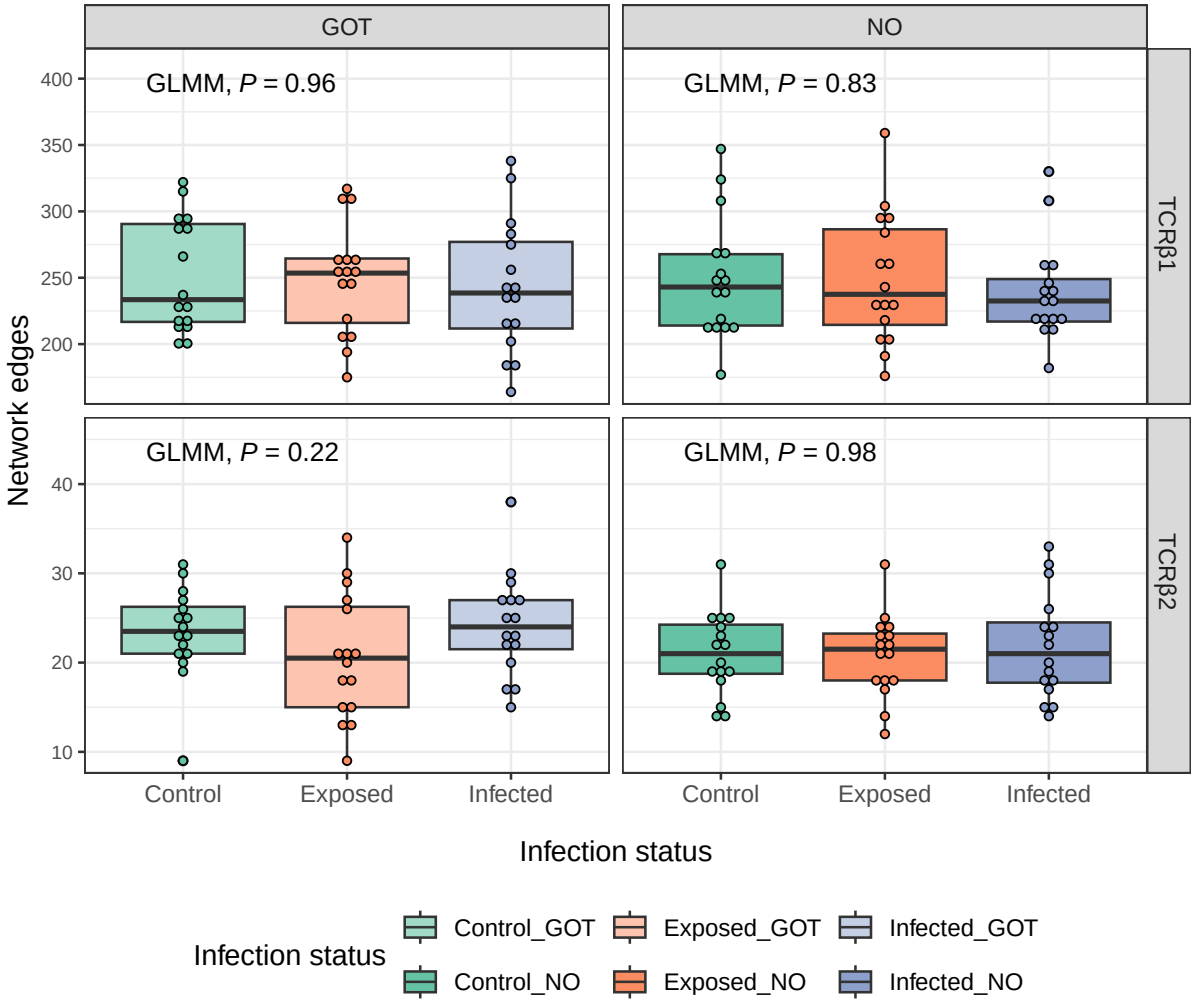


Figure B.17: Average similarity network connectivity levels of the TCRβ repertoires of individuals grouped by infection status and parasite strain. Shown are the total number of edges from each individual's TCRβ1 **(A)** and TCRβ2 **(B)** repertoire networks. Networks composed of the top 500 and top 250 most abundant TCRβ1 and TCRβ2 clonotypes, respectively. P values correspond to factor infection status from GLMMs with family as second independent variable, and tank as random effect.

