VARIATION IN THE MHC CLASS-II REGION OF THE THREE-SPINED STICKLEBACK: FROM GENOMES TO POPULATIONS

DISSERTATION

in fulfillment of the requirements for the degree "Dr. rer. nat." of the Faculty of Mathematics and Natural Sciences
at Kiel University

submitted by

Malavi Sengupta

Emmy Noether Group for Evolutionary Immunogenomics

Max Planck Institute for Evolutionary Biology

Plön, November 14, 2019

First referee: Dr. Tobias Lenz
Second Referee: Dr. Thorsten Reusch
Date of Oral Examination: 18.12.2019.

Once we were blobs in the sea, and then fishes, and then lizards and rats and then monkeys, and hundreds of things in between. This hand was once a fin, this hand once had claws! In my human mouth I have the pointy teeth of a wolf and the chisel teeth of a rabbit and the grinding teeth of a cow! Our blood is as salty as the sea we used to live in! When we're frightened, the hair on our skin stands up, just like it did when we had fur. We are history! Everything we've ever been on the way to becoming us, we still are.

-Sir Terry Pratchett, A Hat Full of Sky

Table of Contents

Summary

Parasites are ubiquitous in nature and pose a heavy fitness cost to the host. Host populations undergo adaptive genetic changes to resist the parasite, while the parasite population adapts to exploit the host. Such reciprocal, adaptive genetic changes are called host-parasite coevolution. Local adaptation is said to have occurred when a population is better adapted to its own environment than to a remote environment, and host-parasite coevolution can result in the host's local adaptation to its parasite environment. When different environments exert different selective pressures on populations with connected geographic ranges, these populations can adapt locally and diverge from one another, eventually leading to reproductive isolation or ecological speciation. For a trait to play a role in this, it must influence the carrier's ecological interactions, such as salinity tolerance, and gill rakers. This trait may additionally have a pleiotropic effect upon mate choice, and are sometimes called "magic traits".

One excellent model for the genomic basis of ecological speciation under parasite-mediated selection that may be a magic trait is the Major Histocompatibility Complex (MHC) genomic region, which contains the MHC genes. The MHC gene products are cell-surface proteins that bind small peptides- called epitopes- and present them to other parts of the immune system, thus helping distinguish between self and non-self antigens, and forming a major part of the vertebrate's adaptive immunity. MHC genes are divided into 3 classes: Class-I, Class-II, and Class-III. In different parasite environments, populations face different selective pressures, and this causes them to adapt to their own environments, and diverge from each other.

Here, we study the MHC Class-II genomic region in the three-spined stickleback (Gasterosteus aculeatus), a bony fish (teleost) species native to the Northern Hemisphere. Pairs of populations of sticklebacks can be thought to exist in a speciation continuum, from a panmictic population, through partial and reversible isolation, to irreversible ecological speciation. This gives us the opportunity to tease

apart the factors that move population pairs towards either end of this continuum. However, the genomic basis of traits that contribute to reproductive isolation still needs extensive study. The purpose of this thesis was to fill these gaps in our knowledge, leading to a deeper understanding of the genomics of speciation.

The primary objective of this PhD thesis is to create a multi-layered understanding of the MHC region in the genome of the three-spined stickleback. The first layer was the identification and characterization of this region in the context of the stickleback reference genome. It was found that the MHC Class-II region of the three-spined is located at one telomeric end of Chromosome VII, and is split into two parts, a Classical region 0.3 MB from the end, and the Non-Classical region 2.4 MB away, closer to the centromere. Many MHC Class-II antigen presentation pathway genes were found on other Chromosomes such as Chr III, and Chr IV.

The next part of this work was to characterize the enormous amount of allelic and haplotypic variation at the MHC Class-IIB loci. For this, we annotated the genes of the stickleback MHC Class-II region using novel genomic sequences of new haplotypes with different numbers of functional MHC loci, using BAC (Bacterial Artificial Chromosome) library sequence data. After annotating functional genes, we turned our attention to regulatory elements in this region, which tells us about differences in regulation between various MHC alleles. Only one confirmed Immune System gene, Compliment Factor 1, was found in the vicinity of the Classical MHC Class-II loci. The annotation allowed us to define bounding genes for both the Classical and Non-Classical MHC Class-II regions, between which lie functionally similar DNA sequences. These were extracted and compared in the next section.

Previous studies of MHC Class-II sequence variants in different stickleback populations together had revealed that the number of functional MHC alleles that are inherited together varies between haplotypes, indicating potential Copy Number Variation. This led to a more detailed study of synteny in this region, finding blocks of recombination that account for differences in structure between haplotypes. The

annotation and synteny work led to the creation of the stickleback MHC Class-II region map, showing genes, promoters, and collinear blocks in three different haplotypes.

However, we still wanted to understand the process by which these differences were created. So, we studied transposable elements found in this region. We found transposable elements that accounted for indels, as well as, possibly, the number and arrangement of MHC loci. This gave us a model for the origin of the extensive CNV in this region.

Finally, we turned our attention to population-level variation in the stickleback MHC Class-II region. From a previously generated whole-genome-sequencing dataset, we were able to study sequence variation in the assembled region, as well as other genomic regions as controls. We performed tests for selection and neutrality, giving us an insight about the evolutionary history of the stickleback species, with respect to parasite-mediated selection.

The three-spined stickleback has undergone recent, repeated adaptive radiations. In this process, sticklebacks must have encountered many novel parasite environments, and fast adaptation was critical. Because of the genomic organization in the MHC Class-II regions, stickleback were able to make novel haplotypes from old ones quickly, and have excessive allelic and copy number variation. This would allow for populations in different environments to have distinct allele pools. This could be one of the reasons why sticklebacks were able colonize different environments so successfully. As such, the organization of the stickleback MHC Class-II genomic region would be a factor that could move populations quickly along the speciation continuum.

Zusammenfassung

(German translation by Nico Fuhrmann.)

Parasiten sind in der Natur allgegenwärtig und verursachen für den Wirt hohe Fitnesskosten. Wirt-Populationen unterliegen adaptiven genetischen Veränderungen um dem Parasiten zu widerstehen, während sich die Parasitenpopulation anpasst, um den Wirt auszunutzen. Solche reziproken, adaptiven genetischen Veränderungen werden Wirt-Parasit-Koevolution genannt. Von lokaler Adaptation spricht man, wenn eine Population besser an das eigene Habitat als an andere angepasst ist, und Wirt-Parasit-Koevolution kann zu einer lokalen Adaptation des Wirts an die Parasiten Umgebung führen. Wenn unterschiedliche Umgebungen verschiedenne Selektionsdrücke auf Populationen in angrenzenden geografischen Bereichen ausüben, können sich diese Populationen lokal anpassen und voneinander abweichen, was schließlich zu einer reproduktiven Isolierung oder ökologischen Speziesbildung führt. Damit ein Merkmal dabei eine Rolle spielt, muss es die ökologischen Wechselwirkungen des Trägers beeinflussen, wie die Salinitätstoleranz und Kiemenreusendornen. Solche Merkmale können zusätzlich eine pleiotrope Wirkung auf die Partnerwahl haben, manchmal als "magic traits " bezeichnet.

Ein ausgezeichnetes Modell für die genomische Basis der ökologischen Speziation unter Parasiten-vermittelter Selektion, das ein ”magic trait” sein könnte, ist die Major Histocompatibility Complex (MHC) Genom-Region, die die MHC-Gene enthält. Die MHC-Genprodukte sind Zelloberflächenproteine, die kleine Peptide, sogenannte Epitope, binden und sie anderen Teilen des Immunsystems präsentieren, wodurch die Unterscheidung zwischen körpereigene und körperfremde Antigenen erleichtert wird und ein Hauptbestandteil der adaptiven Immunität des Wirbeltiers ist. MHC-Gene sind in drei Klassen unterteilt: Klasse I, Klasse II und Klasse III. In verschiedenen Parasitenumgebungen sind die Bevölkerungsgruppen

unterschiedlichem Selektionsdruck ausgesetzt, und dies führt dazu, dass sie sich an ihre eigenen Umgebungen anpassen und voneinander abgrenzen.

Hier untersuchen wir die genomische MHC-Klasse-II-Region im Dreistachligen Stichling (Gasterosteus aculeatus), einer in der nördlichen Hemisphäre beheimateten knöchernen Fischart (Teleost). Es kann angenommen werden, dass Paare von Stichlingpopulationen in einem Speziationskontinuum existieren, von einer panmiktischen Population über eine teilweise und reversible Isolierung bis hin zu einer irreversiblen ökologischen Speziation. Dies gibt uns die Möglichkeit, die Faktoren auseinanderzuhalten, die Bevölkerungspaare zu beiden Enden dieses Kontinuums bewegen. Die genomische Grundlage von Merkmalen, die zur reproduktiven Isolierung beitragen, muss jedoch noch eingehend untersucht werden. Ziel dieser Arbeit war es, diese Wissenslücken zu schließen und ein tieferes Verständnis der Genomik der Speziation zu ermöglichen.

Das Hauptziel dieser Doktorarbeit ist es, ein vielschichtiges Verständnis der MHC-Region im Genom des Dreistachligen Stichlings zu schaffen. Die erste Schicht war die Identifizierung und Charakterisierung dieser Region im Kontext des Stichling-Referenzgenoms. Es wurde festgestellt, dass sich die MHC-Klasse-II-Region des Dreistachligen Stichlings an einem telomeren Ende von Chromosom VII befindet und in zwei Teile untergliedert ist, eine klassische Region 0,3 MB vom Ende entfernt und die nicht-klassische Region 2,4 MB entfernt näher am Centromer. Viele Gene für den MHC-Klasse-II-Antigenpräsentationsweg wurden auf anderen Chromosomen wie Chr III und Chr IV gefunden.

Der nächste Teil dieser Arbeit bestand darin, die enorme Menge an allelischen und haplotypischen Variationen an den MHC Class-IIB-Loci zu charakterisieren. Zu diesem Zweck haben wir die Gene der Stichling-MHC-Klasse-II-Region mit neuartigen Genomsequenzen neuer Haplotypen mit unterschiedlichen Anzahlen von funktionellen MHC-Loci unter Verwendung von Sequenzdaten der BAC-Bibliothek (Bacterial Artificial Chromosome) annotiert. Nachdem funktionelle Gene annotiert

wurden, haben wir unsere Aufmerksamkeit auf regulatorische Elemente in dieser Region gerichtet, die uns über Unterschiede in der Regulation zwischen verschiedenen MHC-Allelen informieren. In der Nähe der Classical MHC Class-II-Loci wurde nur ein bestätigtes Immunsystem-Gen, Compliment Factor 1, gefunden. Die Annotation erlaubte es uns, Grenzgene sowohl für die klassische als auch für die nichtklassische MHC-Klasse-II-Region zu definieren, zwischen denen funktional ähnliche DNA-Sequenzen liegen. Diese wurden extrahiert und im nächsten Abschnitt verglichen.

Frühere Studien von MHC-Klasse-II-Sequenzvarianten in verschiedenen Stichproben Populationen haben herausgefunden, dass die Anzahl der funktionellen MHC-Allele, die zusammen vererbt werden, zwischen den Haplotypen variieren, was auf eine mögliche Kopienzahlvariation hinweist. Dies führte zu einer detaillierteren Untersuchung der Syntenie in dieser Region, wobei Rekombinationsblöcke gefunden wurden, die Unterschiede in der Struktur zwischen den Haplotypen erklären. Die Annotations- und Syntenieарbeit führte zur Erstellung der Stichling-MHC-Klasse-II-Regionskarte, die Gene, Promotoren und kollineare Blöcke in drei verschiedenen Haplotypen zeigt.

Wir wollten jedoch immer noch verstehen, durch welchen Prozess diese Unterschiede entstanden sind. Also haben wir Transposons untersucht, die in dieser Region gefunden wurden. Wir fanden Transposons, die Indels sowie möglicherweise die Anzahl und Anordnung der MHC-Loci ausmachten. Dies gab uns ein Modell für die Entstehung des ausgedehnten CNV in dieser Region.

Schließlich, richteten wir unsere Aufmerksamkeit auf die Population-Ebenen Variation der Stichling-MHC-Klasse-II-Region gerichtet. Auf Basis eines bereits vorhandenen whole-genome-sequencing Datensatz konnten wir Sequenz Variation in der assemblierten Region und als Kontrolle andere genomischen Regionen studieren. Wir haben Selektions- und Neutralitäts-Test durchgeführt, was uns

Einblicke über die evolutionäre Geschichte der Stichling Spezies ermöglicht, im Hinblick auf Parasiten-vermittelter Selektion.

Der Dreistachelige Stichling erfuhr rezente, wiederholte adaptive Radiation. In diesem Prozess müssen Stichlinge viele neue Parasiten Umgebungen ausgesetzt gewesen sein, weshalb schnelle Adaptation kritisch war. Durch die genomische Organisation in der MHC-Klasse-II-Region, Stichlinge konnten rasch neue Haplotypen aus alten generieren und haben nun exzessive Allel- und Kopienzahlvariation. Dies würde Populationen ermöglichen in verschiedenen Umgebungen charakteristische Allel Pools zu haben. Es könnte ein Grund sein, warum Stichlinge in der Lage waren verschiedene Umgebungen so erfolgreich zu kolonisieren. Darum wäre die Organisation der Stichling-MHC-Klasse-II-Region ein Faktor, welcher Populationen rasch entlang des Speziation Kontinuum bewegen könnte.

Glossary of Terms

Allele: Alternative sequences for the same locus, in the classical sense. Here, this means sequence variant, as genomic locus is not known.

Allele Pool: The sum total of all the alleles carried by all the individuals in a population.

Allopatry: The state of two populations, which are geographically isolated from one another.

Antibody: A protein that binds specifically to any substance.

Antigen: Any molecule that triggers an immune response.

Chromosome/Linkage Group: A sequence of nucleic acid, packaged by proteins, carrying genetic information in the form of genes.

Contig: A continuous sequence of DNA that was assembled from overlapping cloned fragments. Here, contig also refers to a part of a haplotype that is on one continuous DNA sequence.

EST: Expressed Sequence Tag, which is a small portion of a gene that can help identify it.

Fitness: How good a genotype is at contributing to the next generation, relative to other genotypes.

Haplotype: A set of alleles inherited together. In the stickleback MHC Class-II region, each haplotype consists of a set of polymorphic loci that are linked and inherited together. The monomorphic locus is found in all fish examined so far, so it is not considered a part of the haplotype.

Homologs: Genes that are similar due to shared ancestry.

Immune System: An organism's defense system against disease.

Indel: Insertion/Deletion, a type of structural variation.

Intergenic sequence: The sequence between two genes.

Invariant/Monomorphic Locus: A locus that has only one allele.

Linkage: The phenomenon where two DNA sequences are inherited together.

Linkage Disequilibrium: This is the non-random association of alleles at different loci in a population.

Locus: A fixed position in a chromosome where a gene is located.

MHC: Major Histocompatibility Complex.

MHC region: The region of the genome that the MHC loci are in.

MHC Class II locus: The combination of corresponding MHC Class-IIA and B loci.

Orthologs: Genes in different species that are similar due to shared ancestry.

Paralogs: Genes that are related due to duplication within the genome.

Parapatry: The state of two populations whose ranges are adjacent, but do not significantly overlap.

Pleiotropy: The phenomenon when one gene affects many, unrelated traits.

Polygenic trait: A polygenic trait is one whose phenotype is influenced by more than one gene.

Pseudogene: A pseudogene is a sequence with the promoters and start codon of a genic locus, but, due to mutation, has an early termination signal or a truncated exon. As such, it does not code for a functional protein.

Reference Genome: The Three-spined stickleback genome first sequenced and assembled by Jones et. al. 2012, then re-assembled and published by Glazer et. al. in 2015. In this work, "Reference Genome" refers strictly to the version published by Glazer et. al.

Resistance: Genetic, biochemical, or physiological characteristics of the host that inhibit pathogen establishment, survival or development.

Scaffold: These are made by assembling non-overlapping contigs together with gaps of known length.

SNP: Single Nucleotide Polymorphism.

Sympatry: The state of two populations whose ranges overlap.

Teleost: Bony fish.

UTR: UnTranslated Region, the part of a transcript that is not translated.

Variable/Polymorphic Locus: A locus that can be occupied by many alternative alleles.

WGS: Whole Genome Sequencing

Introduction

Host-parasite coevolution

A parasite is an organism that lives in association with, and derives nutrients from, another organism (the host). Parasites are ubiquitous in nature, and impose a significant cost to the survival and reproduction of its host (Karvonen and Seehausen 2012). As such, hosts and parasites have mutually incompatible fitness optima: it is beneficial for the parasite to extract the largest amount of nutrients, while it is beneficial for the host to resist the parasite. So, parasites are constantly evolving to exploit the host better, while hosts are evolving to resist the parasite more effectively. As such, each organism must continue to evolve in order to remain at the same level of fitness, which is called the Red Queen Effect (Van Valen 1973). These reciprocal, adaptive genetic changes in two organisms in a parasitic relationship are called host-parasite coevolution (Woolhouse et al. 2002). At the population level, host-parasite coevolution can drive divergence in sympatric or allopatric population pairs (Buckling and Rainey 2002).

Local adaptation

Local adaptation is said to have occurred when a population is better adapted to its own environment than to a remote environment (Gandon and Michalakis 2002). Divergence in fitness-related traits only occurs between different, connected populations when they experience spatially heterogeneous selective pressures. If selective pressures are the same, gene flow will eliminate differences in these traits. On the other hand, if a trait confers a general selective advantage, it will be fixed in all populations. So, local adaptation in populations connected by gene flow is a sign of ongoing, or recent natural selection. Moreover, local adaptation provides a context in which the interaction between ecological conditions and fitness-conferring traits can be studied. Finally, they are immensely useful in the study of repeatability: for example, do the same ecological conditions select for the same traits in different

populations? (Kawecki and Ebert 2004). As such, model systems undergoing recent, repeated divergence are excellent for the study of local adaptation and its consequences.

In nature, ecosystems have distinct parasite communities, and host species interact repeatedly with the same parasites in their range, creating conditions for host-parasite coevolution to occur. From the host's perspective, this involves optimizing its immune system constantly to a parasite that becomes better and better at evading it. This can be seen as a case of local adaptation of hosts to their parasite environment. Under this regime, gene flow will be reduced between connected populations, as hybrids will not be optimized for either parent's environment. This causes populations to diverge, possibly leading to speciation (Eizaguirre and Lenz 2010).

Ecological Speciation

Speciation is the evolution of genetically distinct populations maintained by reproductive isolation in the case of sexual taxa. Sympatric speciation is the evolution of reproductive isolation in populations with broadly overlapping geographical ranges. In nearly all models of sympatric speciation, a panmictic population is affected by disruptive selection, driving an evolutionary change in mating patterns, and this disruptive selection may have an ecological basis (Bolnick and Fitzpatrick 2007). In other words, Ecological speciation is defined as the evolution of reproductive isolation between populations due to ecologically divergent natural selection. Here, different environments exert different selective pressures on populations with connected geographic ranges. Alleles that are beneficial in one environment become more common the corresponding population, and the same happens for alleles that are beneficial in the other population. This leads to reproductive isolation in ecological speciation (Schluter and Conte 2009).

Ecological speciation has been demonstrated in laboratory settings, and has also been implicated in speciation events in nature. An especially important case is that

of parallel speciation, where reproductive isolation has developed independently in replicated populations, in correlation with the environment. To understand ecological speciation, it must be divided into three components: an ecological source of divergent selection, a mechanism of reproductive isolation, and an underlying genetic basis. The empirical data on the genetic basis of ecological speciation indicates that such genes can vary in number, effect sizes, and dominance and epistatic interactions. The contributions of mutations versus standing variations are also not known. For these reasons, it is important to study ongoing ecological speciation, for which the geographical regions are known (Rundle and Nosil 2005).

For a trait to play a role in speciation, it must influence the carrier's ecological interactions. Some of these may also, have a pleiotropic effect upon mate choice. That is, an individual must have an increased likelihood of choosing another individual with a similar trait value (Bolnick and Fitzpatrick 2007). Such traits are sometimes called "magic traits", which are not rare in theory, but have few demonstrated examples, such as body size, body shape, and color pattern (Servedio et al. 2011).

Major Histocompatibility Complex

In vertebrates, there is an excellent model for the genomic basis of ecological speciation under parasite-mediated selection that may even be a magic trait. It is the Major Histocompatibility Complex (MHC) genomic region, which contains the MHC genes. The MHC gene products are cell-surface proteins that bind small peptides-called epitopes- and present them to be recognized by other parts of the immune system, in a process called antigen presentation. They help distinguish between self and non-self antigens, and therefore form a major part of the vertebrate's adaptive immunity. MHC genes are divided into 3 classes: Class-I, Class-II, and Class-III. MHC Class-I gene products are present on all nucleated cells, and display epitopes derived from intracellular proteins. Class-II gene products are present on specialized immune cells called antigen-presenting cells (APC's), and present epitopes derived from extracellular proteins. The function of the MHC Class-III is less defined, but in

humans it is found between the Class-I and II genomic regions. MHC Class I and II genes also have two types, classical and non-classical. Classical MHC genes have known function, and are highly polymorphic, while non-classical MHC genes are not as polymorphic. The Classical MHC Class-II gene product is a heterodimer of the alpha and beta chains, which are coded by the Class-IIA and IIB genes (Janeway et. al. 2001). MHC genes have been studied extensively for the role they play in natural selection (Fig. 1)(Milinski 2006).

The MHC region is present in all jawed vertebrates, with sharks being the oldest group to have them. In mammals, the MHC Class I, II, and III genes are linked, and found together in the same chromosome. In teleosts, unlike in mammals, the class I and II regions are not linked, and neither are class III genes found nearby(Flajnik and Kasahara 2001)(Flajnik and Kasahara 2001)(Flajnik and Kasahara 2001)(Flajnik and Kasahara 2001)(Flajnik and Kasahara 2001)(Flajnik and Kasahara 2001). This may be the result of a whole genome duplication event in the Euteleost ancestor, and it allows genes of the different classes to evolve independently, under different selection pressures (Flajnik and Kasahara 2001).

Also in contrast with mammals, the MHC region of teleosts consists of an unknown number of MHC genes, including core genes (i.e., genes involved in peptide transport, loading, and presentation) that may not be linked to each other. For example, it was found by searching the Zebrafish draft genome with gene sequences from the human MHC region (Classes I, II, III and extended Classes I and II) that the MHC region of zebrafish consists of a set of core MHC genes found together on Chromosome 19, while others are spread genome-wide. (Sambrook, Figueroa, and Beck 2005). This makes the MHC region of teleosts an especially challenging area of study.

MHC is a highly polymorphic and gene-dense region and plays a crucial role in parasite resistance- and is therefore subject to parasite-mediated selection. As such, MHC can be a basis of local adaptation. In different parasite environments, populations face different selective pressures, which causes them adapt to their own

environments, and diverge from each other (Eizaguirre et al. 2009; Eizaguirre and Lenz 2010; Lenz et al. 2009). Another characteristic of the MHC genes is that they are associated with traits that are sexually selected. MHC-based mate choice allows female individuals to choose males within intermediate genetic distance of their own MHC genotype, in order to maximize the fitness of their offspring (Milinski 2006).

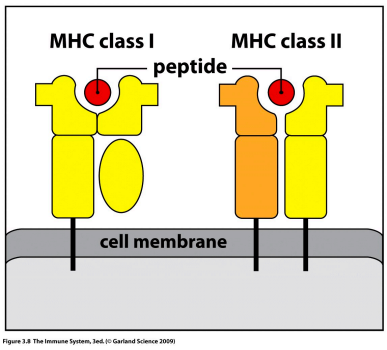
Three-Spined Stickleback

Here, we study the MHC genomic region in the three-spined stickleback *(Gasterosteus aculeatus)*, a bony fish (teleost) species native to the Northern Hemisphere (Fig. 2). It is an excellent model for speciation because many populations of sticklebacks have recently and repeatedly (and independently of each other) undergone adaptive radiation. Towards the end of the last Ice Age, marine sticklebacks colonized the lakes and streams created by melting glaciers. Over the course of ten to fifteen thousand years, the different populations diverged due to different ecological conditions in different habitats. As such, there are now sticklebacks that differ among each other on the basis of many different traits such as body size, skeletal structure, etc., with different levels of reproductive isolation. The recency of stickleback divergence also indicates that the basis of their divergent traits may be simple (Kingsley 2011)(Bell and Foster 1994).

Since marine sticklebacks are likely the ancestors of modern freshwater populations, studying them helped us understand the standing variation and underlying genomic architecture that provided the basis for their rapid adaptive radiation. For this, six individuals from the North Sea were collected and sequenced on the Illumina platform using paired-end and mate-pair libraries. From these individuals it was found that 7% of the genome is polymorphic, with SNP's, indels, SV's, CNV's, and inversions, many of which affect coding regions. This standing variation in the marine populations may have enabled sticklebacks to rapidly diversify and adapt in order to colonize novel freshwater environments (Feulner et al. 2013).

