MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 4896–4914 doi: 10.1111/mec.12447

Targeted sequence capture and resequencing implies a predominant role of regulatory regions in the divergence of a sympatric lake whitefish species pair (*Coregonus clupeaformis*)

FRANCOIS OLIVIER HEBERT,* SÉBASTIEN RENAUT† and LOUIS BERNATCHEZ*

*Département de Biologie, Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Pavillon Charles-Eugènes-Marchand, Québec G1V 0A6, Canada, †Department of Botany, University of British Columbia, 3529-6270 University Blvd, Vancouver BC V6T 1Z4, Canada

Abstract

Latest technological developments in evolutionary biology bring new challenges in documenting the intricate genetic architecture of species in the process of divergence. Sympatric populations of lake whitefish represent one of the key systems to investigate this issue. Despite the value of random genotype-by-sequencing methods and decreasing cost of sequencing technologies, it remains challenging to investigate variation in coding regions, especially in the case of recently duplicated genomes as in salmonids, as this greatly complicates whole genome resequencing. We thus designed a sequence capture array targeting 2773 annotated genes to document the nature and the extent of genomic divergence between sympatric dwarf and normal whitefish. Among the 2728 genes successfully captured, a total of 2182 coding and 10 415 noncoding putative single-nucleotide polymorphisms (SNPs) were identified after applying a first set of basic filters. A genome scan with a quality-refined selection of 2203 SNPs identified 267 outlier SNPs in 210 candidate genes located in genomic regions potentially involved in whitefish divergence and reproductive isolation. We found highly heterogeneous F_{ST} estimates among SNP loci. There was an overall low level of coding polymorphism, with a predominance of noncoding mutations among outliers. The heterogeneous patterns of divergence among loci confirm the porous nature of genomes during speciation with gene flow. Considering that few protein-coding mutations were identified as highly divergent, our results, along with previous transcriptomic studies, imply that changes in regulatory regions most likely had a greater role in the process of whitefish population divergence than protein-coding mutations. This study is the first to demonstrate the efficiency of large-scale targeted resequencing for a nonmodel species with such a large and unsequenced genome.

Keywords: genome scan, lake whitefish, next-generation sequencing, population genomics, sequence capture, speciation

Received 14 March 2013; revision received 3 July 2013; accepted 8 July 2013

Introduction

Speciation, a fundamental process responsible for biological diversity, is the result of neutral and selective processes acting in tandem at the genetic and phenotypic levels (Mitchell-Olds *et al.* 2007; Presgraves 2010). Although many of the intricate mechanisms associated with them still remain unknown (Koonin 2012), recent advances in genetic studies have helped to understand the interactions causing changes to genomic architecture during the process of species divergence (Wu & Ting 2004; Jones *et al.* 2012). Evidence from various studies has shown that ecologically driven genomic divergence

Correspondence: Francois Olivier Hebert, Fax: 1-418-656-7176; E-mail: francois-olivier.gagnon-hebert.1@ulaval.ca

can occur in the face of gene flow (Nosil *et al.* 2008; Cadillo-Quiroz *et al.* 2012; Gagnaire *et al.* 2012a). Molecular analyses of young and diverging lineages in natural hybrid zones reveal highly heterogeneous and complex genome-wide patterns of genetic introgression among loci (Gagnaire *et al.* 2011; Gompert *et al.* 2012). The observed variation in introgression rates among genomic regions can be attributed to the presence of genetic barriers to gene flow, which also limit the rate of introgression of nearby linked loci (Barton 1979; Barton & Bengtsson 1986). Competitive interplay between the intensity of selection and the extent of recombination (Felsenstein 1981) determines the rate of introgression at neutral loci, while advantageous alleles might be slightly delayed by these barriers, depending on migration rate and strength of selection. Consequently, these dynamic and variable introgression rates create semipermeable barriers to gene flow between diverging taxa (Payseur 2010). In a situation of ecological adaptive divergence, natural selection promotes the genetic divergence of loci associated with a higher fitness, while still allowing gene flow in neutral regions (Turner *et al.* 2005; Nosil *et al.* 2009). With increasing time or selection strength, these genomic regions resistant to gene flow, called genomic islands of divergence, are predicted to expand in number and size through genetic hitchhiking (Maynard Smith & Haigh 1974) potentially until both genomes become completely genetically isolated (Wu 2001; Feder *et al.* 2012; Via 2012). The premise is that divergent selection prevents or reduces gene flow between diverging populations in certain genomic regions. This conjointly reduces interpopulation recombination. The localized reduction in effective gene flow at or near selected sites allows the build-up of genetic differentiation around these loci and thus contributes to the expansion of such divergent genomic regions (reviewed in Nosil & Feder 2012). Speciation is a dynamic biological process during which genomic divergence builds up and constantly progresses over time. Consequently, if one wants to investigate how the core processes unfold from the start until the end, it is crucial to identify adaptive genetic changes in the early stages of genome differentiation prior to the completion of the speciation process and before other differences accumulate over time. Much theoretical work has been done in order to develop a conceptual framework for genomic divergence in the context of speciation. Yet, major questions pertaining to the genomic architecture of speciation and the relative importance of various processes facilitating or impeding the spread of divergent loci remain unresolved (Nosil & Feder 2012).

The advent of high-throughput sequencing technologies and new computational developments has opened the possibility of studying such questions on a genome-wide scale. The method predominantly used to find divergent genomic regions consists of scanning large portions of the genome in order to estimate the extent of genetic differentiation among loci (F_{ST}-based genome scans) (Lewontin & Krakauer 1973; Beaumont & Nichols 1996; Excoffier *et al.* 2009). Such approaches have been useful in identifying multiple regions of differentiation with variable degrees of divergence (e.g. Lawniczak *et al.* 2010; Strasburg *et al.* 2012). Moving beyond the identification of outlier genomic regions by characterizing the genes involved and their association with known divergent phenotypes brings compelling evidence of how genome architecture is modelled during ecological speciation (Presgraves 2010). For example, Smadja *et al.* (2012) conducted a large-scale candidate gene approach combining population genomics and QTL methods on the pea aphid system (*Acyrthosiphon pisum*). Their results suggested a restricted effect of hitchhiking around selected loci, giving birth to small islands of divergence, which is similar to the observations made by Nadeau *et al.* (2012) in *Heliconius* butterflies. Conversely, genome-wide patterns of divergence in various species pairs have revealed large islands of divergence rather than small and independent selected regions during the early stages of reproductive isolation (Via & West 2008; Hohenlohe *et al.* 2012; Renaut *et al.* 2012). These apparently contradictory findings ('few large' vs. 'many small' regions of divergence) might reflect different methods for defining regions of divergence, different timing of divergence along the speciation continuum or different demographic dynamics of species (Feder *et al.* 2012). Different processes could be involved in different points on the continuum.

Lake whitefish species pairs (*Coregonus clupeaformis*) offer an ideal situation to study the ongoing process of speciation where they occur in sympatry in several lakes in northeastern North America (Lu & Bernatchez 1999; Bernatchez *et al.* 2010). The 'normal' whitefish, characterized by an epibenthic foraging activity grows faster, becomes larger and reaches maturity at a later age than the 'dwarf' whitefish, which lives in the limnetic zone, grows slower and reaches maturity at an earlier age (reviewed in Bernatchez 2004). Besides striking size difference at sexual maturation, the most discriminating phenotypic trait is the gill-raker apparatus: dwarf whitefish distinctively have more numerous and less separated gill rakers than normal whitefish. Thus, the dwarf whitefish is more efficient at retaining smaller planktonic prey (Bernatchez 2004). The exploitation of the limnetic niche by dwarf whitefish results in a heritable difference in swimming behaviour compared with normal whitefish (Rogers *et al.* 2002). Also, dwarf whitefish have significantly higher metabolic rates than

normal whitefish, which stems from their higher energetic demands (more active swimming) and a lower bio-energetic conversion efficiency (Trudel *et al.* 2001). Considering their pronounced phenotypic differences at many complex traits and their recent divergence on an evolutionary timescale, marked by a phase of allopatry (~60 000 years BP) followed by secondary contact in sympatry about 12 000 years BP (Bernatchez & Dodson 1990; Jacobsen *et al.* 2012), these incipient species of lake whitefish, exploiting distinct ecological niches, are amenable to investigate early mechanisms involved in the process of ecological divergence. The very short period of time characterizing their sympatric divergence (12 000 years) and the extensive gene flow exhibited between morphs in various lakes contribute in creating only a partial reproductive isolation that, according to theory, is considered a young system (e.g. Schluter 2001). Previous work using QTL mapping and common garden experiments has provided evidence for a genetic basis of adaptive traits (swimming behaviour, growth, morphology, gene expression variation) known to differ between both ecotypes (Rogers *et al.* 2002; Rogers & Bernatchez 2007; Derome *et al.* 2008; St-Cyr *et al.* 2008; Whiteley *et al.* 2008; Jeukens *et al.* 2010; Gagnaire *et al.* 2013a). An integrated approach linking QTL mapping and gene expression studies with single-nucleotide polymorphisms (SNPs) analyses also revealed pronounced allele frequency divergence for several key genes differentially expressed between ecotypes (Renaut *et al.* 2010, 2011). Results from this integrative approach suggest that multiple mitochondrial and nuclear genes involved in energy metabolism stand out as potential candidates underlying the ecological divergence of whitefish ecotypes (Renaut *et al.* 2010). Additionally, 16 other candidate genes showed genotype–phenotype associations in relation to four adaptive phenotypes, namely growth, swimming activity, gill rakers and condition factor (see Renaut *et al.* 2011). Overall, these comprehensive studies have demonstrated the role of natural selection in shaping different levels of gene expression, which is thought to be one of the predominant proximate mechanisms responsible for the phenotypic differences among lake whitefish species pairs. As such, they represent an excellent starting point towards a deeper and more precise characterization of the genetic architecture underlying ecological speciation.

Here, the technique of sequence capture (high density DNA microarray allowing the enrichment of targeted genomic regions, see Hodges *et al.* 2007) was used to enrich a large set of genes, get a clearer picture of the extent of genetic differentiation between both ecotypes and quantify the strength of selection on protein-coding divergence. The aim was also to refine and extend the list of candidate genes potentially involved in the process of adaptive divergence. In total, 2773 genes were targeted by the array-based on available EST and cDNA sequences. Thus, regions of interest among the large and complex whitefish genome (~3 Gbp, Animal Genome Size Database) (Jeukens *et al.* 2011), which also underwent a recent duplication event approximately 60 Ma (salmonid duplication event, see Crête-Lafrenière *et al.* 2012), were efficiently and rapidly retrieved and sequenced. Sequencing data from multiple individuals were *de novo* assembled to reconstruct whole gene sequences and document patterns of genetic divergence between dwarf and normal populations through an F_{ST}-based genome scan. Because no salmonid genome is yet available, it proved to be an efficient way of targeting genomic DNA in both coding and adjacent noncoding regions, while circumventing the problems inherent to the sequencing and assembly of massive amounts of repetitive and noncoding elements.