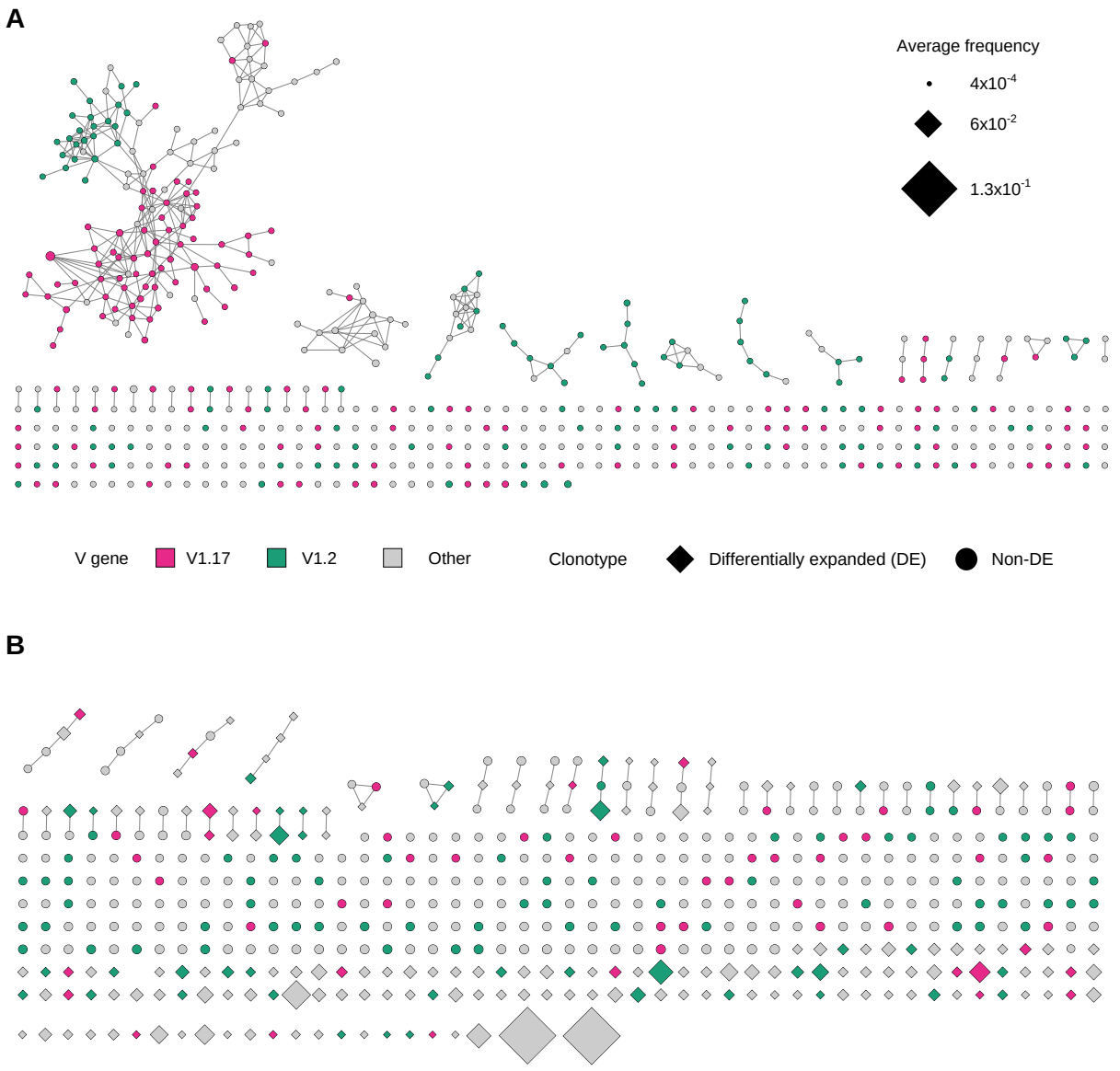


Figure B.18: TCRβ1 public and private clonotypes form distinctly different similarity networks. Networks formed by **(A)** the 500 most widely shared TCRβ1 clonotypes (all 275 public and 225 relatively public) across individuals in our dataset and **(B)** the 500 most abundant private clonotypes, including all previously identified differentially expanded clonotypes. Nodes (clonotypes) were connected by edges defined by a Levenshtein distance of 1. Node size reflects the average frequency of each clonotype across all repertoires in which it was detected. The nodes are colored according to V gene usage. Differentially expanded clonotypes are shown in diamond shape.

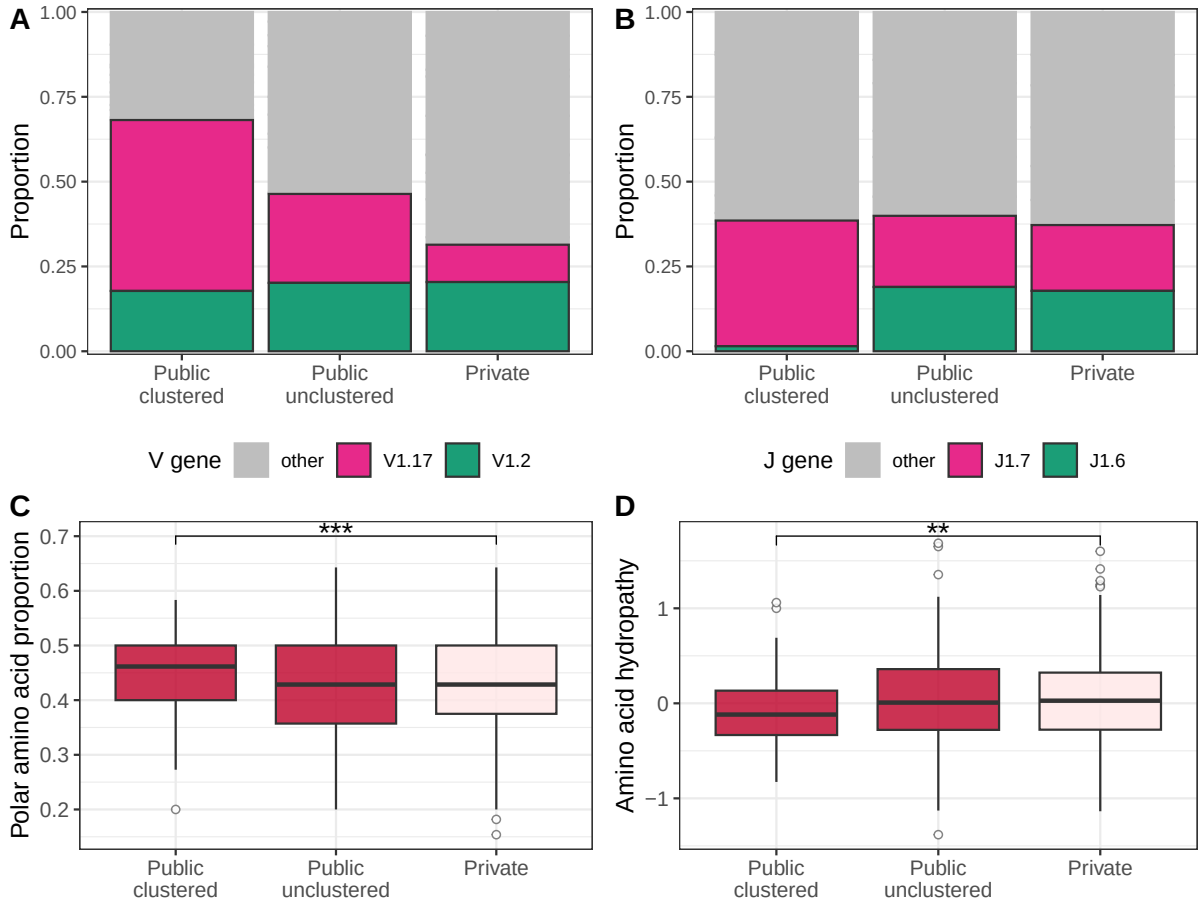


Figure B.19: Distinct properties of TCRβ1 public clustered clonotypes. Proportion of public clustered, public unclustered, and private clonotypes by **(A)** V gene and **(B)** J gene usage. As 'public clustered' were regarded the 135 clonotypes forming the biggest cluster shown in Figure S18A. 'Public unclustered' (n=248) were the clonotypes forming the isolated nodes of Figure S18A, and 'private' (n=500) were all the clonotypes of Figure S18B. **(C)** Comparison of public clustered, public unclustered, and private clonotypes on the basis of average proportion of polar amino acids per clonotype, corrected by clonotype length. Asterisks signify statistical significance inferred from Wilcoxon rank-sum tests (P= 0.00023) with private clonotypes as reference group. **(D)** Comparison of public clustered, public unclustered, and private clonotypes on the basis of average hydropathy levels per clonotype, corrected by clonotype length. Asterisks signify statistical significance inferred from Wilcoxon rank-sum tests (P= 0.002) with private clonotypes as reference group. Non-significant comparisons omitted for visual clarity.