As sticklebacks are relatively easy to collect and breed, they are very suitable for the study of adaptive radiation, trait evolution, etc. Moreover, fully viable hybrids may be generated in the laboratory, and these hybrids are powerful tools for studying the genetic and genomic basis of evolutionary divergence (Kingsley 2011; Bell and Foster 1994).

Pairs of populations of sticklebacks can be thought to exist in a speciation continuum, from no reproductive isolation, through partial and reversible isolation, to irreversible reproductive isolation, and this gives us the opportunity to tease apart the factors that move populations towards either end of this continuum (Fig. 3). However, the genomic basis of traits that contribute to reproductive isolation still needs extensive study. Research on sticklebacks has helped develop, and provide support for, the theory of ecological speciation, which states that divergent adaptations cause the development of reproductive isolation. However, many unanswered questions remain in this field, such as the genetic architecture of traits, and the exact factors, both environmental and genomic, that cause populations to proceed towards, or away from reproductive isolation (Hendry et al. 2013).

1. 

2. 

Figure 1: MHC Gene Products, showing Class-I and Class-II molecules presenting a peptide, and 2: A female Three-Spined stickleback (photo by Ana Teles).

Figure 3: Stickleback Speciation Continuum (adapted from Hendry et. al. 2013), showing two extremes- a panmictic population and complete reproductive isolation- and the space between them that a pair of populations can occupy.

The Northern German Stickleback Population

The stickleback populations in Northern Germany have an enormous amount of allelic variation at the MHC Class-IIB loci. Many sequence variants for the MHC class-II loci have been identified using RSCA typing, which is a PCR-based protocol to differentiate MHC Class-IIB alleles using their exon-2 sequence. Each individual fish was found to have between 1-7 highly divergent alleles, out of which one, called No19, was found in all fish examined so far. However, the Class-IIA alleles are not typed by this technique, and their polymorphism has not been studied so far (Lenz et al. 2009; Eizaguirre et al. 2012). The North American populations have even more alleles. At the population level, the average number of sequence variants per fish is correlated with the parasite diversity of that environment, indicating local adaptation (Eizaguirre et al. 2009), and it follows an optimum(Wegner, Reusch, and Kalbe 2003).

In addition, distinct combinations of these sequence variants are inherited together, that is, they are segregating in haplotypes in the population, but different haplotypes

can also contain different numbers of sequence variants, which suggests that the number of MHC loci in sticklebacks can also vary, that is, there may be copy number variation (CNV) in the MHC Class II locus (Lenz et al. 2009). The genomic basis of this variability, and if CNV is actually present, is not known. This variation found in the MHC region makes this region a fascinating area of study, while also making it difficult to resolve its structure while examining the whole genome.

Stickleback Genome and MHC Class-II

One genome sequence of the three-spined stickleback, and several assemblies were already present. The genome of a stickleback individual from Bear Paw Lake in Alaska was sequenced at BROAD Institute, and was deposited on Ensembl (Jones et al. 2012), which was improved upon in 2015 by placing several of the scaffolds (including two relevant for the MHC) from the Jones et. al. assembly into linkage groups (Glazer et al. 2015). However, it is not known whether this individual was homozygous or heterozygous. The Glazer assembly was used in all steps of this analysis, and it is therefore called the Reference Genome. Chromosome VII (and to a smaller extent Chromosome III) from the Reference Genome was used in our studies.

In the stickleback genome assembly deposited on Ensembl, MHC class II loci were found on Scaffolds 131 and 129, and also on one telomeric end of Chromosome VII. In a more recent assembly, (Glazer et al. 2015) those two scaffolds were placed on the same telomeric end of ChrVII. An Ensembl search of the stickleback genome also showed an MHC Class-IIB locus on Chromosome III, which was also found on the new assembly. So, the locus on the new assembly was also investigated.

Previously, a BAC library was created and a clone containing the MHC class II region was isolated in order to analyze a 99.5kb segment in stickleback genome to find two sets of paralogous MHC class II alpha and beta genes in a tandem arrangement designated Gaac-DAA/DAB and Gaac-DBA/DBB, out of which both are expressed. No association with MHC class II genes of zebrafish was found, suggesting that genome

organization of this region is markedly different between these two species, and possibly within all bony fishes. The genes were found to have been duplicated recently, and evidence for inter-locus gene conversion was also found (Reusch, Schaschl, and Wegner 2004).

It was already known that the linkage between MHC Class I, II, and III genes that is found in mammals is not found in teleosts (Flajnik and Kasahara 2001). MHC antigen presentation pathway genes were found scattered all over the genome in Zebrafish (Sambrook, Figueroa, and Beck 2005), so these genes were sought in the entire Stickleback Reference Genome.

Promoters

After finding the protein-coding sequences and transposable elements in the haplotypes, promoter regions were sought. Promoters are defined as DNA regions that drive the transcription of target regions in response to environmental signals. They are cis-regulatory regions, and they are present upstream of the target sequence. The promoter is divided into three parts; the core promoter, the proximal promoter and the distal promoter. The core promoter is the minimum basic required sequence that can direct the transcription of a gene (Zhang 2007).

The core promoter is recognized by transcription factors that ultimately activate RNA Polymerase II, a multi-subunit enzyme that transcribes DNA into messenger RNA, leading to the start of transcription. The core promoter includes the transcription start site (TSS), as well as upstream or downstream sequences. Some motifs are found in many core promoters, for example, the first eukaryotic promoter discovered, the TATA box, usually located approximately 25 bp upstream of the TSS. However, the TATA box is not universal, neither is any other sequence (Butler and Kadonaga 2002).

Thus, it can be seen that the study of cis-regulatory elements is important for understanding the MHC loci: the presence or absence of regulatory elements upstream from a specific MHC locus indicates whether or not this gene can be transcribed. Differences in promoters between loci can lead to differences in gene regulation.

The promoters of MHC genes are also well-conserved, and are necessary for both constitutive and cytokine-induced gene expression (Ting and Trowsdale 2002). They are called the W/S, X1, X2, and Y-elements, which are found 250 bp upstream of the transcription start site. The Y element is identical to an inverted CCAAT box, another type of promoter (Van Den Elsen et al. 1998). These, and more general promoters were sought upstream of the MHC loci.

Structural Variations

Studying the different MHC Class-II sequence variants in different stickleback populations together revealed that up to 4 of them can be inherited together as one haplotype (Teles, et. al., unpublished data), that is, the number of functional loci varies between haplotypes, indicating potential Structural Variation and Copy Number Variation. Genomic Structural Variations (SV's) are differences of genomic segments larger than 1kb between genomes of individuals of the same species. They include insertions, deletions, duplications, inversions and translocations. If they are in genic or regulatory regions, SV's are mostly maladaptive, but can also be adaptive, as they can then be linked to phenotypic traits (Chain and Feulner 2014), and they can also form a basis of population variation. In addition, they can promote speciation by creating post-zygotic reproductive barriers (Feulner and De-Kayne 2017). A common type of SV's are copy number variation (CNV), when different individuals of a species have different numbers of copies of the same gene, (Sharp et al. 2005), which can facilitate adaptation to novel environments (Fan and Meyer 2014). So, empirical studies on the effect of SV's and CNV's in ecological speciation are of the essence.

Copy Number Variation

Due to the presence of different numbers of functional loci in different haplotypes, it is suspected that there is considerable structural and copy number variation in the stickleback MHC region. Here, we aim to characterise this variation, and also to investigate its probable causes. In order to visualise and understand the CNV, different methods of comparison of synteny or collinearity were used. In a pass of the preliminary annotation pipeline, retroviral reverse transcriptase between the MHC loci had been found, which suggested that transposable elements could have played a role. So, transposable elements were annotated.

Copy number variations are an important force in genome evolution, as well as a cause of human disease (Peng et al. 2015). CNV can be found in many organisms, including humans, mice, fruit flies, and Arabidopsis. CNV's are harder to identify than SNP's but are more likely to affect fitness. The mechanisms of CNV formation are: replication slippage, non-allelic homologous recombination (NAHR), non-homologous end joining (NHEJ), and retrotransposition (Hastings et al. 2009). Duplications can be either tandem, where the daughter copies are close by, or dispersed, where the daughter copies are on different chromosomes or very far on the same chromosome. CNV's are detected with respect to a reference genome. However, the reference genome is that of one individual, so it is difficult to resolve what is an insertion and what is a deletion (Schrider and Hahn 2010). Evidence also shows that genomic regions with increased segmental duplications- large segments of DNA with high sequence similarity- also have increased CNV (Redon et al. 2006).

Aside from slippage and retrotransposition, the other two mechanisms are two different kinds of recombination, homologous (HR) and non-homologous. The former requires extensive homologous sequences, while the latter does not. HR is an accurate way to repair damaged DNA with the same chromosomal position in the sister chromatid, but if sequence other than the sister chromatid is used, it is called

non-allergic or ectopic homologous recombination. This can cause CNV's if the sequence being repaired has direct repeats. On the other hand, any non-homologous recombination can be inaccurate (Hastings et al. 2009).

All of these analyses allowed us to have a comprehensive, annotated synteny map of the three-spined stickleback MHC Class-II region.

Transposable Elements in the Context of CNV

SV's and CNV's can also indicate the presence and activity of Mobile Genetic Elements (MGE's) or Transposable Elements (TE's), and they are also over-represented in the MHC region (van Oosterhout 2009), which are among the most important factors in the shaping of Eukaryotic genomes (Kazazian 2004). TE's are DNA sequences that are able to move independently inside eukaryotic genomes using enzymes that they code for themselves. They have immense sequence diversity, and have been found in nearly all of the eukaryotic genomes examined so far. They make up a significant percentage of eukaryotic genomes (Wicker et al. 2009). Since TE's can modify the genome in a lineage-specific manner, they are thought to play a role in major evolutionary transitions such as speciation (Böhne et al. 2008).

TEs can be divided into two main categories, RNA-based or DNA based. The former, called retrotransposons, propagate using a copy-and-paste mechanism using an RNA intermediate, while the latter propagate using a cut-and-paste mechanism, with no RNA intermediate (Wicker et al. 2009). Retrotransposons usually code for reverse transcriptase (RT), so presence or absence of RT is a useful criterion for classifying transposons- if RT is present, it is a retrotransposon, or else, it is a DNA transposon. DNA transposons propagate by different, unrelated mechanisms, unified only by the lack of an RNA intermediate (Arkhipova 2017).

RNA-based transposons can be further divided into five different categories, on the basis of Open Reading Frames (ORF's) contained, including LTR (Long Terminal

Repeat), and LINE (Long Interspersed Nuclear Element). LINE's lack long terminal repeats, are found in all eukaryotic kingdoms, and are divided into five superfamilies: R2, L1, RTE, I, and Jockey, though they have also historically been classified into clades. To be able to move autonomously in the genome, LINES have to encode at least an RT and a nuclease (Wicker et al. 2009).

TE's have contributed significantly to vertebrate genome evolution in many ways. Firstly, they can cause large-scale structural changes in the genome, such as insertions, deletion, and reversions. Secondly, they can create or disrupt coding regions, as well as modify gene regulation (Muotri et al. 2007). TEs can produce these variants using different mechanisms: insertions mutagenesis, transduction, etc. Particularly, some TEs such as Alu elements and L1, can provide sequence similarities for unequal crossing over (Kazazian 2004) and ectopic recombination (Schrader and Schmitz 2018).

TE's are found in a greater density in the MHC region than in the rest of the genome in many vertebrates. Transposons may be have been found in the MHC region because of genomic features unique to it (van Oosterhout 2009). Previously, these have been found in the MHC class I region of the medaka (*Oryzas latipes*). Here, microsatellites, transposable elements, low-complexity regions and other repeats were found. TEs and their fragments may have played a role in tandem duplications and deletions, and also in medaka-specific genomic rearrangement (Matsuo and Nonaka 2004).

The repertoire of transposable elements in a cell is called its mobilome, and it varies considerably among relatively closely related teleosts such as stickleback, medaka, zebrafish and tetraodon. The stickleback genome is 461.53 Mb in size, out of which TE's constitute 14%. The most common families of DNA transposons are hAT and Tc1/Mariner, and in stickleback they have undergone a recent divergence. Among retrotransposons, sticklebacks have 11 LINE families, and 77 LTR families. Again, retrotransposon activity in sticklebacks has been recent. The clades of L1, L2, RTE,

and Rex-Babar have been subject to substantial expansion in teleost species, and the most predominant one in sticklebacks being L2 (Gao et al. 2016).

Population Genomics

After generating the map, we turned our attention to population-level variation in the MHC Class-II region. When populations split off from each other, divergence along the genome is usually heterogeneous, that is, genome-wide divergence is low, but along some regions, diverge is exceptionally high, and the latter are often called "genomic islands". This pattern can be formed by a combination of mechanisms, out of which stochastic ones- such as drift, and migration- act on the whole genome, whereas deterministic ones- such as selection- occurs on only the part of the genome that confers selected traits. It is not known whether patterns of high and low divergence follow a specific trajectory as populations move along the speciation continuum (Seehausen et al. 2014), so it's important to choose populations that are in the early stages of diverging from one another. If genomic islands are smaller and more localized, then they could contain "speciation genes", that is, genes that can form the basis of reproductive isolation (Wu and Ting 2004).

For this analysis, we had access to a dataset from a previous study: to find wider population-genomic patterns in natural Stickleback populations, Whole Genome Sequence (WGS) data was generated for 66 individuals. These populations are contrasting lake and river populations from Europe and North America, selected so as to find divergent genomic regions along a continuum of population differentiation: i.e. to observe differences among neighboring lake and river populations, as well as between populations on different continents. These populations also represent recent colonizations, as all these water bodies were covered with ice 12,000 years ago (Feulner et al. 2015; Chain et al. 2014).

In the previous studies, it was found that divergence along the genome between different populations is heterogeneous, that is, some regions of the genome are more

diverged than others. These are sometimes called “genomic islands”, and may contain “speciation genes” which promote reproductive isolation and may impose a fitness cost on hybrids. Furthermore, different mechanisms such as genetic drift, hitchhiking, etc. work in concert in to influence genome evolution during population divergence, but local adaptation is also an important contributor to divergence (Feulner et al. 2015).

The same whole-genome sequences showed that a third of young duplicated genes in sticklebacks are copy number variants. Immune genes such as MHC were also found to have more CNV than other genes. In all, the study showed that the stickleback genomic landscape is highly dynamic, and together with sequence divergence, structural variations are important to populations diverging from each other, and therefore, adapting to their local conditions (Chain et al. 2014). So, population genomic analysis needs to be extended to the MHC region, because of its importance in parasite resistance and mate choice.

Thesis Outline

The primary objective of this PhD thesis is to create a multi-layered understanding of the MHC region of the three-spined stickleback (Fig 4). First, we wanted to understand the genomic basis of the possible structural and copy number variation in the MHC region. So, we studied this region in the context of the whole genome, for which we used the reference genome from Glazer et. al. 2015.

Next, we mapped out the genomic structure of the stickleback MHC Class-II region using novel genomic sequences of new haplotypes with different numbers of functional loci, including BAC (Bacterial Artificial Chromosome) library sequence data, as has been used for reconstructing the MHC of other species such as medaka (Nonaka and Nonaka 2010) and crested ibis (Chen et al. 2015) humans (Horton et al. 2004). After functional genes, we turned our attention to regulatory elements in this region, which tells us about functional differences between various genes.

After that, we studied synteny in this region in more detail, finding blocks of recombination that account for differences in structure between haplotypes. However, we still wanted to understand the process by which these differences were created. So, we studied transposable elements found in this region, which give us a clue about the origin of the extensive CNV in this region.

Finally, we turned our attention to population-level variation in the stickleback MHC Class-II region. From a previously generated dataset, we were able to study variation in the assembled region, as well as other genomic regions as controls. We performed tests for selection and neutrality, giving us an insight about the evolutionary history of the stickleback species, with respect to parasite-mediated selection.

Host-parasite coevolution causes stickleback populations to adapt to their local parasite community, and diverge from each other, leading to the formation of cryptic species, which may interbreed in laboratory conditions, but would not do so in natural populations (Eizaguirre et al. 2009). They can be imagined as being in the middle of the speciation continuum, somewhere between a panmictic population, and full reproductive isolation. Different selection regimes such as disruptive or balancing selection can move the populations in either direction, but the underlying genomic structure that allows for this is unknown. This is what this work sets out to explore.

We have already seen that the number and sequence of MHC alleles can provide a basis for population divergence, but what are the salient features of the genome that underlie this vast variation? For example, what mechanism does the stickleback use to optimize its adaptive immune response to a wide variety of parasite communities and form allele pools for different environments such as rivers and lakes? Such a system would have to allow for excessive variability as well as fast adaptability, that is, encode vast standing structural variation to begin with, but would have to include a mechanism for making new combinations as the need arose. As we have already

seen, the genotyping of stickleback MHC Class-II alleles has shown the possibility of copy number variation in the corresponding region- could this, perhaps, be such a mechanism?

Furthermore, genomic islands of reproductive isolation have been studied in stickleback have been studied, but what is their relationship with immunity? The purpose of this thesis was to fill these gaps in our knowledge, leading to a deeper understanding of the genomics of speciation.

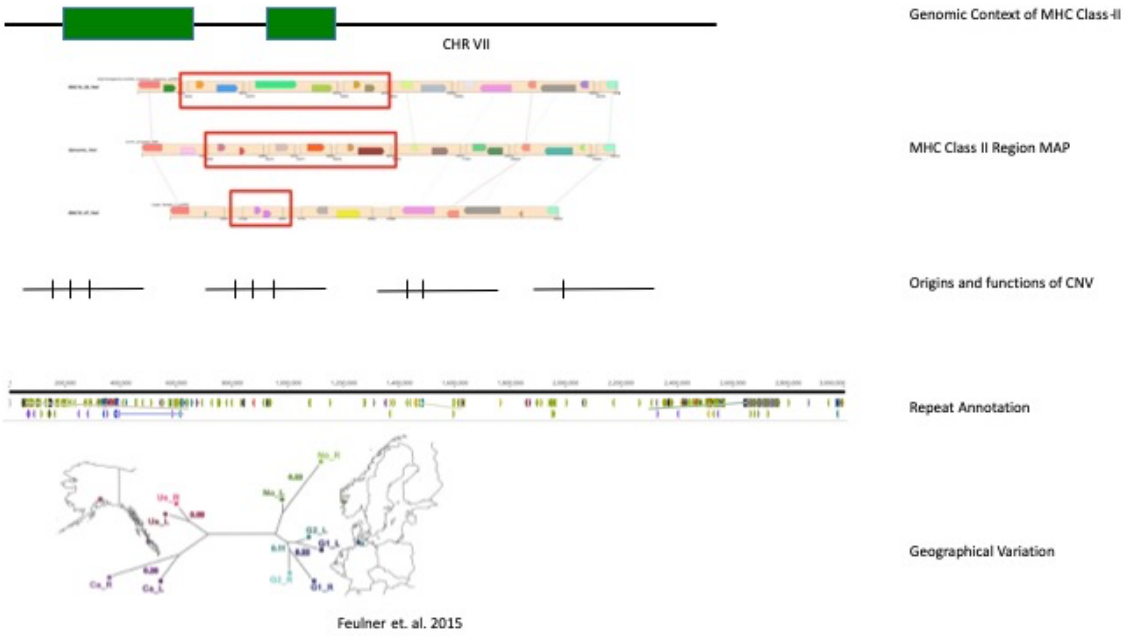


Figure 4: Layers of Understanding the Stickleback MHC Class-II Region, starting with the genomic region, followed by the Synteny map, the origins and functions of CNV, annotation of TE's, and finally, geographic variation.

Materials and Methods

MHC Class-II in the context of the whole genome

1. Finding the position of MHC loci in the stickleback genome

We determined where the MHC loci were present with respect to each other, as well as in the Reference Genome. The genome of the Three-spined stickleback was initially sequenced and assembled at BROAD Institute, from a partially- inbred female individual from Bear Paw Lake in Alaska. The assembly was generated with 9x coverage in paired-end Sanger sequence reads from multiple insert size libraries, and has a total gapped size of 463 MB. Scaffolds were placed on linkage groups in a genetic cross (113 anchored and 1822 unanchored scaffolds), and annotation was performed using the Ensembl evidence-based pipeline (Jones et al. 2012). Later, a Genotype-by-Sequencing (GBS) approach was used to further improve this assembly. Here, 78 previously unanchored scaffolds (including scaffolds 131 and 129) were anchored, 40 scaffolds from the previous assembly were reoriented, and scaffolds were rearranged in 4 new locations. The final scaffold map placed a total of 186 scaffolds, which made up 94.6% of the genome. In addition, recombination rates were found to be high on both ends of Chromosome VII, which was metacentric, i.e., the centromere was in the middle, and both arms were of approximately equal length (Glazer et al. 2015). In this work, the Glazer et. al. assembly is used throughout.

In order to determine the MHC region, we performed a BLAST search for a MHC class IIB exon-2 probe generated by Amplicon Express (No05_L2, see Supplementary Materials, probe sequence 1) against the Reference Genome using command line BLASTn(Zhang et al. 2000) with blastn matrix 1 -2, gap existence penalty 0 and gap extension penalty 2.5. The BLAST results showed four matches in the genome: three loci between 0-3 MB of Chromosome VII, and one locus on Chromosome III. Using an exon-3 probe, Chromosome VII was found to have four matches. This was assumed to indicate the presence of a locus with a truncated exon-2. The region on ChrVII was

assumed to be the MHC region- as it had more loci, and 0-3 MB of ChrVII was extracted for annotation.

2. Finding Pseudogenes in the variable region

In the previous step, we got a clue that one of the polymorphic MHC class-IIB loci in the classical region of the Reference Genome was a pseudogene, with a severely truncated exon2, and an early stop codon. So, the exon 2's of all the Classical MHCII-B genes were extracted and aligned, showing that one was indeed truncated (Supplementary Fig. 4).

3. Genome-wide distribution of MHC Class-II antigen presentation genes

The protein sequences of the zebrafish MHC Class-II antigen presentation pathway genes from Sambrook et. al. (Sambrook, Figueroa, and Beck 2005) were searched for on Genbank and downloaded, and a tBLASTn (Gerts et al. 2006) search was performed for each of them in the Reference Genome. From the BLAST results, the genomic loci of the MHC antigen presentation pathway genes were found.

4. Investigation of the MHC Class-IIB locus on Chromosome III

The region around the MHC Class-IIB antigen gene was extracted, and both genic and repetitive sequences were found. A colleague did the former (Ceron-Noriega, et. al. unpublished data), but the analysis was performed as a part of this project.

The Stickleback MHC Class-II map

1. Cloning of fish

For this study, two MHC Class-II haplotypes from two lab-bred sticklebacks descended from the Northern German stickleback populations in Groesser Ploener See were used. One haplotype was chosen because it had one variable MHC Class-II allele, while the other was chosen because it had 3, and both of them had the invariant Class-II locus. These were chosen because the two previously examined haplotypes had 2 functional alleles. Genomic DNA of both these fish was used to generate BAC libraries. Using RSCA typing, we confirmed that one haplotype (F) has one variable allele, while the other (G) has 3 variable alleles, and both had the invariant locus. RSCA typing only detects the exon-ii of the MHC Class-IIB loci, so at the outset we did not have information about the rest of the B-locus, or the A-locus.

Assembling the MHC Class-II region poses some theoretical challenges: for example, each heterozygous individual has two haplotypes, and sequencing reads have to assembled to the haplotype they are derived from (if the fish that was sequenced was heterozygous, reads might map to haplotype they are not from, this is because of high sequence similarity between different MHC haplotypes); moreover, the repetitive nature of the region makes it difficult to assemble short reads (as reads derived from one region may map to another region that's identical because they are repetitive), so the assembly parameters used to assemble the whole genome would not accurately place reads in the MHC region.

These challenges to sequencing and assembling presented by the MHC region were met in many different ways. First, it was ensured that each fish contained only one MHC haplotype. This was achieved by targeted inbreeding for one of the fish, and gynogenesis (Samonte-Padilla et al. 2011) in the other. In targeted inbreeding, fish from the same family were genotyped and interbred, until a fish was created that had two copies of the same grand-parental MHC haplotype, in this case the F haplotype, containing one variable allele. This individual is called pseudohaploid here, because

it carries two copies of the same chromosomal sequence of the MHC Class-II region. In gynogenesis, during fertilization, sperm were UV-irradiated, and the resulting embryo was treated to heat-shock. This resulted in the sperm losing its genetic material, while the second meiotic division was stopped in the embryo. The resulting fish had solely maternal inheritance, so they would have two copies of its maternal MHC haplotype, in this case the G haplotype, containing 3 variable alleles. The highest yield of gynogen fish was obtained when the sperm were treated with UV radiation for 2 minutes, and then used to activate the eggs. Subsequently, the activated eggs received heat shock at 5 minutes post fertilization at 34°C for 4 minutes (Samonte-Padilla et al. 2011).

