MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 3833–3849 doi: 10.1111/mec.12358

Nonparallelism in MHCIIβ diversity accompanies nonparallelism in pathogen infection of lake whitefish (*Coregonus clupeaformis*) species pairs as revealed by next-generation sequencing

SCOTT A. PAVEY,* MAELLE SEVELLEC,* WILLIAM ADAM,* ERIC NORMANDEAU,* FABIEN C. LAMAZE,* PIERRE-ALEXANDRE GAGNAIRE,*† MARIE FILTEAU,* FRANCOIS OLIVIER HEBERT,* HALIM MAAROUFI*‡ and LOUIS BERNATCHEZ*

**Departement de Biologie, Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Pavillon Charles-Eugene-Marchand, Québec, Québec, Canada G1V 0A6, †Institut des Sciences de l'Evolution – Montpellier (ISEM), Universite Montpellier II, Place Eugene Bataillon, 34095 Montpellier Cedex, France, ‡Plate-forme de Bio-informatique and Institut de Biologie Intégrative et des Systèmes, Université Laval, Pavillon Charles-Eugene-Marchand, Québec, Québec, Canada G1V 0A6*

Abstract

Major histocompatibility (MHC) immune system genes may evolve in response to pathogens in the environment. Because they also may affect mate choice, they are candidates for having great importance in ecological speciation. Here, we use next-generation sequencing to test the general hypothesis of parallelism in patterns of MHCIIβ diversity and bacterial infections among five dwarf and normal whitefish sympatric pairs. A second objective was to assess the functional relationships between specific MHCIIβ alleles and pathogens in natural conditions. Each individual had between one and four alleles, indicating two paralogous loci. In Cliff Lake, the dwarf ecotype was monomorphic for the most common allele. In Webster Lake, the skew in the allelic distribution was towards the same allele but in the normal ecotype, underscoring the nonparallel divergence among lakes. Our signal of balancing selection matched putative peptide binding region residues in some cases, but not in others, supporting other recent findings of substantial functional differences in fish MHCIIβ compared with mammals. Individuals with fewer alleles were less likely to be infected; thus, we found no evidence for the heterozygote advantage hypothesis. MHCIIβ alleles and pathogenic bacteria formed distinct clusters in multivariate analyses, and clusters of certain alleles were associated with clusters of pathogens, or sometimes the absence of pathogens, indicating functional relationships at the individual level. Given that patterns of MHCIIβ and bacteria were nonparallel among dwarf and normal whitefish pairs, we conclude that pathogens driving MHCIIβ evolution did not play a direct role in their parallel phenotypic evolution.

Keywords: ecological divergence, ecological genetics, host–pathogen dynamics, nonparallel evolution, peptide binding region, speciation

Received 30 January 2013; revision received 3 April 2013; accepted 8 April 2013

Introduction

Ecological speciation, or the evolution of reproductive isolation resulting from divergent natural selection (Schluter 2000; Nosil 2012), is currently a major thrust in the study of ecology and evolution. Parallel patterns of divergence in replicated systems are particularly powerful demonstrations of ecological speciation because nonecological drivers of speciation like drift are not expected to result in these parallel patterns (Johannesson 2001). Major histocompatibility complex (MHC) genes have the functional potential to be central for speciation because they interact with ecology as well as

Correspondence: Scott A. Pavey, Fax: 418 656 7176;
E-mail: scott.pavey.1@ulaval.ca

reproduction (Eizaguirre *et al.* 2009). MHC genes are responsible for the identification and presentation of intra (class I)- and extra-cellular (class II) pathogens to initiate an adaptive immune response. They exhibit amazing diversity within species and the allele frequencies of these immune function genes may covary with pathogens in the environment (e.g. Paterson *et al.* 1998) which in turn could influence mate choice (Landry *et al.* 2001) and may lead to or stabilize reproductive isolation (e.g. Milinski *et al.* 2005; Blais *et al.* 2007). This gives MHC genes a potentially pivotal role in ecological speciation, as selection/recombination antagonism disappears if nonrandom mating is directly based on a trait under divergent selection (Gavrilets 2004).

Several, nonexclusive evolutionary processes influence MHC genes as a result of natural selection at a variety of spatial and temporal scales in vertebrates (Bernatchez & Landry 2003; Eizaguirre & Lenz 2010). At the population level, directional selection may result from differences in local pathogen communities. This might decrease MHC diversity within populations as certain alleles are favoured for the local pathogens. However, at a larger spatial scale of the metapopulation, this might increase diversity (Hill *et al.* 1991; Landry & Bernatchez 2001; Miller *et al.* 2001). In contrast, balancing selection may result from mechanisms including variable selection in time and space (Hedrick 2002) as well as negative frequency-dependent selection (Clarke & Kirby 1966; Spurgin & Richardson 2010). For example, if pathogens quickly evolve to overcome the common alleles, rare alleles should be favoured resulting in negative frequency-dependent selection (Clarke & Kirby 1966). These later two processes may have the opposite effect of directional selection, increasing diversity within populations and decreasing divergence in MHC among populations.

At the level of the individual, the extent of heterozygosity may also be influenced by selection (Piertney & Oliver 2006), and this may have emergent properties at the population level. Heterozygote advantage (Doherty & Zinkernagel 1975) can favour increased numbers of alleles and allelic diversity within an individual (Hughes & Nei 1988; Arkush *et al.* 2002; Evans & Neff 2009) and result in balancing selection for the population. This becomes quite complex, because gene duplication is common in MHC, and duplicated loci may share alleles (Chaves *et al.* 2010), resulting in individuals having more than two functional alleles. Given the fact that there is great potential for MHC alleles to duplicate, each individual possesses only a fraction of the population diversity, suggesting that there may be a disadvantage for carrying too many alleles or alleles that are too diverse (Wegner *et al.* 2003a). When an organism initiates a response of the adaptive immune system, there are complex steps involving the maturation of T-cell receptors that first increases diversity, then decreases diversity. The result is lymphocytes that can recognize a diversity of foreign bodies without being too sensitive as to falsely identify self as a foreign body (Allen 1994; Germain 1994). Due to these processes, increased intra-individual diversity of MHC alleles coded in the DNA may not always result in more diversity to present foreign peptides. Also, if an individual's lymphocyte diversity is too high, this may result in autoimmune problems (Nowak *et al.* 1992; Woelfing *et al.* 2009). All of these processes, when taken together, result in an intra-individual number and diversity of alleles that is lower than the intra-population level, and which may represent an optimum (Woelfing *et al.* 2009). While the theoretical framework of this optimum is still debated, the general pattern was confirmed by empirical studies in different taxa (e.g. Wegner *et al.* 2003b; Kalbe *et al.* 2009; Kloch *et al.* 2010; Evans *et al.* 2012a).

Empirical studies support the functional importance of MHC variability in pathogen resistance. Several studies have found temporal variation in MHC allele frequencies within cohorts or between successive cohorts and inferred this to be in response to selection (Westerdahl *et al.* 2004; Charbonnel & Pemberton 2005; Fraser *et al.* 2010; Evans *et al.* 2012a). Matthews *et al.* (2010) found parallel patterns of fewer alleles per individual in limnetic vs. benthic threespine stickleback (*Gasterosteus aculeatus*) and attributed this to environment-specific adaptation to pathogen loads. Evans *et al.* (2010a) found evidence for local adaptation in Chinook salmon (*Oncorhynchus tshawytscha)*, although the above studies did not measure pathogens explicitly. Of studies comparing MHCIIβ variation with explicit measurements of pathogens, there are results that suggest directional selection due to local adaptation (Dionne *et al.* 2007; Eizaguirre *et al.* 2012), and balancing selection due to change in selection through time and space (Dionne *et al.* 2009; Fraser *et al.* 2010), heterozygote advantage (Croisetière *et al.* 2008; Evans & Neff 2009) and negative frequency-dependent selection (Pitcher & Neff 2006; Lenz *et al.* 2009).

Lake whitefish (*Coregonus clupeaformis*) provide striking examples of phenotypic parallel divergence (for review see Bernatchez *et al.* 2010). Many traits display parallel differences among different limnetic (dwarf) and benthic (normal) whitefish sympatric pairs ranging from body size, growth rate, age of maturity, trophic morphology, as well as physiological traits associated with oxygen transport (Evans & Bernatchez 2012; Evans *et al.* 2012b). In the present study, we used next-generation sequencing to characterize both the MHCIIβ diversity and bacterial pathogens of lake whitefish to test the general hypothesis of parallelism in patterns of MHCIIβ

diversity and bacterial infections among five dwarf and normal whitefish sympatric pairs. Thus, if the pathogenic environment in limnetic and benthic environments is similar among lakes, selection acting on MHCIIβ may be parallel as well, and this may be playing a central role in the adaptive divergence of these species pairs. Also, we assessed the functional relationships between specific MHCIIβ alleles and pathogens. Unlike other studied salmonids, for which a single locus has been reported, three MHCIIβ loci have been found in European whitefish (*Coregonus lavaretus*; Binz *et al.* 2001), the sister taxon of lake whitefish, indicating the potential opportunity of a MHCIIβ multilocus system in our study. We screened lake whitefish populations in five study lakes of southern Quebec, Canada, and northern Maine, USA, and characterized both the MHCIIβ diversity within and among individuals and the bacterial community in the kidney tissue. We examined these patterns for parallel divergence in MHCIIβ allele frequency between dwarf and normal whitefish and parallel divergence in the bacterial communities infecting the two species as possible drivers of MHC evolution. Finally, we identified specific associations of alleles with both susceptibility and resistance to specific pathogens.