Appendix C

Exposure to natural parasitic fauna leads to TCR repertoire activation in the three-spined stickleback

C.1 Supplementary Tables

Treatment	**Control**		**Exposed**	
Family\Duration	Four weeks	Eight weeks	Four weeks	Eight weeks
F1	6	6	15	13
F2	6	6	15	16
F3	-	6	-	14
F4	-	5	-	13
F5	-	5	-	14
F6	-	6	-	13
F7	-	6	-	14
F8	-	6	-	15
F9	-	4	-	13
F10	-	4	-	13
Total surviving	12	54 (61)	30 (32)	139* (147)

(*) The family of one fish could not be unambiguously determined.

Table C.1: Numbers of surviving fish used in the study by treatment group, duration of experiment, and fish family. Corresponding total number of deaths in each group are shown in parentheses.

Treatment	Control		Lake-exposed	
Family\Duration	Four weeks	Eight weeks	Four weeks	Eight weeks
F1	5	6	5	12
F2	5	6	5	12
F3	-	6	-	12
F4	-	5	-	12
F5	-	5	-	12
F6	-	6	-	12
F7	-	6	-	12
F8	-	6	-	12
F9	-	3	-	12
F10	-	3	-	12
Total	10	52	10	120

Table C.2: Numbers of fish used for TCRβ sequencing by treatment group, duration of experiment, and fish family.

Parasite load				
Data subset	Explanatory	Df	F value	Pr(>F)
Eight-week-exposed (n=109)	**Host family**	**9**	**2.5887**	**0.011** *
	MHC-II diversity	5	0.3449	0.884
	Sex	**1**	**4.1794**	**0.044** *
	Cage	**2**	**6.4382**	**0.002** **
Number of macroparasite taxa				
Data subset	Explanatory	Df	F value	Pr(>F)
Eight-week-exposed (n=109)	Host family	9	1.8211	0.075
	MHC-II diversity	5	1.1299	0.350
	Sex	**1**	**6.5275**	**0.012** *
	Cage	**2**	**4.6889**	**0.012** *

Table C.3: Effect of family, MHC-II diversity, sex, and cage on the total parasite load of sticklebacks of the eight-week exposure group. Results from generalized linear model (GLM). Statistically significant factors highlighted in bold.

Change in condition factor (CF)				
Data subset	Explanatory	Df	F value	Pr(>F)
Eight-week-exposed (n=104)	**Host family**	**9**	**4.0795**	**2.144e-04** ***
	MHC-II diversity	5	0.5962	0.703
	Parasite load	1	0.0359	0.850
	Sex	1	1.4842	0.226

Table C.4: Effect of family, individual MHC-II diversity, parasite load, and sex on body condition change among sticklebacks of the eight-week exposure group. Results from Type III Wald F tests of generalized linear mixed effect models (GLMMs) using lake cage as random intercept.

Observed diversity					
TCR repertoire	Explanatory	Df	F value	Pr(>F)	
TCRβ1	Host family	9	1.0348	0.41438	
	MHC-II diversity	5	0.5149	0.76478	
	Treatment	1	0.5541	0.45769	
	Duration	1	2.0471	0.15435	
	Sex	**1**	**5.9270**	**0.01596**	*
	Treatment:duration	**1**	**4.7604**	**0.03051**	*
TCRβ2	**Host family**	**9**	**2.9543**	**0.0028**	**
	MHC-II diversity	5	2.1057	0.0670	
	Treatment	1	0.0022	0.9640	
	Duration	**1**	**13.774**	**0.0052**	**
	Sex	**1**	**14.361**	**0.0002**	***
	Treatment:duration	1	0.7777	0.4113	

Table C.5: Effect of various independent variables on the average observed diversity of TCRβ1 and TCRβ2 repertoires, from 10 independent size normalization iterations. For each repertoire the entire datasets were used. Results from Type III F tests of generalized linear mixed effect models (GLMMs). Cage was used as random effect. Statistically significant factors are marked in bold.

Total observed TCR diversity (TCRβ1+TCRβ2)					
Data subset	Explanatory	Df	F value	Pr(>F)	
Control fish (n=58)	Host family	9	1.2634	0.2849	
	MHC-II diversity	4	0.8494	0.5021	
	Duration	1	0.0116	0.9147	
	Sex	1	2.3209	0.1351	
Exposed fish (n=129)	Host family	9	1.1637	0.325209	
	MHC-II diversity	5	0.9471	0.453592	
	Duration	**1**	**10.5314**	**0.001548**	**
	Sex	**1**	**8.6401**	**0.003995**	**

Table C.6: Effect of various independent variables on the average total observed TCR diversity of control and exposed individuals. Diversity values were the mean from 10 independent size normalization iterations. For each repertoire the entire datasets were used. Results from Type III F tests of generalized linear mixed effect models (GLMMs). Cage was used as random effect. Statistically significant factors are marked in bold.

Observed diversity						
Repertoire	Subset	Explanatory	Df	F value	Pr(>F)	
TCRβ1	Eight-week-exposed (n=92)	Host family	9	0.9974	0.4499808	
		MHC-II diversity	5	1.2181	0.3093393	
		Sex	**1**	**11.8363**	**0.0009641**	***
		Parasite load	1	0.1383	0.7110308	
		Cage	2	0.5284	0.5917879	
TCRβ2	Eight-week-exposed (n=93)	Host family	9	1.0086	0.4410672	
		MHC-II diversity	1	1.0950	0.3704563	
		Sex	**1**	**13.3469**	**0.0004806**	***
		Parasite load	1	0.4237	0.5170949	
		Cage	2	2.2782	0.1096124	

Table C.7: Effect of various independent variables on the average observed diversity of TCRβ1 and TCRβ2 repertoires of fish exposed to the lake for eight weeks. Diversity values from 10 independent size normalization iterations. Results from Type III F tests of generalized linear models (GLMs). Statistically significant factors are marked in bold. Results from Type III Wald F tests of generalized linear mixed effect models (GLMMs) using lake cage as random intercept.