2. Creating BAC libraries

Bacterial Artificial Chromosome (BAC) libraries are used to clone large pieces of Eukaryotic DNA, and to propagate them on F-factor based plasmids in *Escherichia coli*, and are useful for cloning large loci along with regulatory elements that are as far as 50kb away, as one unit (O'Connor et. al. 1995). During BAC library creation, DNA is first cut with various concentrations of restriction endonucleases (either BamHI or HindIII), so as to obtain fragments around 150kb in size. This is called a partial digest, where the restriction endonuclease does not cut the DNA at every restriction site, but only at a random subset of sites. These fragments are then cloned into pre-prepared BAC vectors, and then transformed into E. coli that are designed for these vectors. Then, the library is screened for clones containing our regions of interest. Screening can be done in many ways, one of which is to array the BAC clones on a nylon membrane, and find the correct ones using a target sequence-specific labeled DNA probe (Farrar and Donnison 2007). Once the clones are found, they can be isolated, cultured, and sequenced.

From the previous step, we had two MHC Class-II homozygous fish, a pseudohaploid, and a gynogen, from which BAC libraries were created in collaboration with

Amplicon Express (Pullman, USA). High molecular weight DNA from these fish were cut using a restriction endonuclease and cloned into BAC vectors creating two BAC libraries BAC15-29 and BAC15-47. The specifications of the libraries are given in Table 1, below.

Table 1: Specifications of BAC Libraries

Category	BAC15-29	BAC15-47
Number of clones	18,432	21,504
Average Insert Size (kb)	115	105
Restriction endonuclease used	HIND III	HIND III
BAC Cloning Vector	pCC1BAC Epicentre vector	pCC1BAC Epicentre vector
Competent Cell	Invitrogen DH10b Phage Resistant	Invitrogen DH10b Phage Resistant
Average Fold Coverage (Avg. insert size*number of clones)/Length of stickleback genome	~4.6	~4.9

Where the length of the stickleback genome is assumed to be 463.4 MB (Glazer et al. 2015).

3. Selection and sequencing of BAC clones

The new BAC libraries (BAC15-47 for the F haplotype and BAC15-29 for the G haplotype) were then screened by the company for MHC Class-II genes using fluorescence in-situ hybridization, with probes designed to bind to the exon-iii region of the MHC Class-IIB genes using their proprietary techniques. Exon-iii was used because it is conserved among different alleles. From the two BAC libraries, a total of 12 MHC Class IIB positive clones were found. RSCA typing (Lenz et al. 2009) revealed which clones contained which alleles, so this was used to select the clones for sequencing on the PacBio platform, performed in collaboration with Amplicon Express. From BAC15-29, 5 clones were selected: 1 covering the monomorphic (No19) allele, and 4 covering the variable alleles. From BAC15-47, 2 clones were selected, one each for the variable allele and the No19 locus. This was done to get maximum coverage and reduce redundancy. The PacBio platform was chosen for its large sub-read length, which in turn would result in a more accurate assembly of a

region as variable as MHC Class-II. The BAC clones were individually tagged and pooled into one SMRT cell, and standard HGAP assembly was performed (www.pacb.com).

From RSCA typing the 4 clones showing the variable alleles from BAC15-29, we knew the possible order of these alleles. In order to get a larger contig that contained all three loci as well as information from either side of the MHC loci, a multiple sequence alignment of the 4 BAC clone sequences were made using Kalign (Lassmann and Sonnhammer 2005), with gap open penalty 200 and gap extension penalty as 10, as there should not have been gaps, given that it is the same sequence. The consensus sequence of this alignment was used as a larger contig that contained the variable region of the G haplotype. In order to confirm the possible order of the alleles, this larger contig was BLASTN-searched for the exon-II sequences of the three alleles. It was found that the alleles are indeed in the expected order.

During the annotation, it was found that the BAC clone with the F haplotype did not have any genic sequences upstream of the MHC-IIA locus. However, this was necessary because we wanted to annotate and further analyze sequence between the same two genes in both the haplotypes. For this, we sequenced another clone from the BAC15-47 library. A clone contiguous to the MHC region, upstream of the MHC Class-II No05A was found. To do this, first, the "cytochrome P450, family 17, subfamily A, polypeptide 2 gene" locus was chosen, as it was found via annotation to be present in both the other haplotypes, and was far enough away from the MHC locus that we would get a long upstream sequence (See Supplementary Materials, probe sequence 2). Then, this genic sequence was extracted from the Reference Genome and the G haplotype, and aligned. The parts of the sequence with the greatest identity were extracted and used for probe design. Furthermore, 2 Kb of sequence 3Kb upstream from this sequence was also extracted and used to make a probe. Library screening was again performed by the company, and a clone was found with this probe and sequenced as before. Overlaps were found between the new clone and the old clone, and they were manually joined together to make a larger contig.

The Reference Genome and the BAC clone examined by Reusch et. al. in 2004 show two functional MHC-IIB genes each and one pseudogene in the former, which cannot account for the allelic variation observed in the natural populations. Therefore, in this sequencing effort, two fish gave us two BAC libraries, and each library gave us two contigs, one from the classical region, and another from the non-classical region, giving us four in total from this sequencing effort. The classical region of one haplotype (F) has one allele, while the other (G) has 3 MHC-IIB alleles. We also examined the 3MB from the telomeric end of ChrVII, containing the MHC loci (Supplementary Fig. 1, 2 and Table 1). So, we had a total of 5 contigs to annotate- shorter ones from the new sequencing effort, and one 3MB one from the Reference Genome.

4. Annotation of the MHC region

All four contigs and the genomic sequence were structurally and functionally annotated using several different methods. For an initial structural annotation, they were annotated de-novo, that is, without any EST or protein evidence. Here, the sequences were first repeat-masked with RepeatMasker (Tarailo-Graovac and Chen 2009), then genes were predicted using AUGUSTUS(Stanke et al. 2004), all with default settings (results not shown). Since using evidence-based methods increase the accuracy and completeness of annotations (Yandell and Ence 2012), an iterative MAKER (Cantarel et al. 2008; Campbell et al. 2014) workflow with default settings was used to improve the annotations. Two iterations were used: first, a de-novo annotation was made with EST2Genome within MAKER, and then the .gff result of this was used to create a SNAP training file, which was then used to generate the final annotations, in MAKER with the trained SNAP. Two iterations were used because increasing the number of iterations did not improve the annotation, i.e., the resulting annotations were the same for the third iteration. As each run of the MAKER pipeline uses the results of the previous run, more iterations would not give different results, so they were not performed. For EST and protein evidence, the EST and protein

databases of NCBI were searched with the phrase “stickleback” and all the results were downloaded. This gave us 259471 EST’s and 3097 protein sequences.

Functional annotation was performed on the basis of homology to genes and proteins already present in the nr database at NCBI. First, the nucleotide database was searched (using BLASTn) using the mRNA sequence of each predicted gene from the MAKER results, which allowed matches with CDS’s, as they do not contain introns. However, the BLASTn search was not helpful as it only produced the sequence deposited by Reusch et. al. Then, the protein database was searched (using BLASTp) using the protein sequence of each predicted gene, and the first hit was assigned as the gene name. The reciprocal best hit principle (Ward and Moreno-Hagelsieb 2014) was used to confirm the authenticity of the hit, though this confirmation was not possible for some of the genes. In those cases, comparisons were made between haplotypes to confirm gene names. Gene annotations and visualizations were performed in Geneious Prime 2019.0.4 (https://www.geneious.com).

5. Finding regulatory elements in the MHC region

Regulatory elements were found in the MHC region using POSSUM (Lin et al. 2007), with default settings. Then, 400 bp upstream of all MHC II-A and II-B loci were extracted, and multiple sequence alignments were made in order highlight the TATA, GATA, and CCAAT boxes. A manual search for the SXY region was also conducted (Nillson et. al. PhD thesis).

6. Studying Synteny in the variable region

In order to build the map of all the haplotypes, different methods of comparison of synteny or collinearity were used. Classically, synteny was defined as the co-localization of two or more loci along the chromosome, but in the era of next generation sequencing, it is synonymous with collinearity, that is, the conservation

of gene order and orientation. Shared synteny can be broken up by structural variants such as insertions and deletions, so visualising synteny is of the essence (Veltri, Wight, and Crouch 2016).

Annotation also allowed us to determine "bounding genes", that is, the same two genes in every haplotype, between which we could extract the sequence, and compare between segments of DNA that were functionally alike. The bounding genes also allowed us define the classical and non-classical parts of each haplotype. In order to study synteny in this region, three different approaches were used. First, the DNA sequence between the bounding genes (Table 2) was extracted for both the variable and the invariant region, from every contig. Then, structural comparisons were made with three different programs, for the following reasons:

1. Dotplots- Dotplots provide a quick and easy way to visualize similarities and differences between two DNA or protein sequences. The two sequences are put on the x and y axes, and conserved domains, reverse matches, and repeats can be easily seen. This was made using Gepard (Krumsiek, Arnold, and Rattei 2007).
2. Mauve- Shows locally collinear blocks in the sequences, and can visualize exons along with indels and inversions. (Darling, Mau, and Perna 2010; Darling et al. 2004).
3. SimpleSynteny- Shows differences in gene order, used with all default settings except for minimum query coverage cutoff, which was set to 25 (Veltri, Wight, and Crouch 2016).

7. Generation of the Stickleback MHC "map"

The results from the previous steps were used to create a schematic "map" of the MHC region in PowerPoint.

Understanding Copy Number Variation and its Causes

1. Sequences for Annotation

Transposable Elements were annotated all the MHC haplotypes analyzed in this study (two classical and non-classical regions from the current sequencing effort, as well as the MHC-containing 3MB from the Glazer assembly). In addition, we downloaded the sequence deposited by Reusch et. al. in Genbank, Accesion no. AY713945.1 (Reusch, Schaschl, and Wegner 2004). This BAC clone does not contain information about genes upstream of the MHC-IIA gene, but it would give us the intergenic region between the MHC-IIA and B genes. So, it was used. In total, we had 4 classical regions, 3 non-classical regions, and one instance of the sequence between these regions. However, since we were interested in the variability in the MHC Class-II, only the variable regions were used in all analyses. In addition, repeats were identified in the entire Glazer assembly, though this was not annotated, but used to calculate average numbers of transposable elements in the genome.

2. Finding Repetitive Elements
(Library Building Courtesy of Alejandro Ceron-Noriega)

First, TE libraries were generated for sticklebacks using RepeatScout (Price, Jones, and Pevzner 2005) (Ceron-Noriega, unpublished data). Then, these libraries were used to find TE's using RepeatMasker, with CrossMatch as the search engine, and slow speed setting (Tarailo-Graovac and Chen 2009). Using these, TE's were identified in all the sequences. For the MHC haplotypes, the Repeatmasker output files (.out and .gff) were used to annotate the TE's in our contigs. TE annotations, like gene annotations, were performed Geneious Prime 2019.0.4 (https://www.geneious.com). This process gave us a de-novo TE annotation, that is, repeats that were found by aligning the stickleback genome against itself- these are not dependent on repeats annotated in any other species. However a homology-based annotation was sought in order to have an independent confirmation of the repeats found in the previous step. So another, independent round of TE annotation

was performed with CENSOR (Kohany et al. 2006), using the Vertebrate repeat library, as implemented in the package. Finally, Repeatmasker annotations that overlapped with CENSOR annotations were kept in the final annotation table, so that we would have repeats confirmed by multiple independent protocols. After this step, there were still some repeats that were classified as "Unknown". These were extracted, and classified using TE_Class (Abrusán et al. 2009). Simple Repeats and Low-Complexity Regions were excluded from this analysis, as they are not informative.

Population Genomics of the MHC Class-II Region

1. Data Collection

(Performed by and adapted from Chain et. al. 2014 and Feulner et. al. 2015)

The data for the population genomics portion of this work had previously been collected and generated by Chain, Feulner et. al. for the aforementioned papers. The fish were caught from 5 parapatric lake-river population pairs in Canada, U.S., Norway, and two sites in Germany, as well as a marine outgroup. High Molecular Weight DNA was extracted from muscle tissue using a Qiagen Midi Kit following the manufacturer's instructions. From the captured fish, 66 were selected for sequencing on the Illumina platform. For each individual, 2 paired end libraries (100bp reads, average insert size 140bp and 300bp) and 1 mate-pair library (50bp reads, average insert gap of 3KB) were obtained, with an average coverage depth of 26x, and the data was deposited in the European Nucleotide Archive (PRJEB5198) (Feulner et al. 2015; Feulner et al. 2013) .

2. Mapping reads on Chromosome VII

(Read mapping and SNP calling performed by, and methods courtesy of Doko-Miles Jackson Thorburn, unpublished data)

The reads were mapped on to the Reference Genome, although only Chromosome VII was used in this analysis, as we wanted to compare the MHC-encoding and non-MHC parts of the same chromosome. This was done using BWA-MEM (H. Li 2013), and mapped reads were further processed to sort data by coordinate, and verify mate-pair information using Picard toolkit version 2.18.7 (http://broadinstitute.github.io/picard/). Data from different sequencing libraries belonging to the same individual were combined, and duplicated reads were removed using Picard toolkit.

Variant-calling was done using a multi-sample approach, calling all variants simultaneously using GATK version 4.0.8.1 (McKenna, et. al. 2009), following the *best practices* workflow (DePristo, et. al. 2011). Selecting only SNPs, and applying hard filters gave us the final data set. Genome mapping and processing were conducted on the QMUL Apocrita High Performance Computing Cluster (King, Butcher, and Zalewski 2017).

3. Calculation of Summary Statistics on Chromosome VII

The mapping of reads to the Reference Genome Chromosome VII created a vcf file, which was used for to calculate summary statistics using vcftools (Danecek et al. 2011). Genome scans for Tajima's D (Tajima 1984), F_{ST} (Weir and Cockerham 1984), and pi (Nei and Li 1979)(Li and Sadler 1991) along ChrVII, with a window size of 5 kb, in order to get a higher resolution than Feulner et. al., 2015. Tajima's D was chosen as a measure of neutrality, that is, to see if the MHC Class-II region was an outlier compared to the rest of the chromosome. Pi was calculated as a measure of nucleotide diversity, which is the average number of nucleotide differences per site between two randomly chosen DNA sequences (Nei and Li 1979). F_{ST} was calculated to see if lake and river populations diverged more or less than the chromosomal average, as the parasite communities of these environments are different. Another comparison was done between the samples from the North Sea, and European freshwater samples, in order to see if there was divergence in the MHC due to the

colonization of freshwater environments. The rest of ChrVII, served as a control, as most of it does not contain MHC antigen presentation pathway genes (Fig. 6), and is therefore not subjected to parasite-mediated selection in the same specific way that an MHC- region may be.

All data visualization was done in R (R Core Development Team 2015).

Results

MHC Class-II in the context of the whole genome

The Reference Genome was searched for MHC class IIB genes, and most of them were found between 0-3MB of Chromosome VII, and one MHC Class-IIB locus on Chromosome III. Three Classical MHC Class-II loci- the ones that are polymorphic and can have many different alleles- were found at ~0.3MB from the beginning of Chromosome VII, and close to each other, while another one was 2.4MB away, and this one was monomorphic for one allele, No19. Out of the three variable MHC loci, in one of them, the exon-2 of the B locus was truncated, making it a pseudogene. Moreover, each individual fish examined so far shows between 1-7 sequence variants that are variable, and one invariant allele that is found in every fish (Lenz et al. 2009) (Reusch and Langefors 2005), called No19. So, each haplotype must contain variable loci, as well as an invariant locus. (Fig. 5)

It was confirmed that the loci closer to the telomere have the variable sequence variants, while the more interior one is the invariant (No19) locus. In between, there are no immune genes (Supplementary Fig. 1, 2 and Table 1), so the MHC region in stickleback can be divided into two regions, one containing the variable loci, and the other containing the invariant locus. The first was called the "classical" MHC Class-II region, and the second the "non-classical" MHC Class II locus, in accordance with established MHC nomenclature on the basis of polymorphism. In humans, certain MHC class-II loci have many alleles, while others have only a few. The former are called classical MHC Class-II loci, while the latter are called non-classical loci (Horton et al. 2004) (Fig. 5).

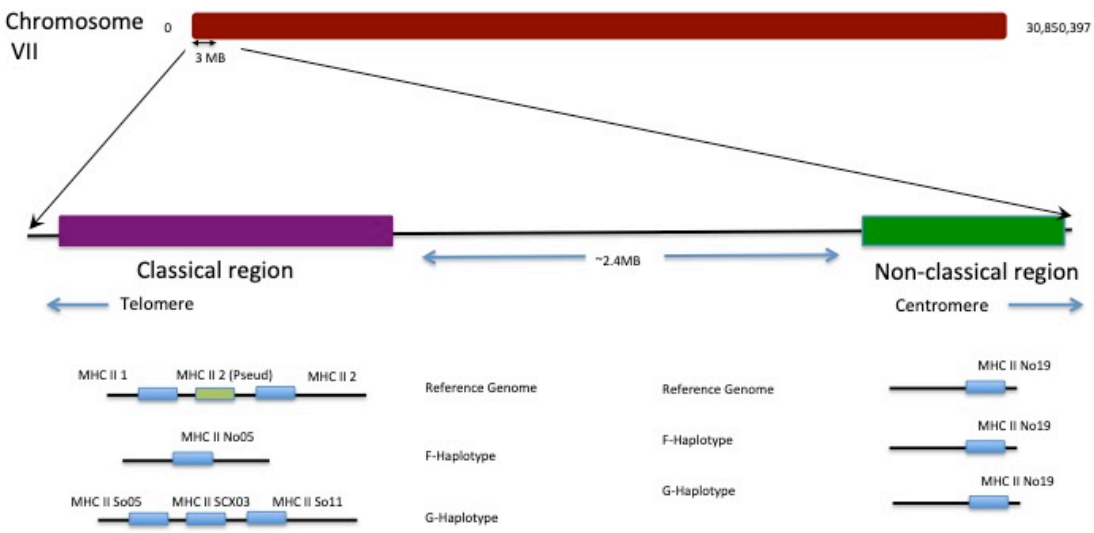


Figure 5: Position of the MHC region in ChrVII, showing the classical and non-classical regions, along with the contigs studied for each region (Not to scale).

In addition, the MHC Class-IIB locus in ChrIII was investigated, and it was found to have no MHC IIA locus in tandem arrangement. For a fully functional MHC locus, the IIA and IIB antigens have to be present and complete in tandem arrangement, which is not the case in ChrIII. Moreover, it was found near several repetitive elements, on clones that end in repeats, thus was hypothesized to be an assembly error. Clones with repeats on either end have more than one possible genomic location (as repeats are in multiple locations), we cannot be sure which location is correct and which are spurious (Supplementary Fig. 3).

In mammals such as humans and rats, MHC Class-I, II, and III genes are linked, but in teleosts they are not. This is an important distinction between the MHC two classes of jawed vertebrates, which may be the result of a whole-genome-duplication followed by pseudogenization in the teleost line (Flajnik and Kasahara 2001). In sticklebacks is only one complement factor gene in the classical MHC region, and the other MHC class II antigen presentation pathway genes are distributed genome-wide, as in Zebrafish (Sambrook, Figueroa, and Beck 2005). Many antigen presentation pathway genes are found on Chromosomes X and XX, while a few others are on III, IV, and XVI. There is no conservation in genomic locations of MHC Class-II pathway genes between stickleback and zebrafish (Fig. 6). This is in line with the observation

that MHC genome organization varies greatly among teleosts(Flajnik andKasahara 2010; Flajnik and Kasahara 2001).

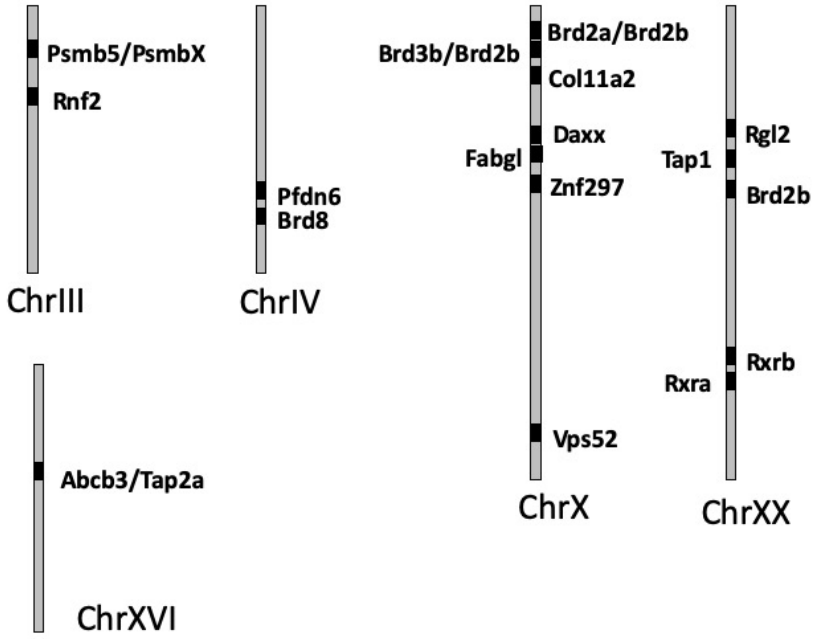


Figure 6: Locations of other MHC class-II antigen presentation pathway genes (Not to scale).

The Stickleback MHC Class-II map

Then gene annotation was performed for of the MHC Class II regions used in this analysis. For this, we used the classical and non-classical MHC Class-II regions of the F and G haplotypes, as well as 0-3MB of the telomeric end of Chromosome VII, as shown in Figure 5 (Supplementary Table 1). This was done for many reasons. Firstly, it showed us the order and direction of the MHC Class-II A and B antigen loci. Secondly, it showed us which genes were present in proximity to the MHC loci. Thirdly, annotation showed us the order and nucleotide positions of the MHC loci within our haplotypes, as well as the other genes in their vicinity. Using the annotation, we selected specific genes as "bounding genes" around the MHC region (Table 2, in red). The regions between the bounding genes from the genomic region

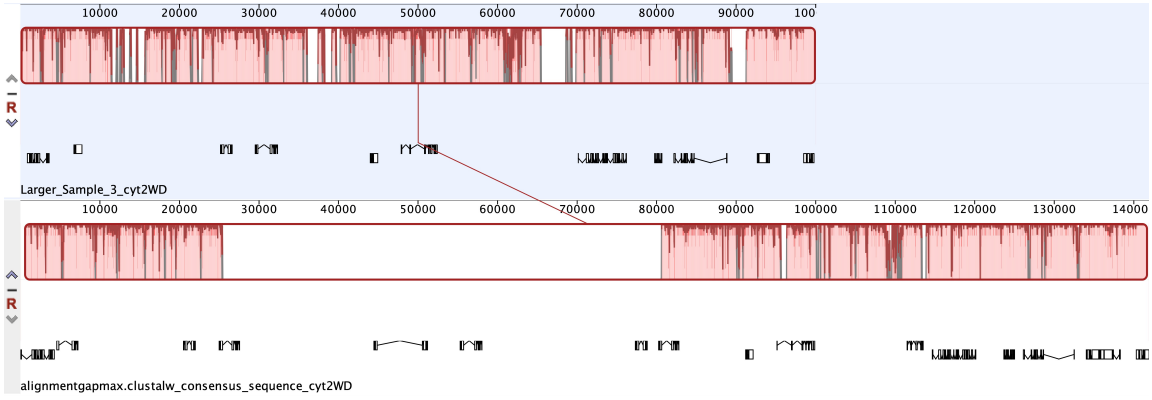
were extracted, and further used as the classical and non-classical MHC Class-II regions of the Reference Genome.

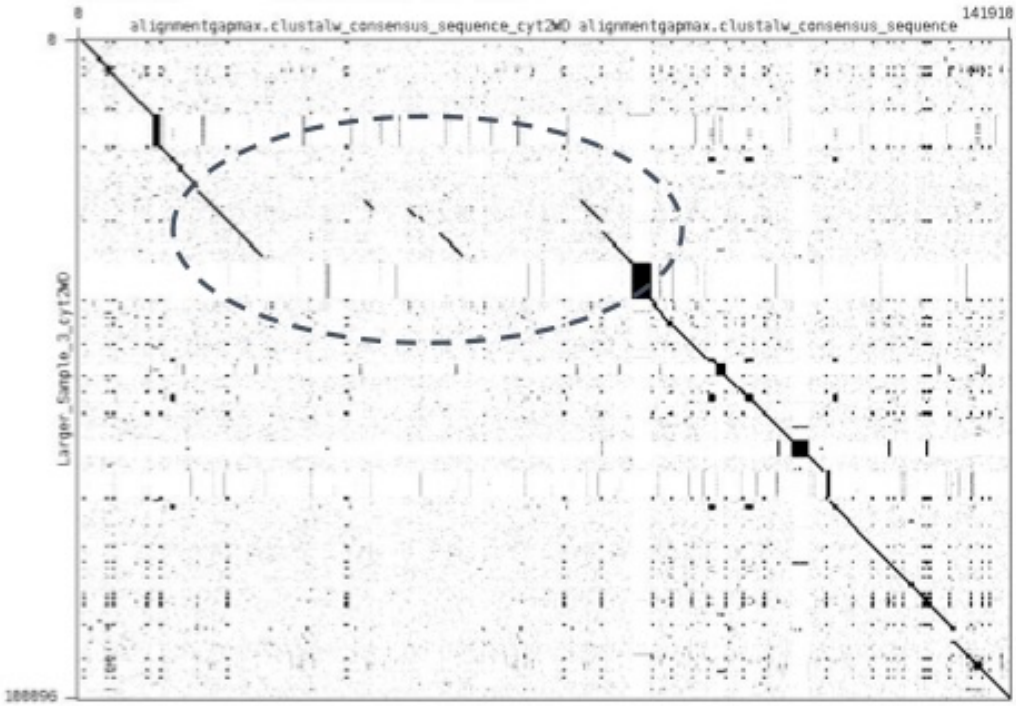
Synteny comparisons were made between the classical regions and the non-classical regions of the F and G haplotypes. The Mauve comparison of the classical region (Fig. 8 (a) i)), shows that the two contigs of the F and G haplotypes are one block on either side of a large indel present in the latter haplotype. This indel accounts for the two extra MHC Class II loci in the former. This, as well as the duplicated MHC genes in tandem arrangement can also be seen in the dotplot (Fig. 8 (a) ii)). This large indel accounts for the observed CNV in the Classical MHC Class-II region.