Materials and methods

Study system and DNA preparation

Fish were collected in 2010 from Cliff Lake (46°23′51″N, 69°15′05″W, St John River drainage, ME, USA), which harbours sympatric populations of normal and dwarf whitefish. Twelve dwarf and twelve normal individuals (48 chromosome sets in total) were randomly chosen among the captured fish. Genomic DNA was extracted from a caudal fin clip using DNeasy tissue kit (Qiagen, Duesseldorf, Germany) according to the manufacturer's protocol. A minimum of 5 μg of unamplified genomic DNA was obtained for each sample.

Selecting exon targets, DNA enrichment and sequencing

NimbleGen capture array technology (Roche, Madison, USA) was used to enrich preselected coding regions from whitefish genes only. All publicly available whitefish expressed sequence tags (ESTs) (cGRASP, http://web.uvic.ca/grasp/) were used as primary data for the first probe design, in addition with cDNA sequences from previous work (454 GS-FLX platform, Renaut *et al.* 2010). A total of 13 516 coding sequences were further processed in order to eliminate redundancy (sequences with BLASTn e-value > 1e-20 were collapsed, and sequences of length <200 bp were discarded), and only selected sequences annotated in either nt, nr or swissprot were kept (BLASTn e-value < 1e-25, BLASTx e-value < 1e-25). Mitochondrial DNA and repetitive sequences, prone to generate an excessive capture compared with other gene targets, were discarded. Following these steps, 3242 unigene sequences were selected

as targets for the first array design. The design was validated and improved by conducting two successive capture tests with three normal and three dwarf individuals using a Roche 454 GS-FLX sequencer at the Plate-forme d'Analyse Génomique (IBIS, Université Laval, Québec, Canada). A small fraction of the initial targets (2.7%), mainly composed of remaining mitochondrial DNA and repetitive elements that had not been discarded through previous cleaning steps, was discarded for capturing more than 90% of the sequenced reads. These optimized targets were sent to NimbleGen bioinformatics service to build the final array spanning the coding sequence (exons) of 2773 whitefish genes. A total of 358 847 tiling DNA oligonucleotides (probes) spanning 100% of the targeted genes were designed.

TruSeq paired-end (2 × 100 base pairs) libraries (Illumina, San Diego, USA) with an insert size of 200–250 bp were prepared for each individual according to the manufacturer's protocol (TruSeq DNA samples prep kit). Each library contained a unique six-bp molecular identifying sequence (MID) included in the commercial kit provided by the manufacturer (Illumina, San Diego, USA). The capture step was performed by the Plate-forme d'Analyse Génomique (IBIS, Université Laval, Québec, Canada) and the captured DNA samples sequenced at the Genome Quebec Innovation Centre (McGill University, Montréal, Canada) on an Illumina HiSeq 2000 platform.

Assembly and sequence analysis

Due to RAM memory limitation, all paired-end reads for the 24 individuals were successively and independently assembled with four different *k-mer* values (*k-mer* values = 27, 37, 47, 57) using ABySS (Simpson *et al.* 2009), a fast and accurate assembler with low RAM requirements. Because ABySS only uses short reads, CLC Genomics Workbench 5.1 (CLC bio, Aarhus, Denmark) was ultimately used to perform a final *de novo* assembly with contigs from the four distinct assemblies previously generated with ABySS (similarity 0.95, overlap 0.5). This final 'meta-assembly' produced final consensus contigs. Because four independent assemblies were conducted with different parameters, contigs that were not identical among the four assemblies were considered suspicious due to potential paralogy and discarded. Contigs of length <200 bp were also discarded.

Given the recently duplicated whitefish genome (Radice *et al.* 1994; Krasnov *et al.* 2005), some contigs may represent chimeras or assembly of different nonspecific genomic DNA fragments being the result of secondary capture (Fu *et al.* 2010). To improve the quality of the assembly and minimize mapping errors, contigs were first blasted against the 3039 original targets, and queries returning no significant hits (BLASTn e-value > 1e-10) were discarded. Redundant contigs (BLASTn e-value < 1e-20) were then merged together using custom Python scripts (v. 2.6.5). These longer contigs represent the complete set of assembled genes, including exons, introns, several gaps of variable length between exons and, in several cases, putative noncoding regions located before the first exon and after the last exon. All paired-end reads used in the assembly step were ultimately mapped back to this final reference set of genes using BWA (v. 0.6.1) with default parameters (Li & Durbin 2009). Because the quality of all the bases sequenced and contained in our reads was above the average quality standard of Q-20 (99% base accuracy), there was no need to trim reads prior to the assembly. Assembly statistics were estimated using custom Python scripts (v. 2.6.5) and R (v. 2.15.1; The R Foundation for Statistical Computing®, 2012, 3-900051-07-0). Functional categories (gene ontology biological functions) were associated with each gene sequence in the final data set with Blast2GO (Conesa *et al.* 2005).

A generalized hidden Markov method implemented in AUGUSTUS (v. 2.6.1, Stanke *et al.* 2004) was used to predict exon positions within the assembled genes. To improve these predictions, a raw assembly of the complete whitefish transcriptome ((Dion-Côté AM, Normandeau E & Bernatchez L., unpublished data) was integrated in the algorithm as a species-specific supplementary resource. The Bayesian model 'mpileup' implemented in SAMTOOLS (v.0.1.18, Li *et al.* 2009) was used to call consensus bases and single-nucleotide polymorphisms (SNPs). Read bases with a Phred quality score <20 or with an insufficient overall depth of coverage (<64 reads) to call genotypes with high confidence (and infer allele frequency differences between populations) were excluded. We performed various tests that showed that the best compromise in order to maximize the number of high-quality SNPs detected in as many individuals as possible, considering the data set quality, was four reads per individual for a minimum of eight individuals for each of the two populations (minimum of 64 reads in total). Results of a recent study show that sequencing more individuals with lower coverage allows a more optimal use of sequencing and sampling effort (Buerkle & Gompert 2012). Given the limited number of individuals that could be analysed in this study due to cost constraints, this threshold was deemed the best compromise in order to extract as much information as possible from the data set. A high coverage threshold (>8000 reads) was also applied to exclude all possible bases located in repetitive regions or in regions of high representation in the genome. Outlier tests were conducted on the subset of remaining loci following all filtering procedures.

Population genetics analyses

To confirm the positions and length of the exons simulated by AUGUSTUS, open reading frame (ORF) for each polymorphic gene were generated (minimum ORF length = 75 nt) using the program *getorf* in EMBOSS (European Molecular Biology Open Software Suite, Rice *et al.* 2000). ORFs were then used to make predictions on any damaging effect of nonsynonymous mutations on phenotypes using an iterative greedy algorithm implemented in POLYPHEN-2 (Adzhubei *et al.* 2010). Allele frequencies were subsequently estimated based on direct read counts from SAMTOOLS output file using custom Python scripts, and absolute allelic frequency divergence ($\delta_{D/N} = |f_{allele1,D} - f_{allele1,N}|$) was finally computed for every SNP according to these frequency estimates. This analysis was conducted as a parallel test and a more exhaustive analysis compared with the values calculated by Renaut *et al.* (2010) using cDNA sequences, which were used as raw material for the sequence capture array used in this study. SNPs with a minor allele count of <2 were discarded, using a filtering step in conjunction with coverage thresholds mentioned above. This final subset of SNPs meant to be conservative towards eliminating as many paralogues as possible from analyses was used to perform the F_{ST}-based genome scan.

To measure the extent of population differentiation, F_{ST} values (Wright 1951) were calculated according to the method of Beaumont & Nichols (1996), an extension to Lewontin-Krakauer test (Lewontin & Krakauer 1973) implemented in LOSITAN Workbench (Antao *et al.* 2008). This model assumes an island migration model, and it also assumes that populations are at mutation–drift selection equilibrium (gene flow between demes), a condition that is respected in this case, based on previous simulations (Campbell & Bernatchez 2004) and on more than 20 years of research on this system (reviewed in Bernatchez *et al.* 2010) that show through various studies the temporal stability of the differentiation between whitefish ecotypes. Using an infinite allele mutational model, assuming a number of two demes with no hierarchical population structure, 500 000 coalescent simulations were performed to obtain the joint distribution of F_{ST} values (FDR = 0.05, confidence interval = 0.99). Expected heterozygosity was also calculated for each locus based on Hardy–Weinberg equilibrium. Loci outside the 99% confidence interval were considered as outliers, based on a 5% false discovery rate threshold.

In order to integrate knowledge from previous studies in this system, genes identified as outliers were classified into the 12 broad functional categories established by St-Cyr *et al.* (2008) in a previous transcriptomic study of whitefish, according to their gene ontology (biological process). Two of these categories were not represented in the final outlier data set (germ-line formation and lipid metabolism), and two additional categories were added: (i) growth and development (bone morphogenesis heart development, fin development, growth, regulation of developmental process) and (ii) nervous system and learning (learning, neural development, cognition).

Results

Sequencing and capture efficiency

DNA enrichment (targeting roughly 10% of the total amount of genes in the whitefish genome) and resequencing yielded more than 841 million short sequence reads (Table 1), with an average of 35.1 million sequence reads per sample (range: 27.8M–50.8M). While only 11.8% of the reads uniquely mapped back to a targeted sequence, 98% of targeted genes (2728) were successfully captured and assembled (Table S1, Supporting information), with a mean read depth of 5392X (Table 1). Average proportion of annotated coding regions (exons) in assembled genes reached 23.8%. More than 95% of targeted genes successfully captured had 60% of their length covered by assembled contigs (average proportion per target: 88.8%, range: 12.9–100%). Among these 2728 assembled genes, 2364 (86.7%) were polymorphic.

