MOLECULAR ECOLOGY

Molecular Ecology (2010), **19** (Suppl. 1), 115–131 doi: 10.1111/j.1365-294X.2009.04477.x

Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. Salmonidae)

SÉBASTIEN RENAUT,* ARNE W. NOLTE† and LOUIS BERNATCHEZ*

**IBIS (Institut de Biologie Intégrative et des Systèmes), Université Laval, QC, Canada G1V 0A6, †Max-Planck-Institute for Evolutionary Biology, August-Thienemann-Str., 2, 24306 Plön, Germany*

Abstract

Next-generation sequencing allows the discovery of large numbers of single nucleotide polymorphisms (SNPs) in species where little genomic information was previously available. Here, we assembled, *de novo*, over 130 mb of non-normalized cDNA using 454 pyrosequencing data from dwarf and normal lake whitefish and backcross hybrids. Our main goals were to gather a large data set of SNP markers, document their distribution within coding regions, evaluate the effect of species divergence on allele frequencies and combine results with previous genomic studies to identify candidate genes underlying the adaptive divergence of lake whitefish. We identified 6094 putative SNPs in 2674 contigs (mean size: 576 bp, range: 101–6116) and 1540 synonymous and 1734 non-synonymous mutations for a genome-wide non-synonymous to synonymous substitution rate ratio (p_N/p_S) of 0.37. As expected based on the young age (<15 000 years) of whitefish species pair, the overall level of divergence between them was relatively weak. Yet, 89 SNPs showed pronounced allele frequency differences between sympatric normal and dwarf whitefish. Among these, SNPs in genes annotated to energy metabolic functions were the most abundant and this, in addition to previous experimental data at the gene expression and phenotypic level, brings compelling evidence that genes involved in energy metabolism are prime candidates explaining the adaptive divergence of lake whitefish species pairs. Finally, we unexpectedly identified 44 contigs annotated to transposable elements and these were predominantly composed of backcross hybrids sequences. This indicates an elevated activity of transposable elements, which could potentially contribute to the reduced fitness of hybrids previously documented.

Keywords: adaptive divergence, hybridization, natural selection, next-generation sequencing, single nucleotide polymorphism, speciation, transcriptomics, transposon

Received 27 May 2009; revision received 18 August 2009; accepted 25 August 2009

Introduction

Next-generation sequencing technologies are rapidly transforming the field of ecology, evolution and genetics (Rokas & Abbot 2009). This avalanche of data promises to answer experimental inquiries ranging from ancient DNA sequencing (Miller *et al.* 2008), sequence variants discovery (Vera *et al.* 2008), microbial ecology (Dinsdale *et al.* 2008) as well as gene expression analysis (Torres *et al.* 2007; Lipson *et al.* 2009). High throughput pyrosequencing developed by 454 Life Sciences (Margulies *et al.* 2005) is of particular interest in ecology and evolution primarily because it yields longer sequencing reads than any other method (up to 600 bp), which allows more accurate *de novo* sequence assemblies often required for non-model organisms. The recent explosion of second- and third-generation sequencing (Branton *et al.* 2008; Shendure & Ji 2008; Metzker 2009) has led some researchers to believe that many technical approaches (e.g. Sanger sequencing, DNA microarrays), which where themselves revolution-

Correspondence: Sébastien Renaut, Fax: +1 418 656 717;
E-mail: sebastien.renaut.1@ulaval.ca

ary a decade or two ago, may already be obsolete today (Ledford 2008). Nevertheless, in order to unleash its full potential, these methods will require careful experimental design, consideration of the techniques' limitations and finally, innovative bioinformatics approaches to process and extract relevant information (Ellegren 2008; Rokas & Abbot 2009).

One of the primary goals of high throughput sequencing projects is to reveal sequence variation such as copy number variants, insertion-deletions (indels) or single nucleotide polymorphisms (SNPs) by sequencing pools of genetically heterogeneous individuals (Barbazuk *et al.* 2007; Vera *et al.* 2008; Wiedmann *et al.* 2008). SNPs are rapidly becoming popular genetic markers in ecology and evolution (Schlötterer 2004; Moen *et al.* 2008; Namroud *et* al. 2008). Their main attraction is that, contrary to most amplified fragment length polymorphisms (AFLP) markers, they can potentially be directly linked to candidate genes of known function and interest. Moreover, as opposed to microsatellites, which may have complex mutations patterns, their genotyping can be highly automated at moderate costs (Schlötterer 2004; Ehrich *et al.* 2005; Shen *et al.* 2005; Van Tassell *et al.* 2008). Lastly, unlike AFLP and microsatellites, SNP data can also easily be standardized across laboratories. Nevertheless, despite their abundance and genotyping automation, SNP markers development may involve several validation steps. Problems with successful SNP locus amplification, low-frequency polymorphisms or gene duplicates render the identification of reliable markers a non-trivial, potentially labour-intensive task (Fredman *et al.* 2004; Hayes *et al.* 2007; Namroud *et al.* 2008).