Family\Experiment	*S. solidus*	Lake exposure
F2 (E)	6	11 (5**)
F7 (B)	6*	6
F10 (C)	6	3
Total	18	20

* TCRβ1 repertoire sequencing from one individual was unsuccessful.

** Number of control individuals from the four-week-control group.

Table C.8: Dataset of individuals used to examine the effects of aging on the stickleback TCR repertoire, grouped by family. Control fish from both the S. solidus experiment (Chapter 2) and the lake exposure experiment (Chapter 3) were used. In parentheses are the corresponding family names used in Chapter 2.

C.2 Supplementary Figures

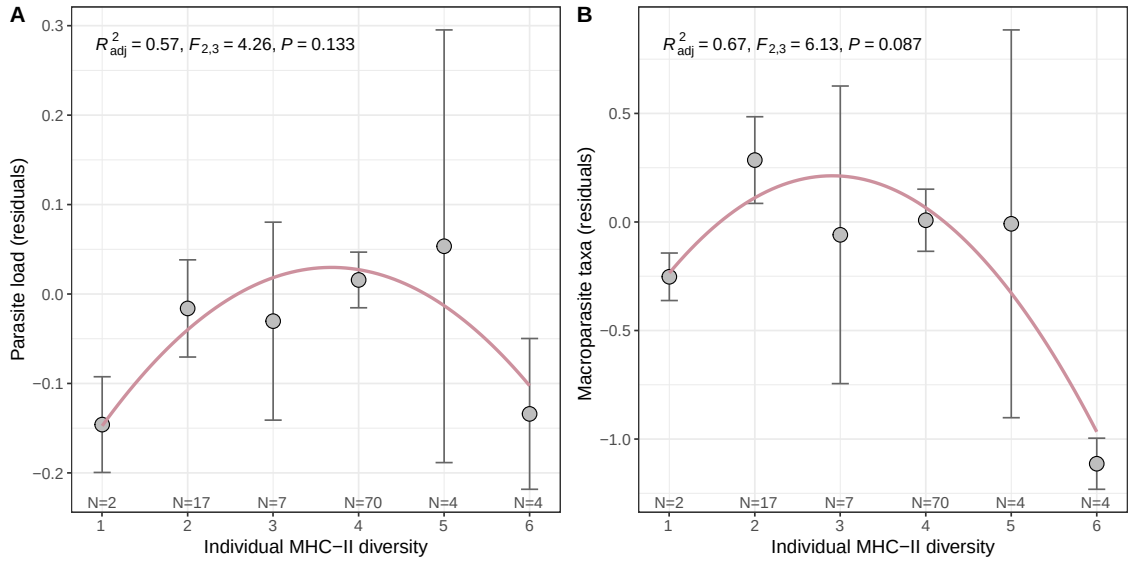


Figure C.1: Effect of individual MHC-II diversity on parasite load and number of macroparasite taxa. Mean **(A)** parasite load residuals ± S.E. and **(B)** number of macroparasite taxa residuals ± S.E. of eight-week-exposed sticklebacks grouped by individual MHC-II diversity. Grey circles indicate the mean value of all individuals possessing the given number of MHC-II alleles. To reduce variation introduced by effects of family, sex, and cage, residuals of a GLM were used for subsequent analysis (Table S3). Accordingly, the GLM for the number of macroparasite taxa included sex and cage as independent variables (Table S3). The displayed quadratic polynomials are fitted to the means. N indicates sample sizes for each diversity level.