A total of 12 597 putative SNP markers were identified, among which 2182 were coding and 10 415 noncoding. After filtering out undesired loci (see Materials and methods), 3021 biallelic markers were retained, with a mean depth of coverage per sample of 31-fold (median = 36X), representing 1104 genes (Table S2, Supporting information). Although the overall coverage greatly varied among assembled genes (Table 1), the depth of coverage for filtered SNP loci among individuals was considerably more uniform (mean = 30.97 ± 12.95). Overall polymorphism rate per gene was relatively low (1.89 SNPs/Kb), with 60 genes (2.2%) showing a number of SNPs per kilobase greater than five (Table 1). These 60 genes were associated with various functional groups, among which nine are significantly over-represented compared with frequencies of functional groups among all genes (Fisher's exact test, Q-value <0.05). Among these over-represented functional groups, three biological processes had most of the hits: haem binding, G-protein binding and hydrogen ion transporting ATP synthase. The majority of the 3021 retained SNPs (Table S2, Supporting information) were located in putative introns, with a total of 2258 noncoding SNPs distributed in 910 different genes, while the remaining 763 SNPs were located in the puta-

Table 1 Summary statistics of captured genes

Total number of reads sequenced (paired-end)	841 559 424
Mean/sample (min. – max.; median)	35 064 976 (27.8M–50.8M; 35.7M)
Percentage of on-target reads	11.8%
Number of captured and assembled genes	2728 (98% of total)
Mean gene length (min. – max.; median)	1361 (200–9002; 1179)
N50*	1633
Mean coverage/gene (min. – max.; median)	5392X (4X–2.8 × 10^6X; 2384X)
Mean coverage/sample (min. – max.; median)	31X (9X–44X; 35X)
Number of exons (num. genes)	6033 (2053)
Mean exons/gene (min. – max.; median)	3.07 (1–18; 2)
Mean exon length (min. – max.; median)	142.9 (49–526; 120)
SNP analysis	
Number of coding SNPs (num. genes)	763 (383)
Number of noncoding SNPs (num. genes)	2258 (910)
Mean SNPs/Kb (min. – max.; median)	1.89 (0–18.7; 1.29)
Mean coding SNPs/Kb (min. - max; median)	0.127 (0–14.3; 0)
Mean noncoding SNPs/Kb (min. - max; median)	0.444 (0–16.9; 0)

*N50 is defined as the longest length for which the collection of all contigs of that length or longer contains at least half of the total of the lengths of the contigs.

tive exons of 383 different genes. We identified 383 ORFs and among the 539 SNPs that fell within these ORFs, 306 were nonsynonymous and 233 synonymous.

Divergent loci and functional analysis

Loci kept after filtering procedures showed a skewed distribution of allelic frequency divergence towards low values, exhibiting a median of 0.09 (Fig. S1, Supporting information), while 271 SNPs had significant divergent allelic frequencies (Q-value <0.05) and 56 were highly divergent ($\delta_{D/N} \geq 0.5$, Table S2, Supporting information). Observed genome-wide level of differentiation was moderate ($F_{ST} = 0.046$), yet considerable heterogeneity was observed among estimated F_{ST} (Table S2, Supporting information). Average multilocus F_{ST} estimate was similar between putative coding and noncoding regions ($F_{ST\text{ - NONCODING}} = 0.041$, $F_{ST\text{ - CODING}} = 0.049$, $P > 0.2$).

Coalescent simulations performed with LOSITAN were used to obtain the distribution of F_{ST} estimates among all loci (Fig. 1). These were plotted against their respective expected heterozygosity (H_e) (Fig. 2 and S2, Supporting information). Most loci were located within the 99% confidence interval (CI) expected under neutrality, but 267 F_{ST} estimates (11.5%) laid outside this envelope and were thus considered outliers. This group of loci represents highly differentiated markers that could harbour potential genes of interest in the study of whitefish divergence and reproductive isolation. Such markers could be under selection or be linked to direct targets of selection and should be considered for further analyses. Even though these loci show significant signs of

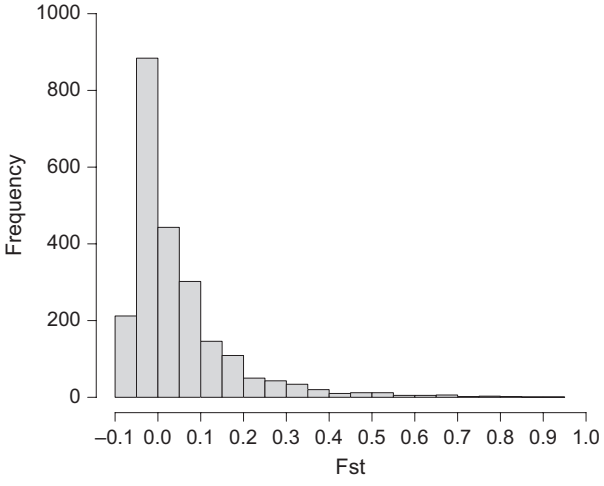

Fig. 1 Distribution of F_{ST} estimates between dwarf and normal whitefish at 2302 single-nucleotide polymorphism (SNP) loci. SNP frequency is plotted against F_{ST} values. Estimates of population divergence were calculated according to the method of Beaumont and Nichols. Average multilocus $F_{ST} = 0.046$.

divergence, their definitive role in the adaptive divergence of whitefish species pairs remains hypothetical and requires further candidate gene analyses. Results indicate that 27 outliers were nonsynonymous substitutions (Tables 2 and S4, Supporting information), compared with 28.9 expected based on the proportion among the whole data set (55.6% of nonsynonymous SNPs expected among all coding SNPs). Similarly, proportions of synonymous and nonsynonymous SNPs among total outliers were not significantly different (Fisher's exact test, $P = 0.5517$, Table 2). In total, 52 out-

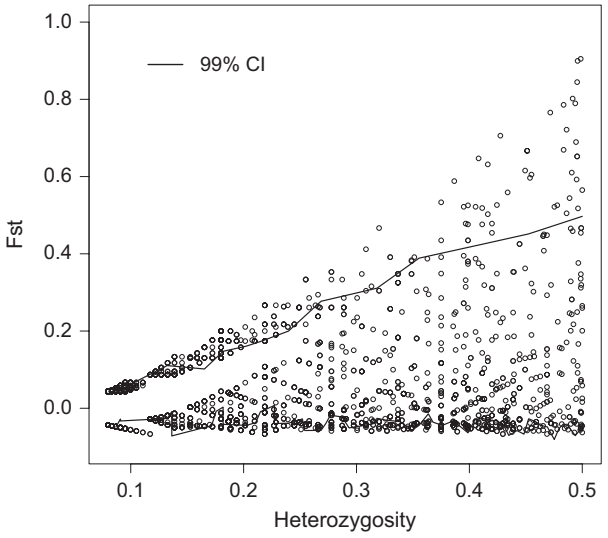

Fig. 2 F_{ST}-based scan for selection using FDIST2 implemented in LOSITAN Workbench. F_{ST} is plotted against expected heterozygosity (H_e), which was estimated using allele counts at 2302 SNP loci given by SAMTOOLS mpileup command. Solid black lines represent the upper and lower limits of the 99% confidence interval (CI). Outlier loci were selected based on the 99% confidence interval and a false discovery rate (FDR) of 5%.

liers were located in putative coding regions (19.5% of the total number of outliers), a similar result compared with the overall proportion of coding SNPs and overall proportion of coding regions within assembled genes (25% and 24%, respectively, Fisher's exact test, $P > 0.1$). Outliers were enriched for neither coding SNP nor nonsynonymous SNPs. Also, proportions of noncoding SNPs were identical in outliers and nonoutliers (Table 2). A significant reduction in observed heterozygosity (H_o) among outlier loci was also observed, but only in the dwarf population (*t*-test, $H_{o(dwarf)\ -\ outliers} = 0.13$ vs. $H_{o(dwarf)\ -\ nonoutliers} = 0.28$ and $H_{o(normal)\ -\ nonoutliers} = 0.25$, $P < 0.05$ in both cases, Fig. 3). Average H_o for outliers was also lower in dwarf than normal population ($H_{o(dwarf)\ -\ outliers} = 0.13$ vs. $H_{o\ (normal)\ -\ outliers} = 0.23$), and this difference was marginally significant (*t*-test, $P = 0.08$). A Pearson correlation test was conducted on normal and dwarf populations to confirm that variability in coverage among loci was not responsible for the observed reduction in heterozygosity (loci with lower coverage could be biased towards a lower observed heterozygosity). There was thus no significant correlation between coverage and H_o either in dwarf (*cor* = 0.0158, $P = 0.447$) or in normal (*cor* = 0.00336, $P = 0.872$) populations. In addition, there is no statistical difference in coverage values between outliers vs. nonoutliers in the dwarf population (*t*-test, $P = 0.0663$) and in the normal population (*t*-test, $P = 0.437$). Also, coverage in the dwarf population for

Table 2 Outlier single-nucleotide polymorphism (SNP) results for annotated coding and noncoding regions. Proportions in parentheses are the proportion among all analysed SNPs

Synonymous SNPs	233 (10.1%)
Outliers	25 (1.1%)
Nonoutliers	208 (9.0%)
Proportion of synonymous outliers among outlier SNPs*	10.3%†
Nonsynonymous SNPs	306 (13.3%)
Outliers	27 (1.2%)
Nonoutliers	279 (12.2%)
Proportion of nonsynonymous outliers among outlier SNPs*	9.2%†
Noncoding SNPs	1752 (76.4%)
Outliers	215 (9.3%)
Nonoutliers	1537 (67.0%)
Proportion of noncoding outliers among outlier SNPs	80.5%†

*Fisher's exact test comparing proportions of synonymous outliers vs. nonsynonymous outliers, $P = 0.5517$.
†Neither of these SNP categories show an enrichment within outlier group, Fisher's exact test, $P > 0.05$.

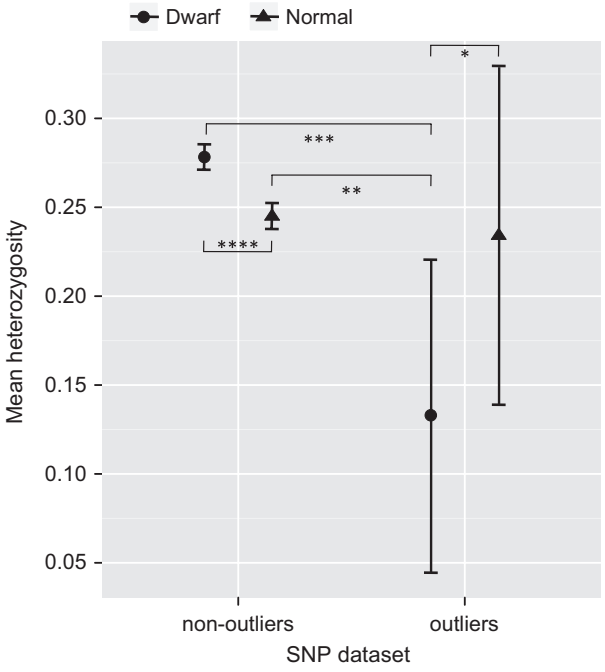

Fig. 3 Observed heterozygosity (H_o) for outlier and nonoutlier loci for both dwarf and normal populations. *t*-tests comparing mean heterozygosity for each single-nucleotide polymorphism (SNP) data set. Mean heterozygosities in nonoutlier group: 0.278 ± 0.007 and 0.245 ± 0.007 for dwarf and normal, respectively. Mean heterozygosities in outlier group: 0.133 ± 0.088 and 0.234 ± 0.095 for dwarf and normal respectively. Coverage did not significantly differ among populations and SNP categories ($P > 0.05$, data not shown). *$P = 0.08$, **$P = 0.021$, ***$P = 0.0068$, ****$P < 0.0001$.

Table 3 Functional groups over-represented* among outlier genes compared with their proportion among all genes assembled.