In the structural collinearity comparison in the non-classical region, it can be seen that all the different contigs are one collinear block, with small indels compared to others (Fig. 8(b)).

In addition to Mauve and Dotplots, SimpleSynteny was also used to generate synteny maps, in which the Classical region of the Reference Genome was included. This showed the same information, with one additional result: the MHC region from the Glazer genome shows a large indel between a Class-IIA gene and a IIB gene, which happens to be the pseudogenized copy (Fig. 8 (a) iii)).

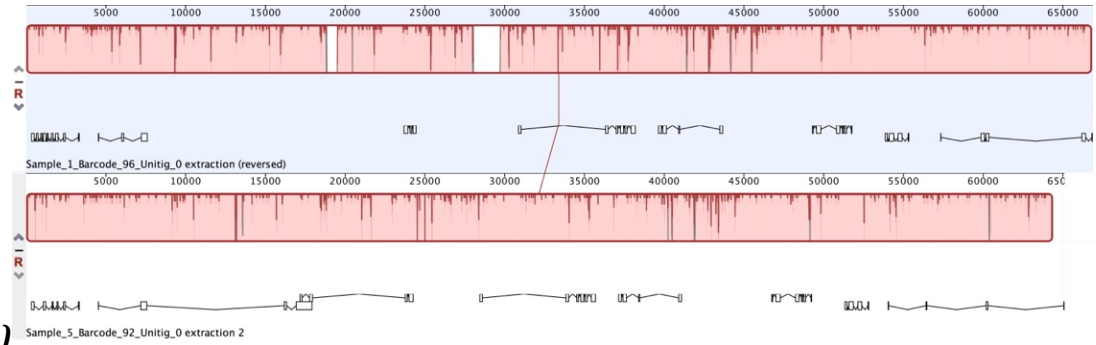
a)i)





ii)

iii)



(b)

Figure 8: Structural Comparisons(a) Classical region i)using Mauve, with the F haplotype on the top, and G haplotype on the bottom ii)using dotplot with the

F haplotype on the vertical axis, and the G haplotype on the horizontal axis, showing the large indel that accounts for the differences between them iii) using SimpleSynteny (MHC genes bound in red blocks) with the F haplotype, the Classical region of the Reference Genome, and the G haplotype, showing the different collinear blocks, especially the A and B antigen genes being on separate blocks in the reference genome (b) Non-classical region, using Mauve, with the F haplotype on the top, and G haplotype on the bottom, showing overall collinearity between the two haplotypes.

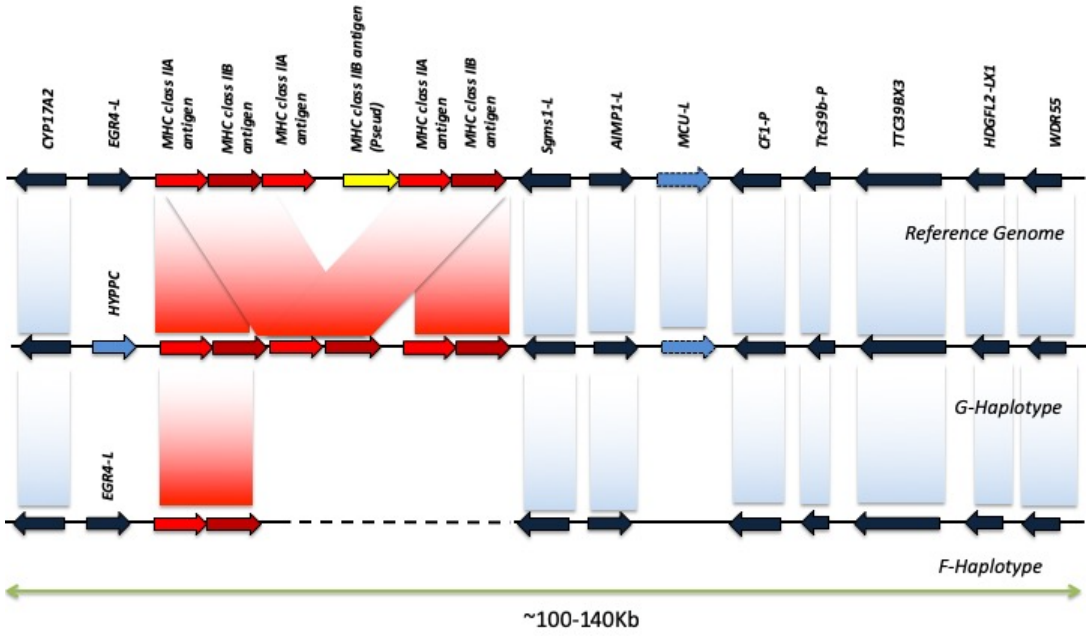
The classical region of the Reference Genome contained two functional loci, and another pseudogenized MHC locus between the functional loci with a truncated exon2 of the B locus. In the classical region of the G haplotype, 3 functional loci were found, in the order predicted from RSCA typing of the individual sub-clones, and in the H haplotype, one Classical MHC gene was found. In the Classical region, the Class IIA and IIB genes were always in the same orientation. It was also found that there is only one immune gene in their vicinity, a complement factor in the classical region. This shows that in stickleback, like in other teleost species, MHC Class-I, II, Class-III, and antigen presentation pathway genes are unlinked and scattered throughout the genome. The different haplotypes of the non-classical region contained one copy each of the MHC Class-IIA and B antigen genes (the latter carrying the No19 allele found by RSCA typing in all individuals so far), in opposing orientations. None had any truncated exons. (Fig.9, Table 2).

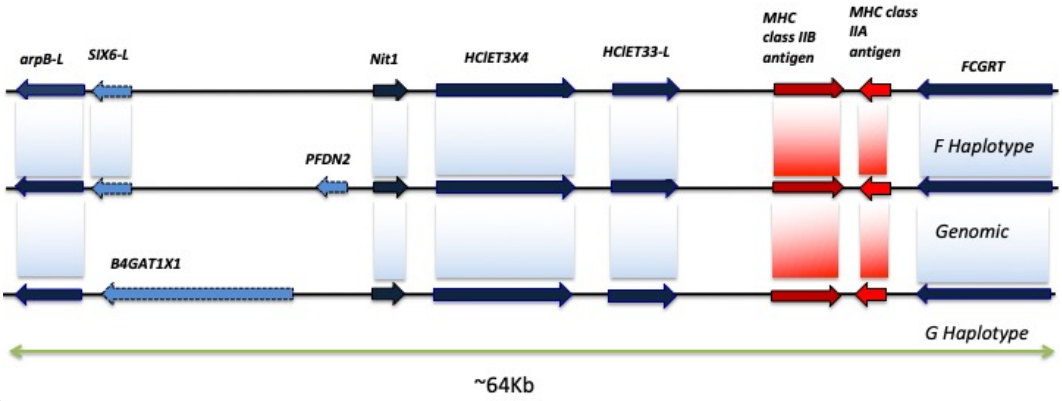
From the annotated map, of the classical region, we can see that different haplotypes have different numbers of polymorphic MHC Class-II loci. That is, copy number variation can be seen at the genomic level, for the first time in sticklebacks. Intraspecific variation was studied in the MHC Class-I and II of the medaka(Tsukamoto et al. 2005; Bannai and Nonaka 2013), but did not find CNV in the same region as we have here. At first glance, it may seem that there is variability in both the classical and non-classical regions, outside the MHC loci, with some genes

only being found in some haplotypes. However, the exons were the same, with variation generated due to small differences interpreted by MAKER. In the MHC loci, Mauve and MAKER corroborated the number of exons and genes, indicating that these annotations are more reliable (Fig. 8 and 9).

The POSSUM search for promoter regions did not find any TATA, GATA or CCAAT boxes in the orientation of the MHC genes. However, inverse promoters were found. Particularly, the inverse CCAAT box that is the Y element was found. The promoter regions for the Non-classical MHC Class-IIA alleles were reversed as the gene is in the reverse direction. A manual search for the S and X regions was conducted, but they were not found. It is possible that the screen was not thorough enough, because SXY motifs are crucial for MHC gene expression(Van Den Elsen et al. 1998) (Fig.10).

(a)





(b)

Figure 9: Map of the MHC region (a) Classical region (b) Non-classical region Color scheme: red- MHC gene (bright red, MHC Class-IIA locus, dark red, MHC Class-IIB locus), blue- other gene, present in all haplotypes, light blue- different in one haplotype. Gene names are given in Table 2. The dotted line represents the indel in the other two haplotypes with respect to the F-haplotye.

Table 3: Gene names of Non-MHC genes genes, (a) Classical,(b) Non-classical

(a)

Number	Name	Abbreviation	Gene name from	Orientation
1.	cytochrome P450, family 17, subfamily A, polypeptide 2	CYP17A2	Genbank	reverse
2.	early growth response protein 4-like	EGR4-L	Uniprot	forward
3.	hypothetical protein, conserved	HYPPC	Assigned here	forward
4.	phosphatidylcholine:ceramide cholinephosphotransferase 2-like	Sgms1-L	Uniprot	reverse

5.	aminoacyl tRNA synthase complex-interacting multifunctional protein 1-like	AIMP1-L	Uniprot	forward
6.	calcium uniporter protein, mitochondrial-like	MCU-L	Uniprot	forward
7.	complement factor I, partial	CF1-P	Uniprot	reverse
8.	tetratricopeptide repeat protein 39B-like, partial	Ttc39b-P	Uniprot	reverse
9.	tetratricopeptide repeat protein 39B isoform X3	TTC39BX3	Uniprot	reverse
10	hepatoma-derived growth factor-related protein 2-like isoform X1	HDGFL2-LX1	Uniprot	reverse
11.	WD repeat-containing protein 55	WDR55	Uniprot	reverse

(b)

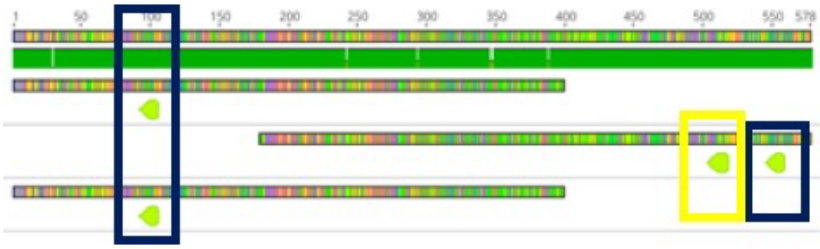
Number	Name	Abbreviation	Gene name from	Orientation
1.	actin-related protein 2-like	arpB-L	Uniprot	reverse
2.	homeobox protein SIX6-like	SIX6-L	Uniprot	reverse
3.	beta-1,4-glucuronyltransferase 1 isoform X1	B4GAT1X1	Uniprot	reverse
4.	prefoldin subunit 2	PFDN2	Uniprot	reverse
5.	nitrilase homolog 1	Nit1	Uniprot	forward

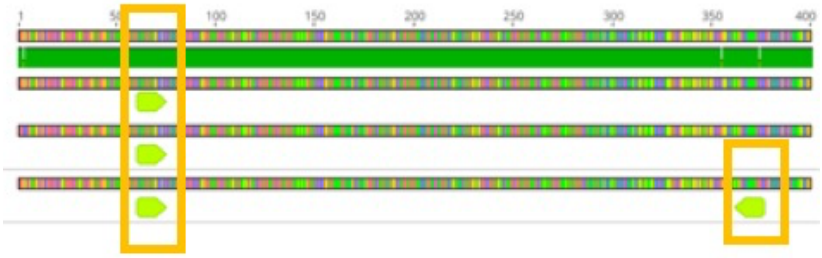
6.	H(+)/Cl(-) exchange transporter 3 isoform X4	HClET3X4	Assigned here	forward
7.	H(+)/Cl(-) exchange transporter 3-like	HClET33-L	Assigned here	forward
10.	Fc-Receptor like	FCGRT	GenBank	reverse

(a)

(b)

(c)





(d)

Figure 10: Promoter Regions of the MHC genes (a) Classical A-antigen genes (b) Classical B-antigen genes (c) Non-classical A-antigen genes (d) Non-Classical B-antigen genes. CCAAT boxes are boxed in yellow, GATA in blue, and TATA in orange.

Understanding Copy Number Variation and its Causes

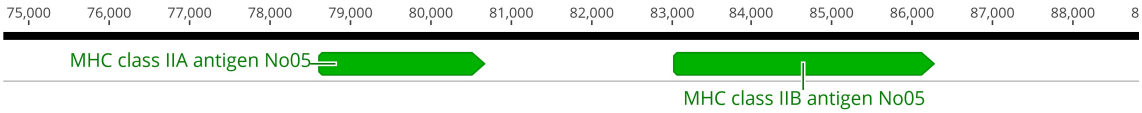
Thorough analysis of the TEs in the Classical MHC Class-II region of the Reference Genome showed an average of 590 repeats/MB, as opposed to the genomic average of 533/MB. This shows that repetitive elements might be overrepresented in the MHC region. Each classical region also contained a different number of repeats (Table 4).

A closer look suggested many intriguing patterns, including one that could potentially account for the CNV. Firstly, it showed that the large intergenic region upstream of the pseudogenized B locus consists mainly of transposable elements of the LTR/ERV family. In the 3 MB from the telomeric end of Chromosome VII, 1725 repetitive elements were found, and each classical MHC Class-II haplotype contained different numbers of repeats (Table 4)(Supplementary Table 2). One element, LINE-L1/Tx1, seemed to be present in front of every copied MHC locus, in a distinctive pattern. If the number of MHC loci is n, the number of L1 present in the intergenic region was always n-1 (Fig. 11).

Table 4: Numbers of Transposable Elements in the Classical MHC Class-II Region

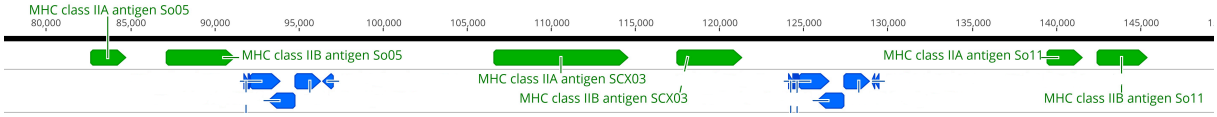
Variable Region Contig	Number of Transposable Elements
Reference Genome	88
G Haplotype	86
F Haplotype	22

a)

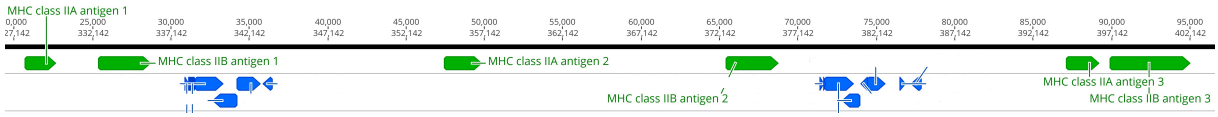


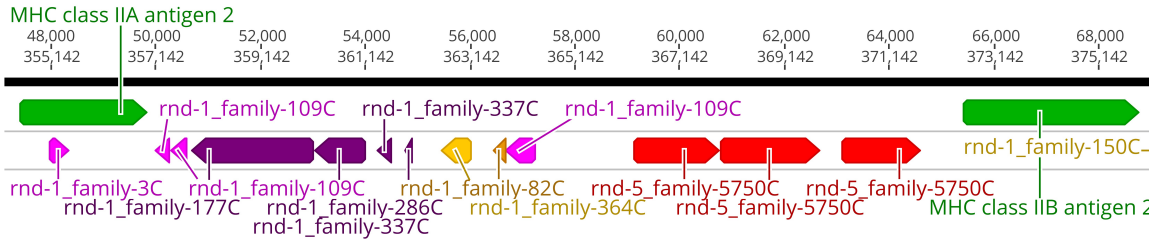
b)

c)



d)





e)

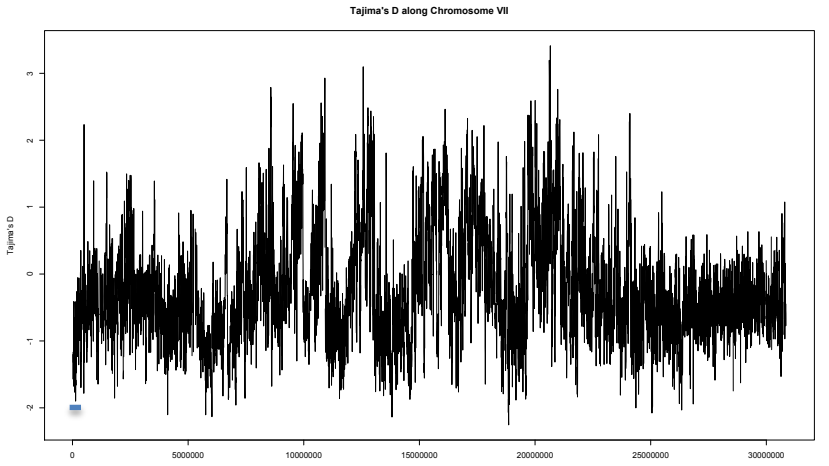
Figure 11: Presence of Line L1/Tx-1 in the vicinity of the MHC loci in a) F-haplotype b) Reusch et al. clone c) G-haplotype d) Reference Genome, showing the pattern of alternated MHC loci and L1 e) Repeats in the expanded intergenic region between the A and pseudogenized B locus on the Reference Genome.

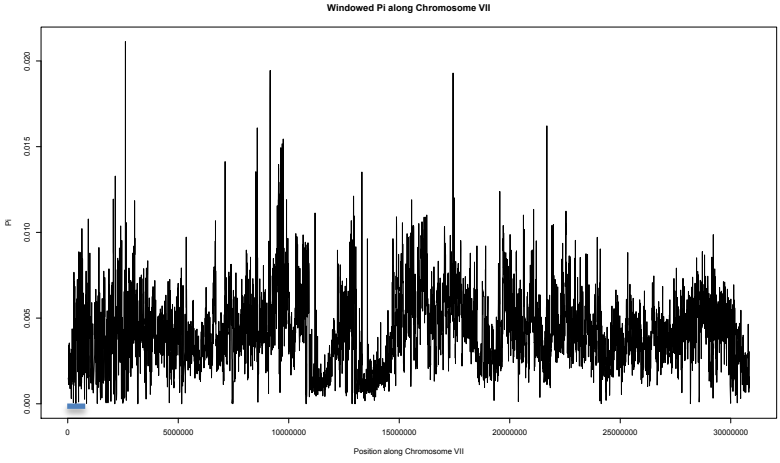
Population Genomics of the MHC Class-II Region

Population-level nucleotide sequence variation was studied in the stickleback MHC Class-II region, in comparison with the entire Chromosome VII, to understand population-level genetic variation in this region. First, Tajima's D was calculated as a test of Neutrality, and Pi was calculated as a measure of nucleotide diversity. The Tajima's D graph shows that while there are many peaks on Chromosome VII, the biggest ones are not between 0-0.3 MB (Fig. 12(a)).

To study nucleotide diversity in and around the MHC region, a chromosome-wide scan for pi was conducted with 5 kb windows. This also showed no peaks larger than others in the MHC region, nor did it show that the MHC region is subject to different selective pressures than the neighboring regions.

(a)



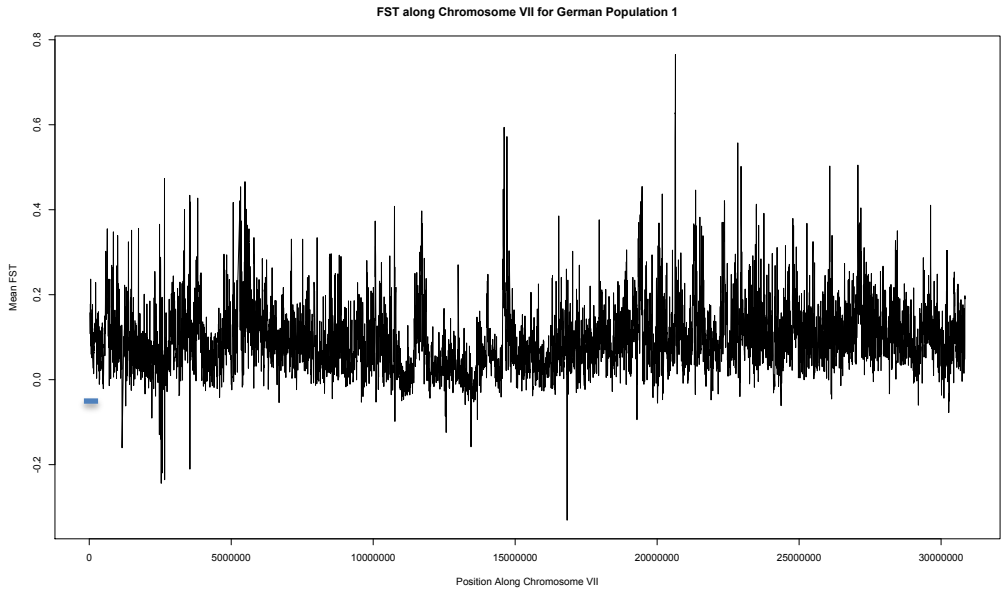


(b)

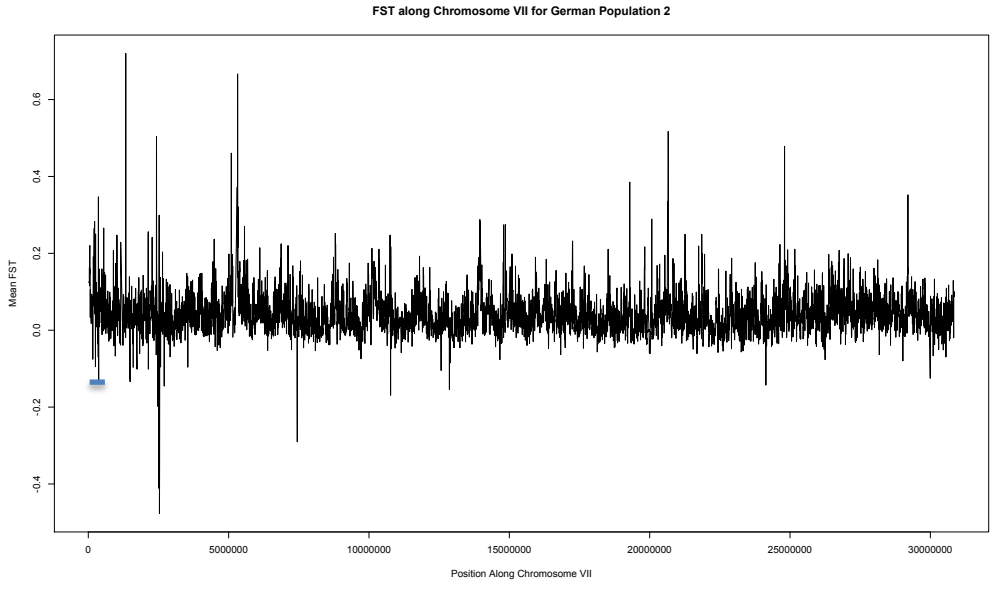
Figure 12: (a) Tajima's D (b) Windowed Pi along Chromosome VII (area of interest highlighted with blue line, not to scale)

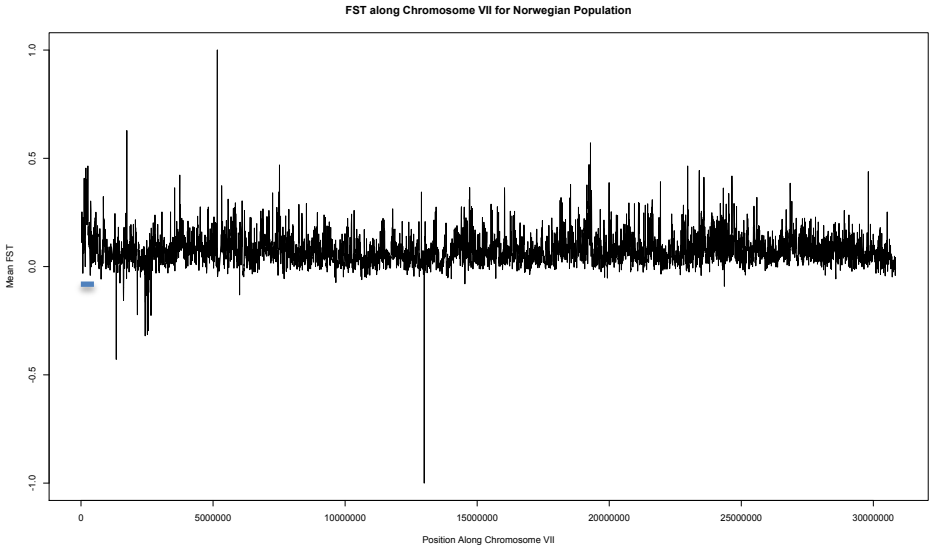
After this, pairwise FST comparisons between river and lake fish population pairs were made for each of the five river-lake population pairs (2 German pairs, 1 Norwegian, 1 Canadian, 1 from USA), as well as between the Danish Marine Population, and all European Freshwater populations.