Methods

Field sampling

Fish in five different lakes (Cliff, Webster, Indian, East, and Témiscouata) of the St. John River drainage were each sampled during the warmest months of the year between June and August 2010. Gillnets of mesh size ranging between 0.5- and 1-inch mesh were set and checked regularly. Each fish was euthanized if necessary by means of cervical vertebral breakage, and the body cavity was carefully opened as to not severe the digestive tract. All protocols used when handling the fish were approved by the Animal Care and Use Committee of Université Laval. The loose organs were removed from the abdomen, and the kidney tissue was sampled with a separate set of instruments that were cleaned thoroughly between each fish with 10% bleach solution and further sterilized using a blow torch. The kidney tissue was placed in a sterile cyrotube and immediately frozen on liquid N_2. A fin clip was collected from the same individual for genomic DNA extraction. All samples were brought back from the field, and the kidney samples were transferred to a −80 °C freezer.

Pyrosequencing MHC

Total DNA was extracted from about 20 mg fin tissue using the salt-extraction protocol from Aljanabi & Martinez (1997). The amplicon library preparation followed the method for the Roche© GS FLX titanium series 454 sequencing as recommended by the manufacturer. This protocol involves four different steps: amplicon preparation, purification, library construction and 454 sequencing. Each genomic DNA sample (50–100 ng) was used to amplify the entire exon 2 252-bp region in MHCIIβ (Pavey *et al.* 2011) using Platinum® Taq DNA polymerase high fidelity (Invitrogen, Carlsbad, CA, USA). All PCRs were performed in a final volume of 50 μL with the following protocol, using each primer at a final concentration of 0.4 μM. Primers used for PCR amplifications were described by Pavey *et al.* (2011) and have been shown to efficiently amplify the exon 2 of the MHCIIβ locus of numerous salmonidae. A total of 151 different primers were designed, and the common amplicon-specific forward primer sequence SaCo_F was appended to the 3′ end of 151 unique MID sequence tags which were appended to the universal A primer sequence (Table S1, Supporting information). The reverse primer sequence SaCo_R was appended to the universal B primer sequence (Table S1, Supporting information). PCR conditions were preceded by a 2 min of an initial denaturation step, followed by 35 cycles with a denaturation step at 94 °C for 15 s, an annealing step at 55 °C for 45 s and an extension step at 68 °C for 1 min. The final extension step was performed at 68 °C for 8 min. Following amplification, samples were visualized with a 1% agarose gel to verify tight, strong bands at the expected size. The products were purified using AMPure© beads (Beckman Coulter Genomics, Danvers, MA, USA) with a 96-well plate. The manufacturer's protocols were followed except that the entire remaining reaction (45 μL) was used, and we did not remove the 96-well plate from the magnetic ring during the 70% EtOH wash steps. Samples were quantified with Picogreen© (Invitrogen) reagent on a Fluoroskan Ascent FL (Thermo Labsystems, Helsinki, Finland). Samples were combined in an equimolar fashion for a final DNA concentration between 30 and 50 ng/μL in three different libraries each containing 151 samples. Then, libraries were sent to the Plateforme d'Analyses Biomoléculaires (Institut de Biologie Intégrative et des Systèmes, Université Laval) to perform the pyrosequencing with a 454 GS-FLX DNA Sequencer with the Titanium Chemistry (Roche, Penzberg, Germany) using the procedure described by the manufacturer. Each library was pyrosequenced on 1/8 of a plate.

Haplotype inference and genotyping from high-throughput amplicon sequencing data

We developed a bioinformatic method for haplotype inference and genotyping with fusion primers. An

analysis pipeline, including an R script, which uses an iterative procedure in three successive steps, was developed. The first step generates putative allele sequences for each individual. The second step combines and strengthens these into the global alleles for all individuals. Then, the third step genotypes each individual. The input files are from the 454 output with the adaptors removed and split into one file per individual based on the barcode. Each file then goes through a iterative cleaning algorithm and alignment with MUSCLE (Edgar 2004). The pipeline as well as the current documentation is available at this link: github.com/enormandeau/ngs_genotyping.

The first step of haplotype detection begins with a hierarchical clustering analysis based on the pairwise distance matrix calculated from the aligned sequence reads of a given individual. The dendrogram produced by hierarchical clustering is then split at its deepest node to generate two clusters. If the proportion of the total reads contained in a cluster is below a cluster size threshold value, this cluster is removed from the analysis as it probably contains only low-frequency sequencing errors. This procedure is reiterated on each remaining cluster, while its deepest branch length is higher than an internal branch length threshold value, which reflects the accepted level of dissimilarity between a real allele and related reads with sequencing errors. In the present study, we set the minimal proportion threshold to 0.04, so after splitting at the deepest node, only clusters with more than 4% of the original total sequences for that individual were retained. The internal branch length threshold value was set to 0.20, so when the distance between the two clusters diminished to this value or below, no further iterations of splitting were performed. The result of this step is a list of putative consensus alleles for each individual. Most pyrosequencing errors are filtered out during the process, but chimeric sequences may still remain.

Before the second step, all individual consensus sequences obtained from the first step were merged together and realigned with MUSCLE. Twenty Sanger sequences of clone inserts from a pilot project of the Cliff populations were included in this alignment. The Sanger sequences are assumed to be free of the types of errors commonly found in 454 sequencing, and gaps in the resulting alignment were removed from all sequences. Phase two the script was then performed with a minimum internal branch length of 0.08 (which enabled us to distinguish alleles that differed by a single SNP), and a minimum number of two sequences to define a cluster. The result of this step represents the putative global consensus alleles for all populations. However, common chimeric sequences that were present in two or more individuals may still remain. True alleles are differentiable from artifactual alleles in that even rare alleles present in only a few individuals should represent a substantial proportion of the sequences in these individuals (Babik *et al.* 2009).

For the third step of the pipeline, the input FASTA files (from step 1) were BLASTed to the global consensus alleles (output of step 2), resulting in a count for each individual of the number of reads that blasted to each global consensus allele. A histogram was then produced for each allele, and individuals that received a blast hit to that allele were ordered in the histogram based on number of sequences blasting to that allele. Examining each allele in this fashion, we noticed there was a natural break in this relationship (as in the study Babik *et al.* 2009) and a minimal threshold of number of sequences per individual was established to call the allele for that individual. For most of the alleles, this threshold was 50 sequences, but for alleles 1, 4, 7 and 13 (see Results), the threshold for the individual assignment of an allele was set to 70, 20, 25 and 38 respectively, as the natural break in the plots occurred at these numbers of sequences. As the mean number of sequences per individual was 515.4 (range 212–1000), a putative rare allele is likely to be an artefact if it does not represent a substantial number of sequences in any individuals where it is found. Thus, we went through all alleles in all individuals a second time, and if an allele did not blast to more than 70 sequences for any single individual, it was excluded and removed from all individuals as a likely artifactual allele.

MHC sequence analysis

As MHC alleles tend to exhibit recombination and gene conversion (Reusch *et al.* 2004) in addition to point mutations, gene trees do not accurately represent historical processes (Huson & Kloepper 2005). A phylogenetic network illustrates many trees simultaneously, enabling a more accurate depiction of diversity generated by gene conversion as well as recombination among alleles. To visualize relationships among alleles, a haplotype network was constructed with the program SplitsTree4 (Huson & Kloepper 2005) using the NeighborNet algorithm and the Jukes–Cantor (JC) model of mutation. An additional haplotype network on the translated sequences from the present study combined with the 20 alleles previously identified in European whitefish (Binz *et al.* 2001) was created with the NeighborNet algorithm.

For the set of consensus alleles of all populations, D_N/D_S was calculated in MEGA 5 (Tamura *et al.* 2011) separately for the putative peptide binding region (PBR) as well as for the putative non-PBR residues as defined by alignment with the Human leucocyte

antigen class II exon 2 (Brown *et al.* 1993). The Nei–Gojobori's modified method was performed. The JC model of mutation was assumed with a calculated gamma distribution of shape $\alpha = 7.95$ and transition/transversion bias ratio of 0.5.

Protein 3D structure

The most common allele (allele 1; see Results) was chosen to predict the three dimensional protein structure using the modelling software MODELLER (Eswar *et al.* 2008) from the crystallographic structures 1uvq_B and 3c5j_B, of human HLA class II histocompatibility antigen (Brown *et al.* 1993). The model quality was assessed by analysis of Ramachandran plot through PROCHECK (Laskowski *et al.* 1993).

Population allele frequencies

Pairwise F_{ST} values between dwarf and normal whitefish were calculated in two different ways. First, the program ARLEQUIN 3.5.1.2 was used considering each base in each allele as a locus and again assuming the JC model of mutation. The haplotype data were coded as population-level allele frequencies. Second, because the gametic phase of this multilocus gene was unknown, the program SPAGeDi (Hardy & Vekemans 2002) was used to generate F_{ST} values as this program can treat alleles assigned to individuals with an unknown number of loci. We computed a measure of allelic diversity (π) for each individual and each population using the 'within group mean distance' in MEGA, again assuming the JC model of mutation with gamma set to 7.9. This represents the average polymorphisms per site within individual or population. MHCIIβ F_{ST} values were compared with genome-wide F_{ST} values generated from three previous studies of the same populations using three types of markers (microsatellites, AFLP, SNP; Lu & Bernatchez 1999; Campbell & Bernatchez 2004; Renaut *et al.* 2011).

Residue-specific selection

OmegaMap (Wilson & McVean 2006) was used to perform a Bayesian determination of D_N/D_S ratios in our sampled populations. OmegaMap generates residue-specific posterior probabilities of both D_N/D_S ratio (ω) and recombination rate (ρ). The omega_model was set to independent, which allows each residue to vary independently in ω. The rho_model was set to variable with the rho_block set to 3. To conserve computational resources, we made two independent random subsamples of 200 alleles from all populations. Five runs for each subsample were performed with >1 500 000 steps. The first 100 000 were discarded as the burn-in period. Residues with >95% posterior probability in both independent subsample run sets were considered to be under balancing selection. A visual representation of the protein model displaying the concordance of putative PBR and non-PBR residues with balancing selection found in this study was created with PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC).