Functional groups†	GO terms	Number of genes‡
Behaviour	Feeding behaviour (GO:0007631), adult behaviour (GO:00030534), adult locomotory behaviour (GO:0008344), behavioural interaction between organisms (GO:0051705), visual behaviour (GO:0007632), locomotory behaviour (GO:0007626)	2
Blood and transport	Sodium ion export (GO:0071436), sodium ion transmembrane transport (GO:0035725), regulation of vasoconstriction (GO:0019229)	3
Cell cycle regulation	Interphase of mitotic cell cycle (GO:0051329), interphase (GO:0051325)	15
Cell structure	Cellular component biogenesis (GO:0044085), extracellular structure organization (GO:0043062), cellular macromolecular complex subunit organization (GO:0034621), regulation of cell shape (GO:0008360)	17
Energy metabolism	Hexose metabolic process (GO:0019318), UDP-glucose metabolic process (GO:0005996), glucose metabolic process (GO:0006006), nucleotide-sugar metabolic process (GO:0009225), UDP-glucuronate biosynthetic process (GO:0006065), UDP-glucuronate metabolic process (GO:0046398), hexose biosynthetic process (GO:0019319)	10
Growth and development	Developmental growth (GO:0048589), regulation of anatomical structure size (GO:0090066), developmental growth involved in morphogenesis (GO:0060560), heart development (GO:0007507), bone development (GO:0060348), bone morphogenesis (GO:0060349), fin development (GO:0033333), regulation of developmental process (GO:0050793)	11
Immunity	Platelet degranulation (GO:0002576), acute-phase response (GO:0006953), antigen processing (GO:0002504), inflammatory response (GO:0006954)	10
Iron homeostasis	Cellular metal ion homeostasis (GO:0006875), metal ion homeostasis (GO:0055065), iron ion homeostasis (GO:0055072), ion homeostasis (GO:0050801)	4
Muscle contraction	Positive regulation of muscle contraction (GO:0045933), regulation of smooth muscle contraction (GO:0006940)	1
Nervous system development and learning	Associative learning (GO:0008306), learning (GO:0007612), cognition (GO:0050890), learning or memory (GO:0007611), regulation of neurogenesis (GO:0050767), regulation of nervous system development (GO:0051960)	6
Protein catabolism	Regulation of peptidase activity (GO:0052547), skeletal muscle tissue regeneration (GO:0043403), peptide cross-linking (GO:0018149), positive regulation of peptidase activity (GO:0010952)	6
Protein synthesis	Skeletal muscle fibre development (GO:0048741), protein complex subunit organization (GO:0071822)	1

*Fisher's exact test, Q-value <0.05.
†Functional groups as defined in the Materials and methods section.
‡Each gene can be associated with multiple functional groups.

outliers is not significantly different than that of outliers in the normal population (t-test, $P = 0.096$).

According to a function enrichment analysis performed by BLAST2GO on these 210 outlier genes, 78 biological processes classified into 12 general functional groups (see Materials and methods) and associated with 72 different genes were over-represented among outliers compared with their frequencies among all assembled genes (Fisher's exact test, Q-value <0.05, Tables 3 and S5, Supporting information). Of particular interest among outliers, two nuclear genes (prostaglandin-E synthase 2 like, average $F_{ST} = 0.81$ and glutathione peroxydase, $F_{ST} = 0.91$), involved in energy metabolism, were the two most differentiated genes between dwarf and normal whitefish. More specifically, three high-quality SNPs were identified in prostaglandin-E synthase 2 like (Fig. 4). Two of them are nonsynonymous mutations: the first one, L21V, located in putative exon 1, changes a leucine into a valine ($F_{ST} = 0.84$) and the second one, M115V, located in putative exon 4, changes a methionine into a leucine ($F_{ST} = 0.67$). None of these two substitutions are predicted to have a deleterious impact on protein function, according to the functional analysis performed using PolyPhen-2 (PolyPhen scores of 0.013 and 0.001, respectively, mutations are predicted to be benign). The third mutation was located in putative intron 4 ($F_{ST} = 0.90$). The second most divergent gene, glutathione peroxydase, showed one high-quality SNP located in putative exon 2 (Fig. 5). This was a synonymous substitution in the codon GTG associated with a

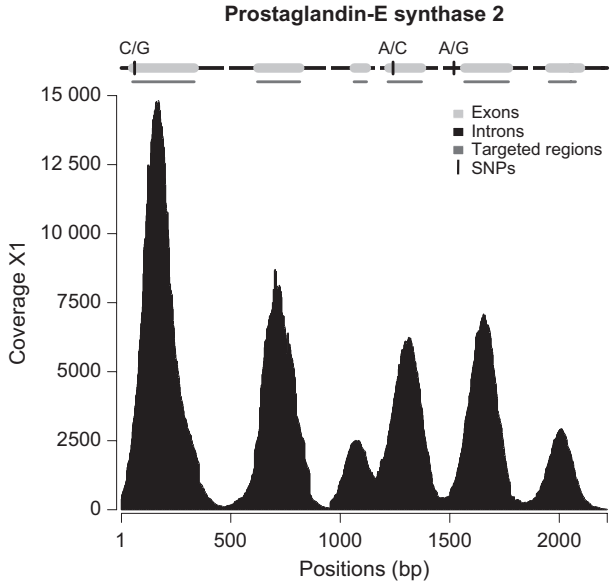

Fig. 4 Assembled prostaglandin-E synthase 2 (PGE-2 synthase) with single-nucleotide polymorphism (SNP) and probe positions. Depth of coverage (X1) is plotted against positions (base pairs) on the gene sequence. Above the graph is a representation of the assembled genomic sequence of PGE-2 synthase, with noncoding regions (black), exons (light grey), regions targeted by the probes (dark grey) and SNPs (vertical black lines). Incomplete black lines between annotated exons indicate the presence of gaps in putative introns.

valine for the alternative codon GTC, a less common codon used for the same amino acid (Codon Usage Database, http://kazusa.or.jp/codon/).

Comparison with previous whitefish studies

Other studies have also identified several genes potentially involved in the process of adaptive divergence of lake whitefish. Among 267 outlier loci identified in this study, 32 of them, representing the same number of candidate genes, have been labelled as divergent in previous work (Table S6, Supporting information). Seven of these genes were also previously identified by Renaut *et al.* (2011) in a genome scan-based approach using F_{ST} estimates from natural populations in multiple lakes. Although these loci that have been identified as outliers, absolute F_{ST} estimates differed between studies (Renaut *et al.* 2011) (Table 4).

Discussion

In a research context in which new technologies are constantly developed, sequencing costs are dramatically decreasing and theoretical frameworks are getting refined, it is now feasible to perform more in-depth analyses of patterns of genetic differentiation between related taxa spanning the speciation continuum. New research avenues taking advantage of this progress are required towards deciphering the respective roles of evolutionary processes involved in each phase of speciation (Nosil & Feder 2012). The lake whitefish system offers this possibility with a series of lakes harbouring sympatric populations of partially isolated ecotypes. Moreover, these lakes occupy different positions on the continuum of morphological and genomic differentiation (Renaut *et al.* 2011, 2012; Gagnaire *et al.* 2013b). Here, by conducting a large-scale targeted enrichment genome scan in the lake showing the most divergent species pair in the study system, we observed that neither synonymous nor nonsynonymous mutations were over-represented among outlier SNP loci, while more than half of the divergent SNPs were noncoding mutations. Based on numerous previous studies focusing on morphology, QTL mapping and gene expression differences among whitefish species pairs, observations reported in this study suggest that genetic divergence driven by selection might be more oriented towards noncoding and perhaps regulatory changes, compared with structural changes, as discussed below.

Sequence capture applied to a salmonid genome

We successfully applied the technique of sequence capture to enrich sequence and assemble several thousand genomic regions of the whitefish genome in a *de novo* context. Despite excellent results in terms of proportion of genes captured (98%), percentage of on-target reads was low (11.8%) compared with other studies (33.5% and 54%, respectively, in *Heliconius* and pea aphid, Smadja *et al.* 2012; Nadeau *et al.* 2012). A plausible explanation for this result is the use of a nonspecific DNA blocker solution prior to the hybridization step, combined with the complexity of whitefish genome. In the case of targeted enrichment of human DNA, preparations of C_0t1 DNA comprising short fragments (50–300 bp) of human placental DNA enriched for repetitive elements are added to the hybridization reaction in order to improve capture specificity (Mamanova *et al.* 2010). Because no whitefish-specific DNA blocker was available, solutions of human C_0t1 and PlantCaptureEnhancer (Roche) were used. Such blockers might not have been efficient enough in preventing secondary and nonspecific capture on the array. In the context of a large, duplicated salmonid genome, the use of a nonspecific blocker solution might have resulted in lower capture efficiency, compared with smaller and less repetitive genomes that are fully sequenced, like *Heliconius* and *Acyrthosiphon pisum* genomes. However, it could be possible to design a custom-built blocker solution for any species for which a sufficient database of

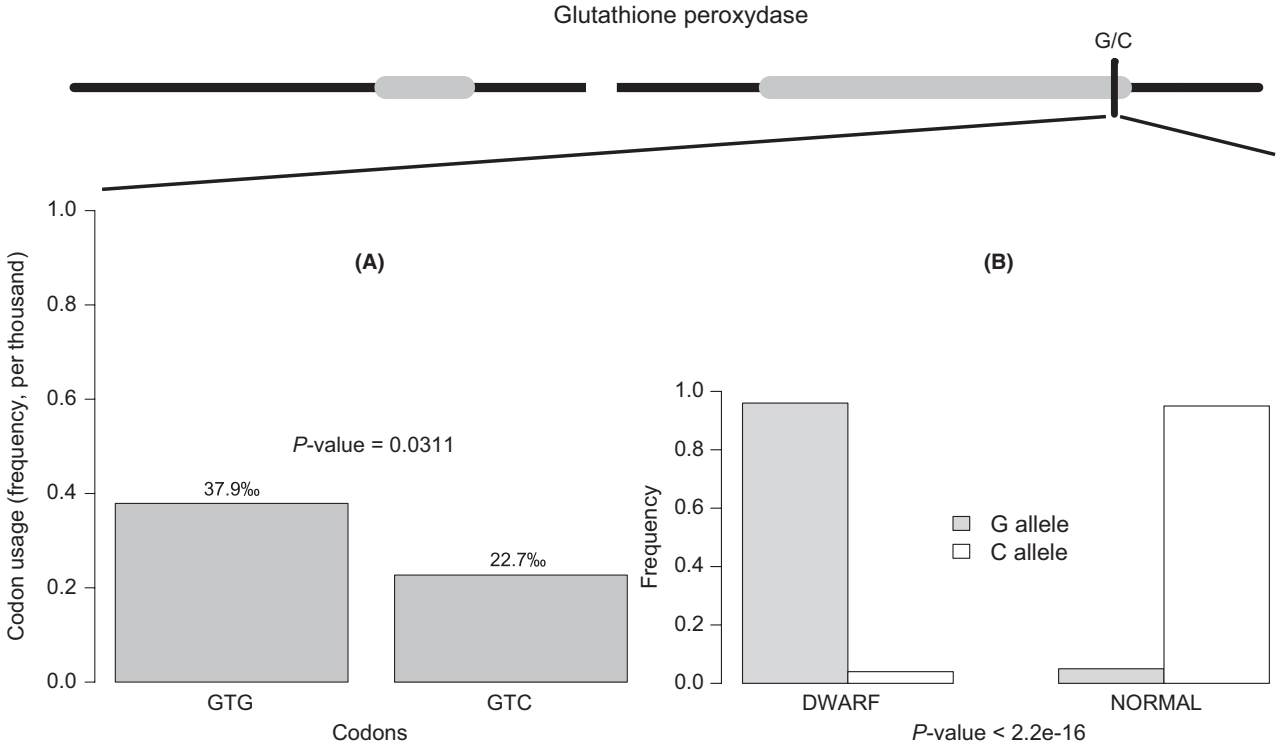

Fig. 5 Assembled sequence (partial) for glutathione peroxidase. Figure shows the first two exons in light grey, incomplete noncoding regions in black (with gap in white) and an outlier synonymous SNP in exon 2. (A) Codon usage found in *Coregonus clupeaformis* (Codon Usage Database) associated with both codons generated by each allele. (B) Allele frequencies in each population.