(a)

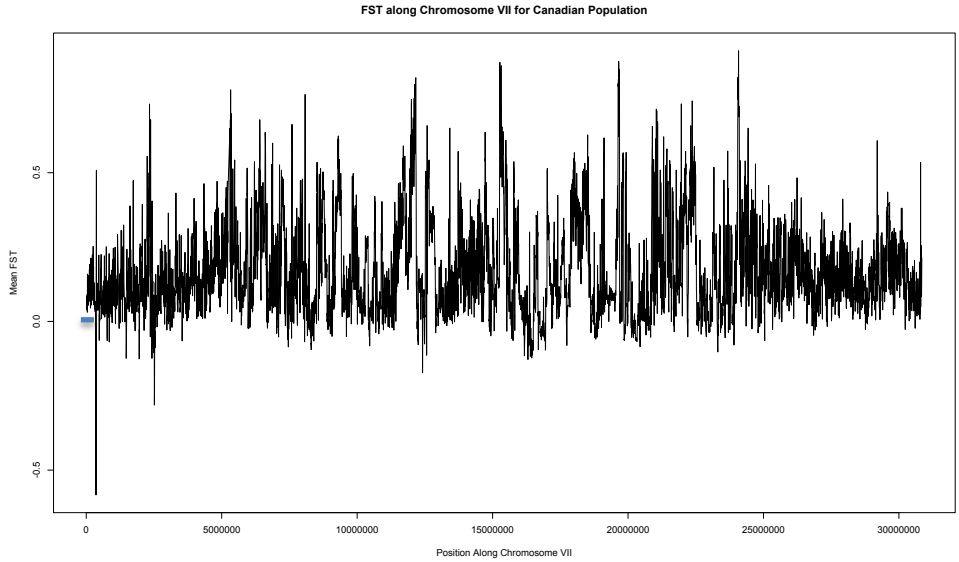


(b)

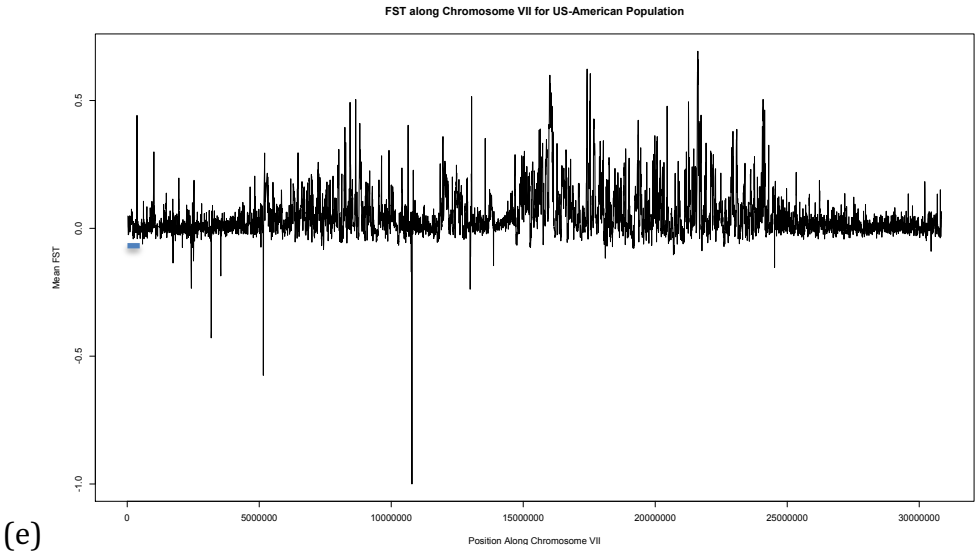



FST along Chromosome VII for Norwegian Population
Mean FST
1.0
0.5
0.0
-0.5
-1.0
0
5000000
10000000
15000000
20000000
25000000
30000000
Position Along Chromosome VII

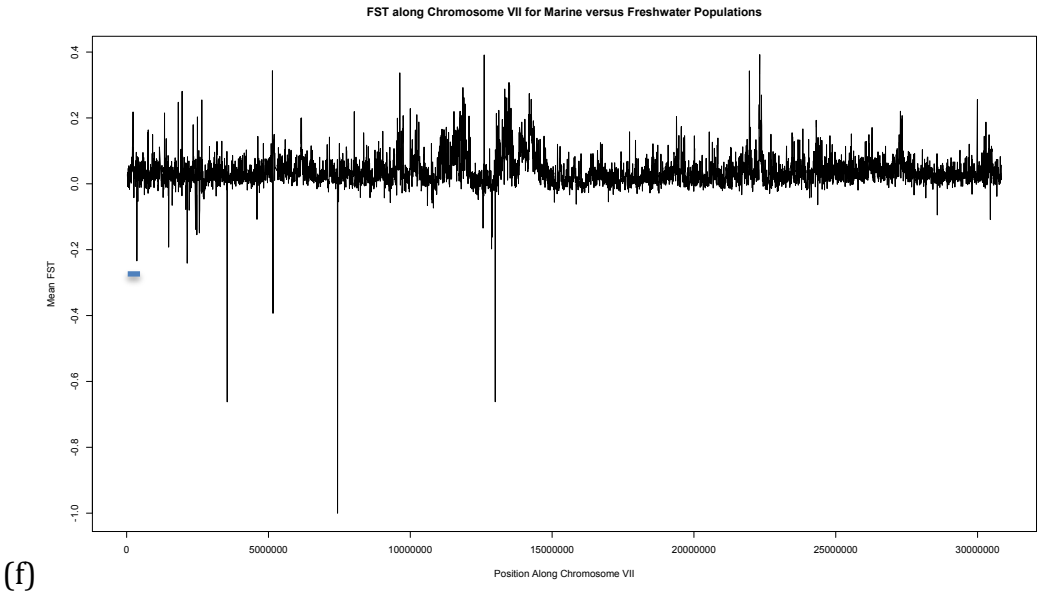
(c)


FST along Chromosome VII for Canadian Population
Mean FST
0.5
0.0
-0.5
0
5000000
10000000
15000000
20000000
25000000
30000000
Position Along Chromosome VII

(d)



(e)



(f)

Figure 13: FST along Chromosome VII for (a) German Population 1 (b) German Population 2 (c) Norwegian Population (d) Canadian Population (e) US-American Population (f) Marine and freshwater populations (area of interest highlighted with blue line, not to scale)

The FST calculations showed many regions of divergence in Chromosome VII, with no outliers in the MHC Class- II region. In the German population pairs, FST varied greatly along Chr VII, with no obvious outliers. On the other hand, the Norwegian population pair had 3 large peaks larger than all the others, with slightly elevated FST at the beginning of the chromosome. The Canadian population pair again gave highly variable FST values, but with 7 peaks. The US-American population pair again

gave us fairly uniform FST along the chromosome, but with 3 outliers. After these pairs, The Danish North Sea Marine, samples were compared to European Freshwater populations (so, the German and Norwegian river and lake population pairs pooled together), and this gave us fairly uniform FST values, but with 4 clear peaks. The smaller FST values indicate that allele frequencies are similar between each pair (river-lake, marine-freshwater), but in places where the FST is larger, this means that allele frequencies at those loci are different- this could indicate that those places have been subjected to diversifying selection (Holsinger and Weir 2015). In the case if FST, since 0 indicates panmixis, and 1, total differentiation (Weir and Cockerham 1984), negative FST values can be converted to zero (Hider et al. 2013). This indicates that the MHC regions have 0 or small differentiation.

Discussions

The synteny map generated in this work is the first direct proof of MHC CNV outside humans, to our knowledge. Before this work, it was known that this region contain excessive allelic variation for the MHC loci, and that these alleles are inherited together as haplotypes (Lenz et al. 2009). It was also known that the number of MHC class-IIB alleles per fish follows an optimum: that is, the mean residual number of parasite species is high in both fish with few alleles, as well as those with many alleles (Wegner, Reusch, and Kalbe 2003). Moreover, the average number of alleles per fish was found to be different depending on its origin. Lake fish have more alleles than river fish, coinciding with the fact that lakes have a higher parasite diversity than rivers (Eizaguirre et al. 2011). Also, lake and river stickleback also have different allele pools (Lenz et al. 2009; Eizaguirre et al. 2011).

In both the stickleback genome on Ensembl, and in the BAC clone examined by Reusch, et. al., there are two functional gene loci of the MHC Class-II A and B genes. However, studies of stickleback populations had shown that the high allelic diversity found in nature that could not be accounted for by two functional loci: some haplotypes were shown to have three alleles, whereas others had just one. This indicated that there might be copy number variation in this region, whose genomic basis was unknown. This was the motivation for creating two further BAC libraries, this time from fish with haplotypes containing different numbers of alleles. When haplotypes were chosen for sequencing, one with one allele, and one with three were picked.

Usually, the "genome" of a species refers to the genome of one individual of that species. In loci that are monomorphic, this is not an issue, but for one as polymorphic as the MHC, it is. The MHC region differs between individuals, and necessarily so, since it plays such an important role in adaptive immunity and mate choice. So, for further studies of the stickleback MHC, it is vital to have a comprehensive map that accounts for at least some of the variability in this region- this is the purpose of the

map generated (Fig 9). This map shows copy number variation, and can be used as a reference to map reads for population genomic studies.

In order to facilitate sequencing and assembly, homozygous fish were used as the source of DNA. One homozygous fish was produced by targeted inbreeding, while other was produced by gynogenesis, where an unfertilized egg was treated with UV radiated sperm and heat, upon which it developed into an adult fish (Samonte-Padilla et al. 2011). Genomic DNA was then extracted from these fish to make BAC libraries. Before BAC library creating, RSCA typing was performed on the fish, and it was known that one of them contained one variable allele, and the other contained three, and both contained the invariant locus, with the allele called No19. The variation in the MHC Class-II genomic region makes it problematic to assemble during the process of whole genome sequencing, so producing a BAC library, and isolating and sequencing BAC clones containing the MHC alleles was done to facilitate the study of this region. One BAC library was generated from each fish, giving us two in total from this sequencing effort.

Using a probe for the MHC locus, BAC clones containing MHC alleles were found. One clone from each library containing the non-classical region, and one clone with the classical region of the F haplotype were sequenced. For the G haplotype, 4 clones were sequenced, spanning all three loci. RSCA typing was done on these clones to confirm which alleles they contained. The four clones were assembled into the haplotype by performing a multiple sequence alignment, and using the consensus from this alignment (Supplementary Fig. 1). The alignment had the largest possible gap penalties; the logic being that since they are from the same sequence, there should not be any gaps. Later, it was necessary to extend the No05-containing contig, so as to have sequence information upstream of the MHC locus. All in all, we had two contigs containing variable alleles, and two containing No19. These were then annotated to determine “bounding genes”, so that we could select the homologous sequence between the same genes for all haplotypes, including the Reference

Genome. These sequences were then analyzed for shared synteny and structural and copy number variation.

The indel between the F and G haplotypes was first visualized using a dotplot. In addition, the three MHC loci in tandem arrangement could also be seen. To further elucidate the large indel between the two sequences, Mauve was used. In the classical region, it can be clearly seen that most of the MHC region forms one collinear block, except for the indel that accounts for the CNV. It appears that the first MHC locus on the G-haplotype, and the only locus on the H-haplotype are part of the collinear block, whereas the last two are part of the indel. This suggests that the sequence variants in the loci that are shared by the two haplotypes are alleles in the classical sense. An effort was made to assign loci to known alleles, which was not successful.

In addition, Mauve comparisons were also carried out in the non-classical region. Here, it was seen that the differences in annotated genes between the various haplotypes are not, indeed structural variation. In fact, the pattern of exons and introns on them are largely the same. So, the differences in annotated genes are due to small changes in nucleotides that are interpreted differently by the MAKER pipeline.

Finally, we wanted to study the conservation of gene order between the three haplotypes. So, a SimpleSynteny figure was generated. From there, it can be seen that there is collinearity between all the other genes in the region except for the MHC genes. Also, while the MHC Class-II A and B antigen genes of each locus are collinear for the functional loci, in the Reference Genome, they are on two separate blocks as shown in red rectangles in Fig. 8 (a)(iii). Another noticeable feature is that while the intergenic region between the A and the B antigens in all the functional genes are all relatively short, the intergenic region in front of the pseudogenized B locus is relatively large- which could perhaps have played a role in pseudogenization. Theoretically, if the Classical A and B loci share promoters, the transcription machinery would be more likely to dissociate from the DNA in the larger intergenic

region, and the B locus would not be transcribed, alleviating selective pressure, and allowing it to pseudogenize.

From the study of promoters, it was found that most MHC loci have the Y region, as well as either an inverse TATA or GATA box, or both. This shows that transcription machinery can at least recognize these loci, and that they can be transcribed. One Classical B locus (the first one on the Reference Genome) does not have any promoters at all. This can mean either of two things, that it shares the promoter of the A locus, or that it's not transcribed, and the gene product of the corresponding A locus forms a dimer with a different B locus gene product. Alternatively, the screening was not comprehensive enough to pick up a promoter. Moreover, for the classical region, there was generally more homology among the A locus promoters than the B loci, suggesting that the B loci are transcribed with their corresponding A loci from the same promoters. In the non-classical region, this is not possible, as the No19 A and B loci have opposite orientations. Here, we can see that not only are both promoter regions conserved, each has at least one promoter box. These can then be transcribed independently (Fig. 10).

The map generated in this work is the first comprehensive map of the MHC region of the three-spined stickleback. It is the first proof of the CNV that was suspected in this region, as well as the first proof of there being two sub-regions within the MHC region. One part of this region was found to be monomorphic, with the same allele- No19- found in all fish examined so far. This region was therefore called the non-classical region, following Horton et. al.'s system of classification. The other region had alleles that varied between fish, and it was called the classical-region. The MHC alleles in the classical region code for proteins that present antigens, but the function of the non-classical MHC gene product is unknown. In humans, Non-classical MHC have various different functions in the innate and adaptive immune system, such as activating natural killer cells (Braud, Allan, and McMichael 1999), so, the stickleback Non-Classical locus may also have some function.

So, we were able to see the pattern of CNV's in the MHC region, but that still left the question of how this pattern was formed. A pass of the preliminary annotation pipeline had showed viral reverse transcriptase genes between the MHC loci, which provided a clue that viral-like sequences, such as transposable elements, could have played a role. So, transposable elements were analyzed the MHC-Class-II region.

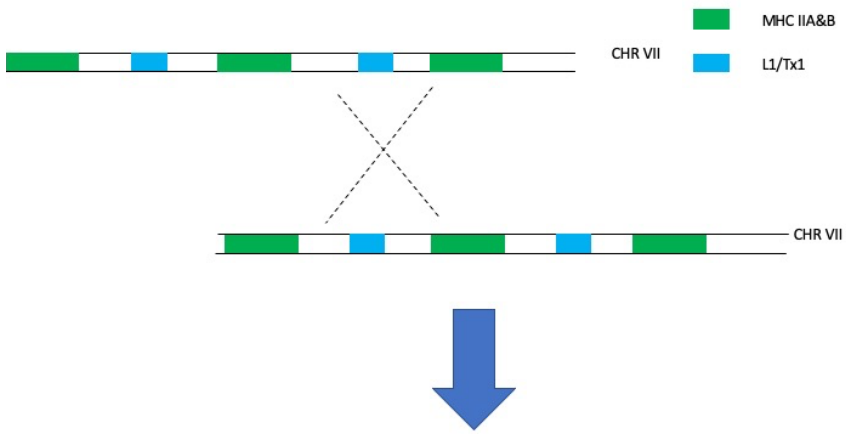
The L1 LINE is an interesting TE for many different reasons. It is the largest contributor to mammalian genome size and makes up 35% of the human genome (Babushok and Kazazian 2007). A full-length human L1 is approx. 6 Kb in length, and consists of a 5' untranslated region (UTR), two open reading frames (ORF1 and ORF2, both required for retrotransposition), separated by a linker, and a 5' UTR that contains a variable-length poly-A tail, and is flanked by short direct repeats. Existing genomic L1's can recombine to cause insertions, deletions, and disruptions of genic or regulatory regions. (Babushok and Kazazian 2007). It is also capable of endonuclease-independent retrotransposition, that is, it can restore non homologous end joining in cells that are deficient in endonuclease (Morrish et al. 2002). Unequal homologous crossing over caused by L1 has already been implicated in human disease (Burwinkel and Kilimann 1998).

In this work, I posit two possible mechanisms for speeding up the formation of CNV in the stickleback MHC region (Fig. 14): Non-Allelic Homologous Recombination (NAHR), and Break-Induced Recombination (BIR). In these models, the presence of L1 in tandem arrangement with the MHC loci suggests the following model:

1. An L1 invades the MHC region in an ancestral stickleback.
2. The L1 acts as a substrate for unequal crossing over (NAHR), or strand invasion (BIR).
3. This leads to the formation of different CNV haplotypes.
4. This also causes the specific pattern of MHC and L1 being present in n, and n-1.
5. The haplotypes that contain only MHC locus were never invaded by L1.

In these models, a diploid fish starts out with two identical haplotypes (i.e. a homozygote), but after the NAHR and BIR, it has two different haplotypes with different numbers of gene copies (i.e. a heterozygote), thereby increasing the amount of variation. Of course, the initial copies of the MHC genes have already been made, by some other mechanism such as slippage. However, the amount of variation in the population is still smaller. Invasion of an L1 increases the probability of recombination to make haplotypes with fewer or more MHC loci, as the probability of NAHR is increased when repeats larger than 1 kb in length exist close by (Hurles 2004), and the L1 in the vicinity of the MHC loci are ~ 5 kb in length. Finally, since recombination splits the L1, the MHC loci emerge unscathed. This allows for much faster CNV formation within the same population.

An endogenous retrovirus (ERV) accounts for the large gap between the MHC Class-IIA and pseudogenised B locus in the Reference Genome. It likely invaded this region in an ancestor of this haplotype- and not the others seen here- and made many copies, extending the intergenic region (Fig. 11 e)).



(a)

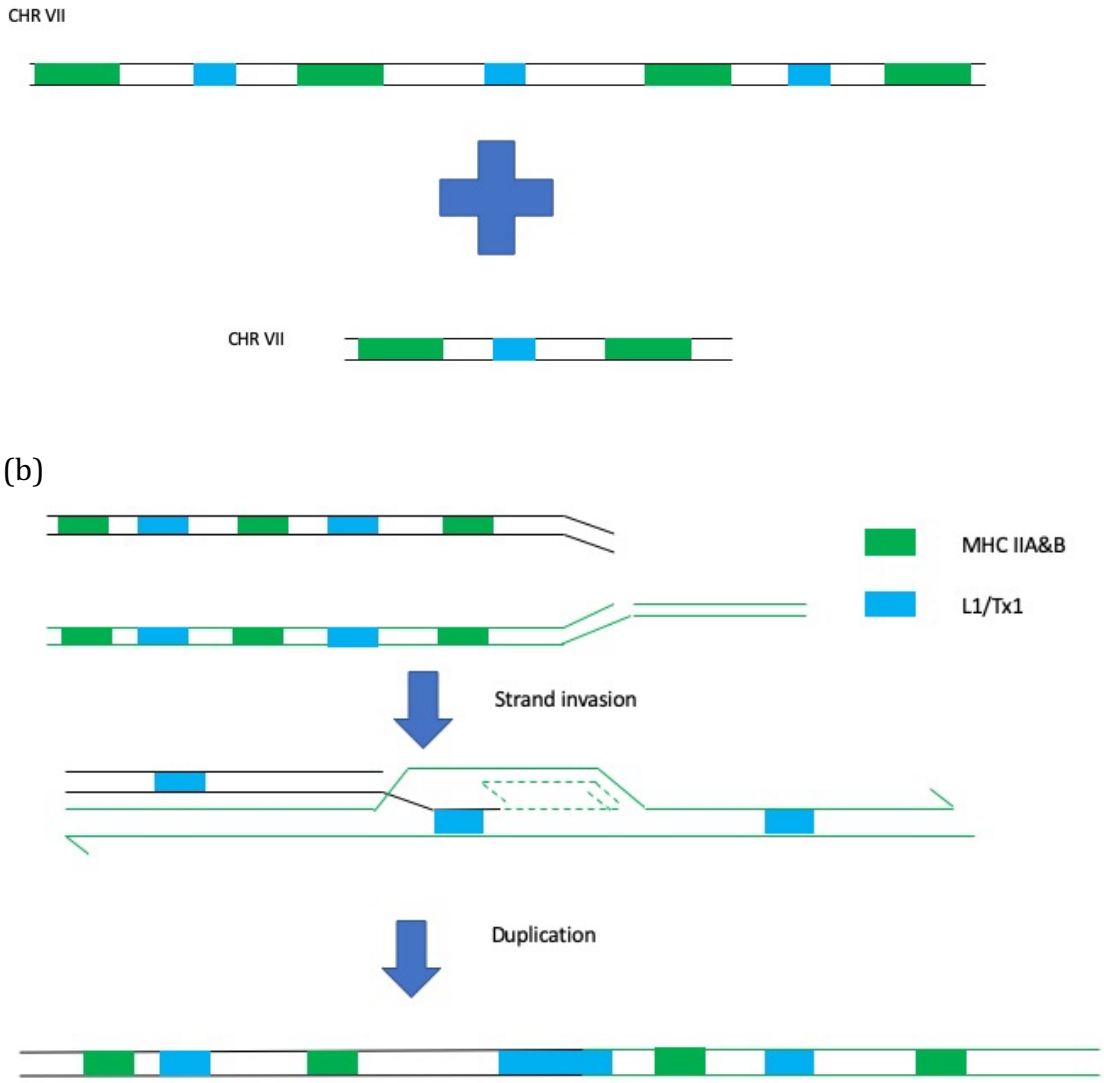


Figure 14: Possible mechanism of CNV formation (a) Unequal Crossing Over (b) Break- induced recombination, adapted from (Hastings et al. 2009). Both these mechanisms result in an alternating pattern of MHC and L1.

After looking at the extent of haplotypic variation in the MHC Class-II region, and how it is generated, we wanted to understand its consequences at the population level, over evolutionary timescales. First, we performed a test of neutrality to compare the MHC region to the rest of the Chromosome it is on, Chromosome VII. For this, we conducted a chromosome scan with window size of 5 kb. Our expectation was that because the former is subject to parasite pressures, and since this region shows local

adaptation, it would show a signature of selection, that is, the value of Tajima's D in this region would be a peak compared to its surrounding regions. This was not observed. Chromosome VII was found to contain regions of very high and very low Tajima's D, with the MHC region not being an outlier (Fig. 12 (a)). The bin containing the MHC genes was found to have a value of -0.44347, which is not significant.

Next we calculated a measure of nucleotide diversity, pi. However, this did not show any evidence for selective sweeps or balancing selection at the MHC region (Fig. 12(b)). Finally, we calculated FST for 5 river-lake population pairs, and one marine-freshwater pair. The MHC region did not have significantly different FST from its neighbors. Two German populations showed highly variable FST values, with no obvious outliers. The Norwegian population pair had 3 large outliers, which means that they had 3 highly differentiated regions, just not in our region of interest. The Canadian and US-American population pairs gave us 3 and 7 highly differentiated regions, though not in the MHC. The Marine versus Freshwater showed us regions that might have diverged due to the colonization of freshwater environments, which did not include the MHC. However, the polymorphisms found in these scans could also be uninformative, for which additional tests can be performed, such as, examining the allele frequency distribution of these sites (Roesti, Salzburger, and Berner 2012).

It was found that the MHC Class-II region is not an outlier when it comes to measures of neutrality, diversity, and divergence, and did not show obvious signatures of selective sweeps or increased balancing selection. This was contrary to our expectations, but could also be a result of the methods used. For example, in the future derived allele frequency with respect to distance from the MHC loci could be calculated as in (Lenz et al. 2016). However, this work could provide a jumping-off point while studying variation in natural populations of the three-spined stickleback MHC Class-II region.