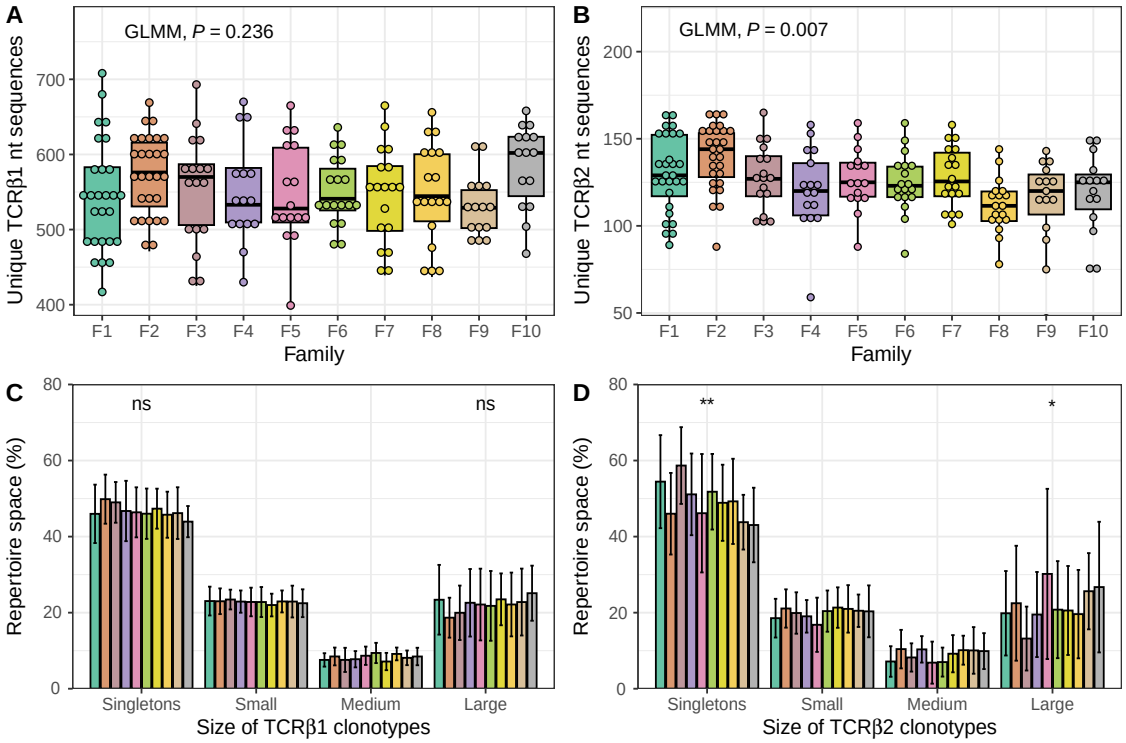


Figure C.2: Intra-individual repertoire diversity (unique nt sequences) of the TCRβ1 (A) and TCRβ2 (B) repertoires was compared between stickleback families. P value for family was calculated using GLMM with family, sex, treatment, duration, and treatment:duration as fixed effects; cage was used as random effect. **C, D) Clonal space homeostasis showing the average percent of clonal space occupied by clones of a given size.** Singletons, small, medium and large clones of the TCRβ1 repertoire (C) were defined as corresponding to 1, 2-4,5-7 and >=8 UMIs, respectively. For the TCRβ2 repertoire (D), the same categories correspond to 1, 2-3,4-5 and >=6 UMIs, respectively. Comparisons were made between families. For singletons and large clonotypes, GLMMs were fitted with family, sex, duration, and treatment as fixed effects and cage as random effect. Asterisks on top of each category indicate P value significance levels for family (**P < 0.01, * P < 0.05, ns non-significant, P > 0.05).

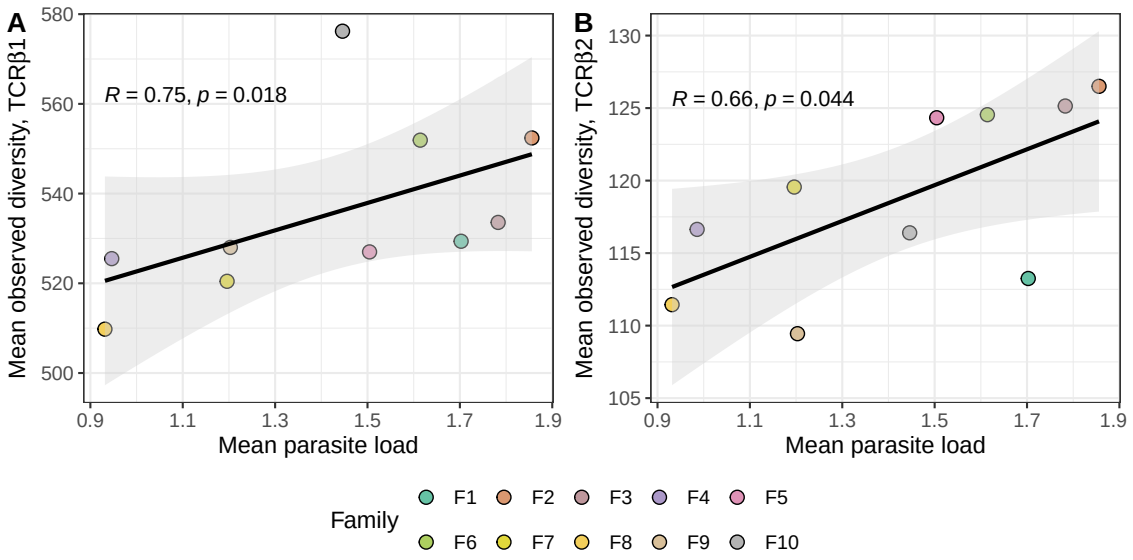


Figure C.3: Correlation between mean TCR repertoire diversity and mean individual parasite load per stickleback family among eight-week-exposed sticklebacks. Analysis was performed separately for the (A) TCRβ1 and (B) TCRβ2 repertoire. Mean per family repertoire diversity (number of unique nt sequences) was calculated based on each individual's average from 10 independent size normalization iterations. Spearman's R and P values depicted on each graph. 95% confidence intervals are indicated by the grey shaded areas.

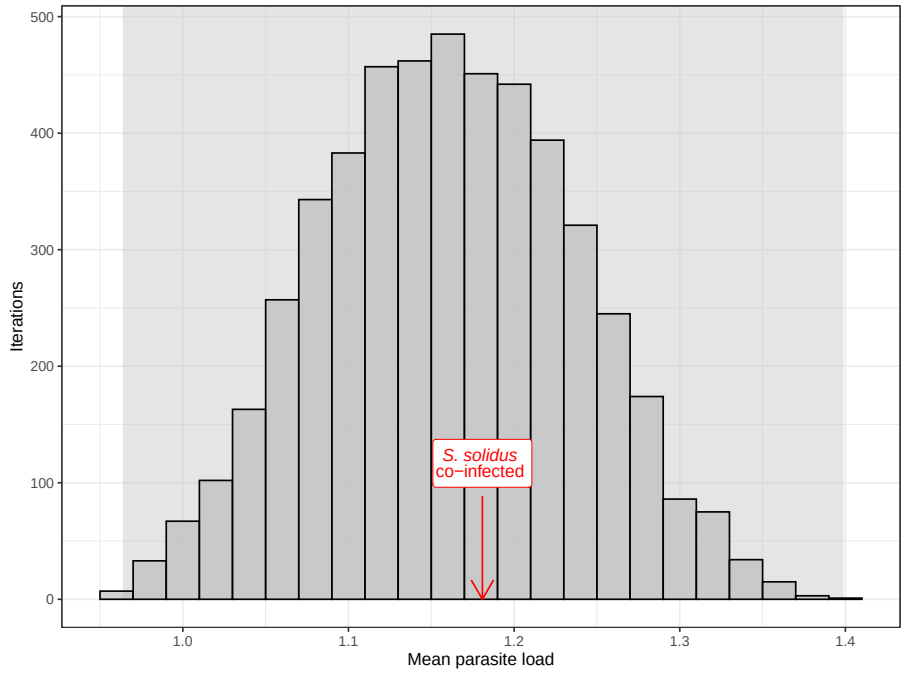


Figure C.4: Comparison of parasite load between eight-week exposed fish who are either co-infected with Schistocephalus solidus or not. Histograms show the distribution of mean parasite load of randomly sub-sampled individuals not co-infected with S. solidus over 5,000 iterations. Grey area marks the 95% confidence interval. Family and cage were controlled so that for each iteration the sub-sampled group (n=9) was matching the S. solidus co-infected group (n=9). Red arrow indicates the mean parasite load of the S. solidus co-infected fish.