Table 4 Outlier genes also labelled as outliers in a previous genome scan[†]

Gene product	F_{ST} - SNP genotyping study	F_{ST} - Sequence capture[‡]
Fibrinogen beta chain	0.47	0.17
Ubiquitin carboxyl-terminal hydrolase isozyme L1	0.93*	0.20
Cyclin I	0.53***	0.18
Heat-shock protein HSP 90-beta	0.44	0.55
Red protein	0.76**	0.69
Sodium/potassium transporting ATPase subunit alpha-1	0.71**	0.05
Sodium/potassium transporting ATPase subunit beta-233	0.27****	0.42

Gene products in bold = no significant difference in F_{ST} estimates between both studies (Fisher's exact test, $P > 0.1$).
P-values calculated as explained in Renaut *et al.* 2011 *$P < 0.05$, **$P < 0.1$, ***$P < 0.2$
[†]Genome scan based on genotype information at 96 SNP loci, Renaut *et al.* (2011).
[‡]$P < 0.05$ for all genes, calculated by LOSITAN as described in Materials and methods.

repeated sequences is available. A home-made DNA solution analogous to the C_0t1 DNA preparation can be produced using this sequence information, which could help improve the on-target percentage. Nonetheless, other than requiring more sequencing effort, decreased capture efficiency was not immensely problematic as 98% of the genes were captured and assembled, mean coverage per gene was very high, and the depth of coverage for the filtered SNPs was uniform among individuals.

Genetic differentiation between sympatric populations

Using predefined criteria for selecting SNP loci with high-confidence genotypes for a sufficient number of individuals in each population, we identified a lower rate of polymorphism than expected according to previous work based on 454 sequencing (Renaut *et al.* 2010). Here, highest mean number of SNPs per kilobase was 18.7, whereas Renaut *et al.* (2010) identified a maximum of 44.8 SNPs per kilobase using cDNA. Difference in level of polymorphism is not surprising considering the discrepancy between data sets, assembly parameters and SNP detection methods used. Renaut *et al.* (2010)

used a partial non-normalized transcriptome assembly to call SNP genotypes with a significantly lower coverage threshold (minimum of 6X compared with 64X in this study). The array developed here spans a larger number of genes and is considerably more precise, resulting in a high-quality assembly that was optimized for paralogue detection (see Materials and methods). Genes with highest level of polymorphism in Renaut *et al.* (2010) were mainly ribosomal genes, which have a high probability of being paralogous sequences. Difference between both studies also stems from the fact that we used genomic DNA to perform the assembly and to find and call SNP genotypes instead of cDNA. First, sequencing errors committed during cDNA library construction could explain these biased estimates. It has also been suggested that differential expression of paralogues could be responsible for biased sequence polymorphism estimates when using cDNA data. Such a difference in levels of polymorphism was observed by Gagnaire *et al.* (2012b) in *Anguilla rostrata* (American eel) where cDNA yielded significantly higher estimates of levels of polymorphism and genetic differentiation compared with gDNA. Similarly, a recent study on sockeye salmon (*Oncorhynchus nerka*) also reported abnormally high levels of polymorphism detected in cDNA sequences, due to the presence of paralogous sequence variants (Everett *et al.* 2011). Our results thus re-emphasize that measures of levels of polymorphism and population differentiation based on cDNA genotyping are likely to be highly biased and should be interpreted with caution.

Population differentiation estimated for each locus (F_{ST} values) revealed an overall lower level of genetic differentiation compared with previous findings (Renaut *et al.* 2011). Moreover, distribution of F_{ST} values for heterozygosities ranging from 0 to 0.25 showed grouped data separated by possible missing values (Fig. 2). This could be due to presence of false homozygous genotypes that were mislabelled due to low individual coverage. The use of SAMTOOLS program mpileup should correct genotype calls for low coverage data (Li & Durbin 2009), but some true heterozygous genotypes might have been missed, although using a different method to estimate heterozygosity and changing basic SNP filters to eliminate these problematic loci did not seem sufficient enough to close the gap (Fig. S2, Supporting information). Such a trend was also shown by Smadja *et al.* (2012), who used the same capture technology and a similar genome scan method (see Fig. 1 in Smadja *et al.* 2012). Here, estimated average multilocus F_{ST} was 0.046, whereas mean F_{ST} values calculated in Cliff Lake based on two different SNP data sets, AFLPs and microsatellites were, respectively, 0.28, 0.22, 0.22 and 0.26 (Lu & Bernatchez 1999; Campbell & Bernatchez 2004; Renaut *et al.* 2011; Gagnaire *et al.* 2013b). Yet, F_{ST} values calculated in this study exhibited a very wide range, and some loci showed a clear sign of high genetic differentiation (F_{ST} up to 0.90, Fig. 1), concordant with previous findings (Renaut *et al.* 2011). A considerable proportion of loci (91%) were nonetheless mildly or weakly genetically differentiated ($F_{ST} < 0.2$, Fig. 1). Observed differences in overall F_{ST} values with previous studies can partly be explained by the use of different types of markers. Microsatellites are selected based on their high level of polymorphism, which reflects high mutation rates. Moreover, AFLPs, microsatellites and RAD-SNPs are three types of markers that largely consist of noncoding DNA (Ellegren 2004; Meudt & Clarke 2007; Hodgkinson & Eyre-Walker 2011; Gagnaire *et al.* 2013b). As such, they could diverge at a faster rate, as opposed to SNPs identified in or near coding regions, which are under stronger purifying selection. With respect to the SNP genotyping study of Renaut *et al.* (2011), who used a limited number of 100 SNPs and calculated an F_{ST} of 0.28, their markers were partly biased because they were chosen based on polymorphism rate and also on predicted putative role in explaining divergence. As such, we argue that the mean overall divergence we observe here may be more reflective of the true patterns of genetic differentiation observed in nature at coding genes. Given the recent divergence of whitefish sympatric species pairs, we hypothesize that balancing selection might also play a significant role in maintaining a lower level of differentiation at coding genes underlying the expression of numerous traits still common to both dwarf and normal whitefish, compared with noncoding regions. Although the action of balancing selection in this particular case remains speculative, this is an idea that deserves further investigation. More importantly, as we discuss below, the highly heterogeneous pattern of genetic divergence confirms the highly porous nature of genomes during speciation, where locally adapted populations of normal and dwarf whitefish can exchange genes across much of the genome while remaining morphologically and ecologically distinct.

Detection of outlier loci

Previous efforts to document 'genome-wide' patterns of functional divergence in whitefish were based on patterns of gene expression (Derome *et al.* 2008; St-Cyr *et al.* 2008; Whiteley *et al.* 2008; Renaut *et al.* 2009; Jeukens *et al.* 2010). Here, our results corroborate previous interpretations based on transcriptomics whereby a vast array of genes associated with various biological processes, each of small effect, are involved in divergence between dwarf and normal whitefish. Because no

reference genome is yet available in any salmonid species, those genes could not be positioned on chromosomes. Therefore, at this point, available information on genes labelled as outliers suggests that they might be part of genomic regions of reduced gene flow between whitefish ecotypes, although specific information on the number and size of 'islands' or 'continents' has to be confirmed. Yet, among this global set of outliers accounting for almost 8% of targeted genes, 78 biological processes were over-represented (Table S5, Supporting information). Using QTL mapping, Rogers & Bernatchez (2007) and Gagnaire *et al.* (2013a) have associated several genomic regions with key phenotypes contributing to whitefish adaptive divergence. These loci were distributed in many different linkage groups, suggesting that several genomic regions could be under the influence of divergent selection. In line with this knowledge, we thus propose that outliers identified here fall in some of these various genomic locations that might be part of islands of divergence, displaying variable levels of differentiation.

Among all five lakes in this whitefish system, Cliff Lake, from which samples have been collected for this study, harbours the most differentiated sympatric populations (Lu & Bernatchez 1999; Bernatchez *et al.* 2010; Renaut *et al.* 2011). Here, we confirm that the reduction in heterozygosity for outlier loci in the dwarf population observed by Renaut *et al.* (2011) also applies to many other genes not previously analysed. Their results strongly suggest an absence of ancestral bottleneck in the dwarf population of Cliff Lake and pointed out the fact that locus-specific reduction in heterozygosity is often related to the action of positive selection. Thus, both of these studies make a strong point regarding the role of directional selection acting predominantly on standing genetic variation of dwarf whitefish in driving adaptive divergence between whitefish ecotypes. Although we might have identified significant statistical outlier genes that are included in major genomic regions of differentiation and that might be directly or indirectly selected, selection acting on standing genetic variation and on polygenic phenotypes could also be responsible for soft or incomplete sweeps that remain undetected (Pritchard & Di Rienzo 2010; Pritchard *et al.* 2010; Le Corre & Kremer 2012). Therefore, outliers as identified in this study could only represent a small fraction of all the genes involved early in the divergence process.