Conclusions

Parasite-mediated divergent selection can occur in a population when three prerequisites are met:

1. There must be differences in parasite exposure among individuals in the host populations. This is common in allopatry, but can also occur in sympatry or parapatry due to heterogeneity in parasite species (i.e. the makeup of the parasite community) within the host population.
2. These differences in host infection should remain reasonably constant through time in order for selection to occur.
3. Finally, parasite infections have to impose a strong fitness cost upon the host, and also overrule fitness benefits from any other antagonistic mode of selection. That is, if there is a different mode of selection acting because of some other ecological factor- say, food availability- the effect of this selection is small compared to the effect of parasite-mediated selection.

Parasite-mediated selection has been theorized (Summers et al. 2003) to speed up speciation for both the host and the parasite, and while differences in parasitism between different ecotypes of stickleback has been found, it is not known how applicable these findings are for other systems. The mechanisms of speciation are still under study: if speciation is the result of a reproductive barrier between populations, how does this barrier emerge? In a system where there are two connected populations with different parasite communities, there are three mechanisms that can work either alone or in some combination:

1. Direct Natural Selection, where hybrid viability is directly reduced by parasites.
2. Pleiotropy, where genes that confer parasite resistance also play a role in mate choice.
3. Ecologically-based Sexual Selection, where parasite-mediated divergent selection can result in populations being isolated from one another.

While the theoretical considerations of this problem have been studied to great depth, empirical evidence still needs to be collected. In the metaphor of the speciation continuum, while the two ends have been characterized, the continuum between them needs to be scrutinized- for example, what role do parasites play in this continuum (Karvonen and Seehausen 2012)?

While parasites can be thought of as external forces acting on a population, speciation can also occur due to internal or genetic factors. The genetic basis- though not necessarily the cause- of reproductive isolation can be broadly divided into three types: extra-chromosomal (such as transposable elements), chromosomal (such as polyploidy), and genic, that is, incompatibilities between the genes of the diverging species. In the genic view, speciation genes are those that cause isolation to physiological behavioral, or ecological conditions, such as hybrid infertility. To understand genic speciation, we have to find out what the speciation genes are, what their normal function is, and how this function diverges between populations (Wu and Ting 2004).

The main motivation behind this work was to improve our understanding of various factors that move populations along the speciation continuum. Specifically, what are the genomic factors- that is, speciation genes- that interact with an ecological selective force to move populations either towards or away from reproductive isolation? Furthermore, what do these different modes of selection do to the host's genome organization? To answer these questions, the three-spined stickleback MHC Class-II system was ideal in many ways.

This system has all the three components of incipient reproductive isolation by parasite-mediated selection: extracellular parasites are the ecological source of divergent selection, mate choice occurs to maximize parasite resistance (Andreou et al. 2017) and selection against hybrids or migrants (Kaufmann et al. 2014), are the mechanisms of (cryptic) reproductive isolation, and the underlying genomic basis is MHC (Eizaguirre et al. 2009). Moreover, MHC are immune genes that have a

pleiotropic effect on mate choice (Milinski 2006). If we consider MHC to be speciation genes, we know their normal function, which is antigen presentation. We also know how this function diverges between natural populations- different MHC haplotypes provide resistance to different sets of parasites, leading to local adaptation for different parasite environments.

Because the effect of parasite-mediated selection on allele frequency, selection against migrants and mate choice are known, we were able to focus on the underlying genomic basis of this phenomenon. Another reason why the MHC Class-II region was chosen was because the gene products of the MHC Class-II loci interact with extra-cellular parasites, which are in consideration here.

Copy Number Variation has been shown to be an important force in evolution, and their role in human evolution has been very well documented. When copy number variations form, phenotypes can be altered, and this provides a substrate for selective forces to act on. CNV can change phenotype in many ways that can act alone or in concert:

1. Change coding regions of genes
2. Create paralogs that can take on new functions (neofunctionalization)
3. Altering gene expression patterns

Since most changes to genic or regulatory DNA sequences are maladaptive, they are removed through purifying selection, but if the change confers an advantage, it will be selected for. However, variants in non-functional DNA sequences will evolve under neutrality. This is as true of CNV as it is of SNP's, though the former affect larger numbers of bases at a time. CNV's have been shown to occur in many genes in the human genome, and play a significant role in evolution. Some genes whose CNV's that have been implicated evolutionary transitions under different selection regimes are salivary amylase (AMY1) and alpha-globin (Iskow, Gokcumen, and Lee 2012).

CNV's account for about 5 times as many variable base pairs than SNP's, so if they are in coding regions, they are likely to have a larger effect. CNV's may increase in allele frequency because of local adaptation, and would do so differently in different populations (Saitou and Gokcumen 2019). As such, the role of CNV's in the stickleback MHC Class-II was a very interesting candidate for study.

The synteny map generated in this work is the first comprehensive schematic of the MHC Class-II region of the three-spined stickleback. It is also the first genomic evidence of MHC CNV outside humans to our knowledge, as previous works only estimated CNV (Minias et al. 2019)(Iskow, Gokcumen, and Lee 2012). This work is also proof of there being two sub-regions within the MHC class II region. One of these subregions harbored only one MHC class II locus, which was found to be invariant, with the same allele- No19- found in all fish examined so far. This region was therefore called the non-classical region, following Horton et. al.'s system of classification. The other subregion had alleles that varied among fish, and it was called the classical region. The MHC alleles in the classical region code for proteins that present antigens, but the function of the non-classical MHC gene product is unknown.

Beyond confirming the presence of CNV, this work also explores its possible causes. As shown above, the invasion of transposable elements may have formed a basis for different mechanisms to generate CNV, including slippage or non-allelic homologous recombination. This might have led to the formation of haplotypes with various number of loci – 2, 3, 4, etc. Haplotypes that were not invaded by this TE contain only one gene copy. This invasion- and the formation of CNV- likely happened at timescales larger than the differentiation of sticklebacks into the Northern German populations, as this pattern is also seen in the Reference Genome, which is of an individual from Alaska.

There also many aspects of the stickleback MHC that are expected for a teleost- the general absence of other immune genes around the MHC class-II loci, as well as the

genome-wide distribution of MHC antigen presentation pathway genes. This is true of Zebrafish (Sambrook, Figueroa, and Beck 2005), and it has now been shown to be true of sticklebacks.

The three-spined stickleback has undergone recent, repeated adaptive radiations (Bell and Foster 1994; Hendry et al. 2013). In this process, sticklebacks must have encountered many novel parasite environments, and fast adaptation was critical. Because of the genomic organization in the MHC Class-II regions, stickleback were able make novel haplotypes from old ones, as well as recombine two new haplotypes within two generations to have novel genotypes that could generate excessive allelic and copy number variation. This would allow sticklebacks to optimize the number of alleles for their parasite environment, and ultimately allow for populations to have distinct allele pools- which, in the absence of CNV could occur by divergent selection on one locus. This could be one of the reasons why sticklebacks were able colonize different environments so successfully. As such, the organization of the stickleback MHC Class-II genomic region would be one of the factors that could move populations quickly along the speciation continuum.

Bibliography

Abrusán, György, Norbert Grundmann, Luc Demester, and Wojciech Makalowski. 2009. "TEclass - A Tool for Automated Classification of Unknown Eukaryotic Transposable Elements." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp084.

Andreou, Demetra, Christophe Eizaguirre, Thomas Boehm, and Manfred Milinski. 2017. "Mate Choice in Sticklebacks Reveals That Immunogenes Can Drive Ecological Speciation." *Behavioral Ecology* 00: 1–9. https://doi.org/10.1093/beheco/arx074.

Arkhipova, Irina R. 2017. "Using Bioinformatic and Phylogenetic Approaches to Classify Transposable Elements and Understand Their Complex Evolutionary Histories." *Mobile DNA*. https://doi.org/10.1186/s13100-017-0103-2.

Babushok, Daria V., and Haig H. Kazazian. 2007. "Progress in Understanding the Biology of the Human Mutagen LINE-1." *Human Mutation*. https://doi.org/10.1002/humu.20486.

Bannai, Hidemi P., and Masaru Nonaka. 2013. "Comprehensive Analysis of Medaka Major Histocompatibility Complex (MHC) Class II Genes: Implications for Evolution in Teleosts." *Immunogenetics* 65 (12): 883–95. https://doi.org/10.1007/s00251-013-0731-8.

Barth, Julia M. I., David Villegas-Ríos, Carla Freitas, Even Moland, Bastiaan Star, Carl André, Halvor Knutsen, et al. 2019. "Disentangling Structural Genomic and Behavioural Barriers in a Sea of Connectivity." *Molecular Ecology*, no. June 2018: 1394–1411. https://doi.org/10.1111/mec.15010.

Bell, Michael A., and Susan A. Foster. 1994. "The Evolutionary Biology of the Threespine Stickleback."

Böhne, Astrid, Frédéric Brunet, Delphine Galiana-Arnoux, Christina Schultheis, and Jean Nicolas Volff. 2008. "Transposable Elements as Drivers of Genomic and Biological Diversity in Vertebrates." *Chromosome Research* 16 (1): 203–15. https://doi.org/10.1007/s10577-007-1202-6.

Bolnick, Daniel I., and Benjamin M. Fitzpatrick. 2007. "Sympatric Speciation: Models

and Empirical Evidence." *Annual Review of Ecology, Evolution, and Systematics* 38 (1): 459–87. https://doi.org/10.1146/annurev.ecolsys.38.091206.095804.

Braud, Veronique M., David Sj Allan, and Andrew J. McMichael. 1999. "Functions of Nonclassical MHC and Non-MHC-Encoded Class I Molecules." *Current Opinion in Immunology* 11 (1): 100–108. https://doi.org/10.1016/S0952-7915(99)80018-1.

Buckling, Angus, and Paul B. Rainey. 2002. "The Role of Parasites in Sympatric and Allopatric Host Diversification." *Nature* 420 (6915): 496–99. https://doi.org/10.1038/nature01164.

Burwinkel, Barbara, and Manfred W. Kilimann. 1998. "Unequal Homologous Recombination between LINE-1 Elements as a Mutational Mechanism in Human Genetic Disease." *Journal of Molecular Biology* 277 (3): 513–17. https://doi.org/10.1006/jmbi.1998.1641.

Butler, Jennifer E.F., and James T. Kadonaga. 2002. "The RNA Polymerase II Core Promoter: A Key Component in the Regulation of Gene Expression." *Genes and Development* 16 (20): 2583–92. https://doi.org/10.1101/gad.1026202.

Campbell, Michael S., Carson Holt, Barry Moore, and Mark Yandell. 2014. *Genome Annotation and Curation Using MAKER and MAKER-P*. *Current Protocols in Bioinformatics*. Vol. 2014. https://doi.org/10.1002/0471250953.bi0411s48.

Cantarel, Brandi L., Ian Korf, Sofia M C Robb, Genis Parra, Eric Ross, Barry Moore, Carson Holt, Alejandro Sánchez Alvarado, and Mark Yandell. 2008. "MAKER: An Easy-to-Use Annotation Pipeline Designed for Emerging Model Organism Genomes." *Genome Research* 18 (1): 188–96. https://doi.org/10.1101/gr.6743907.

Chain, Frédéric J.J., Philine G.D. Feulner, Mahesh Panchal, Christophe Eizaguirre, Irene E. Samonte, Martin Kalbe, Tobias L. Lenz, et al. 2014. "Extensive Copy-Number Variation of Young Genes across Stickleback Populations." *PLoS Genetics* 10 (12): 655–63. https://doi.org/10.1371/journal.pgen.1004830.

Chain, Frédéric J J, and Philine G D Feulner. 2014. "Ecological and Evolutionary Implications of Genomic Structural Variations." *Frontiers in Genetics*, 2014. https://doi.org/10.3389/fgene.2014.00326.

Chen, Li-Cheng, Hong Lan, Li Sun, Yan-Li Deng, Ke-Yi Tang, and Qiu-Hong Wan. 2015. “Genomic Organization of the Crested Ibis MHC Provides New Insight into Ancestral Avian MHC Structure.” *Scientific Reports* 5: 7963. https://doi.org/10.1038/srep07963.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, et al. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics* 27 (15): 2156–58. https://doi.org/10.1093/bioinformatics/btr330.

Darling, Aaron C E, Bob Mau, Frederick R. Blattner, and Nicole T. Perna. 2004. “Mauve: Multiple Alignment of Conserved Genomic Sequence with Rearrangements.” *Genome Research* 14 (7): 1394–1403. https://doi.org/10.1101/gr.2289704.

Darling, Aaron E., Bob Mau, and Nicole T. Perna. 2010. “Progressivemauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement.” *PLoS ONE* 5 (6). https://doi.org/10.1371/journal.pone.0011147.

Eizaguirre, C., and T. L. Lenz. 2010. “Major Histocompatibility Complex Polymorphism: Dynamics and Consequences of Parasite-Mediated Local Adaptation in Fishes.” *Journal of Fish Biology* 77 (9): 2023–47. https://doi.org/10.1111/j.1095-8649.2010.02819.x.

Eizaguirre, Christophe, Tobias L. Lenz, Martin Kalbe, and Manfred Milinski. 2012. “Rapid and Adaptive Evolution of MHC Genes under Parasite Selection in Experimental Vertebrate Populations.” *Nature Communications* 3: 621. https://doi.org/10.1038/ncomms1632.

Eizaguirre, Christophe, Tobias L. Lenz, Ralf D. Sommerfeld, Chris Harrod, Martin Kalbe, and Manfred Milinski. 2011. “Parasite Diversity, Patterns of MHC II Variation and Olfactory Based Mate Choice in Diverging Three-Spined Stickleback Ecotypes.” *Evolutionary Ecology* 25 (3): 605–22. https://doi.org/10.1007/s10682-010-9424-z.

Eizaguirre, Christophe, Tobias L. Lenz, Arne Traulsen, and Manfred Milinski. 2009. “Speciation Accelerated and Stabilized by Pleiotropic Major Histocompatibility Complex Immunogenes.” *Ecology Letters* 12 (1): 5–12.

https://doi.org/10.1111/j.1461-0248.2008.01247.x.

Elsen, Peter J. Van Den, Ad Peijnenburg, Marja C.j.a. Van Eggermond, and Sam J.p. Gobin. 1998. “Shared Regulatory Elements in the Promoters of MHC Class I and Class II Genes.” *Immunology Today* 19 (7): 308–12. https://doi.org/10.1016/S0167-5699(98)01287-0.

Fan, Shaohua, and Axel Meyer. 2014. “Evolution of Genomic Structural Variation and Genomic Architecture in the Adaptive Radiations of African Cichlid Fishes.” *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2014.00163.

Farrar, Kerrie, and Iain S Donnison. 2007. “Construction and Screening of BAC Libraries Made from Brachypodium Genomic DNA.” *Nature Protocols* 2 (7): 1661–74. https://doi.org/10.1038/nprot.2007.204.

Feulner, P. G. D., and R. De-Kayne. 2017. “Genome Evolution, Structural Rearrangements and Speciation.” *Journal of Evolutionary Biology*. https://doi.org/10.1111/jeb.13101.

Feulner, Philine G.D., Frédéric J.J. Chain, Mahesh Panchal, Christophe Eizaguirre, Martin Kalbe, Tobias L. Lenz, Marvin Mundry, et al. 2013. “Genome-Wide Patterns of Standing Genetic Variation in a Marine Population of Three-Spined Sticklebacks.” *Molecular Ecology* 22 (3): 635–49. https://doi.org/10.1111/j.1365-294X.2012.05680.x.

Feulner, Philine G D, Fr??d??ric ????J J Chain, Mahesh Panchal, Yun Huang, Christophe Eizaguirre, Martin Kalbe, Tobias L. Lenz, et al. 2015. “Genomics of Divergence along a Continuum of Parapatric Population Differentiation.” *PLoS Genetics* 11 (2): 1–18. https://doi.org/10.1371/journal.pgen.1004966.

Flajnik, Martin F., and Masanori Kasahara. 2001. “Comparative Genomics of the MHC: Glimpses into the Evolution of the Adaptive Immune System.” *Immunity* 15 (3): 351–62. https://doi.org/10.1016/S1074-7613(01)00198-4.

Gandon, Sylvain, and Y. Michalakis. 2002. “Local Adaptation, Evolutionary Potential and Host-Parasite Coevolution: Interactions between Migration, Mutation, Population Size and Generation Time.” *Journal of Evolutionary Biology* 15 (3): 451–62. https://doi.org/10.1046/j.1420-9101.2002.00402.x.

Gao, Bo, Dan Shen, Songlei Xue, Cai Chen, Hengmi Cui, and Chengyi Song. 2016. “The

Contribution of Transposable Elements to Size Variations between Four Teleost Genomes." *Mobile DNA* 7 (1): 4. https://doi.org/10.1186/s13100-016-0059-7.

Gerts, E. Michael, Yi Kuo Yu, Richa Agarwala, Alejandro A. Schäffer, and Stephen F. Altschul. 2006. "Composition-Based Statistics and Translated Nucleotide Searches: Improving the TBLASTN Module of BLAST." *BMC Biology* 4: 1–14. https://doi.org/10.1186/1741-7007-4-41.

Glazer, Andrew M., Emily E. Killingbeck, Therese Mitros, Daniel S. Rokhsar, and Craig T. Miller. 2015. "Genome Assembly Improvement and Mapping Convergently Evolved Skeletal Traits in Sticklebacks with Genotyping-by-Sequencing." *G3: Genes|Genomes|Genetics* 5 (7): 1463–72. https://doi.org/10.1534/g3.115.017905.

Hastings, P. J., James R. Lupski, Susan M. Rosenberg, and Grzegorz Ira. 2009. "Mechanisms of Change in Gene Copy Number." *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2593.

Hendry, Andrew P., Catherine L. Peichel, Blake Matthews, Janette W. Boughman, and Patrik Nosil. 2013. "Stickleback Research: The Now and the Next." *Evolutionary Ecology Research* 15 (2): 111–41.

Hider, Jessica L., Rachel M. Gittelman, Tapan Shah, Melissa Edwards, Arnold Rosenbloom, Joshua M. Akey, and Esteban J. Parra. 2013. "Exploring Signatures of Positive Selection in Pigmentation Candidate Genes in Populations of East Asian Ancestry." *BMC Evolutionary Biology* 13 (1). https://doi.org/10.1186/1471-2148-13-150.

Holsinger, Kent E, and Bruce S Weir. 2015. "Genetics in Geographically Structured Populations: Defining, Estimating and Interpreting F ST." *Nature Reviews. Genetics* 10 (9): 639–50. https://doi.org/10.1038/nrg2611.

Horton, Roger, Laurens Wilming, Vikki Rand, Ruth C. Lovering, Elspeth A. Bruford, Varsha K. Khodiyar, Michael J. Lush, et al. 2004. "Gene Map of the Extended Human MHC." *Nature Reviews Genetics* 5 (12): 889–99. https://doi.org/10.1038/nrg1489.

Hurles, Matthew. 2004. "Gene Duplication: The Genomic Trade in Spare Parts." *PLoS*

Biology 2 (7). https://doi.org/10.1371/journal.pbio.0020206.

Iskow, RC, O Gokcumen, and C Lee. 2012. "Exploring the Role of Copy Number Variants in Human Adaptation." *Trends in Genetics* 28 (6): 245–57. https://doi.org/10.1016/j.tig.2012.03.002.Exploring.

Jones, Felicity C., Manfred G. Grabherr, Yingguang Frank Chan, Pamela Russell, Evan Mauceli, Jeremy Johnson, Ross Swofford, et al. 2012. "The Genomic Basis of Adaptive Evolution in Threespine Sticklebacks." *Nature* 484 (7392): 55–61. https://doi.org/10.1038/nature10944.

Karvonen, Anssi, and Ole Seehausen. 2012. "The Role of Parasitism in Adaptive Radiations-When Might Parasites Promote and When Might They Constrain Ecological Speciation?" *International Journal of Ecology* 2012. https://doi.org/10.1155/2012/280169.

Kaufmann, Joshka, Tobias L. Lenz, Manfred Milinski, and Christophe Eizaguirre. 2014. "Experimental Parasite Infection Reveals Costs and Benefits of Paternal Effects." *Ecology Letters* 17 (11): 1409–17. https://doi.org/10.1111/ele.12344.

Kawecki, Tadeusz J., and Dieter Ebert. 2004. "Conceptual Issues in Local Adaptation." *Ecology Letters* 7 (12): 1225–41. https://doi.org/10.1111/j.1461-0248.2004.00684.x.

Kazazian, Haig H. 2004. "Mobile Elements: Drivers of Genome Evolution." http://science.sciencemag.org/.

King, Thomas, Simon Butcher, and Lukasz Zalewski. 2017. "Apocrita - High Performance Computing Cluster For Queen Mary University Of London," 3–4. https://doi.org/10.5281/ZENODO.438045.

Kingsley, David. 2011. "Sequencing the Genome of Threespine Sticklebacks (Gasterosteus Aculeatus)." *National Human Genome Research Institute White Paper.*, no. 650: 1–14.

Kohany, Oleksiy, Andrew J. Gentles, Lukasz Hankus, and Jerzy Jurka. 2006. "Annotation, Submission and Screening of Repetitive Elements in Repbase: RepbaseSubmitter and Censor." *BMC Bioinformatics* 7 (October). https://doi.org/10.1186/1471-2105-7-474.

Krumsiek, Jan, Roland Arnold, and Thomas Rattei. 2007. "Gepard: A Rapid and

Sensitive Tool for Creating Dotplots on Genome Scale." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btm039.

Lassmann, Timo, and Erik Ll Sonnhammer. 2005. "Kalign-an Accurate and Fast Multiple Sequence Alignment Algorithm." https://doi.org/10.1186/1471-2105-6-298.

Lenz, T. L., C. Eizaguirre, J. P. Scharsack, M. Kalbe, and M. Milinski. 2009. "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: 2122–42. https://doi.org/10.1111/j.1095-8649.2009.02410.x.

Lenz, Tobias L., Victor Spirin, Daniel M. Jordan, and Shamil R. Sunyaev. 2016. "Excess of Deleterious Mutations around HLA Genes Reveals Evolutionary Cost of Balancing Selection." *Molecular Biology and Evolution* 33 (10): 2555–64. https://doi.org/10.1093/molbev/msw127.

Lenz, Tobias L, Christophe Eizaguirre, Sven Becker, and Thorsten B H Reusch. 2009. "RSCA Genotyping of MHC for High-Throughput Evolutionary Studies in the Model Organism Three-Spined Stickleback Gasterosteus Aculeatus." *BMC Evolutionary Biology* 9: 57. https://doi.org/10.1186/1471-2148-9-57.

Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM" 00 (00): 1–3. http://arxiv.org/abs/1303.3997.

Li, W. H., and L. A. Sadler. 1991. "Low Nucleotide Diversity in Man." *Genetics* 129 (2): 513–23.

Lin, Jane M., Patrick J. Collins, Nathan D. Trinklein, Yutao Fu, Hualin Xi, Richard M. Myers, and Zhiping Weng. 2007. "Transcription Factor Binding and Modified Histones in Human Bidirectional Promoters." *Genome Research* 17 (6): 818–27. https://doi.org/10.1101/gr.5623407.

Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibuls, David Altshuler & Mark J Daly. 2011. "A Framework for Variation Discovery and Genotyping Using Next- Generation

DNA Sequencing Data" 43 (5): 491–98. https://doi.org/10.1038/ng.806.A.

Martin Flajnik and Masanori Kasahara. 2010. "Origin and Evolution of the Adaptive Immune System: Genetic Events and Selective Pressures." *Nat Rev Genet* 11 (1): 47–59. https://doi.org/10.1038/nrg2703.Origin.

Matsuo, Megumi Y., and Masaru Nonaka. 2004. "Repetitive Elements in the Major Histocompatibility Complex (MHC) Class I Region of a Teleost, Medaka: Identification of Novel Transposable Elements." *Mechanisms of Development* 121 (7–8): 771–77. https://doi.org/10.1016/j.mod.2004.03.014.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A. 2009. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." *Proceedings of the International Conference on Intellectual Capital, Knowledge Management & Organizational Learning* 20: 254–60. https://doi.org/10.1101/gr.107524.110.20.

MICHAEL O'CONNOR,* MARK PEIFER, WELCOME BENDER. 1995. "Construction of Large DNA Segments in Escherichia Coli," 5419–28.

Milinski, M. 2006. "The Major Histcompatibility Complex, Sexual Selection, and Mate Choice." *Annual Review of Ecology, Evolution, and Systematics* 37 (2006): 159–86. https://doi.org/10.2307/annurev.ecolsys.37.091305.30000008.

Milinski, Manfred. 2006. "The Major Histocompatibility Complex, Sexual Selection and Mate Choice." *Annu. Rev. Ecol. Evol. Syst.* 37 (May): 159–86. https://doi.org/10.2307/annurev.ecolsys.37.091305.30000008.