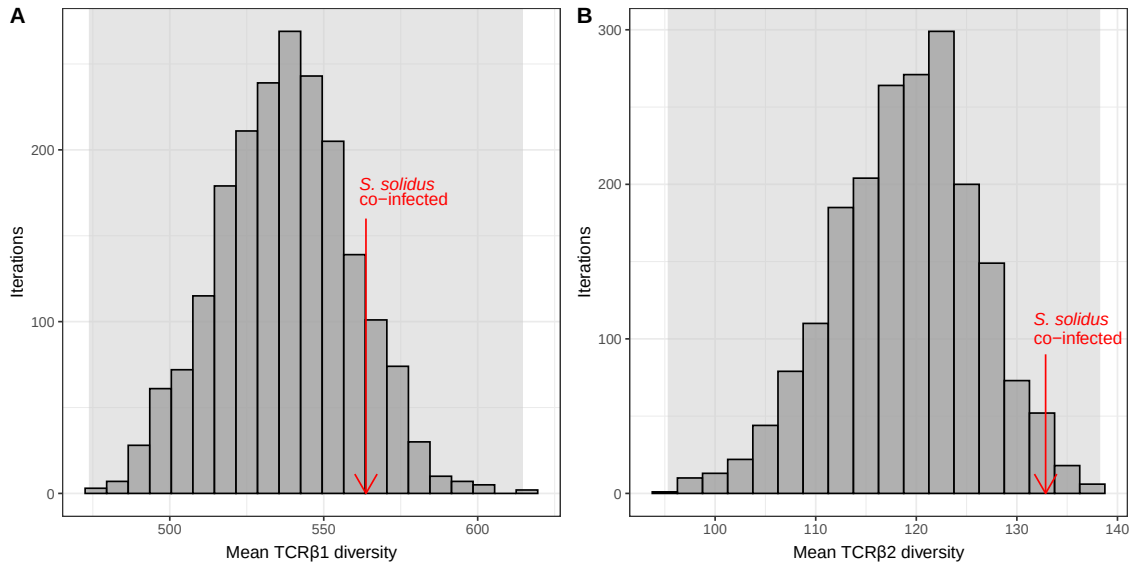


Figure C.5: Comparison of intra-individual repertoire diversity (unique nt sequences) of the TCRβ1 (A) and TCRβ2 (B) repertoires between eight-week exposed fish who are either co-infected with Schistocephalus solidus or not. Histograms show the distribution of mean repertoire diversity of randomly sub-sampled individuals not co-infected with S. solidus over 2,000 iterations. Grey area marks the 95% confidence interval. Sex was controlled so that for each iteration the sub-sampled group (n=7) were matching those of the S. solidus co-infected group (n=7). Red arrows indicate the mean diversity of the S. solidus co-infected fish in each repertoire.

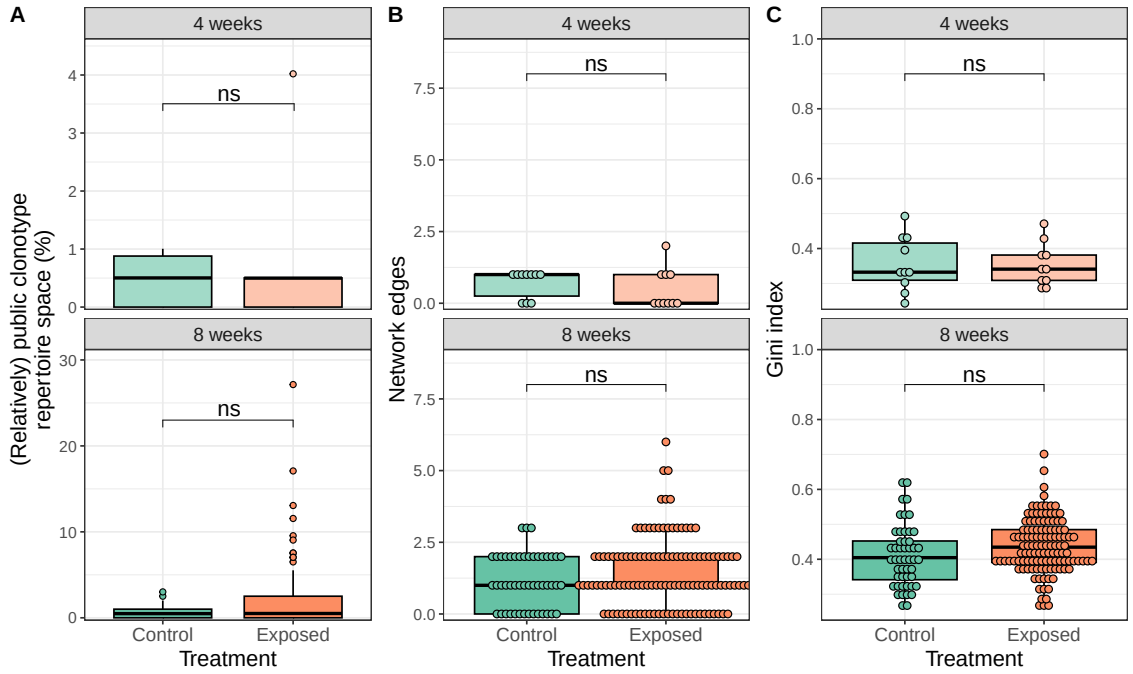


Figure C.6: TCRβ2 repertoire of eight-week-exposed individuals include more highly shared clonotypes. textbf(A) Comparison of the occupied repertoire space (*i.e.*, cumulative frequency) of public and relatively public (n>24) clonotypes among the entire down-sampled TCRβ1 repertoires of control and lake-exposed sticklebacks, by duration of exposure. textbf(B) Comparison of the number of edges connecting nodes in each individual network between control and lake-exposed sticklebacks, by duration of exposure. Networks were constructed using the top 50 most abundant clonotypes of each individual. textbf(C) Comparison of the Gini index (a measure for repertoire evenness) between control and lake-exposed sticklebacks, by duration of exposure. Asterisks in all panels signify P value significance levels for factor treatment, as calculated using GLMMs for each duration group separately, with family, sex, and treatment as fixed effects, and cage as random effect (ns: non-significant P > 0.05).