Integration of previous knowledge & functional analysis

Integrating our genome scan results with data from previous studies based on QTL, gene expression, phenotypic information and preliminary SNP data refines and strengthens our understanding of the genetic mechanisms involved early in the speciation process. Gene expression studies have demonstrated that adaptation to the benthic (normal ecotype) and limnetic (dwarf ecotype) niches could be linked with differences in expression of genes involved in many biological functions (St-Cyr *et al.* 2008; Jeukens *et al.* 2009; Nolte *et al.* 2009; Renaut *et al.* 2009). Here, after performing an enrichment analysis for over-representation of biological functions among outliers, 72 genes (27% of outliers) belonged to 10 of 12 global functional groups previously identified by St-Cyr *et al.* (2008) as differentially expressed in a microarray experiment conducted on liver tissues. These functional groups also reflect numerous potentially adaptive physiological processes between whitefish ecotypes, including energy metabolism, cell cycle regulation, growth and development, muscle contraction, immunity, protein synthesis and behaviour (Table 3 and S5, Supporting information). In total, 30 genes showing outlier patterns of genetic differentiation between dwarf and normal whitefish had been previously identified in 14 studies combining gene expression, eQTL and genome scan analyses (Table S6, Supporting information), a proportion significantly greater than expected by chance (chi-square test, $P < 2.2e\text{-}16$). Again, the same functional categories were represented among those genes, namely energy metabolism (six genes), cell cycle regulation (nine genes), protein synthesis (seven genes) and immunity (five genes). The fact that we have identified genes showing divergence both in terms of sequence polymorphism (albeit mainly in noncoding regions), as evidenced in this study, and patterns of gene expression strengthens their role as strong candidate genes.

In particular among outliers, two genes distinguished themselves as the most divergent ones in terms of differential allele frequencies and F_{ST} values: prostaglandin-E synthase 2 and glutathione peroxydase (Figs 4 and 5). In the first case, prostaglandin-E synthase 2 (PGE_2 synthase) produces an enzyme involved in the production of prostaglandin-E 2 (PGE_2). In fishes, prostaglandins are found in many different cells and tissues, like macrophages, red blood cells and oocytes (Stacey & Goetz 1982; Cagen *et al.* 1983; Pettitt *et al.* 1991). A study conducted on copper rockfish (*Sebastes caurinus*) showed that one of the main target tissues of PGE_2 is the liver, where it efficiently stimulates glycogenolysis and gluconeogenesis, suggesting direct metabolic and endocrine roles for this prostanoid (Busby *et al.* 2002). Data also suggest that high levels of endogenous glucose in fish liver are due to a lack of regulation of gluconeogenesis by dietary carbohydrates (Enes *et al.* 2009). The product of PGE_2 synthase

thus helps to maintain this high endogenous glucose production in fish liver by controlling glycogenolysis and gluconeogenesis, two essential physiological processes involved in energy production. Here, even though both nonsynonymous mutations identified in the coding region of PGE_2 synthase did not have any deleterious phenotypic effect or any change in the global polarity of the molecule, they might be an example of changes with relatively small, yet significant effect. They might also be linked with neighbouring loci that are the direct targets of selection and that have not been identified in this study. Considering previously identified trade-offs in energy allocation between dwarf (high metabolic rates) and normal whitefish (low metabolic rates, Rogers & Bernatchez 2007; St-Cyr *et al.* 2008), this gene represents a good candidate for further investigation involving functional analyses at the protein level and in-depth tri-dimensional modelling of the protein.

The second most divergent gene, glutathione peroxydase, is involved in the protection of the organism from oxidative damage (Grim *et al.* 2011). Here, one SNP located in the coding region showed a high F_{ST} value (0.91), although it did not change the amino acid sequence (Fig. 5). Despite the fact that synonymous SNP do not alter coding sequences, data in many different species have suggested strong codon usage bias (Sharp & Li 1987; Gu *et al.* 2004; Dass & Sudandiradoss 2012). For example, it has been demonstrated in humans that silent polymorphism can alter the function of a gene. Kimchi-Sarfaty *et al.* (2007) hypothesized that the replacement of a frequent codon by a less frequently used codon can affect the timing of cotranslational folding of the protein, resulting in an altered function. Here, reference allele for the only nonsynonymous SNP found in the last exon of glutathione peroxydase encodes the most frequent codon for valine (GTG), which has a usage frequency of 3.79% in lake whitefish according to the codon usage database (http://kazusa.or.jp/codon/). Alternate allele for this SNP encodes a less frequently used codon (GTC) for valine (frequency of 2.27%, http://kazusa.or.jp/codon/). Interestingly, in dwarf whitefish, frequency of the most frequent codon was 0.96, whereas in the normal population, it drops to 0.05. Assuming that there is a functional link between timing of cotranslational folding and codon usage (Kimchi-Sarfaty *et al.* 2007), we hypothesize that the major difference in allele frequencies for glutathione peroxydase between dwarf and normal whitefish is responsible for a slight change in translational timing: a less frequently used codon is harder to recruit and thus lowers translation rate. Once again, this reflects the role of changes with small, yet meaningful, effects on phenotypes.

Implications of noncoding variation

Based on the type of changes in the most differentiated genes, on the fact that there are few nonsynonymous outlier SNPs (only 3 nonsynonymous mutations in the top 40 most differentiated SNP loci, Tables 2 and S4, Supporting information) and on previous studies focusing on whitefish transcriptomics, results suggest a greater role of noncoding variation, possibly associated with the regulation of gene expression (St-Cyr *et al.* 2008; Whiteley *et al.* 2008; Jeukens *et al.* 2010). Noncoding changes would thus better explain divergence between dwarf and normal whitefish than functional mutations. Interestingly, glutathione peroxydase and two prostaglandin genes have been previously identified as differentially expressed in gene expression studies (see Table S1, Supporting information in Bernatchez *et al.* 2010). Previous results on cytosolic malate dehydrogenase (MDH1) in whitefish also showed that a noncoding mutation located in the regulatory region of the gene was more likely to be involved in metabolic divergence between dwarf and normal ecotypes than mutations in protein-coding regions (Jeukens & Bernatchez 2012). Concurrently, the predominance of noncoding and synonymous outlier mutations argues that many changes of small effects have a significant importance during early speciation rather than large effect mutations (Fisher's infinitesimal model, reviewed in Bulmer 1980). Considering the direct impact of nonsynonymous mutations on amino acid composition and their potential to generate adaptive functional changes within proteins (McDonald & Kreitman 1991), we would expect to find a greater proportion of these changes within outlier group compared with noncoding and synonymous mutations. This prediction is valid if population divergence mostly depends upon functional changes in protein sequences. However, despite the functional importance of nonsynonymous SNPs and the fact that our capture efforts were directly and specifically focused on coding regions, we did not find any over-representation of nonsynonymous substitutions or any coding substitutions among outlier SNPs. A recent study using sequence capture to enrich 50 human exomes found, among nearly 20 000 genes, that the only mutation apparently responsible for adaptation to high altitude is located in an intron (Yi *et al.* 2010). Their results support the idea that small changes (involving in this case a noncoding mutation) can have significant effect in particular adaptations, which seems to be the most likely scenario in whitefish. Several theoretical models recently developed and confirmed by empirical data also suggest that evolution and adaptive changes most often occur via many different small-effect polygenes (Rockman 2011; Burke 2012). This

could explain how numerous and widespread changes of small effect, observed at the gene level in our data, contribute to the adaptation and evolution of species. An alternative hypothesis to explain these results is that in Cliff Lake, the most advanced lake in our system in terms of divergence between ecotypes, it is harder to detect outliers as 'islands of divergence' have become 'continents of divergence' (Gagnaire *et al.* 2013b).

Complex genomes found in species like salmonid fishes are characterized by massive amounts of noncoding DNA, once referred to as 'junk DNA'. Noncoding regions, comprising 5′ and 3′ UTRs, introns and intergenic regions, are nonetheless essential elements in gene regulation and expression profiles, which confers them major evolutionary roles (Barrett *et al.* 2012). For example, most significant changes observed between humans and other primates are found in noncoding regions (King & Wilson 1975; Pollard *et al.* 2006). Increasing research attention is being directed towards the characterization of noncoding elements. The ENCODE project reflects this shift of perspective by aiming at identifying all functional elements in the human genome, including regions of transcription, transcription factor association, chromatin structure and histone modification (ENCODE 2012). The hypothesis proposed is that few changes affecting regulatory processes have the potential to generate enormous variations, which is a faster and easier way to induce significant phenotypic changes than the accumulation of mutations in protein-coding regions (Gibson & Weir 2005; Stern & Orgogozo 2008). A recent study documenting genome-wide patterns of divergence in threespine stickleback, for which previous research efforts have largely focused on the identification of genes with major effect (Colosimo *et al.* 2005; Shimada *et al.* 2011), revealed that most of the genetic differentiation observed between marine and freshwater forms was located in noncoding regions (Jones *et al.* 2012). Their results suggest that repeated evolution of freshwater sticklebacks is best explained by regulatory changes, which are predominant over coding mutations.

Relevance of sequence capture in nonmodel organisms

Understanding patterns of genomic divergence in an ecological context requires precise sequence information on coding regions. These genomic regions are very informative in terms of amino acid sequence and thus protein structure and function, although they only represent 1–3% of the common eukaryotic genome (Lynch 2007). On the other hand, despite the known influence of protein-coding regions in producing various phenotypes, much work is needed to elucidate the nature of genomic regions involved in population differentiation during speciation. Explicit comparisons between coding and noncoding regions are thus needed (Nosil & Feder 2012). Our results demonstrate that sequence capture can simultaneously address these questions by efficiently targeting a very small and informative fraction of a large and complex genome with minimum a priori information. By targeting protein-coding regions, it is possible to retrieve sequence information in noncoding regions with sufficient depth of coverage to identify SNP markers that will allow significant and insightful comparisons between coding and noncoding regions, as discussed above. Detailed genomic information can thus be gained in any nonmodel organism without a reference genome, given that some preliminary information is available. Even though array design, development and optimization can be time-consuming, this technology significantly reduces the complexity of analysing and comparing massive genomic data sets (Grover *et al.* 2012).

Although whole-genome resequencing is increasingly feasible for species with small and simple genomes (e.g. Burke *et al.* 2010; Jones *et al.* 2012), it will remain a challenging and tedious task for many species. After more than five years of ongoing efforts on two species (*Salmo salar* and *Oncorhynchus mykiss*), we still do not have at this date a complete reference genome for any salmonid species (Davidson *et al.* 2010; Palti *et al.* 2011; Bernardi *et al.* 2012). Consequently, it is very unlikely that whole-genome resequencing will soon be widely used for such species. On the other hand, recent high-throughput techniques such as RAD-tag sequencing can potentially challenge the relevance of sequence capture. However, one of the main benefits of sequence capture over RAD tag in any evolutionary study is the potential to massively enrich one or many genomic regions of known identity and also in unknown flanking regions as well, while RAD tag provides information on random and anonymous regions. Thus, RAD tag will allow the identification of a widely distributed set of anonymous SNP markers (but also a small proportions of coding genes) generated by a particular restriction enzyme at a high density throughout the genome without any a priori information. In addition of the fact that sequence capture needs some a priori sequence information (e.g. ESTs, annotated transcriptome), it does not offer such a wide distribution of markers. Nonetheless, it has the potential to retrieve in-depth information on many known loci; it all depends on the research objectives. Both techniques can even be complementary as they achieve very distinct goals. Another limitation in this study is the restricted number of targets initially available for probe design. When the array was designed, no complete

whitefish transcriptome was available, which is one of the prerequisites to achieve exon capture and to conduct a complete genome-wide SNP survey. Further array designs could take advantage of increased transcriptome coverage when it becomes available. Lastly, the in-solution hybrid-capture approach seems to perform better than two other methods, namely array-based (used in this study) and PCR-based enrichment (Day-Williams *et al.* 2011). The main advantage offered by an in-solution capture approach is the hybridizing time, which is significantly shorter than array-based capture, resulting in less nonspecific hybridization and less secondary capture (Day-Williams *et al.* 2011). Because this technology became available shortly after our final array was designed, the opportunity was not available to increase capture specificity by using an in-solution approach.