Characterizing bacterial pathogens

We characterized a subset of 234 individuals with putative pathogenic genera from the data set described by Boutin *et al.* (2012). Putative pathogenic genera were retained based on genera described by Austin & Austin (2007) which is a recent account of pathogens specific to fish. The methods are fully described elsewhere (Boutin *et al.* 2012). Briefly, DNA was extracted from each kidney sample under a laminar flow hood using aseptic techniques. A double-nested PCR was performed with three sets of 16S ribosomal RNA primers. If a band was present after the third PCR (indicating bacteria were present in kidney tissue; potentially infected individual), the third PCR was performed again with sequence tag–labelled primers. The resulting libraries were then sequenced on 3/4 of a 454 pyrosequencing run. Sequences were cleaned, trimmed and filtered *in silico* with the application MOTHUR (Schloss *et al.* 2009). Sequences were classified to the level of genus, and only genera previously described in the literature as containing pathogenic species for fish were considered for this study, as we wanted to reduce the potential for false positives. Pathogenic genera only represented in a single individual were excluded, again to reduce the potential for false positives.

MHC pathogen associations

To determine whether certain alleles were associated with pathogenic bacteria found within the kidney tissue, we used both univariate and multivariate approaches, as MHC alleles have both general and specific susceptibility and resistance relationships with pathogens (Croisetière *et al.* 2008; Dionne *et al.* 2009; Fraser & Neff 2010). First, mixed-model logistic regressions were performed for each allele separately to determine whether certain alleles were more or less likely to be found in individuals where pathogenic bacteria were also found. The dependent variable was infection status and with ecotype as a fixed factor and lake as random factor. Second, to determine whether the number of alleles of a single individual, or the allelic diversity (π) of that individual, was predictive of the presence of

pathogenic bacteria, two separate mixed-model logistic regressions were performed. In the first, the fixed factors were number of alleles, ecotype and their interaction, with lake as a random factor. In the second, the fixed factors were π, ecotype and their interaction, and the random factor was lake. For all of the above mixed-model logistic regressions, we used the 'lmer' function in R (2.15.0). Third, to determine whether individual allelic diversity was predictive of individual bacterial diversity, we tested a third mixed model, as above, except the Shannon index of diversity was the dependent variable instead of infection status, with the same independent variables as the second model (above). For this last model, we used the R package 'lme'.

For the multivariate approach, a co-inertial analysis (COIA) was performed (Tollenaere *et al.* 2008; Evans & Neff 2009; Evans *et al.* 2010a). First, we produced two separate principal component analyses (PCAs) based on presence/absence binary encoded matrices of both pathogens and MHC alleles. In the COIA, the separate PCAs were rotated to a comparable alignment and normalized co-inertial loadings were obtained for each allele and pathogenic genus. For the COIA, we used the R package 'ade4TkGUI'.

To test for parallelism between dwarf and normal across lakes, we performed a separate PCA of the presence/absence of both the MHC and pathogen matrices using the 'prcomp' command in R. Then, we performed two type II MANOVAs with the matrices of principle components as the dependent variables and lake, ecotype and their interaction as the independent variables. We used the command 'MANOVO' from the 'car' library, and we used the *F* statistics estimated by the (more conservative) Pillai method. A nonsignificant interaction term coupled with a significant ecotype term would indicate a parallel difference.

Results

MHC haplotypes

A total of 367 whitefish individuals were assigned MHCIIβ alleles for a 249-bp fragment, as we were not able to consistently resolve the first three base pairs (Table 1). For quality control, 27 individuals were run twice, and in all cases, we found the same alleles in replicated pairs. After filtering out 12 likely artifactual alleles, we found a total of 17 alleles in the five study lakes (Figs 1 and S1, Supporting information). These 17 alleles were characterized by high degrees of nucleotide (80/249 sites segregating, $\pi = 0.095$) and amino acid (42/83 sites segregating, $\pi = 0.209$ average substitutions per site) polymorphism. Allele 1 was the most common allele and the only one present in both ecotypes of all five lakes (Fig. 1). Alleles 1–5 and 15 had an insertion of 3 bp compared with the other 11 alleles, corresponding to an additional amino acid residue (number 59 of our 83). The haplotype network depicted two groups of alleles: those containing the 3 bp insertion on the right and those without the insertion on the left (Fig. 2). Allele 10 had an 11 bp deletion resulting in a frame shift mutation.

Table 1 Summary table of the sample size for each ecotype genotyped at the MHCIIβ gene per lake, the number of MHCIIβ alleles per ecotype within each lake, the number of fish used for the characterization of the bacterial pathogens and the number of pathogenic bacteria genera found within each ecotype within each lake

Lake	Ecotype	Sample size MHC	Number of alleles in population (mode allele number/individual)	Allelic diversity	Sample size pathogen screen	Number of pathogenic genera
Cliff	Dwarf	29	1 (1)	0	28	7; A, B, D, F, H, J, L
	Normal	26	5 (3)	0.068	24	9; A, B, C, D, E, G, H, J, K
Webster	Dwarf	44	12 (2)	0.092	24	8; A, C, D, E, F, G, H, J
	Normal	50	11 (1)	0.057	25	8; A, B, D, E, F, G, H, J
Indian	Dwarf	33	12 (3)	0.097	29	6; A, B, D, E, H, M
	Normal	29	8 (2)	0.106	21	11; A, B, C, D, E, F, G, H, I, J, L
East	Dwarf	34	9 (3)	0.091	20	8; A, D, E, F, G, H, J, K
	Normal	33	6 (3)	0.098	25	8; A, C, D, E, F, G, H, I
Témiscouata	Dwarf	56	9 (3)	0.088	26	5; A, D, E, H, J
	Normal	33	10 (3)	0.089	13	1; J
All groups		367	17 (3)	0.092	235	13

MHC, major histocompatibility.
'A' for *Acinetobacteria*, 'B' for *Aeromonas*, 'C' for *Chryseobacterium*, 'D' for *Clostridium*, 'E' for *Corynebacterium*, 'F' for *Flavobacterium*, 'G' for *Micrococcus*, 'H' for *Pseudomonas*, 'I' for *Straphylococcus*, 'J' for *Shewanella* 'K' for *Oxalobacter*, 'L' for *Janthinobacterium*, and 'M' for *Citrobacter*.

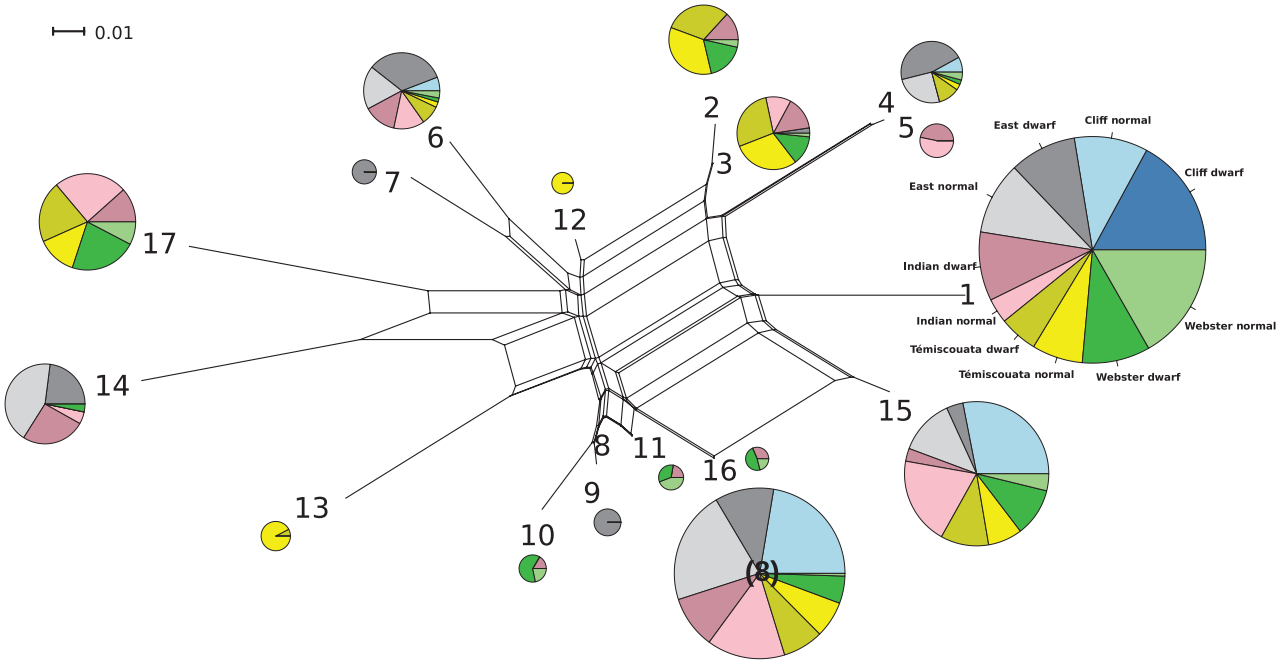

Fig. 1 NeighborNet haplotype network of the 17 alleles observed in lake whitefish. The size of each circle represented the global log-transformed frequency of each allele, whereas the individual slices of each pie represents how that allele is distributed in the populations. The scale is Jukes–Cantor maximum likelihood (ML) distance.

The alleles of lake whitefish and European whitefish did not segregate by species and were all part of the same network (Fig. S2, Supporting information), as both species had alleles with and without the additional amino acid residue. However, no identical alleles were shared between the two species. The number of alleles per population varied greatly among lakes and between ecotypes (Table 1). The mode of the number of alleles per individual was three with a maximum of four, most likely indicating the existence of two loci (Fig. S3, Supporting information). D_N/D_S ratios were significantly >1 for the residues putatively identified in the PBR region ($D_N/D_S = 3.162$; $P = 0.002$). The non-PBR residues also had D_N/D_S ratios > 1, but this was not significant (D_N/D_S 1.642; $P = 0.102$).