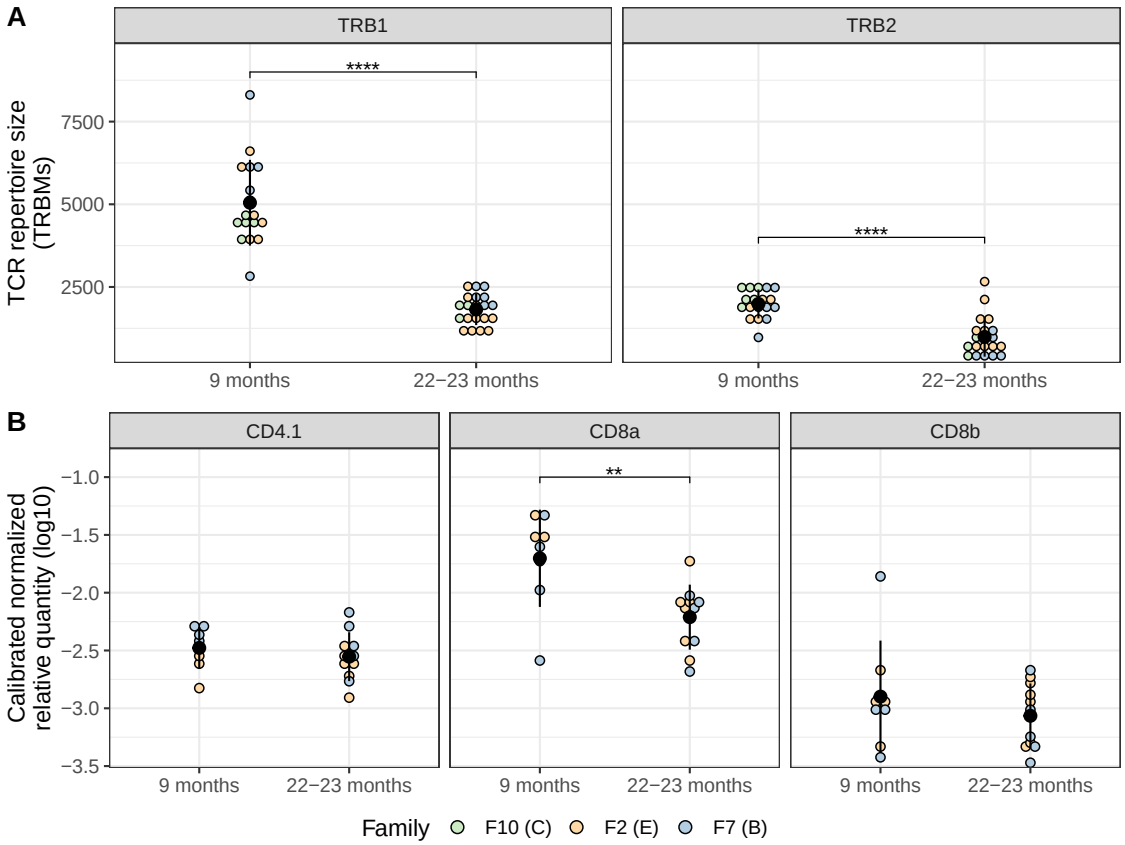


Figure C.7: Effect of aging on the relative size of the TCR repertoire, *i.e.*, size within a given amount of spleen RNA. (A) Comparison of the size of TCRβ1 (left panel) and TCRβ2 (right panel) non-normalized repertoires (*i.e.*, total number of unique functional CDR3 nt sequences) between 9-month-old and 22- to 23-month-old sticklebacks. Each dot represents a single individual, and fill color indicates the individual's family. **(B)** Comparison of relative expression levels of CD4.1, CD8α and CD8β genes between 9-month-old and 22- to 23-month-old sticklebacks. Each dot represents a single individual, and fill color indicates the individual's family. In all panels, asterisks indicate P value significance levels for variable age, as calculated with GLMs with family, sex, and age as independent variables (**** $P<10^{-4}$, **$P < 0.01$).

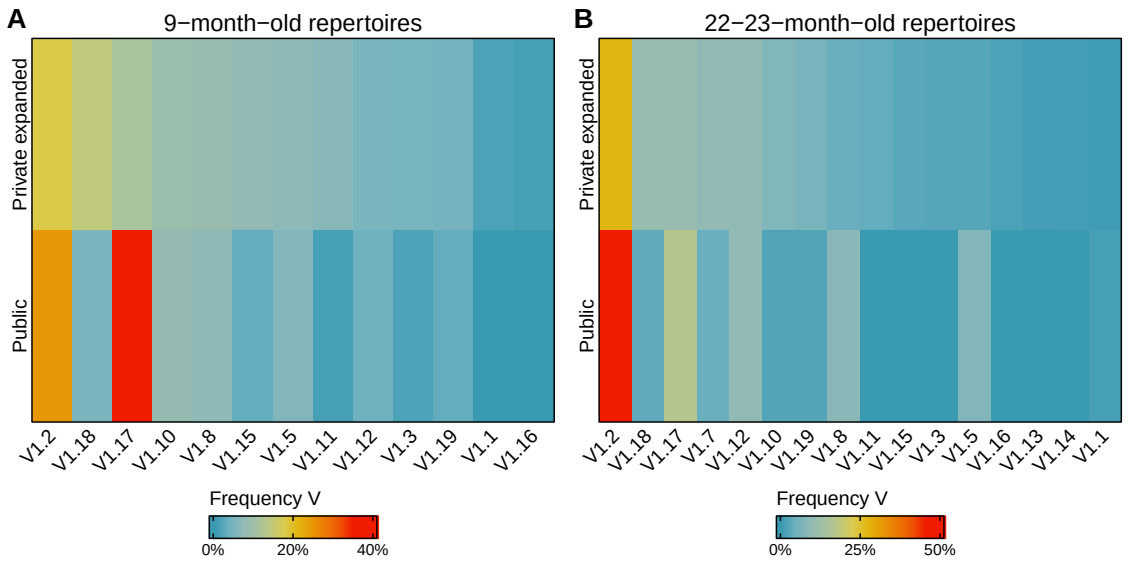


Figure C.8: Heatmaps of frequency of usage of V gene segments within the private expanded and (relatively) public clonotypes of the TCRβ1 repertoires of (A) young and (B) old sticklebacks. Private clonotypes were regarded as expanded if their average intra-repertoire frequency exceeded 0.002. This filtering resulted in 689 private expanded clonotypes among the repertoires of young individuals, and 714 private expanded clonotypes among the repertoires of old ones.