Conclusion

In conclusion, the technique of sequence capture applied to a salmonid genome allowed the identification of numerous coding and noncoding regions significantly differentiated between incipient species of lake whitefish. These loci represent numerous biological functions, as expected, based on the many phenotypic traits that differ between whitefish ecotypes. The aim of this study was to document the extent of sequence divergence specifically in coding regions using an efficient enrichment technique. Despite focused efforts on the identification of divergence in protein-coding regions, we found very few significant functional changes. The absence of a clear over-representation of nonsynonymous SNPs in outlier genes in conjunction with previous transcriptomic and phenotypic studies confirms in whitefish (i) the importance of many small changes of measurable effect at the gene level and (ii) the likely predominance of regulation of gene expression acting in the early process of adaptive divergence. While such genome scan results should be interpreted with caution (Vilas *et al.* 2012), integrated information on the number and identity of divergent loci showing various levels of genetic differentiation combined with accurate data on phenotypes and diverging time between populations will undoubtedly help to understand how genomes are moulded and modified during the process of diversification of life (Koonin 2012; Nosil & Feder 2012).

Acknowledgements

We thank P-A Gagnaire, S. Pavey and C. Sauvage for their precious comments and suggestions about analyses; E. Normandeau for incredible help with bioinformatics; and B. Boyle and J. St-Cyr for conducting the capture experiment at Service de Séquencage de l'IBIS. We are also grateful to Associate Editor C. Schlotterer and three anonymous referees for their critical inputs. This work was supported by postgraduate scholarships from the Natural Science and Engineering Research Council of Canada (NSERC) and the Fonds de Recherche Nature et Technologies Québec (FRNTQ) to F.O.H, an NSERC postdoctoral fellowship to S.R. and an NSERC Discovery grant and a Canadian Research Chair to L.B. This study is a contribution of the research programme of Québec-Océan.

References

Adzhubei IA, Schmidt S, Peshkin L *et al.* (2010) A method and server for predicting damaging missense mutations. *Nature Methods*, **7**, 248–249.

Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G (2008) LOSITAN: a workbench to detect molecular adaptation based on a Fst-outlier method. *BMC Bioinformatics*, **9**, 323.

Barrett LW, Fletcher S, Wilton SD (2012) Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cellular and molecular life sciences: CMLS*, **69**, 3613–3634.

Barton NH (1979) Gene flow past a cline. *Heredity*, **43**, 333.

Barton N, Bengtsson BO (1986) The barrier to genetic exchange between hybridising populations. *Heredity*, **57**(Pt 3), 357–376.

Beaumont M, Nichols R (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings Of The Royal Society Of London. Series B*, **263**, 1619–1626.

Bernardi G, Wiley EO, Mansour H *et al.* (2012) The fishes of Genome 10K. *Marine Genomics*, **7**, 3–6.

Bernatchez L (2004) *Ecological Theory of Adaptive Radiation An Empirical Assessment from Coregonine Fishes (Salmoniformes)* (eds Stearns SC, Hendry AP). Oxford University Press, New York, USA.

Bernatchez L, Dodson J (1990) Allopatric origin of sympatric populations of lake whitefish (*Coregonus clupeaformis*) as revealed by mitochondrial-DNA restriction analysis. *Evolution*, **44**, 1263–1271.

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

Buerkle AC, Gompert Z (2012) Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology*, **22**, 3028–3035. doi:10.1111/mec.12105.

Bulmer MG (1980) *The Mathematical Theory of Quantitative Genetics* (ed.Bulmer MG). Oxford University Press, USA.

Burke MK (2012) How does adaptation sweep through the genome? Insights from long-term selection experiments. *Proceedings Of The Royal Society B-Biological Sciences*, **279**, 5029–5038.

Burke M, Dunham J, Shahrestani P, Thornton K (2010) Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature*, **467**, 587–592.

Busby ER, Cooper GA, Mommsen TP (2002) Novel role for prostaglandin E2 in fish hepatocytes: regulation of glucose metabolism. *The Journal of Endocrinology*, **174**, 137–146.

Cadillo-Quiroz H, Didelot X, Held NL *et al.* (2012) Patterns of gene flow define species of thermophilic Archaea(ed.Barton NH). *PLoS biology*, **10**, e1001265.

Cagen LM, Qureshi Z, Nishimura H (1983) Synthesis of prostaglandin E2 and prostaglandin F2 alpha by toadfish red blood cells. *Biochemical and Biophysical Research Communications*, **110**, 250–255.

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.

Colosimo P, Hosemann K, Balabhadra S *et al.* (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science*, **307**, 1928–1933.

Conesa A, Gotz S, Garcia-Gomez JM *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Crête-Lafrenière A, Weir LK, Bernatchez L (2012) Framing the salmonidae family phylogenetic portrait: a more complete picture from increased taxon sampling. *PLoS ONE*, **7**, e46662.

Dass JFP, Sudandiradoss C (2012) Insight into pattern of codon biasness and nucleotide base usage in serotonin receptor gene family from different mammalian species. *Gene*, **503**, 92–100.

Davidson WS, Koop BF, Jones SJM *et al.* (2010) Sequencing the genome of the Atlantic salmon (*Salmo salar*). *Genome Biology*, **11**, 403.

Day-Williams AG, McLay K, Drury E *et al.* (2011) An evaluation of different target enrichment methods in pooled sequencing designs for complex disease association studies (ed.Song Y-Q). *PLoS ONE*, **6**, e26279.

Derome N, Bougas B, Rogers SM *et al.* (2008) Pervasive sex-linked effects on transcription regulation as revealed by expression quantitative trait loci mapping in lake whitefish species pairs (*Coregonus* sp., *Salmonidae*). *Genetics*, **179**, 1903–1917.

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics*, **5**, 435–445.

ENCODE (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **488**, 57–74.

Enes P, Panserat S, Kaushik S, Oliva-Teles A (2009) Nutritional regulation of hepatic glucose metabolism in fish. *Fish Physiology and Biochemistry*, **35**, 519–539.

Everett MV, Grau ED, Seeb JE (2011) Short reads and nonmodel species: exploring the complexities of next-generation sequence assembly and SNP discovery in the absence of a reference genome. *Molecular Ecology Resources*, **11**(Suppl 1), 93–108.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298.

Feder JL, Gejji R, Yeaman S, Nosil P (2012) Establishment of new mutations under divergence and genome hitchhiking. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, **367**, 461–474.

Felsenstein J (1981) Skepticism towards Santa Rosalia, or why are there so few kinds of animals? *Evolution*, **35**, 124–138.

Fu Y, Springer NM, Gerhardt DJ *et al.* (2010) Repeat subtraction-mediated sequence capture from a complex genome. *The Plant Journal: for Cell and Molecular Biology*, **62**, 898–909.

Gagnaire P-A, Minegishi Y, Zenboudji S *et al.* (2011) Within-population structure highlighted by differential introgression across semipermeable barriers to gene flow in *Anguilla marmorata*. *Evolution*, **65**, 3413–3427.

Gagnaire PA, Normandeau E, Bernatchez L (2012a) Comparative genomics reveals adaptive protein evolution and a possible cytonuclear incompatibility between European and American Eels. *Molecular Biology and Evolution*, **29**, 2909–2919.

Gagnaire P-A, Normandeau É, Côté C, Møller Hansen M, Bernatchez L (2012b) The genetic consequences of spatially varying selection in the panmictic American eel (*Anguilla rostrata*). *Genetics*, **190**, 725–736.

Gagnaire P-A, Normandeay É, Pavey SA, Bernatchez L (2013a) Mapping phenotypic, expression and transmission ratio distortion QTL using RAD markers in the Lake Whitefish (*Coregonus clupeaformis*). *Molecular Ecology*, **22**, 3036–3048.

Gagnaire P-A, Normandeay É, Pavey SA, Bernatchez L (2013b) The genetic architecture of reproductive isolation during speciation-with-gene-flow in Lake Whitefish species pairs assessed by Rad sequencing. *Evolution*, doi:10.1111/evo.12075.

Gibson G, Weir B (2005) The quantitative genetics of transcription. *Trends in Genetics: TIG*, **21**, 616–623.

Gompert Z, Parchman TL, Buerkle CA (2012) Genomics of isolation in hybrids. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, **367**, 439–450.

Grim JM, Hyndman KA, Kriska T, Girotti AW, Crockett EL (2011) Relationship between oxidizable fatty acid content and level of antioxidant glutathione peroxidases in marine fish. *The Journal of Experimental Biology*, **214**, 3751–3759.

Grover CE, Salmon A, Wendel JF (2012) Targeted sequence capture as a powerful tool for evolutionary analysis. *American Journal of Botany*, **99**, 312–319.

Gu W, Zhou T, Ma J, Sun X, Lu Z (2004) The relationship between synonymous codon usage and protein structure in *Escherichia coli* and *Homo sapiens*. *Bio Systems*, **73**, 89–97.

Hodges E, Xuan Z, Balija V *et al.* (2007) Genome-wide in situ exon capture for selective resequencing. *Nature Genetics*, **39**, 1522–1527.

Hodgkinson A, Eyre-Walker A (2011) Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics*, **12**, 756–766.

Hohenlohe PA, Bassham S, Currey M, Cresko WA (2012) Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **367**, 395–408.

Jacobsen MW, Hansen MM, Orlando L *et al.* (2012) Mitogenome sequencing reveals shallow evolutionary histories and recent divergence time between morphologically and ecologically distinct European Whitefish (*Coregonus* spp.). *Molecular Ecology*, **21**, 2727–2742.

Jeukens J, Bernatchez L (2012) Regulatory versus coding signatures of natural selection in a candidate gene involved in the adaptive divergence of whitefish species pairs (*Coregonus* spp.). *Ecology and Evolution*, **2**, 258–271.

Jeukens J, Bittner D, Knudsen R, Bernatchez L (2009) Candidate genes and adaptive radiation: insights from transcriptional adaptation to the limnetic niche among coregonine fishes (*Coregonus* spp., *Salmonidae*). *Molecular Biology and Evolution*, **26**, 155–166.