Allele frequencies

Frequencies of MHCIIβ alleles were significantly different between dwarf and normal whitefish in four of five lakes, and the rank order of F_{ST} was the same among populations in MHC compared with the neutral markers (Table 2). Dwarf and normal whitefish of Cliff Lake were substantially more differentiated at MHC than neutral markers mainly due to fixation of allele 1 in Cliff dwarf, whereas the patterns observed in Webster, Indian and East were similar to the signal found with neutral markers. In contrast, in Témiscouata, the ecotypes were not significantly differentiated in MHC, and the F_{ST} value was lower than those observed at other markers. Population allelic diversity (π) ranged from 0 in Cliff dwarf, which was monomorphic for allele 1, to 0.109 in Indian normal, which had eight alleles (Fig. 2, Table 1).

Signature of balancing selection

The OmegaMap analysis revealed that 23 of 83 residues are under balancing selection (>95% posterior probability agreement in both independent runs). These positions lined-up with only 12 of the 22 putative PBR residues as defined by the alignment to the human HLA-DR1 (Fig. S1, Supporting information; Brown *et al.* 1993). In addition, 11 of the putative non-PBR residues had a signal of balancing selection. These residues were mapped on the 3D protein model, which was found to be of high quality in PROCHECK; 96.4% (80/83) of all residues were in favoured (98%) regions and 100.0% (85/85) of all residues were in allowed (>99.8%) regions (Fig. S4, Supporting information). Posterior credibility intervals for the recombination parameter in the omegaMap analysis were above 0 for the entire sequence, rising further at the 3′ end of the sequence (Fig. S5, Supporting information), meaning there is a small but credibly greater than zero signal of within locus recombination throughout the entire sequence.

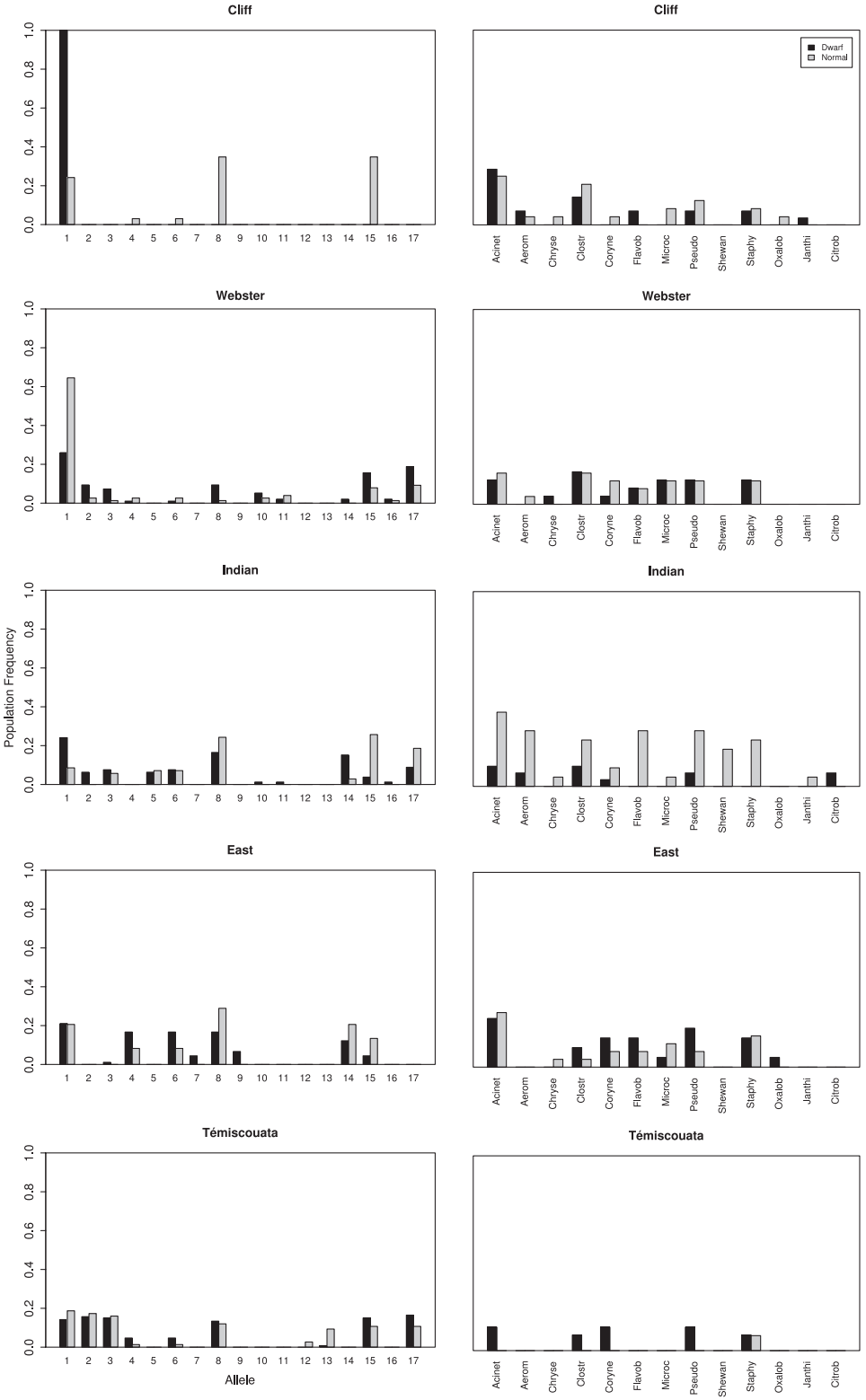

Fig. 2 Left column: Allele frequencies of all 17 major histocompatibility (MHC) alleles in all dwarf and normal ecotypes for all populations. Right column: Pathogen frequencies of all lakes. The 13 genera of pathogenic bacteria are abbreviated by the first six letters of that genus as follows: 'Acinet' for *Acinetobacteria*, 'Aeromo' for *Aeromonas*, 'Chryse' for *Chryseobacterium*, 'Clostr' for *Clostridium*, 'Coyrne' for *Corynebacterium*, 'Flavob' for *Flavobacterium*, 'Microc' for *Micrococcus*, 'Pseudo' for *Pseudomonas*, 'Staphy' for *Straphylococcus*, 'Shewan' for *Shewanella* 'Oxalob' for *Oxalobacter*, 'Janthi' for *Janthinobacterium*, and 'Citrob' for *Citrobacter*.

Table 2 F_{ST} between dwarf and normal ecotypes of lake whitefish of three types of neutral markers from other studies (range represents 95% confidence interval; Lu & Bernatchez 1999; Campbell & Bernatchez 2004; Renaut *et al.* 2011) as well as the two different methods of measuring F_{ST} of MHCIIβ allele frequencies in the present study. Method 1 (meth 1) is calculated from population allele frequencies of the sequences in ARLIQUIN, and method 2 (meth 2) is calculated from the program SPAGEDI which accepted genotypes of unknown gametic phase

	Microsatellites	AFLP	SNP	MHC (meth 1)	MHC (meth 2)
Cliff	0.256 (0.147–0.339)	0.220 (0.194–0.248)	0.278 (0.206–0.329)	0.445	0.455
Webster	0.140 (0.020–0.302)	0.172 (0.151–0.195)	0.110 (0.076–0.145)	0.099	0.105
Indian	0.084 (0.012–0.170)	0.042 (0.033–0.053)	0.060 (0.038–0.082)	0.023	0.051
East	0.058 (0.013–0.110)	0.114 (0.094–0.136)	0.022 (0.0098–0.036)	0.011	0.024
Témiscouata	0.041 (0.006–0.083)	NA	0.010 (0.0028–0.0190)	−0.0005	0.0023

MHC, major histocompatibility.

Pathogenic bacteria

We found the presence of 13 putatively pathogenic bacteria genera (*Acinetobacteria*, *Aeromonas*, *Chryseobacterium*, *Clostridium*, *Corynebacterium*, *Flavobacterium*, *Micrococcus*, *Pseudomonas*, *Straphylococcus*, *Shewanella*, *Oxalobacter*, *Janthinobacterium* and *Citrobacter*) in kidney tissues of 88 of the 234 individuals assayed (37.6% infection rate, Fig. 2, Table 1). We excluded five putatively pathogenic genera that were each present in a single individual because in such a case, it was too likely that the results could be due to sequencing error. All of these had <10 reads. We excluded a single outlier individual that had 630 genera of bacteria. Number of pathogens per lake was not correlated with number of individuals assayed ($P = 0.23$). These genera represent only a portion of the full microbiome data set, which will be presented in details elsewhere (M. Sevellec, S. A. Pavey, S. Boutin, M. Filteau, W. Adam, N. Derome, & L. Bernatchez., unpublished data). *Acinetobacter* was the most frequent genus across all populations (average frequency 20.1%), followed by *Clostridium*, *Pseudomonas* and *Staphylococcus* (12.8%, 12.0% and 10.7%, respectively). The distributions of these genera were not partitioned by dwarf vs. normal lake whitefish in a parallel fashion among lakes by visual inspection Fig. 2.

MHC and pathogen associations

The mixed models for each allele's association with infection status resulted in two significant and one suggestive susceptibility alleles (more likely to be found in an infected individual, allele 5: $P = 0.023$, allele 15: $P = 0.023$ and allele 17: $P = 0.057$). One allele was suggestive but not significant for resistance (less likely to be found in an infected individual, allele 2: $P = 0.096$). The individual allelic diversity was positively associated with the presence of pathogenic bacteria in kidney tissue ($Z = 2.197$, $P = 0.028$; Fig. S6, Supporting information). Number of alleles trended in the same positive relationship (slope = 0.2373), but this was not significant ($Z = 1.596$, $P = 0.111$; Fig. S7, Supporting information). In both of the latter two mixed-model logistic regressions, the interaction terms were not significant and were removed from the models. For the linear mixed model, the interaction term was significant and retained ($t = 2.279$, $P = 0.0236$), but the diversity of alleles was not associated with the Shannon index of pathogenic bacteria ($t = 0.093$, $P = 0.926$).