Minias, Piotr, Ewa Pikus, Linda A. Whittingham, and Peter O. Dunn. 2019. "Evolution of Copy Number at the MHC Varies across the Avian Tree of Life." *Genome Biology and Evolution* 11 (1): 17–28. https://doi.org/10.1093/gbe/evy253.

Morrish, Tammy A., Nicolas Gilbert, Jeremy S. Myers, Bethaney J. Vincent, Thomas D. Stamato, Guillermo E. Taccioli, Mark A. Batzer, and John V. Moran. 2002. "DNA Repair Mediated by Endonuclease-Independent LINE-1 Retrotransposition." *Nature Genetics* 31 (2): 159–65. https://doi.org/10.1038/ng898.

Muotri, Alysson R., Maria C N Marchetto, Nicole G. Coufal, and Fred H. Gage. 2007. "The Necessary Junk: New Functions for Transposable: Elements." *Human*

Molecular Genetics. https://doi.org/10.1093/hmg/ddm196.

Nei, Masatoshi, and Wen-Hsiung Li. 1979. "Mathematical Model for Studying Genetic Variation in Terms of Restriction Endonucleases (Molecular Evolution/Mitochondrial DNA/Nucleotide Diversity)." *Genetics*. Vol. 76.

Nonaka, Mayumi I., and Masaru Nonaka. 2010. "Evolutionary Analysis of Two Classical MHC Class i Loci of the Medaka Fish, Oryzias Latipes: Haplotype-Specific Genomic Diversity, Locus-Specific Polymorphisms, and Interlocus Homogenization." *Immunogenetics* 62 (5): 319–32. https://doi.org/10.1007/s00251-010-0426-3.

Oosterhout, C van. 2009. "Transposons in the MHC: The Yin and Yang of the Vertebrate Immune System." *Heredity* 103 (3): 190–91. https://doi.org/10.1038/hdy.2009.46.

Peng, Zhen, Weichen Zhou, Wenqing Fu, Renqian Du, Li Jin, and Feng Zhang. 2015. "Correlation between Frequency of Non-Allelic Homologous Recombination and Homology Properties: Evidence from Homology-Mediated CNV Mutations in the Human Genome." *Human Molecular Genetics*. https://doi.org/10.1093/hmg/ddu533.

Price, Alkes L., Neil C. Jones, and Pavel A. Pevzner. 2005. "De Novo Identification of Repeat Families in Large Genomes." *Bioinformatics* 21 (SUPPL. 1): 351–58. https://doi.org/10.1093/bioinformatics/bti1018.

R Core Development Team. 2015. *R: A Language and Environment for Statistical Computing*. Vol. 2. https://doi.org/10.1007/978-3-540-74686-7.

Redon, Richard, Shumpei Ishikawa, Karen R. Fitch, Lars Feuk, George H. Perry, T. Daniel Andrews, Heike Fiegler, et al. 2006. "Global Variation in Copy Number in the Human Genome." *Nature*. https://doi.org/10.1038/nature05329.

Reusch, Thorsten B H, and ??sa Langefors. 2005. "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): 531–41. https://doi.org/10.1007/s00239-004-0340-0.

Reusch, Thorsten B H, Helmut Schaschl, and K Mathias Wegner. 2004. "Recent Duplication and Inter-Locus Gene Conversion in Major Histocompatibility Class

II Genes in a Teleost, the Three-Spined Stickleback." *Immunogenetics* 56 (6): 427–37. https://doi.org/10.1007/s00251-004-0704-z.

Roesti, Marius, Walter Salzburger, and Daniel Berner. 2012. "Uninformative Polymorphisms Bias Genome Scans for Signatures of Selection." *BMC Evolutionary Biology* 12 (1). https://doi.org/10.1186/1471-2148-12-94.

Rundle, H.D., and Patrick Nosil. 2005. "Ecological Speciation." *Ecology Letters*. Tohoku University. https://doi.org/10.18960/seitai.66.3_561.

Saitou, Marie, and Omer Gokcumen. 2019. "An Evolutionary Perspective on the Impact of Genomic Copy Number Variation on Human Health." *Journal of Molecular Evolution*, no. 0123456789. https://doi.org/10.1007/s00239-019-09911-6.

Sambrook, Jennifer G, Felipe Figueroa, and Stephan Beck. 2005. "A Genome-Wide Survey of Major Histocompatibility Complex (MHC) Genes and Their Paralogues in Zebrafish." *BMC Genomics* 6: 152. https://doi.org/10.1186/1471-2164-6-152.

Samonte-Padilla, Irene E, Christophe Eizaguirre, Jörn P Scharsack, Tobias L Lenz, and Manfred Milinski. 2011. "Induction of Diploid Gynogenesis in an Evolutionary Model Organism, the Three-Spined Stickleback (Gasterosteus Aculeatus)." *BMC Developmental Biology* 11 (1): 55. https://doi.org/10.1186/1471-213X-11-55.

Schluter, Dolph, and Gina L Conte. 2009. "Genetics and Ecological Speciation." *Proceedings of the National Academy of Sciences of the USA* 106 Suppl: 9955–62. https://doi.org/10.1073/pnas.0901264106.

Schrader, Lukas, and Jürgen Schmitz. 2018. "The Impact of Transposable Elements in Adaptive Evolution." *Molecular Ecology*, no. May. https://doi.org/10.1111/mec.14794.

Schrider, Daniel R., and Matthew W. Hahn. 2010. "Gene Copy-Number Polymorphism in Nature." *Proceedings of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rspb.2010.1180.

Seehausen, Ole, Roger K Butlin, Irene Keller, Catherine E Wagner, Janette W Boughman, Paul A Hohenlohe, Catherine L Peichel, and Glenn-peter Saetre.

2014. “Genomics and the Origin of Species.” *Nature Publishing Group* 15 (3): 176–92. https://doi.org/10.1038/nrg3644.

Servedio, Maria R, G Sander Van Doorn, Michael Kopp, Alicia M Frame, and Patrik Nosil. 2011. “Magic Traits in Speciation : ‘ Magic ’ but Not Rare ?” 26 (8). https://doi.org/10.1016/j.tree.2011.04.005.

Sharp, Andrew J, Devin P Locke, Sean D Mcgrath, Ze Cheng, Jeffrey A Bailey, Rhea U Vallente, Lisa M Pertz, et al. 2005. “Segmental Duplications and Copy-Number Variation in the Human Genome,” 78–88.

Stanke, Mario, Rasmus Steinkamp, Stephan Waack, and Burkhard Morgenstern. 2004. “AUGUSTUS: A Web Server for Gene Finding in Eukaryotes.” *Nucleic Acids Research* 32 (WEB SERVER ISS.): 309–12. https://doi.org/10.1093/nar/gkh379.

Summers, Kyle, Sea McKeon, Jon Sellars, Mark Keusenkothen, James Morris, David Gloeckner, Corey Pressley, Blake Price, and Holly Snow. 2003. “Parasitic Exploitation as an Engine of Diversity.” *Biological Reviews of the Cambridge Philosophical Society* 78 (4): 639–75. https://doi.org/10.1017/S146479310300616X.

Tajima, Fumio. 1984. “Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism.” *Pharmatherapeutica* 3 (9): 607–12.

Tarailo-Graovac, Maja, and Nansheng Chen. 2009. “Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences.” *Current Protocols in Bioinformatics*, no. SUPPL. 25: 1–14. https://doi.org/10.1002/0471250953.bi0410s25.

Ting, Jenny Pan Yun, and John Trowsdale. 2002. “Genetic Control of MHC Class II Expression.” *Cell*. Cell Press. https://doi.org/10.1016/S0092-8674(02)00696-7.

Tsukamoto, Kentaro, Shinpei Hayashi, Megumi Y Matsuo, Mayumi I Nonaka, Mariko Kondo, Akihiro Shima, Shiuchi Asakawa, Nobuyoshi Shimizu, and Masaru Nonaka. 2005. “Unprecedented Intraspecific Diversity of the MHC Class I Region of a Teleost Medaka, Oryzias Latipes.” *Immunogenetics* 57 (6): 420–31. https://doi.org/10.1007/s00251-005-0009-x.

Valen, Leigh Van. 1973. “A NEW EVOLUTIONARY LAW.” *Evol. Theory.* 1: 1–30.

Veltri, Daniel, Martha Malapi Wight, and Jo Anne Crouch. 2016. “SimpleSynteny: A Web-Based Tool for Visualization of Microsynteny across Multiple Species.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkw330.

Ward, Natalie, and Gabriel Moreno-Hagelsieb. 2014. “Quickly Finding Orthologs as Reciprocal Best Hits with BLAT, LAST, and UBLAST: How Much Do We Miss?” *PLoS ONE* 9 (7): 1–6. https://doi.org/10.1371/journal.pone.0101850.

Wegner, K. Mathias, T. B H Reusch, and M. Kalbe. 2003. “Multiple Parasites Are Driving Major Histocompatibility Complex Polymorphism in the Wild.” *Journal of Evolutionary Biology* 16 (2): 224–32. https://doi.org/10.1046/j.1420-9101.2003.00519.x.

Weir, B. S., and C. Clark Cockerham. 1984. “Estimating F-Statistics for the Analysis of Population Structure Author (s): B . S . Weir and C . Clark Cockerham Published by : Society for the Study of Evolution.” *Evolution* 38 (6): 1358–70. https://doi.org/10.2307/2408641.

Wicker, Thomas, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, et al. 2009. “Reply: A Unified Classification System for Eukaryotic Transposable Elements Should Reflect Their Phylogeny.” *Nature Reviews Genetics* 10 (4): 276–276. https://doi.org/10.1038/nrg2165-c4.

Woolhouse, Mark E.J., Joanne P. Webster, Esteban Domingo, Brian Charlesworth, and Bruce R. Levin. 2002. “Biological and Biomedical Implications of the Co-Evolution of Pathogens and Their Hosts.” *Nature Genetics* 32 (4): 569–77. https://doi.org/10.1038/ng1202-569.

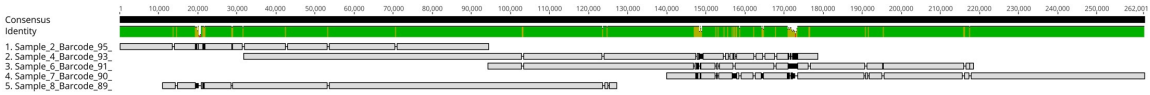
Wu, Chung I., and Chau Ti Ting. 2004. “Genes and Speciation.” *Nature Reviews Genetics* 5 (2): 114–22. https://doi.org/10.1038/nrg1269.

Yandell, M., and D. Ence. 2012. “A Beginner’s Guide to Eukaryotic Genome Annotation.” *Nature Rev Genet* 13 (5): 329–42. https://doi.org/10.1038/nrg3174.

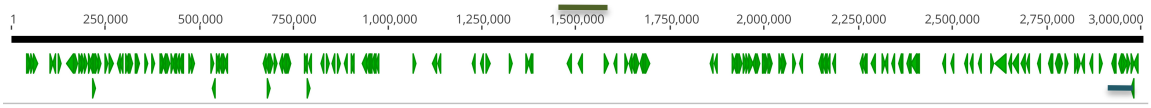
Zhang, Michael Q. 2007. “Computational Analyses of Eukaryotic Promoters.” *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-8-S6-S3.

Zhang, Zheng, Scott Schwartz, Lukas Wagner, and Webb Miller. 2000. “A Greedy Algorithm for Aligning DNA Sequences.” *Journal of Computational Biology* 7 (1–2): 203–14. https://doi.org/10.1089/10665270050081478.

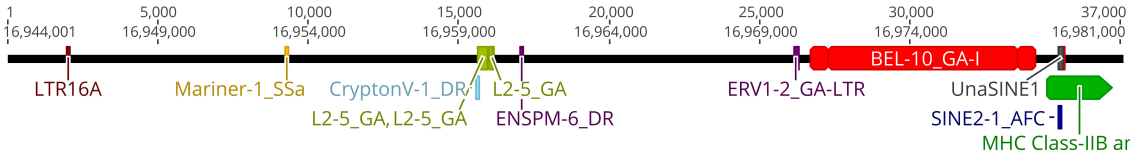
Supplementary Materials



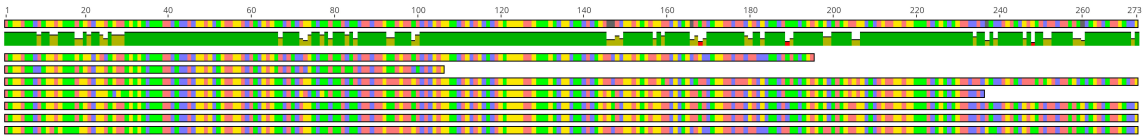
Supplementary Figure 1: Alignment of four clones to make a contig of the G haplotype.



Supplementary Figure 2: Annotated 3MB of ChrVII, the green line indicates the classical region, and the blue, the non-classical region.



Supplementary Figure 3: MHC Class-IIB locus on ChrIII.



Probe sequences:

Supplementary Figure 4: Multiple Sequence Alignments of the exon2's of all classical MHC Class-IIB alleles. The second sequence from the top shows the severely truncated exon2 of the pseudogene.

Probe sequence 1. For the MHC region:

>No05_L2

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~GAGTTCATCGACTCGTATTACTTCAACAAG