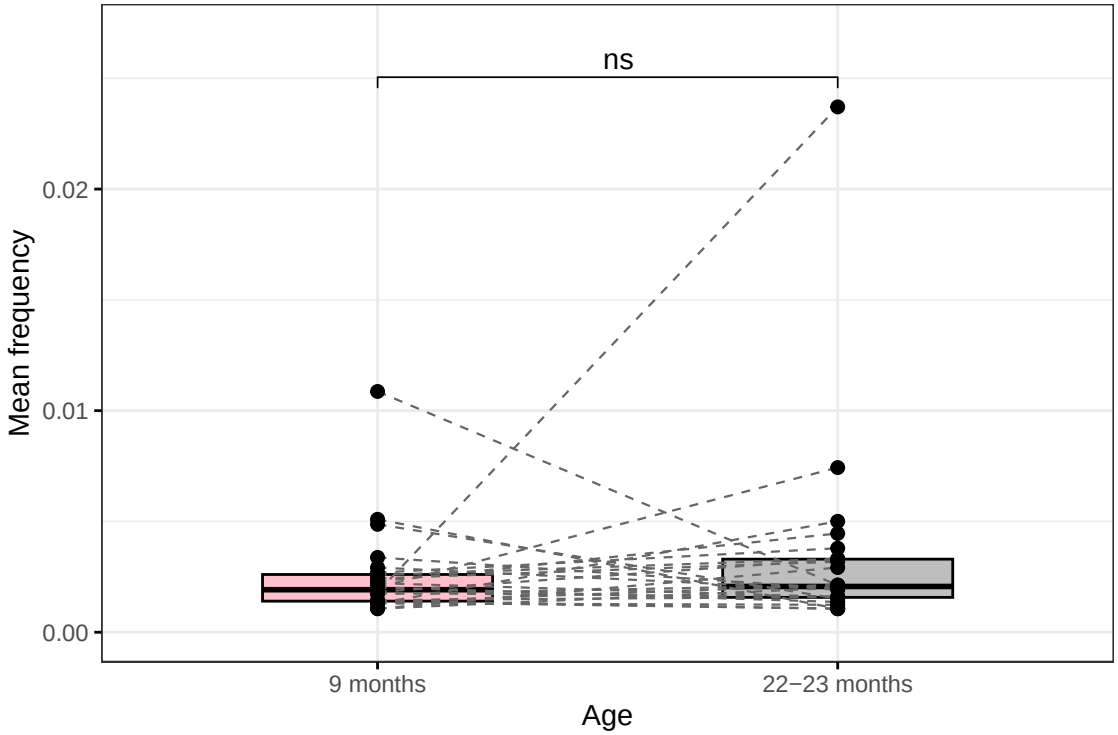


Figure C.9: Average intra-repertoire frequency of the 24 relatively public and public TCRβ1 clonotypes common to both young (nine-month-old) and old (23-month-old) sticklebacks. Each dot represents a single clonotype, and dotted lines connect the same clonotypes across the two groups. P value from the Wilcoxon signed-rank test (P= 0.39).

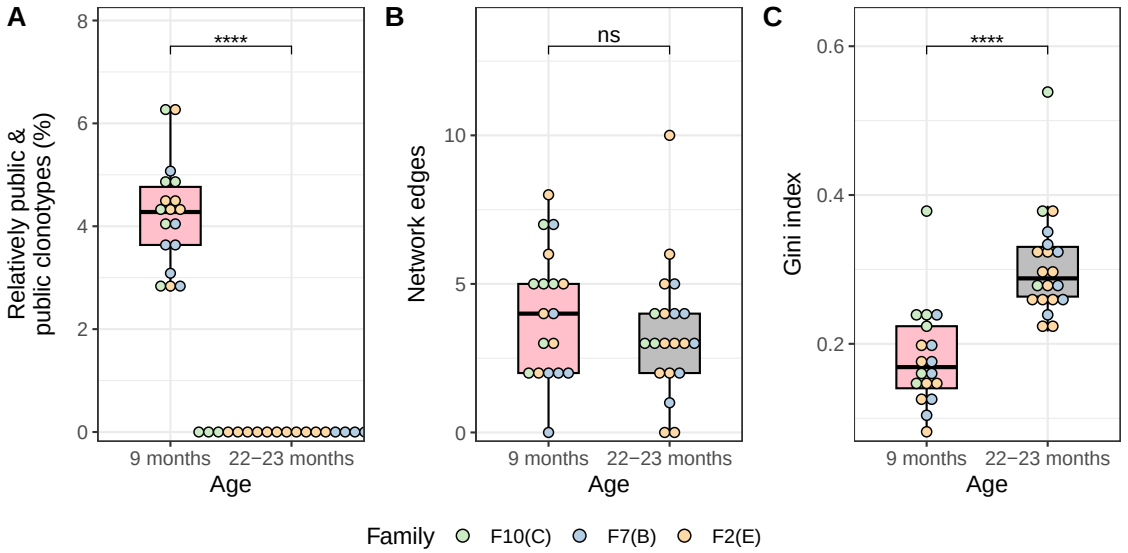


Figure C.10: Effect of aging on the stickleback TCRβ2 repertoire. **(A)** Comparison of the proportion of relatively public and public clonotypes among the entire TCRβ2 repertoires of young (nine-month-old) and old (23-month-old) sticklebacks. Clonotypes were regarded as relatively public and public if present in more than one quarter of the individuals in each group. **(B)** Comparison of the number of edges connecting nodes in individual similarity networks of young and old sticklebacks. Networks were built using the 100 most frequent CDR3 aa sequences expressed in down-sampled TCRβ2 repertoires. **(C)** Comparison of the Gini index (a measure for repertoire evenness) between young and old sticklebacks. Asterisks in all panels signify P value significance levels for factor age, as calculated using GLMs with family, sex, and age as fixed effects (**** $P < 10^{-4}$, ns non-significant). Each dot represents a single clonotype, and dotted lines connect the same clonotypes across the two groups. P value from the Wilcoxon signed-rank test (P= 0.39).