In the MHC PCA, the first two axes explained 47.1% of the variance (Fig. 3). Axis 1 was structured primarily with alleles 1, 4, 5, 6, 7 and 9. Axis 2 was structured primarily with alleles 2, 3, 8, 13, 14 15 and 16. In the bacterial PCA, the first two axes explained 51.0% of the variance (Fig. 3), Axis 1 was structured primarily by *Acinetobacteria*, *Pseudomonas*, *Flavobacterium* and *Micrococcus*, whereas axis 2 was structured by *Aeromonas*, *Chryseobacterium* and *Citrobacter*.

In the COIA, the global RV value was 0.061 (Fig. 3), indicating that there was little global, overall correlation between the MHC and pathogen matrices. The first two axes explained 60.0% of the variance. Most of the bacterial genera grouped into two orthogonally loading groups, and several alleles were associated with these groups. Specifically, alleles 5–7 were positively associated with a bacterial group including *Corynebacterium*, *Citrobacter*, *Pseudomonas* and *Clostridium*. The second most frequent allele overall, allele 8, was positively associated with *Acinetobacteria*, *Aeromonas*, *Chryseobacterium*, *Flavobacterium* and *Oxalobacter*. Allele 17 was associated with *Staphylococcus*. Alleles 1–3 and 13 were associated with the absence of the detected pathogenic bacteria.

In the MANOVA to test for parallelism of MHC, both independent variables were significant (lake: $F = 9.64$, $P < 0.001$; ecotype: $F = 4.54$, $P < 0.001$) as well as the interaction ($F = 4.55$, $P < 0.001$). In the test for parallelism for pathogens, lake was significant ($F = 1.36$,

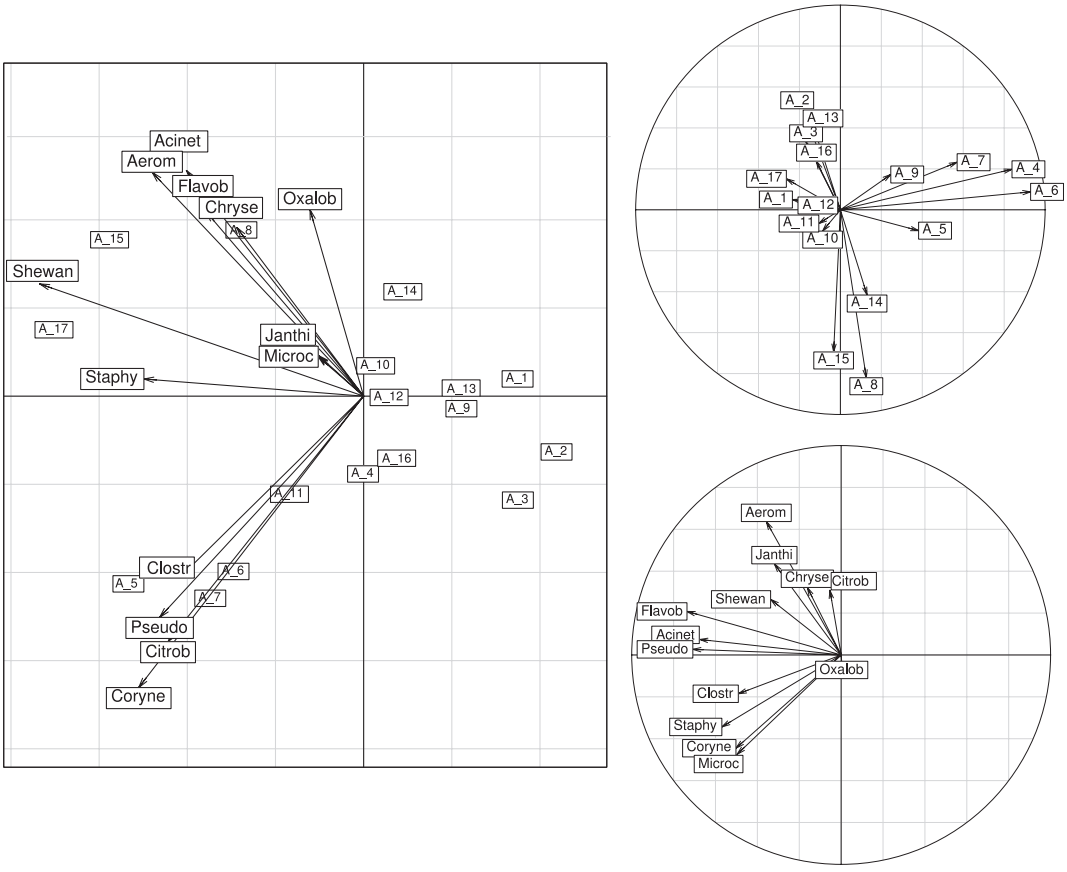

Fig. 3 Loadings of the principal component analyses (PCAs) of 17 alleles (lower right), 13 genera of pathogenic bacteria (upper right), and the co-inertial analysis (COIA) that combines the two PCAs (left). In the PCAs, the *x*-axis represents PC1 and the *y*-axis represents PC2. In the COIA, the *x*-axis represents co-inertial axis 1 (COA 1; 38.8% of variance explained) and the *y*-axis represents COA 2 (21.1%). The 17 alleles are indicated numerically following 'A_'. The 13 genera of pathogenic bacteria are abbreviated by the first six letters of that genus as listed in the legend of Fig. 2. When a label was on top of another, both were moved slightly so both were visible.

$P = 0.049$), form was not significant ($F = 1.03$, $P = 0.421$), and the interaction was significant ($F = 1.37$, $P < 0.042$). These results indicate that the patterns between dwarf and normal ecotypes were nonparallel among lakes and that for both MHC and pathogens, lake effect was more pronounced than ecotype effect and also that the extent of differences between dwarf and normal varied among lakes. These nonparallel patterns are also visually obvious from the inspection of Fig. 2.

Discussion

In this study, we tested whether dwarf and normal sympatric ecotypes of lake whitefish experienced parallel pathogenic bacterial infections across lakes and whether this leads to parallel patterns of divergence in the MHCIIβ gene. Our results indicate that this is not the case. Although there are certainly overlapping bacterial genera and MHCIIβ alleles between forms and among lakes, the patterns of divergence in both pathogenic bacteria and MHCIIβ alleles were clearly nonparallel among lakes.

MHC alleles

The total number of alleles was comparable to the findings of many other studies of salmonids. Although similar to European whitefish alleles (Binz *et al.* 2001), these 17 alleles had a nucleotide and amino acid diversity substantially more than found in many other salmonids (Binz *et al.* 2001; Landry & Bernatchez 2001; Dionne *et al.* 2007; Evans & Neff 2009). Also similar to discoveries made in other salmonids is the presence of a 3 bp polymorphic indel. This same indel occurs among species of Pacific salmon (*Oncorhynchus* sp.; Miller & Withler 1996) and also within species of Atlantic salmon (*Salmo salar*), European whitefish (*Coregonus lavaretus*) and Chinook salmon (*Oncorhynchus tshawytscha*;

Grimholt *et al.* 1994; Binz *et al.* 2001; Evans *et al.* 2010b). Interestingly, this indel polymorphism cannot be readily associated with allele segregation between different MHCIIβ loci because some of those species, namely Chinook and Atlantic salmon, possess a single locus. Thus, the transpecific maintenance of this indel polymorphism and the complete overlap of the alleles of lake whitefish and European whitefish suggest there is long-term balancing selection maintaining this structural polymorphism and it has functional relevance in host–pathogen dynamics.

In the NeighborNet haplotype network (Fig. 1), the omegaMap results, as well as the visual inspection of the sequences themselves (Fig. S1, Supporting information), it is apparent that many alleles arose via gene conversion and/or recombination between alleles (Spurgin *et al.* 2011). Clearly then, although the haplotypes that include the deletion are on the right of the network and the haplotypes that exhibit the deletion are on the left, it would be inaccurate to consider these as two clades of alleles. The overall D_N/D_S ratios being significantly >1 are consistent with a signal of historical long-term balancing selection, which was especially pronounced in regions orthologous to the human PBR. This is also comparable to other salmonids as well as many other species (Binz *et al.* 2001; Landry & Bernatchez 2001; Evans *et al.* 2010b). In humans, where the PBR sites are known through X-ray crystallography, it has been shown that PBR residues are under positive selection for the maintenance of diversity, whereas non-PBR residues tend to be under negative, purifying selection (Klein *et al.* 1993). The nonconcordance of balancing selection of specific putative PBR residues is notable, as 10 putative PBRs have no signal of balancing selection in our populations. Also, in the 11 putative non-PBR residues with a signal of balancing selection, seven are nonadjacent to a putative PBR residue. Other species of salmonids consistently show a signal of balancing selection in particular at residues 18 and 45 (HLA residues 26 and 53 in Brown *et al.* (1993). Our study supports that, in salmonids, these residues are also likely to be PBR (Aguilar & Garza 2007; Evans *et al.* 2010b). Taken together, these findings support the idea of substantial differences in PBR sites in fish in general and salmonids in particular compared with humans.

Population-specific selection and drift

Two populations of 10 substantially differed from the mode of three alleles per individual. Most notably, Cliff dwarf individuals are all monomorphic for the globally most common allele 1. The other substantially different population is Webster normal (mode = 1 allele per individual). Similar to Cliff dwarf, the Webster normal population has a low allelic diversity of 0.057, the second lowest of all 10 populations. Although 11 alleles are found in Webster normal, the distribution is highly skewed towards allele 1, which represents 65% of all alleles in the population. In our COIA, the most ubiquitous allele (allele 1) was found to be negatively associated with all pathogenic genera of bacteria. This raises the hypothesis that episodic strong outbreaks of bacteria could result in strong directional selection favouring allele 1, independent of the ecotypes illustrated, for instance, in Cliff dwarf and Webster normal whitefish. Dionne *et al.* (2009) found the allele frequency of a susceptibility allele decrease by 5% in a population over a 2-month time period, demonstrating the speed at which pathogen-driven selection can change allele frequencies.