TTAGAATACACGAGGTTCAGCAGCTCAGTGGGGAAGTTTGTCGGCTTCACTGAGTACGGA

GTGAGGAACGCTGAATACTGGAACAACGACGCTTCATTTCTGAGTGCTATGAGAGCTCAG

AAGGAGGTTTACTGTCTGAACCACGTCCCGGTCTATTACAACAATGTGCTGACTAAGTCC

G

>TB_29-7_TB7F

---------ACCTCGGGGTAACGACTGGAGTAAGTACCTGTGTGTCTACATACCGGTCTG

CCTGCCTCTCCACCTGTCTCTCTGCTAACCCGTCCCCCTGTCTGTCCTCAGACCCGTCCA

TGCCCGAGTCTGAGAGGAACAAAGTCGCCATCGGAGCCTCAGGACTGATCCTGGGTCTGA

CTTTGTCTCTGGCTGGATTCATCTACTACAAGAGGAAAGCCAGAGGTCAGAACACACCGA

TGTCCTCTGATGGGTCAGATTGGTCCAGAATGAGGGGAGACCTTCAGCAGGTCCACCAGA

GTCTGGACCTGGTTCACCTGCTCCCTCTGACA

>TB_29-7_TB7R

TGCTGGTTTGTGTGTCAGCAGGTCTGGAGAGAAGATCTCGTGTGTGGTGGAGCACATCAG

TCTGAGTAAACCTCTGGTTACCGACTGGAGTAAGTACCTGTCTGTCTGCATACCGGTCTG

CCTGCCTCTCCACCTGTCTCTCTGCTAACCCGTCCCCCTGTCTGTCCTCAGACCCGTCCA

TGCCCGAGTCTGAGAGGAACAAAGTCGCCATCGGAGCCTCAGGACTGATCCTGGGTCTGA

CTTTGTCTCTGGCTGGATTCATCTACTACAAGAGGAAAGCCAGAGGTCAGAACACACCGA

TGTCCTCTGATGGGTCAGATTGGT

>TB_47-7_TB7F

~~~~~~----------

---------ACCTCTGGTTACCGACTGGAGTAAGTACCTGTCTGTCTGCATACCGGTCTG

CCTGCCTCTCCACCTGTCTCTCTGCTAACCCGTCCCCCTGTCTGTCCTCAGACCCGTCCA

TGCCCGAGTCTGAGAGGAACAAAGTCGCCATCGGAGCCTCAGGACTGATCCTGGGTCTGA

CTTTGTCTCTGGCTGGATTCATCTACTACAAGAGGAAAGCCAGAGGTCAGAACACACCGA

TGTCCTCTGATGGGTCAGATCGGTCCAAAATAAGAGGAGACCTTCAGCAGGTCCACCAAA

GTCTGGACCTGGTTCACCTNTNCCCTCTGACA

>TB_47-7_TB7R

GCTGTTTTTGTGTGTCAGCAGGTCTGGAGAGAAGATCTCGTGTGTGGTGGAGCACATCAG

TCTGAGTAAACCTCTGGTTACCGACTGGAGTAAGTACCTGTCTGTCTGCATACCGGTCTG

CCTGCCTCTCCACCTGTCTCTCTGCTAACCCGTCCCCCTGTCTGTCCTCAGACCCGTCCA

TGCCCGAGTCTGAGAGGAACAAAGTCGCCATCGGAGCCTCAGGACTGATCCTGGGTCTGA

CTTTGTCTCTGGCTGGATTCATCTACTACAAGAGGAAAGCCAGAGGTCAGAACACACCGA

TGTCCTCTGATGGGTCAGATCGGTCCAGAATAA

Probe sequence 2. For cytochrome P450, family 17, subfamily A, polypeptide 2

>Upstream_sequence:

ACGTCGCCATGGAGACCATCACAGCCGCTGCCACCTGCCTGGCGGCGTCCGAGGGTACGGTCAACACTCGGGGCCCCGATGCCGAGGACT
CGCCGGTCGTGGAGGTGGTCGTTATGGGGGCGGCCTCTCCTGACTCAGAGGGGTCTCTTTCTGTGACTACATCGCCAATGAAGCAACAGT
CCAAGCCGGTTCTGCTTCCCAGTCCAAGGGGCGTTGCCCCACGGCGCCCCTCCATTCAGACCGTGGGCGCCGAGGCTCTGGATCGGGGCCC
CGGCACTTCTGACAGAAGCCTCGAGCAGACGAGAACATCCGGCACGGCCGATGATCCTTGTGGTCCAAACACCATACCGGGTCCAGACGC
TAGCAGGGCCTCAGTCACAGGCCCTACTAGCGTCTCAACCCCGGGCCCAAGTCCTAGTCCGGTGCCAATTCCAGCCTGTCTGAATCCAAC
CCCATCTCCAAAACCAGTACCTTCAAACCCAACCCACGCTTCCACCCGAGTACCTTCAAACCTTAAACAAGCTCCAGAACCAGTACCCAA
ACACCTAATCTCTGTACCGGACCCAATACTTACAAACCTTACCCACGTGCCAAAGCAACACTCTACAATTACCACGGTATCAGACTGTGA
AAAGAGCCCACCTCCAAAACCTGCCTATTTCAACCGTTGCAAGTTTACAGAACCGTCAGCAGGGTCAGATGAAGTGCTGGTTCCTCCCGG
CCCGATGCCAGGCTCACCCTTCTGCCTGGGTTCTGTTGACCTTTTTGTATCTGGTCCAGATTTGGCTCATTCTGAGGTCCCGACTCTATC
ACCGGACCTTGTCAATGCATCGGCGTCCAATAGAGCTGGATTGAGTTCTGGTTCTGGATCAGGTCCAGCGCTCACTTTCACATCACCTTC
GCCCTCCCACCCGGACATAACCATCAAGTCCGAGTCCTCCTCCGGCCTCCACCTTTCCTCGCCTCCTTCCCTCCTGGCCCCGGCTCTCCTGC
TCGATCCCCTTGCCGCTCTGCACCAGTCGAACAAAGGCGGGCCTCCCCCCGGCCACGCCTCCTCCCTGTCCCCGTCGCCTCGGGCCACCCA
GTCCCCCCCCGAAGCAGACCCTCTTCTCCCCTTGCGTGGACGTCTTTGAGCCGGGACCTCCCAGCTGGGCGGTCGGCGAGGAGGAGCGGGA
GGAGGACGAGGAGGAAGAGGAAGACATTGGAGCCGACGAGAGCCAGTACCGACATCGCCGACTCACCGGCGACTCGGGCATCGAGGTCT
GCCGGTGTCGGGTGGAGGAGGAGGAGGAGGAAGAGGAAGAGGAAGAGGAGAGGGATATGAAAGGAGGCGGAGACACTGAACTCCACGA
CAGTGCGGACTGCATCGCCAGAGGTCACACGACCGTGGGGGAGGGCCTCAGGCCGCGTAGCTCCGCCTCCACTGCCACGCCCGAGGACGC
TGGCAAAGTTGTTGTTGTCGTCGTGGAGACCGTGTGATCGGCCTGAACCGGGTGGATCAAACCAATCCCACCAAGCAGTGGCAGTGAAG

GTAACTACTACTACTACAGTGAAGGCTACTACGGTTAGCATCCGTTTGTCGGGCGTTAGCAAGAGATTCTACGGGAAGCCTCAGCTCAA
CTCTACGGGGAGTATTAGCTTGATTTTACGGGAGGCTGTAGCGCGATTCCGTGGGAGGCGTTTTAGACGTGGTTGCACCACTAAACCAA
CGGCTGGAAGAAGTATGGGAAATTGTCAATCCGGTGATTTAAAGGGTTGAACAGAAATTAAACCCACGGCCTCGATTTGTAATTTGTCA
ACTCCAAACTTGGTAGAATACCAATAGTTTCTTTCAACCTGCATCATCTGCCTCCATATTTGCGGGCTGGCCGTCGTTGGCCTCCTCTGT
TGATCCGTCATCAGTAACATTTGTGTTAAACAACAAATGCCCTCACCCTTTAAAGCAGGACAACAGGTAATAAAGCTATTCTTTAGCGA
TTAAATAGATTTAAACCTCGCTAAT

Supplementary Table 1: Annotations Table for (a) 3MB of ChrVII of the Reference Genome (bounding genes in red) (b) G Haplotype, Classical region(c) F Haplotype, classical region (d) F Haplotype, Non-Classical (e) G Haplotype, Non-Classical

(a)

Name	Minimum	Maximum	Length	Direction
phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit alpha isoform-like	40273	44116	3844	forward
eukaryotic translation initiation factor 4E type 2-like	45551	49447	3897	forward
guanylyl cyclase GC-E-like	50174	57980	7807	forward
retinal guanylyl cyclase 2-like isoform X1	59510	68411	8902	forward
sodium/potassium-transporting ATPase subunit beta-2-like	102216	111302	9087	forward
high mobility group protein B2-like	112669	116321	3653	reverse
CST complex subunit CTC1	124145	130465	6321	forward
disks large homolog 4-like	145412	164723	19312	reverse
7SK snRNA methylphosphate capping enzyme	168544	173654	5111	forward
ephrin type-B receptor 4-like, partial	178275	184260	5986	forward
proteasome subunit beta type-6-like	185676	187220	1545	reverse
trafficking protein particle complex subunit 1	188102	188937	836	forward
voltage-gated potassium channel subunit beta-3-like, partial	192121	198691	6571	forward
C-type lectin domain family 10 member A-like	202955	207441	4487	forward
uncharacterized protein CXorf57-like isoform X2	212118	215118	3001	reverse
intraflagellar transport protein 46 homolog	215157	218391	3235	reverse
mitochondrial 2-oxoglutarate/malate carrier protein	217317	221970	4654	forward
Profilin-2	222675	227468	4794	forward
beta-enolase	228503	233344	4842	forward
calmodulin-binding transcription activator 2 isoform X2	248224	252782	4559	forward
serrate RNA effector molecule homolog	260412	270385	9974	forward
NF-X1-type zinc finger protein NFXL1	279553	287054	7502	reverse

cyclic nucleotide-gated channel rod photoreceptor subunit alpha-like	291541	296927	5387	reverse
WW domain-binding protein 1-like	301692	306352	4661	forward
cytochrome P450, family 17, subfamily A, polypeptide 2	307143	311118	3976	reverse
early growth response protein 4-like	314548	317527	2980	forward
MHC class IIA antigen 1	327832	329792	1961	forward
MHC class IIB antigen 1	332529	335845	3317	forward
MHC class IIA antigen 2	354562	356951	2390	forward
MHC class IIB antigen 2	372560	375893	3334	forward
MHC class IIA antigen 3	394262	396312	2051	forward
MHC class IIB antigen 3	397066	402121	5056	forward
phosphatidylcholine:ceramide cholinephosphotransferase 2-like	407928	410583	2656	reverse
aminoacyl tRNA synthase complex-interacting multifunctional protein 1-like	411777	416088	4312	forward
Coiled-coil domain-containing protein 109B	425604	428308	2705	forward
complement factor I, partial	428300	434451	6152	reverse
tetratricopeptide repeat protein 39B-like, partial	437507	439132	1626	reverse
tetratricopeptide repeat protein 39B-like, partial	441928	447355	5428	reverse
hepatoma-derived growth factor-related protein 2-like isoform X2	448601	451506	2906	reverse
WD repeat-containing protein 55	453997	456152	2156	reverse
zinc finger protein 436-like isoform X1	472937	474905	1969	forward
uncharacterized protein LOC108874780	478692	486287	7596	forward
unnamed protein product, partial	530424	536607	6184	forward
fibrinogen beta chain	536481	540795	4315	reverse
putative toll-like receptor 2, partial	543084	544346	1263	forward
putative toll-like receptor 2, partial	545244	546817	1574	forward
protein FAM111A-like	548915	550420	1506	reverse
RING finger protein 175 isoform X2	554624	557407	2784	reverse
tripartite motif-containing protein 2-like	564494	569448	4955	forward
L-amino-acid oxidase-like	571196	574411	3216	reverse
uncharacterized protein LOC109953029	673787	674489	703	reverse
protein FAM160A1-like	679916	683576	3661	reverse
hypothetical protein HMPREF9081_1219	680335	682995	2661	forward
putative uncharacterized protein	686743	688440	1698	forward
SH3 domain-containing protein 19	697003	702934	5932	forward
transmembrane protein 184C	718712	720867	2156	reverse
endothelin B receptor-like	723451	728734	5284	reverse
tetratricopeptide repeat protein 29	732306	734978	2673	forward

sodium/bile acid cotransporter 7 isoform X1	735329	740476	5148	forward
OTU domain-containing protein 4	778481	779792	1312	forward
ATP-binding cassette sub-family E member 1 isoform X1	780137	785697	5561	reverse
anaphase-promoting complex subunit 10 isoform X1	785259	788058	2800	forward
F-box/WD repeat-containing protein 7-like	788489	796726	8238	reverse
protein phosphatase 1K, mitochondrial-like	822962	827907	4946	reverse
protein phosphatase methylesterase 1-like isoform X1	835545	843257	7713	forward
type II inositol 3,4-bisphosphate 4-phosphatase-like isoform X3	852242	860301	8060	reverse
ATP-binding cassette sub-family G member 2-like	864248	873043	8796	forward
tyrosine-protein kinase TXK	889638	892303	2666	reverse
tyrosine-protein kinase Tec isoform X2	902635	904551	1917	forward
myelin-oligodendrocyte glycoprotein-like	909244	910882	1639	reverse
putative bifunctional E2/E3 enzyme	933420	941656	8237	reverse
Butyrophilin subfamily 1 member A1	948494	951349	2856	reverse
cytochrome C biosynthesis protein	954272	954800	529	reverse
cytochrome C biosynthesis protein	957376	957904	529	reverse
cytochrome C biosynthesis protein	960544	964240	3697	reverse
ABC transporter ATP-binding protein	967262	967789	528	reverse
acidic leucine-rich nuclear phosphoprotein 32 family member B-like	969554	973840	4287	reverse
protein canopy homolog 3	1065302	1067672	2371	forward
retinal guanylyl cyclase 2-like	1119348	1127856	8509	reverse
ras-like protein family member 11A	1132581	1138557	5977	reverse
40S ribosomal protein S4, X isoform	1227739	1231180	3442	reverse
histone deacetylase 8	1251358	1254684	3327	reverse
phosphorylase b kinase regulatory subunit alpha, skeletal muscle isoform isoform X4	1259472	1273295	13824	forward
H(+)/Cl(-) exchange transporter 5 isoform X1	1321173	1327239	6067	forward
G2/mitotic-specific cyclin-B3-like	1367165	1374442	7278	forward
serine protease 33	1379378	1381828	2451	reverse
serine protease 27-like	1383385	1385415	2031	reverse
uncharacterized protein LOC102079667 isoform X2	1476339	1487063	10725	reverse
unnamed protein product	1506255	1515524	9270	reverse
collagen alpha-6(IV) chain-like	1575194	1587025	11832	forward
uncharacterized protein LOC104958540	1603583	1607310	3728	reverse
selection and upkeep of intraepithelial T-cells protein 1	1629385	1632548	3164	forward
crossover junction endonuclease MUS81	1642546	1647252	4707	reverse
putative transcription factor Ovo-like 1	1648964	1654781	5818	reverse

N-acylneuraminate cytidylyltransferase	1658970	1661895	2926	forward
triokinase/FMN cyclase	1674181	1676205	2025	reverse
triokinase/FMN cyclase	1676259	1680215	3957	reverse
exocyst complex component 6B	1683744	1693796	10053	forward
uncharacterized protein LOC106097725	1856406	1862845	6440	reverse
uncharacterized protein LOC10887368	1871383	1874634	3252	reverse
Zinc finger protein 638	1915640	1919454	3815	forward
transmembrane protein 55B-A-like	1921346	1924390	3045	reverse
C-type mannose receptor 2-like	1924944	1927277	2334	reverse
zinc finger protein RFP-like	1928243	1930664	2422	reverse
E3 ubiquitin-protein ligase TRIM21-like isoform X2	1934301	1937978	3678	forward
E3 ubiquitin-protein ligase TRIM39-like	1943123	1947476	4354	forward
E3 ubiquitin-protein ligase TRIM39-like	1951658	1954859	3202	forward
E3 ubiquitin-protein ligase TRIM39-like	1958336	1969426	11091	forward
receptor-interacting serine/threonine-protein kinase 3-like isoform X2	1973003	1977817	4815	reverse
protein KHNYN-like	1981615	1987842	6228	forward
adenylate cyclase type 4-like	1993417	1999256	5840	forward
adenylate cyclase type 4-like	2000128	2002209	2082	forward
ubiquitin carboxyl-terminal hydrolase CYLD-like	2003231	2007500	4270	forward
ER membrane protein complex subunit 4	2010911	2013098	2188	forward
lysophospholipid acyltransferase LPCAT4 isoform X1	2015098	2020096	4999	forward
solute carrier family 12 member 6 isoform X5	2037970	2042430	4461	forward
TOX high mobility group box family member 4 isoform X2	2045674	2047390	1717	forward
TOX high mobility group box family member 4 isoform X2	2048980	2050839	1860	forward
E3 ubiquitin/ISG15 ligase TRIM25-like	2076119	2081383	5265	forward
Astrocytic phosphoprotein PEA-15	2097209	2100121	2913	reverse
E3 ubiquitin-protein ligase TRIM21-like	2149624	2151645	2022	reverse
uncharacterized protein LOC104925763	2155131	2157605	2475	reverse
nuclear factor 7, ovary-like isoform X1	2161706	2163063	1358	reverse
nuclear factor 7, brain-like	2163421	2165065	1645	reverse
E3 ubiquitin-protein ligase TRIM39-like	2166588	2173290	6703	reverse
Purkinje cell protein 4-like protein 1	2181866	2188998	7133	reverse
putative ferric-chelate reductase 1	2256925	2260108	3184	reverse
uncharacterized protein LOC109995170 isoform X3	2264235	2270208	5974	reverse
uncharacterized protein LOC103375002 isoform X4	2271533	2274159	2627	reverse
protein NDRG2	2283734	2295410	11677	reverse

uncharacterized protein LOC104941075 isoform X1	2311693	2314850	3158	forward
poly [ADP-ribose] polymerase 14-like	2325520	2327551	2032	reverse
solute carrier family 12 member 3-like	2338240	2345865	7626	reverse
solute carrier family 12 member 3-like	2361129	2365307	4179	reverse
solute carrier family 12 member 3-like, partial	2365423	2367721	2299	reverse
leukotriene B4 receptor 1-like	2380656	2385959	5304	reverse
tripartite motif-containing protein 47-like	2388998	2390392	1395	reverse
uncharacterized protein LOC110972738	2391554	2402711	11158	reverse
GTPase IMAP family member 7-like	2407116	2408817	1702	reverse
GTPase IMAP family member 7-like isoform X1	2411214	2412654	1441	reverse
GTPase IMAP family member 7-like	2480155	2481678	1524	reverse
GTPase IMAP family member 7-like	2500156	2501460	1305	reverse
GTPase IMAP family member 7-like	2539173	2540443	1271	reverse
GTPase IMAP family member 7-like	2551467	2552440	974	reverse
GTPase IMAP family member 7-like	2573641	2575028	1388	reverse
vang-like protein 2 isoform X1	2602723	2608407	5685	forward
Fc receptor-like protein 5	2611847	2641328	29482	reverse
obscurin-like	2649860	2655914	6055	reverse
Fc receptor-like A	2661441	2673271	11831	reverse
obscurin-like	2680461	2692142	11682	reverse
obscurin-like	2698711	2704112	5402	reverse
Fc receptor-like A	2724819	2732317	7499	reverse
Fc receptor-like protein 5	2756825	2763285	6461	reverse
uncharacterized protein LOC108883612 isoform X5	2765792	2767502	1711	reverse
ATP-sensitive inward rectifier potassium channel 10-like	2773607	2780329	6723	reverse
purine nucleoside phosphorylase-like	2784278	2786640	2363	forward
splicing factor 3B subunit 2 isoform X1	2800573	2806472	5900	forward
phosphofurin acidic cluster sorting protein 1	2823665	2826525	2861	forward
kinesin light chain 2	2832927	2837869	4943	forward
alpha-actinin-3	2844967	2850973	6007	reverse
copper chaperone for superoxide dismutase	2867487	2870773	3287	reverse
galactose-3-O-sulfotransferase 3	2889965	2892718	2754	forward
actin-related protein 2-like	2926626	2929986	3361	reverse
homeobox protein SIX6-like	2931624	2934175	2552	reverse
prefoldin subunit 2	2945943	2947270	1328	reverse
nitrilase homolog 1	2949202	2950983	1782	forward
H(+)/Cl(-) exchange transporter 3 isoform X4	2951668	2962237	10570	forward

H(+)/Cl(-) exchange transporter 3-like	2963362	2968393	5032	forward
MHC class IIB antigen	2973665	2979151	5487	forward
MHC class IIA antigen	2978654	2980255	1602	reverse
high affinity immunoglobulin epsilon receptor subunit alpha-like isoform X1 (FC receptor-like)	2981953	2991293	9341	reverse

(b)

Name	Minimum	Maximum	Length	Direction
Mitochondrial sodium/hydrogen exchanger 9B2	8943	13609	4667	reverse
atrial natriuretic peptide-converting enzyme, partial	22348	29748	7401	reverse
NF-X1-type zinc finger protein NFXL1-like	38189	41231	3043	reverse
cyclic nucleotide-gated channel rod photoreceptor subunit alpha-like	47347	52827	5481	reverse
flocculation protein FLO11-like	57352	59052	1701	forward
cytochrome P450, family 17, subfamily A, polypeptide 2	62179	66448	4270	reverse
hypothetical protein, conserved NOT RECIPROCAL FIRST HIT	66759	69401	2643	forward
MHC class IIA antigen So05	82654	84707	2054	forward
MHC class IIB antigen So05	87142	91157	4016	forward
MHC class IIA antigen SCX03	106588	114539	7952	forward
MHC class IIB antigen SCX03	117456	121301	3846	forward
MHC class IIA antigen So11	139462	141523	2062	forward
MHC class IIB antigen So11	142411	145407	2997	forward
phosphatidylcholine:ceramide cholinephosphotransferase 2-like	153382	156053	2672	reverse
aminoacyl tRNA synthase complex-interacting multifunctional protein 1-like	157256	162082	4827	forward
calcium uniporter protein, mitochondrial-like	173303	175683	2381	forward
complement factor I, partial	176439	182428	5990	reverse
tetratricopeptide repeat protein 39B-like, partial	185522	187181	1660	reverse
tetratricopeptide repeat protein 39B isoform X3	188036	194843	6808	reverse
hepatoma-derived growth factor-related protein 2-like isoform X1	195632	200517	4886	reverse
WD repeat-containing protein 55	201934	204089	2156	reverse
zinc finger protein 436-like isoform X2	222486	224438	1953	forward
hypothetical protein ASZ78_015611, partial	226264	231029	4766	forward
uncharacterized protein LOC108874780	232064	239870	7807	forward
lecithin retinol acyltransferase-like	240529	243346	2818	forward

(c)

Name	Minimum	Maximum	Length	Direction
mitochondrial sodium/hydrogen exchanger 9B2-like	103	3347	3245	reverse
NF-X1-type zinc finger protein NFXL1	26100	33375	7276	reverse
cyclic nucleotide-gated channel rod photoreceptor subunit alpha-like	37885	43287	5403	reverse
mucin-5AC-like	48058	51197	3140	forward
cytochrome P450, family 17, subfamily A, polypeptide 2	53510	57120	3611	reverse
early growth response protein 4-like	60139	61630	1492	forward
MHC class IIA antigen	78618	80677	2060	forward
MHC class IIB antigen	83026	86271	3246	forward
phosphatidylcholine:ceramide cholinephosphotransferase 2-like	97418	100183	2766	reverse
aminoacyl tRNA synthase complex-interacting multifunctional protein 1-like	101388	105935	4548	forward
complement factor I, partial	123654	129840	6187	reverse
tetratricopeptide repeat protein 39B-like	132175	134512	2338	reverse
tetratricopeptide repeat protein 39B isoform X3	135442	142496	7055	reverse
hepatoma-derived growth factor-related protein 2-like isoform X2	145898	147698	1801	reverse
WD repeat-containing protein 55	151457	153606	2150	reverse

(d)

Name	Minimum	Maximum	Length	Direction
actin-related protein 2-like	31503	34904	3402	reverse
homeobox protein SIX6-like	36027	39094	3068	reverse
nitrilase homolog 1	54902	56690	1789	forward
H(+)/Cl(-) exchange transporter 3 isoform X4	57352	69711	12360	forward
H(+)/Cl(-) exchange transporter 3-like	71086	75577	4492	forward
MHC class IIB antigen	80783	83306	2524	forward
MHC class IIA antigen	85056	86842	1787	reverse
Fc receptor-like A	88373	98377	10005	reverse

(e)

Name	Minimum	Maximum	Length	Direction
actin-related protein 2-A-like	7148	10547	3400	reverse
beta-1,4-glucuronyltransferase 1 isoform X1	11667	25060	13394	reverse
nitrilase homolog 1	24320	31631	7312	forward
H(+)/Cl(-) exchange transporter 3 isoform X4	32304	42851	10548	forward
H(+)/Cl(-) exchange transporter 3-like	44227	48648	4422	forward
MHC class IIB antigen	53883	56416	2534	forward
MHC class IIA antigen	58176	59971	1796	reverse
uncharacterized protein LOC110960817 NOT RECIPROCAL FIRST HIT	61192	72227	11036	reverse
obscurin-like	75647	85072	9426	reverse
microfibrillar-associated protein 3-like	90717	95666	4950	reverse
kynurenine/alpha-aminoadipate aminotransferase, mitochondrial	97634	102484	4851	reverse
INO80 complex subunit B isoform X2	102768	105710	2943	forward

Supplementary Table 2: Percentages of repetitive elements:

1. 3MB of ChrVII

==========

file name: chrVII_unmasked_3MB.fasta

sequences: 1

total length: 3000000 bp (2669633 bp excl N/X-runs)

GC level: 47.31 %

bases masked: 499766 bp (16.66 %)

==========

	number of elements*	length occupied	percentage of sequence
SINEs:	0	0 bp	0.00 %
ALUs	0	0 bp	0.00 %
MIRs	0	0 bp	0.00 %

LINEs: 221 86194 bp 2.87 %
LINE1 1 722 bp 0.02 %
LINE2 127 43725 bp 1.46 %
L3/CR1 0 0 bp 0.00 %

LTR elements: 35 18834 bp 0.63 %
ERVL 0 0 bp 0.00 %
ERVL-MaLRs 0 0 bp 0.00 %
ERV_classI 9 5157 bp 0.17 %
ERV_classII 0 0 bp 0.00 %

DNA elements: 123 56782 bp 1.89 %
hAT-Charlie 58 36368 bp 1.21 %
TcMar-Tigger 2 952 bp 0.03 %

Unclassified: 1200 223451 bp 7.45 %

Total interspersed repeats: 385261 bp 12.84 %

Small RNA: 0 0 bp 0.00 %

Satellites: 0 0 bp 0.00 %
Simple repeats: 2200 103783 bp 3.46 %
Low complexity: 188 10779 bp 0.36 %
==========

* most repeats fragmented by insertions or deletions
have been counted as one element

The query species was assumed to be homo
RepeatMasker Combined Database: Dfam_Consensus-20170127

run with cross_match version 1.090518
The query was compared to classified sequences in ".../GaacLDB-families.fa"

2. BAC15_29 Variable Region:

==========
file name: alignmentgapmax.clustalw_consensus_sequence.fasta
sequences: 1
total length: 262001 bp (262001 bp excl N/X-runs)
GC level: 47.68 %
bases masked: 71763 bp (27.39 %)
==========

	number of elements*	length occupied	percentage of sequence
SINEs:	0	0 bp	0.00 %
ALUs	0	0 bp	0.00 %
MIRs	0	0 bp	0.00 %
LINEs:	38	15944 bp	6.09 %
LINE1	0	0 bp	0.00 %
LINE2	22	4397 bp	1.68 %
L3/CR1	0	0 bp	0.00 %
LTR elements:	11	7379 bp	2.82 %
ERVL	0	0 bp	0.00 %
ERVL-MaLRs	0	0 bp	0.00 %
ERV_classI	1	661 bp	0.25 %

ERV_classII 0 0 bp 0.00 %

DNA elements: 17 3915 bp 1.49 %
hAT-Charlie 1 47 bp 0.02 %
TcMar-Tigger 0 0 bp 0.00 %

Unclassified: 109 22332 bp 8.52 %

Total interspersed repeats: 49570 bp 18.92 %

Small RNA: 0 0 bp 0.00 %

Satellites: 0 0 bp 0.00 %
Simple repeats: 282 20953 bp 8.00 %
Low complexity: 15 1240 bp 0.47 %
==========

* most repeats fragmented by insertions or deletions
have been counted as one element

The query species was assumed to be homo
RepeatMasker Combined Database: Dfam_Consensus-20170127

run with cross_match version 1.090518
The query was compared to classified sequences in ".../GaacLDB-families.fa"

3. BAC15_47 Variable Region

==========

file name: Larger_Sample_3.fasta

sequences: 1

total length: 153816 bp (153816 bp excl N/X-runs)

GC level: 47.98 %

bases masked: 30980 bp (20.14 %)

==========

	number of elements*	length occupied	percentage of sequence
SINEs:	0	0 bp	0.00 %
ALUs	0	0 bp	0.00 %
MIRs	0	0 bp	0.00 %
LINEs:	12	6247 bp	4.06 %
LINE1	1	1582 bp	1.03 %
LINE2	10	4397 bp	2.86 %
L3/CR1	0	0 bp	0.00 %
LTR elements:	0	0 bp	0.00 %
ERVL	0	0 bp	0.00 %
ERVL-MaLRs	0	0 bp	0.00 %
ERV_classI	0	0 bp	0.00 %
ERV_classII	0	0 bp	0.00 %
DNA elements:	5	947 bp	0.62 %
hAT-Charlie	1	47 bp	0.03 %
TcMar-Tigger	0	0 bp	0.00 %
Unclassified:	54	8886 bp	5.78 %

Total interspersed repeats: 16080 bp 10.45 %

Small RNA: 0 0 bp 0.00 %

Satellites: 0 0 bp 0.00 %
Simple repeats: 171 14088 bp 9.16 %
Low complexity: 11 812 bp 0.53 %
==========

* most repeats fragmented by insertions or deletions
have been counted as one element

The query species was assumed to be homo
RepeatMasker Combined Database: Dfam_Consensus-20170127

run with cross_match version 1.090518
The query was compared to classified sequences in ".../GaacLDB-families.fa"

Acknowledgements

Firstly, I would like to thank my supervisor, Dr. Tobias L. Lenz. His kindness, compassion, and understanding were as important to me as his scientific guidance, and for that I will always be grateful. I would like to thank my Thesis Advisory Committee, Dr. Thorsten Reusch, and Dr. Tal Dagan, for their help. I'd like to thank my co-authors, the reason for this project: Doko-Miles Jackson Thorburn, Alejandro Ceron-Noriega, Dr. Mahesh Panchal-Binzer, Dr. Irene E. Samonte-Padilla, Dr. Frederic Chain, Dr. Philine Feulner, Dr. Christophe Eizaguirre, Dr. Martin Kalbe, Dr. Erich Bornberg-Bauer, Dr. Reusch (again), Dr. Manfred Milinski, and Dr. Lenz (again). I would like to thank the International Max Planck Research School for selecting and funding me, along with the Emmy Noether foundation. Of course, I would also like to thank the MPI for having me, and giving me all the resources and opportunities a young researcher could ask for.

I would like to thank my family, especially my parents, Surajit and Sarmistha Sengupta, for raising me nerdy, and my little sister, Rajanya for being tiny and cute. Thank you, Sudipto, for understanding me the way no one else can. Thanks also to Soumashree, for her endless optimism, especially on the days when I had none of my own. Any acknowledgements section would be incomplete without Sreepu, Sam, Sohini, Tinni and Maddy- most people grow apart from their childhood friends, but I've had the privilege of growing with you. Thanks also to my pseudo childhood friend, Achin, or rather Dr. Achintya Prahlad.

I would be remiss not to mention my friends and teachers at the Indian Institute of Technology, Bombay. I'm grateful to Dr. Kiran Kondabagil, for letting me pursue my interest in evolution, and to Dr. Amrutraj Zade for mentoring me, and all of Saathi-IIT Bombay for their acceptance. Also, I'd like to thank Riddhi, Purna, Padmaja, Amita, Disha, and Suman, for helping me through exams, for never having to explain Harry Potter references, for the jokes, and for being my home away from home.

I've been very lucky to have excellent role models, and for this I'd like to thank Dr. Rohini Balakrishnan, Dr. Swati Patankar and Dr. Rama Govindarajan. Meeting Dr. Karthik "Bittu" Kondaiah when I did also had a huge role in the direction both my career and my personal convictions took, and I would like him to know that, now. I'd also like to thank Jayshree for her advice that has stuck with me for 10+ years: "whatever you do, make sure it's foolproof." Well, I try.

The best part of Plön is friends, and I don't even know where to start, but I'm going to try. First, the 2Cellos Appreciation Society, Neva and Federica: thanks for all the good times, but also for checking up on me during the bad times, and for the fact that our love of 2Cellos always brings us together. Fede, from the moment we said the same (unpopular) opinion, I knew it was going to be something special. Thanks also to Dominik, who figured out early on the best way to get in my good books, which is feeding me. Luka, Natasha and Dushka, thanks for all the Indo-Balkan cultural exchange! Thanks also to Ana Banana for being the life of every gathering, Maria, for our musical exchanges and the fish, to Lole for being the best Spanish teacher, to Anuradha, Devika, Samer, Noemie, Ela, Filipa, Gillian, Juan, Karen, Goekce, Alice, and Ana: I'm so glad you exist! However, there's someone who I'd like to thank over and above the others: Ezgi, for way too many things to count, but mostly for being my "mom friend" and for fulfilling my lifelong dream of going to Istanbul.

Of course, I'd like to thank the people I saw every day, current and past members of the Evolutionary Immunogenomics Group: Artemis, Alejandro, Onur, Clinton, Reem, Leo, Wei, Pryce, Jamie, Arka, Marc, and Jatin.

Finally, I'd like to thank my doctor, Dr. Antje Denmert and the entire setup at the ZIP, which helped me when I needed it the most.

Contributions in Thesis

Thesis title: Variation in the MHC Class-II Region of the three-spined Stickleback: From Genomes to Populations
Student's name: Malavi Sengupta

The entire thesis was written solely by the candidate Malavi Sengupta. Below is a detailed listing of individual contributions to the specific topics covered in this thesis.

Topic I - MHC Class-II In the context of the whole genome
TLL conceived the study. **MS** collected the data and performed the bioinformatic analysis. **MS** analyzed the data with input from TLL.

Topic II - The Stickleback MHC Class II Map
TLL, TBHR, and MM conceived the study. CE, FC, IESP, MK, **MS** and TLL performed the lab work. **MS** analyzed the data with input from MP, TLL and CE**.**

Topic III - Understanding Copy Number Variation and its Causes
MS and TLL conceived the study. **MS** collected and analyzed the data with input from TLL. ACN contributed to analysis of repetitive elements. **MS** and TLL interpreted the results.

Topic IV - Population Genomics of the MHC Class-II Region
MS designed the study. FC, MP, and PGDF and CE generated the data. DMJT performed the read mapping and SNP calling. **MS** performed the population genomics analysis, generated the results, and interpreted them with input from TLL.

Resulting manuscript for publication:

MS and TLL are writing a manuscript incorporating most of the above data and results, with input from ACN, DMJT CE, EBB, FC, PGDF, MP, MST, MM, and TBHR.

Author Names in Alphabetical order

ACN: Alejandro Ceron-Noriega, CE: Christophe Eizaguirre, DMJT: Doko-Miles Jackson Thorburn, EBB: Erich Bornberg-Bauer, FC: Frederic Chain, IESP: Irene E. Samonte-Padilla, PGDF: Philine G. D. Feulner, MP: Mahesh Panchal, MK: Martin Kalbe, MMnfred Milinski, **MS: Malavi Sengupta,** MST: Monica Stoll, TBHR: Thorsten B. H. Reusch, TLL: Tobias L. Lenz.

Plön, 12.11.2019

Declaration

Hereby I declare that:

i. apart from my supervisor's guidance, the content and design of this dissertation is the product of my own work. The coauthors' contributions to specific chapters are listed in the Author Contributions section of this thesis;

ii. this thesis has not already been submitted either partially or wholly as part of a doctoral degree to another examination body

iii. the preparation of the thesis has been subjected to the Rules of Good Scientific Practice of the German Research Foundation.

iv. I have not had any academic degree withdrawn

Plön, November 14, 2019.

Malavi Sengupta