Jeukens J, Renaut S, St-Cyr J, Nolte AW, Bernatchez L (2010) The transcriptomics of sympatric dwarf and normal lake whitefish (*Coregonus clupeaformis* spp., *Salmonidae*) divergence as revealed by next-generation sequencing. *Molecular Ecology*, **19**, 5389–5403.

Jeukens J, Boyle B, Kukavica-Ibrulj I *et al.* (2011) BAC library construction, screening and clone sequencing of lake whitefish (*Coregonus clupeaformis*, *Salmonidae*) towards the elucidation of adaptive species divergence. *Molecular Ecology Resources*, **11**, 541–549.

Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55–61.

Kimchi-Sarfaty C, Oh JM, Kim IW *et al.* (2007) A "Silent" polymorphism in the MDR1 gene changes substrate specificity. *Science*, **315**, 525–528.

King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. *Science*, **188**, 107–116.

Koonin EV (2012) A half-century after the molecular clock: new dimensions of molecular evolution. *EMBO Reports*, **13**, 664–666.

Krasnov A, Koskinen H, Afanasyev S, Mölsä H (2005) Transcribed Tc1-like transposons in salmonid fish. *BMC Genomics*, **6**, 107.

Lawniczak MKN, Emrich SJ, Holloway AK *et al.* (2010) Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science*, **330**, 512–514.

Le Corre V, Kremer A (2012) The genetic differentiation at quantitative trait loci under local adaptation. *Molecular Ecology*, **21**, 1548–1566.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

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.

Lynch M (2007) *The Origins of Genome Architecture* (ed. Lynch M). Sinauer Associates Inc., Sunderland, USA.

Mamanova L, Coffey AJ, Scott CE *et al.* (2010) Target-enrichment strategies for next-generation sequencing. *Nature Methods*, **7**, 111–118.

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. *Genetical Research*, **23**, 23–35.

McDonald J, Kreitman M (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature*, **351**, 652–654.

Meudt HM, Clarke AC (2007) Almost forgotten or latest practice? AFLP applications, analyses and advances. *Trends in Plant Science*, **12**, 106–117.

Mitchell-Olds T, Willis JH, Goldstein DB (2007) Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nature Reviews Genetics*, **8**, 845.

Nadeau NJ, Whibley A, Jones RT *et al.* (2012) Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **367**, 343–353.

Nolte AW, Renaut S, Bernatchez L (2009) Divergence in gene regulation at young life history stages of whitefish (*Coregonus* sp.) and the emergence of genomic isolation. *BMC Evolutionary Biology*, **9**, 59.

Nosil P, Feder JL (2012) Genomic divergence during speciation: causes and consequences. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **367**, 332–342.

Nosil P, Egan S, Funk D (2008) Heterogeneous genomic differentiation between walking-stick ecotypes:" Isolation by adaptation" and multiple roles for divergent selection. *Evolution*, **62**, 316–336.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375–402.

Palti Y, Genet C, Luo M-C *et al.* (2011) A first generation integrated map of the rainbow trout genome. *BMC Genomics*, **12**, 180.

Payseur BA (2010) Using differential introgression in hybrid zones to identify genomic regions involved in speciation. *Molecular Ecology Resources*, **10**, 806–820.

Pettitt TR, Barrow SE, Rowley AF (1991) Thromboxane, prostaglandin and leukotriene generation by rainbow trout blood. *Fish & Shellfish Immunology*, **1**, 71–73.

Pollard KS, Salama SR, King B *et al.* (2006) Forces shaping the fastest evolving regions in the human genome. *Plos Genetics*, **2**, e168.

Presgraves DC (2010) The molecular evolutionary basis of species formation. *Nature Reviews Genetics*, **11**, 175–180.

Pritchard JK, Di Rienzo A (2010) Adaptation - not by sweeps alone. *Nature Reviews Genetics*, **11**, 665–667.

Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, **20**, R208–R215.

Radice AD, Bugaj B, Fitch DH, Emmons SW (1994) Widespread occurrence of the TC1 transposon family: Tc1-like transposons from teleost fish. *Molecular & General Genetics*, **244**, 606–612.

Renaut S, Nolte AW, Bernatchez L (2009) Gene expression divergence and hybrid misexpression between Lake Whitefish species pairs (*Coregonus* spp. *Salmonidae*). *Molecular Biology and Evolution*, **26**, 925–936.

Renaut S, Nolte AW, Bernatchez L (2010) Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. *Salmonidae*). *Molecular Ecology*, **19**(Suppl 1), 115–131.

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.

Renaut S, Maillet N, Normandeau E *et al.* (2012) Genome-wide patterns of divergence during speciation: the lake whitefish case study. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **367**, 354–363.

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

Rockman M (2011) The Qtn program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution*, **66**, 1–17.

Rogers SM, Bernatchez L (2007) The genetic architecture of ecological speciation and the association with signatures of selection in natural lake whitefish (*Coregonas* sp., *Salmonidae*) species pairs. *Molecular Biology and Evolution*, **24**, 1423–1438.

Rogers S, Gagnon V, Bernatchez L (2002) Genetically based phenotype-environment association for swimming behavior in lake whitefish ecotypes (*Coregonus clupeaformis* Mitchill). *Evolution*, **56**, 2322–2329.

Schluter D (2001) *The Ecology of Adaptive Radiation*(edsHarvey H, May RM). Oxford University Press, New-York, USA.

Sharp PM, Li WH (1987) The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, **15**, 1281–1295.

Shimada Y, Shikano T, Merila J (2011) A high incidence of selection on physiologically important genes in the three-spined stickleback, *Gasterosteus aculeatus*. *Molecular Biology and Evolution*, **28**, 181–193.

Simpson JT, Wong K, Jackman SD *et al.* (2009) ABySS: a parallel assembler for short read sequence data. *Genome Research*, **19**, 1117–1123.

Smadja CM, Canbäck B, Vitalis R *et al.* (2012) Large-scale candidate gene scan reveals the role of chemoreceptor genes in host plant specialization and speciation in the pea aphid. *Evolution*, **66**, 2723–2738.

Stacey NE, Goetz FW (1982) Role of prostagladins in fish reproduction. *Canadian Journal Of Fisheries And Aquatic Sciences*, **39**, 92–98.

Stanke M, Steinkamp R, Waack S, Morgenstern B (2004) AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Research*, **32**, W309–W312.

St-Cyr J, Derome N, Bernatchez L (2008) The transcriptomics of life-history trade-offs in whitefish species pairs (*Coregonus* sp.). *Molecular Ecology*, **17**, 1850–1870.

Stern DL, Orgogozo V (2008) The loci of evolution: how predictable is genetic evolution? *Evolution*, **62**, 2155–2177.

Strasburg JL, Sherman NA, Wright KM *et al.* (2012) What can patterns of differentiation across plant genomes tell us about adaptation and speciation? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **367**, 364–373.

Trudel M, Tremblay A, Schetagne R, Rasmussen JB (2001) Why are dwarf fish so small? An energetic analysis of polymorphism in lake whitefish (*Coregonus clupeaformis*). *Canadian Journal of Fisheries And Aquatic Sciences*, **58**, 394–405.

Turner T, Hahn M, Nuzhdin S (2005) Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology*, **3**, 1572–1578.

Via S (2012) Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **367**, 451–460.

Via S, West J (2008) The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Molecular Ecology*, **17**, 4334–4345.

Vilas A, Pérez-Figueroa A, Caballero A (2012) A simulation study on the performance of differentiation-based methods to detect selected loci using linked neutral markers. *Journal of Evolutionary Biology*, **25**, 1364–1376.

Whiteley AR, Derome N, Rogers SM *et al.* (2008) The phenomics and expression quantitative trait locus mapping of brain transcriptomes regulating adaptive divergence in lake whitefish species pairs (*Coregonus* sp.). *Genetics*, **180**, 147–164.

Wright S (1951) The genetical structure of populations. *Annals of Human Genetics*, **15**, 323–354.

Wu C (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–865.

Wu C, Ting C (2004) Genes and speciation. *Nature Reviews Genetics*, **5**, 114–122.

Yi X, Liang Y, Huerta-Sanchez E *et al.* (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, **329**, 75–78.

F.O.H. designed the sequence capture chip, carried out the *de novo* assembly, scripted the bioinformatic tools, performed molecular genetic and statistical analyses and drafted the manuscript. S.R. performed molecular genetic analyses and helped to draft the manuscript. L.B. conceived of the study, and participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Data accessibility

Complete sequences of captured and assembled genes, including annotations and SNP genotypes for all 24 individuals, are available in Supplementary Tables. Data also accessible through Dryad: complete assembly for 2728 genes (doi:10.5061/dryad.3nc54) and custom Python scripts specifically designed for data treatment and analysis (doi:10.5061/dryad.3nc54).

Supporting information

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

Fig. S1 Distribution of allelic frequency divergence between dwarf and normal based on a selection of 2302 SNPs. Polymorphic markers with an allelic frequency divergence above 0.5 and showing a *Q*-value >0.05 were considered as highly divergent. Vertical red dotted line = median (0.09). Allelic frequency divergence = $|\text{frequence}_{\text{A1-Dwarf}} - \text{frequence}_{\text{A1-Normal}}|$.

Fig. S2 F_{ST}-based scans for selection using the method of Beaumont & Nichols (1996) implemented in LOSITAN Workbench (extension to Lewontin-Krakauer test). Estimates of heterozygosity are calculated according to Weir and Cockerham (1984) as used in Beaumont & Nichols (1996). Different SNP filters based on depth of coverage and number of individuals respecting coverage threshold in each population were used. This set of outlier tests comprises two additional lanes of Illumina

sequencing in order to increase overall depth of coverage and to improve the detection of heterozygosity. CT = per individual coverage threshold. IND = number of individuals per population respecting the coverage threshold. a) CT = 4X – IND = 8, b) CT = 6X – IND = 8, c) CT = 8X – IND = 8, d) CT = 4X – IND = 10, e) CT = 6X – IND = 10, f) CT = 8X – IND = 10.

Table S1 List of all the genes successfully captured and assembled with detailed information on their genetic architecture, including exons positions, SNP positions, SNP genotypes and complete captured gene sequence.

Table S2 Complete information on analysed SNPs. Each row represents one SNP locus and contains information about SNP position within the gene sequence, coding status, genotype, synonymy, frequencies within both populations, allelic frequency divergence (with corresponding *Q*-value) and F_{ST} value.

Table S3 List of SNP genotypes for all 24 individuals (dwarf: 12, normal: 12). Reference and alternate alleles are given for each SNP.

Table S4 Complete information on outlier SNP loci. Excel spreadsheet is organized as in Table S2

Table S5 List of genes exhibiting biological processes over-represented among outlier group. Each gene is associated with its broad functional group(s) and corresponding biological processes (GO terms).

Table S6 List of genes identified as divergent in previous studies on sympatric whitefish species pairs.