The overall diversity in both populations of Cliff is lower than the other lakes, with only five alleles with both ecotypes combined. While we cannot exclude any role that genetic drift played in these patterns, it is not likely to be the predominant process. First of all, heterozygosity and allelic diversity at other types of markers is similar to the populations in other lakes (Lu & Bernatchez 1999). For instance, the point estimates of heterozygosity observed at microsatellites (Lu & Bernatchez 1999) were comparable in Cliff normal (H_E = 0.050) and Cliff dwarf (H_E = 0.057), to the average of all 10 populations (H_E = 0.057). It is also striking that the same allele (allele 1) predominates in abundance in all lakes, which is not expected under strong random genetic drift effects in each lake. Admittedly, it is unknown whether a bottleneck occurred after colonization of the lake. If such a bottleneck was not strong enough to result in contemporary reduction in neutral genetic diversity, it could still have reduced the diversity in MHCIIβ, as can happen in conjunction with selection (Ejsmond & Radwan 2011; Sutton *et al.* 2011). As stated above, it is likely that directional selection favouring allele 1 resulted in the fixation of this allele in Cliff dwarf. In Cliff normal, episodic periods of strong directional selection sequentially favouring different alleles are most likely to have resulted in the observed reduced diversity. Regardless of the causes, the nonparallel nature of these patterns is striking, as the reduced number of alleles per individual and the increased frequency of allele 1 occurred in different ecotypes in the most divergent lakes.

F_{ST} values for MHCIIβ compared with neutral markers were similar between the ecotypes in East, Indian and Webster. However, they were more divergent in Cliff, whereas less divergent in Témiscouata. In these cases, this pattern also holds true for both microsatellites and SNPs. Thus, the complete spectrum of selection patterns is found among these five lakes with this comparison, resulting in a signal of directional selection

in Cliff, balancing selection in Témiscouata and no departure from neutral expectations in East, Webster and Indian. However, we do not readily take this as evidence that selection is not important in these latter three lakes, as five of the six populations have the highest five MHCIIβ allelic diversities of all 10 study populations and Webster normal is skewed towards allele 1, the same allele that was driven to fixation in Cliff dwarf. A mix of directional and balancing selection could easily produce results that do not depart from neutral expectation (Bernatchez & Landry 2003).

Pathogenic bacteria

We found a greater percentage of individuals with putative pathogenic bacteria (37.6%) than was found in similar studies of natural populations of salmonids, which measured all bacterial genera (12.1% in Atlantic salmon and 20.2% in Chinook salmon; Dionne *et al.* 2009; Evans & Neff 2009). This could be due to the high sensitivity of our double-nested PCR detection technique prior to library preparation (Boutin *et al.* 2012). This technique is likely much more sensitive than culture methods and other PCR-based methods. It is possible that this technique is sensitive enough to detect the presence of pathogens that the immune system is in the process of successfully purging, as there is a delay between initial infection, generation of pathogen-specific antibodies and complete purging of the pathogen from the host (Frank 2002).

Bacterial communities infecting fishes in nature are poorly described. The majority of current knowledge is the result of microbiological characterization following aquaculture disease outbreaks. Fortunately, other salmonids are frequently used in aquaculture; thus, there is a certain amount of information available about pathogens that may potentially affect lake whitefish. For example, there are several well-known fish diseases in the genus *Staphylococcus* (Austin & Austin 2007), including *Staphylococcus warneri*, which was present in kidney tissue of rainbow trout (Gil *et al.* 2000). *Clostridium botulinum* was documented in a fish farm outbreak (Cann & Taylor 1982). Also, *Micrococcus luteus* was found in internal organs including kidney in rainbow trout (Austin & Stobie 1992). *Aeromonas salmonicida* is the causative bacteria for furonculosis that can cause severe mortality in juvenile salmonids. Five (*Acinetobacter*, *Aeromonas*, *Flavobacterium*, *Pseudomonas* and *Clostridium*) of our 13 genera are known to dominate intestinal microbiota in freshwater fish (Gomez & Balcazar 2008). Pathogens may migrate from the GI track to other internal organs by penetrating the commensal bacteria, mucus and gut-associated lymphoid tissue (Perez *et al.* 2010). Four (*Acinetobacter*, *Corynebacterium*, *Pseudomonas* and *Staphylococcus*) of 13 pathogenic bacterial genera were found in common with a study of wild Chinook salmon kidney tissue (Evans & Neff 2009). However, as our technique only identified potential pathogens to the genus level, we do not know which exact species were infecting the lake whitefish. Species within a genus can vary greatly in their pathogenicity so we cannot evaluate the pathogenicity of the genera we identified in this study; however, we chose the genera because they are known to contain species that are pathogenic to fish.

The distribution of pathogenic genera between dwarf and normal was not parallel among the studied lakes. Although we sampled at the warmest time of the year, our study lacks temporal replication; thus, our 'snapshot' of pathogenic bacteria may differ from temporal replicates in different seasons or life stages. It is nevertheless noteworthy because we sampled both forms in each lake at the same time and observed contrasting patterns in the difference between dwarf and normal infection among lakes, suggesting an environment-specific effect. Our study contrasts with documented parallel patterns of internal and external macroparasites in benthic vs. limnetic threespine stickleback (MacColl 2009), as well as macroparasites in lava vs. mud benthic habitats of stickleback (Natsopoulou *et al.* 2012). However, parasites are longer-lived than bacteria, and the single sampling event of macroparasites is more likely to represent the long-term pattern than pathogenic bacteria. Nonetheless, in the present study, the nonparallel patterns of pathogenic infections suggest that the distribution of bacterial niches do not correspond with the benthic and limnetic niches clearly defining the distribution of dwarf and normal whitefish. This underscores the need for whole-lake distribution studies of bacteria in general and pathogenic bacteria in particular (Stendera *et al.* 2012).

MHC and pathogen associations

We found no evidence for heterozygote advantage. In fact, our results rather suggest that there is a homozygote advantage. Indeed, our mixed models revealed a significant positive association with the diversity of alleles and bacterial infection. In other words, individuals with more diverse alleles were more likely to be infected. The number of alleles trended in the same direction, although this was not significant. This is not so surprising, because many other studies have also found no evidence for heterozygote advantage (Fraser & Neff 2010; Froeschke & Sommer 2012). Moreover, this heterozygote disadvantage pattern was also found in a controlled infection study of mice (Ilmonen *et al.* 2007), in transplanted Chinook salmon (Evans *et al.* 2010a) and in wild populations of Atlantic salmon (Dionne

et al. 2009). One possible explanation is a dosage effect whereby one allele provides the biggest advantage in a given environment and further diversity only dilutes this positive effect (Dionne *et al.* 2009). While this is certainly a possible explanation in this study, there are other explanations that pertain to experimental design. First, our pathogenic bacteria detection technique is potentially much more sensitive than previous techniques. Studies of wild individuals using single nested PCR (Dionne *et al.* 2009) and culture techniques (Evans & Neff 2009) may be more likely to detect only the most severe infections. The signal of heterozygote advantage may be most apparent when considering only severe infections or when assessing the level of infection severity (Paterson *et al.* 1998) and may possibly be obscured in our study, with the extreme sensitivity of detection technique. In the present study, our detection technique may be detecting bacteria in the process of a successful immune response because of diverse alleles. Another nonexclusive reason for this pattern is a consequence of sampling adults in wild populations. A large portion of the pathogenic-induced mortality may occur when the fish are very young (de Eyto *et al.* 2007; Evans *et al.* 2010a), and these individuals are not included in the study. Thus, we may be sampling the excellent (no pathogens were detected) and marginal (pathogens present, but fish survived long enough to be sampled by adult size gill nets) genotypes, but not the extremely poor genotypes that result in mortality of young individuals. This is a problem that any study in nature will encounter, and without performing experimental manipulations of pathogen levels, it will thus always be difficult to draw firm conclusions on pathogen–MHC genotype interactions.

Similar to other studies, we found a handful of allele-specific pathogen associations in our univariate test (Croisetière *et al.* 2008; Dionne *et al.* 2009; Fraser & Neff 2010), namely three susceptibility alleles (two significant) and one nonsignificant resistance allele. Alleles associated with specific pathogens may be susceptibility alleles, or they may protect the individual from severe infections, imparting quantitative resistance (Westerdahl *et al.* 2011). These findings are in concordance with the COIA, as these alleles loaded strongest in association with groups of pathogens for the susceptibility alleles, and strongly away from groups of pathogens in the case of the resistance allele. In the multivariate PCAs and COIA, both pathogens and MHC alleles did not exhibit a random distribution of loadings, but rather formed 3–4 groups. Moreover, groups of MHC alleles do not coincide with close allelic similarity. In parallel, groups of pathogens do not correspond to phylogenetic similarity. Specifically, alleles 5–7 are positively associated with *Corynebacterium*, *Citrobacter*, *Pseudomonas* and *Clostridium*, allele 8 was positively associated with *Acinetobacteria*, *Aeromonas*, *Chryseobacterium*, *Flavobacterium* and *Oxalobacter*, and allele 17 was associated with *Staphylococcus*. Alleles 1–3 and 13 were associated with the absence of the detected pathogenic bacteria. To summarize, we found functional association between groups of pathogenic genera of bacteria and certain MHCIIβ alleles at the level of the individual, but these patterns were not parallel at the population level.

Nonparallel patterns of MHC and pathogens

The over-arching finding of this study is that we did not find parallel patterns of divergence of MHCIIβ or pathogens in dwarf and normal lake whitefish species pairs. This study was the first to measure both internal pathogens in recently diverged species pairs and MHCIIβ genetic diversity. There is a striking lack of parallel patterns in alleles per individual, allele frequencies, pathogen frequencies and neutral vs. MHCIIβ diversity between ecotypes among lakes. This contrasts with a similar study of threespine stickleback, which found parallel patterns in both number of alleles per individuals and allele frequencies of MHCIIβ in two replicate lakes (Matthews *et al.* 2010), as well as ecotype-specific external macro parasites in arctic charr in two replicate sampling sites (Frandsen *et al.* 1989). However, another study of threespine stickleback saw a weaker pattern of parallelism with more than two replicates (Natsopoulou *et al.* 2012). In the present study with five replicate lakes, patterns of pathogenic bacteria infecting kidney tissue of lake whitefish are not parallel between dwarf and normal individuals of different lakes, so it is not surprising that MHCIIβ allele frequencies are also not parallel. Infection rates can vary wildly in both space and time (Dionne *et al.* 2009), and depending on the temporal and spatial patterns of pathogen-imposed selection, both balancing and directional selection appear to have resulted in the complex and nonparallel pattern of MHCIIβ allele frequencies in our five study lakes. Our data nevertheless revealed associations of MHC class II genes and pathogens that indicate functional relationships, but this does not overlap with the benthic and limnetic environments and therefore is not driving parallel evolution in dwarf and normal lake whitefish. Thus, we conclude that pathogens driving MHCIIβ evolution did not play a direct role in the parallel phenotypic evolution divergence of dwarf and normal whitefish.

Acknowledgements

We wish to thank S. Boutin for help with the microbial identification and analysis. We thank Dr. Brian Boyle at the Plateforme d'Analyses Génomiques de l'Université Laval for

valuable advice on library preparation and 454 sequencing. We thank M. Evans for field assistance and valuable comments on the manuscript. We are also grateful to AE Dr. Jon Slate and three anonymous referees for their insightful comments and suggestions. We thank Guillaume Côté for field assistance. SAP was supported through a fellowship from Fonds de la recherche en santé du Québec. FL was supported from the Collaborative Research and Training Experience (CREATE) program. MF was supported by Quebéc Ocean. LB research programme is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Aguilar A, Garza JC (2007) Patterns of historical balancing selection on the salmonid major histocompatibility complex class II beta gene. *Journal of Molecular Evolution*, **65**, 34–43.

Aljanabi SM, Martinez I (1997) Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Research*, **25**, 4692–4693.

Allen PM (1994) Peptides in positive and negative selection: a delicate balance. *Cell*, **76**, 593–596.

Arkush KD, Giese AR, Mendonca HL *et al.* (2002) Resistance to three pathogens in the endangered winter-run chinook salmon (*Oncorhynchus tshawytscha*): effects of inbreeding and major histocompatibility complex genotypes. *Canadian Journal of Fisheries and Aquatic Sciences*, **59**, 966–975.

Austin B, Austin DA (2007) *Bacterial Fish Pathogens: Diseases of Farmed and Wild Fish*. Praxis Publishing Ltd, Chichester.

Austin B, Stobie M (1992) Recovery of *Micrococcus luteus* and presumptive planococcus from moribund fish during an outbreak of rainbow trout, *Oncorhynchus mykiss* (Walbaum), fry syndrome in England. *Journal of Fish Diseases*, **15**, 203–206.

Babik W, Taberlet P, Ejsmond MJ, Radwan J (2009) New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. *Molecular Ecology Resources*, **9**, 713–719.

Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *Journal of Evolutionary Biology*, **16**, 363–377.

Bernatchez L, Renaut S, Whiteley AR *et al.* (2010) On the origin of species: insights from the ecological genomics of whitefish. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **365**, 1783–1800.

Binz T, Largiader C, Muller R, Wedekind C (2001) Sequence diversity of Mhc genes in lake whitefish. *Journal of Fish Biology*, **58**, 359–373.

Blais J, Rico C, van Oosterhout C *et al.* (2007) MHC adaptive divergence between closely related and sympatric african cichlids. *PLoS ONE*, **2**, e734.

Boutin S, Sevellec M, Pavey SA, Bernatchez L, Derome N (2012) A fast, highly sensitive double-nested PCR-based method to screen fish immunobiomes. *Molecular Ecology Resources*, **12**, 1027–1039.

Brown JH, Jardetzky TS, Gorga JC *et al.* (1993) 3-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature*, **364**, 33–39.

Campbell D, Bernatchez L (2004) Generic scan using AFLP markers as a means to assess the role of directional selection in the divergence of sympatric whitefish ecotypes. *Molecular Biology and Evolution*, **21**, 945–956.

Cann DC, Taylor LY (1982) An outbreak of botulism in rainbow trout, *Salmo Gairdneri Richardson*, farmed in Britain. *Journal of Fish Diseases*, **5**, 393–399.

Charbonnel N, Pemberton J (2005) A long-term genetic survey of an ungulate population reveals balancing selection acting on MHC through spatial and temporal fluctuations in selection. *Heredity*, **95**, 377–388.

Chaves LD, Faile GM, Krueth SB, Hendrickson JA, Reed KM (2010) Haplotype variation, recombination, and gene conversion within the Turkey MHC-B locus. *Immunogenetics*, **62**, 465–477.

Clarke B, Kirby DRS (1966) Maintenance of histocompatibility polymorphisms. *Nature*, **211**, 999–1000.

Croisetière S, Tarte P, Bernatchez L, Belhumeur P (2008) Identification of MHC class IIβ resistance/susceptibility alleles to *Aeromonas salmonicida* in brook charr (*Salvelinus fontinalis*). *Molecular Immunology*, **45**, 3107–3116.

Dionne M, Miller KM, Dodson JJ, Caron F, Bernatchez L (2007) Clinal variation in MHC diversity with temperature: evidence for the role of host-pathogen interaction on local adaptation in Atlantic salmon. *Evolution*, **61**, 2154–2164.

Dionne M, Miller KM, Dodson JJ, Bernatchez L (2009) MHC standing genetic variation and pathogen resistance in wild Atlantic salmon. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **364**, 1555–1565.

Doherty PC, Zinkernagel RM (1975) Enhanced immunilogical surveillance in mice heterozygous at the H-2 gene complex. *Nature*, **256**, 50–52.

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 1–19.

Eizaguirre C, Lenz TL (2010) Major histocompatibility complex polymorphism: dynamics and consequences of parasite-mediated local adaptation in fishes. *Journal of Fish Biology*, **77**, 2023–2047.

Eizaguirre C, Yeates SE, Lenz TL, Kalbe M, Milinski M (2009) MHC-based mate choice combines good genes and maintenance of MHC polymorphism. *Molecular Ecology*, **18**, 3316–3329.

Eizaguirre C, Lenz TL, Kalbe M, Milinski M (2012) Divergent selection on locally adapted major histocompatibility complex immune genes experimentally proven in the field. *Ecology Letters*, **15**, 723–731.

Ejsmond MJ, Radwan J (2011) MHC diversity in bottlenecked populations: a simulation model. *Conservation Genetics*, **12**, 129–137.

Eswar N, Eramian D, Webb B, Shen M-Y, Sali A (2008) Protein structure modeling with MODELLER. In: *Structural Proteomics* (eds Kobe B, Guss M, Huber T), pp. 145–159. Humana Press, Totowa, New Jersey.

Evans ML, Bernatchez L (2012) Oxidative phosphorylation gene transcription in whitefish species pairs reveals patterns of parallel and nonparallel physiological divergence. *Journal of Evolutionary Biology*, **25**, 1823–1834.

Evans ML, Neff BD (2009) Major histocompatibility complex heterozygote advantage and widespread bacterial infections in populations of Chinook salmon (*Oncorhynchus tshawytscha*). *Molecular Ecology*, **18**, 4716–4729.

Evans ML, Neff BD, Heath DD (2010a) MHC-mediated local adaptation in reciprocally translocated Chinook salmon. *Conservation Genetics*, **11**, 2333–2342.

Evans ML, Neff BD, Heath DD (2010b) MHC genetic structure and divergence across populations of Chinook salmon (*Oncorhynchus tshawytscha*). *Heredity*, **104**, 449–459.

Evans ML, Dionne M, Miller KM, Bernatchez L (2012a) Mate choice for major histocompatibility complex genetic divergence as a bet-hedging strategy in the Atlantic salmon (*Salmo salar*). *Proceedings of the Royal Society of London B: Biological Sciences*, **279**, 379–386.

Evans ML, Praebel K, Peruzzi S, Bernatchez L (2012b) Parallelism in the oxygen transport system of the lake whitefish: the role of physiological divergence in ecological speciation. *Molecular Ecology*, **21**, 4038–4050.

de Eyto E, McGinnity P, Consuegra S *et al.* (2007) Natural selection acts on Atlantic salmon major histocompatibility (MH) variability in the wild. *Proceedings of the Royal Society of London B: Biological Sciences*, **274**, 861–869.

Frandsen F, Malmquist HJ, Snorrason SS (1989) Ecological parasitology of polymorphic arctic charr, *Salvelinus alpinus* (L) in Thingvallavatn, Iceland. *Journal of Fish Biology*, **34**, 281–297.

Frank SA (2002) *Immunology and Evolution of Infectious Disease*. Princeton University Press, Princeton, New Jersey.

Fraser BA, Neff BD (2010) Parasite mediated homogenizing selection at the MHC in guppies. *Genetica*, **138**, 273–278.

Fraser BA, Ramnarine IW, Neff BD (2010) Temporal variation at the MHC classIIB in wild populations of the guppy (*Poecilia reticulata*). *Evolution*, **64**, 2086–2096.

Froeschke G, Sommer S (2012) Insights into the complex associations between MHC class II DRB polymorphism and multiple gastrointestinal parasite infestations in the striped mouse. *PLoS ONE*, **7**, e31820.

Gavrilets S (2004) *Fitness Landscapes and the Origin of Species*. Princeton University Press, Princeton, New Jersey.

Germain RN (1994) MHC-dependent antigen processing and peptide presentation: providing ligands for T-lymphocyte activation. *Cell*, **76**, 287–299.

Gil P, Vivas J, Gallardo CS, Rodriguez LA (2000) First isolation of *Staphylococcus warneri*, from diseased rainbow trout, *Oncorhynchus mykiss* (Walbaum), in Northwest Spain. *Journal of Fish Diseases*, **23**, 295–298.

Gomez GD, Balcazar JL (2008) A review on the interactions between gut microbiota and innate immunity of fish. *FEMS Immunology and Medical Microbiology*, **52**, 145–154.

Grimholt U, Olsaker I, Lindstrom CD, Lie O (1994) A study of variability in the MHC class IIb1 and class 1a2 domain exons of Atlantic salmon *Salmo salar* L. *Animal Genetics*, **25**, 147–153.

Hardy OJ, Vekemans X (2002) SPAGEDI: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, **2**, 618–620.

Hedrick PW (2002) Pathogen resistance and genetic variation at MHC loci. *Evolution*, **56**, 1902–1908.

Hill AVS, Allsopp CEM, Kwiatkowski D *et al.* (1991) Common west African HLA antigens are associated with protection from severe malaria. *Nature*, **352**, 595–600.

Hughes AL, Nei M (1988) Patterns of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, **335**, 167–170.

Huson DH, Kloepper TH (2005) Computing recombination networks from binary sequences. *Bioinformatics*, **21**, 159–165.

Ilmonen P, Penn DJ, Damjanovich K *et al.* (2007) Major histocompatibility complex heterozygosity reduces fitness in experimentally infected mice. *Genetics*, **176**, 2501–2508.

Johannesson K (2001) Parallel speciation: a key to sympatric divergence. *Trends in Ecology & Evolution*, **16**, 148–153.

Kalbe M, Eizaguirre C, Dankert I *et al.* (2009) Lifetime reproductive success is maximized with optimal major histocompatibility complex diversity. *Proceedings of the Royal Society of London B: Biological Sciences*, **276**, 925–934.

Klein J, Satta Y, Ohuigin C, Takahata N (1993) The molecular descent of the major histocompatibility complex. *Annual Review of Immunology*, **11**, 269–295.

Kloch A, Babik W, Bajer A, Sinski E, Radwan J (2010) Effects of an MHC-DRB genotype and allele number on the load of gut parasites in the bank vole *Myodes glareolus*. *Molecular Ecology*, **19**, 255–265.

Landry C, Bernatchez L (2001) Comparative analysis of population structure across environments and geographical scales at major histocompatibility complex and microsatellite loci in Atlantic salmon (*Salmo salar*). *Molecular Ecology*, **10**, 2525–2539.

Landry C, Garant D, Duchesne P, Bernatchez L (2001) 'Good genes as heterozygosity': the major histocompatibility complex and mate choice in Atlantic salmon (*Salmo salar*). *Proceedings of the Royal Society of London B: Biological Sciences*, **268**, 1279–1285.

Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, **26**, 283–291.

Lenz TL, Eizaguirre C, Scharsack JP, Kalbe M, Milinski M (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**, 2122–2142.

Lu G, Bernatchez L (1999) Correlated trophic specialization and genetic divergence in sympatric lake whitefish ecotypes (*Coregonus clupeaformis*): support for the ecological speciation hypothesis. *Evolution*, **53**, 1491–1505.

MacColl ADC (2009) Parasite burdens differ between sympatric three-spined stickleback species. *Ecography*, **32**, 153–160.

Matthews B, Harmon LJ, M'Gonigle L, Marchinko KB, Schaschl H (2010) Sympatric and allopatric divergence of MHC genes in threespine stickleback. *PLoS ONE*, **5**, e10948.

Milinski M, Griffiths S, Wegner KM *et al.* (2005) Mate choice decisions of stickleback females predictably modified by MHC peptide ligands. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 4414–4418.

Miller KM, Withler RE (1996) Sequence analysis of a polymorphic Mhc class II gene in Pacific salmon. *Immunogenetics*, **43**, 337–351.

Miller KM, Kaukinen KH, Beacham TD, Withler RE (2001) Geographic heterogeneity in natural selection on an MHC locus in sockeye salmon. *Genetica*, **111**, 237–257.

Natsopoulou ME, Palsson S, Olafsdottir GA (2012) Parasites and parallel divergence of the number of individual MHC alleles between sympatric three-spined stickleback *Gasterosteus aculeatus* morphs in Iceland. *Journal of Fish Biology*, **81**, 1696–1714.

Nosil P (2012) *Ecological Speciation*. Oxford Press, Oxford.

Nowak MA, Tarczyhornoch K, Austyn JM (1992) The optimal number of major histocompatibility complex molecules in an individual. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 10896–10899.

Paterson S, Wilson K, Pemberton JM (1998) Major histocompatibility complex variation associated with juvenile survival

and parasite resistance in a large unmanaged ungulate population (*Ovis aries* L.). *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 3714–3719.
Pavey SA, Lamaze FC, Garant D, Bernatchez L (2011) Full length MHC II beta exon 2 primers for salmonids: a new resource for next generation sequencing. *Conservation Genetics Resources*, **3**, 665–667.
Perez T, Balcazar JL, Ruiz-Zarzuela I *et al.* (2010) Host-microbiota interactions within the fish intestinal ecosystem. *Mucosal Immunology*, **3**, 355–360.
Piertney SB, Oliver MK (2006) The evolutionary ecology of the major histocompatibility complex. *Heredity*, **96**, 7–21.
Pitcher TE, Neff BD (2006) MHC class IIB alleles contribute to both additive and nonadditive genetic effects on survival in Chinook salmon. *Molecular Ecology*, **15**, 2357–2365.
Renaut S, Nolte AW, Rogers SM, Derome N, Bernatchez L (2011) SNP signatures of selection on standing genetic variation and their association with adaptive phenotypes along gradients of ecological speciation in lake whitefish species pairs (*Coregonus* spp.). *Molecular Ecology*, **20**, 545–559.
Reusch TB, Schaschl H, Wegner KM (2004) Recent duplication and inter-locus gene conversion in major histocompatibility class II genes in a teleost, the three-spined stickleback. *Immunogenetics*, **56**, 427–437.
Schloss PD, Westcott SL, Ryabin T *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, **75**, 7537–7541.
Schluter D (2000) *The Ecology of Adaptive Radiation*. Oxford University Press, Oxford.
Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proceedings of the Royal Society of London B: Biological Sciences*, **277**, 979–988.
Spurgin LG, van Oosterhout C, Illera JC *et al.* (2011) Gene conversion rapidly generates major histocompatibility complex diversity in recently founded bird populations. *Molecular Ecology*, **20**, 5213–5225.
Stendera S, Adrian R, Bonada N *et al.* (2012) Drivers and stressors of freshwater biodiversity patterns across different ecosystems and scales: a review. *Hydrobiologia*, **696**, 1–28.
Sutton JT, Nakagawa S, Robertson BC, Jamieson IG (2011) Disentangling the roles of natural selection and genetic drift in shaping variation at MHC immunity genes. *Molecular Ecology*, **20**, 4408–4420.
Tamura K, Peterson D, Peterson N *et al.* (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, **28**, 2731–2739.
Tollenaere C, Bryja J, Galan M *et al.* (2008) Multiple parasites mediate balancing selection at two MHC class II genes in the fossorial water vole: insights from multivariate analyses and population genetics. *Journal of Evolutionary Biology*, **21**, 1307–1320.
Wegner KM, Kalbe M, Kurtz J, Reusch TBH, Milinski M (2003a) Parasite selection for immunogenetic optimality. *Science*, **301**, 1343.
Wegner KM, Reusch TBH, Kalbe M (2003b) Multiple parasites are driving major histocompatibility complex polymorphism in the wild. *Journal of Evolutionary Biology*, **16**, 224–232.
Westerdahl H, Hansson B, Bensch S, Hasselquist D (2004) Between-year variation of MHC allele frequencies in great reed warblers: selection or drift? *Journal of Evolutionary Biology*, **17**, 485–492.
Westerdahl H, Asghar M, Hasselquist D, Bensch S (2012) Quantitative disease resistance: to better understand parasite-mediated selection on major histocompatibility complex. *Proceedings of the Royal Society of London B: Biological Sciences*, **279**, 577–584.
Wilson DJ, McVean G (2006) Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics*, **172**, 1411–1425.
Woelfing B, Traulsen A, Milinski M, Boehm T (2009) Does intra-individual major histocompatibility complex diversity keep a golden mean? *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **364**, 117–128.

S.A.P. performed the MHC genotyping laboratory work, conducted the data analysis and wrote the manuscript. M.S. performed the pathogenic bacterial identification laboratory work and microbial bioinformatic and data analysis. W.A. led the team of field collections and co-conceived the study. E.N. developed the pipeline for the MHC genotyping and greatly contributed in the data analysis. F.C.L. cowrote the manuscript. P.A.G. conceived and wrote the R script used in the genotyping pipeline and substantially revised the manuscript. M.F. contributed to the data analysis and substantially revised the manuscript. F.-O.G.-H. conducted the pilot study of MHC genotyping that laid the foundation for this manuscript. H.M. performed the protein modelling. L.B. conceived the study and substantially revised the manuscript.

Data accessibility

Raw MHCIIβ 454 reads have been deposited in NCBI SRA (SRA073466) and the Dryad entry for this manuscript (doi: 10.5061/dryad.ft94b). MHC and pathogen matrices and databases, putative pathogen sequences, input files, Splitstree networks, R scripts, and MHC allele sequences have been deposited in the Dryad repository for this manuscript (doi: 10.5061/dryad.ft94b). The pipeline used to determine and assign alleles to the individuals is available at this link: github.com/enormandeau/ngs_genotyping.

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Seventeen MHC alleles in lake whitefish including nucleotide and amino acid sequences.

Fig. S2 NeighborNet haplotype network of the overlapping translated 16 alleles (nonfunctional allele 10 excluded) observed in lake whitefish together with 20 alleles from European whitefish (as indicated with 'EW'; Binz *et al.* 2001).

Fig. S3 Histograms of alleles per individual in dwarf and normal lake whitefish for all five lakes.

Fig. S4 Results of the omegaMap analysis mapped onto our 3D protein structure prediction of allele 1.

Fig. S5 Posterior probability for the rho parameter for the MHCIIβ sequences generated with omegaMap.

Fig. S6 Spineplot representing the infection status vs. within individual MHCIIβ diversity.

Fig. S7 Spineplot representing the infection status vs. number of MHCIIβ alleles within an individual.

Table S1 454 multiplex primers.