Identifying sequence variants in transcribed regions of the genome is of primary interest in an attempt to characterize the effects of selection on protein evolution. Sequence polymorphisms within a gene have different impacts depending on their exact genomic location (intron, exon, untranslated region). Mutations within coding regions are especially insightful as their effect on amino acid composition and therefore protein functionality can be easily assessed. Similar to d*n*/d*s* ratios, the rate of accumulation of non-synonymous polymorphism (p_N) scaled by the rate of synonymous polymorphism (p_S) provides a glimpse on the selective forces driving the evolution of a protein-coding sequence. Thus, genes with a high p_N/p_S (i.e. >1) ratio are likely to be evolving under the influence of positive selection (McDonald & Kreitman 1991; Axelsson *et al.* 2008; Ellegren 2008). Furthermore, if this is associated with phenotypically distinct populations, either through *de novo* mutations or sorting of standing genetic variation, such genes may represent candidates potentially involved in an adaptive divergence event.

Lake whitefish species pairs represent excellent model species to study the early onset of reproductive isolation and its effect on genomic divergence (Lu & Bernatchez 1998; Bernatchez 2004; Rogers & Bernatchez 2006; Nolte *et al.* 2009; Renaut *et al.* 2009). Geographic isolation during the Pleistocene caused genetic divergence between whitefish populations inhabiting distinct glacial refugia but without distinctive phenotypic divergence between glacial races in allopatry (Bernatchez & Dodson 1990, 1991). Secondary contact of these evolutionary lineages subsequently occurred 12 000 years BP and has led to the parallel evolution of two morphologically and ecologically divergent sympatric whitefish species in several lakes of northeastern North America: benthic *Normal* and limnetic *Dwarf* whitefish (Bernatchez & Dodson 1990, 1991; Pigeon *et al.* 1997). As expected from a recent divergence event, the overall level of genetic differentiation between species pairs is relatively weak (Bernatchez *et al.* 1999; Campbell & Bernatchez 2004) and hybrids can be found in nature (Lu *et al.* 2001; Falush *et al.* 2007). At the same time, it has been shown that intrinsic (genetic) and extrinsic (ecological) post-zygotic isolation mechanisms lead to a fitness decrease in hybrids (Lu & Bernatchez 1998; Rogers & Bernatchez 2006; Whiteley *et al.* 2009) and this is partially caused by gene deregulation (Renaut *et al.* 2009).

Genome scan studies using anonymous AFLP markers as well as markers linked to qualitative trait loci (QTLs) suggest that a small proportion of the whitefish genome (~1–2%) might be under the effect of directional selection in the process of adaptive population divergence (Campbell & Bernatchez 2004; Rogers & Bernatchez 2005, 2007). Identifying such key islands of genomic divergence and isolation (*sensu* Wu 2001) and, more specifically, candidate genes showing evidence of reduced gene flow may represent a daunting task, yet it offers priceless information to pinpoint the causative variations responsible for reproductive isolation and speciation (Wu & Ting 2004; Turner *et al.* 2005; Schluter 2009). Our ongoing research programme on the ecological functional genomics of whitefish adaptive divergence and speciation involves a combination of both gene mapping and genome scan aiming at identifying more precisely genomic region evolving under the effect of divergent selection in dwarf and normal whitefish. To this end, we herein sequenced the transcriptome of two sympatric dwarf and normal species of lake whitefish and backcross hybrids with four specific objectives; to gather a large data set of candidate SNP markers; secondly, to look at the distribution of these markers within coding regions; thirdly to evaluate the effect of species divergence on allele frequencies and fourthly, as an *a posteriori* objective, to evaluate rates of transposon activity among normal, dwarf and hybrid

whitefish. Our ultimate goal, linking all this information to previous genomic studies in this system (QTL, eQTL, genome scan and gene expression) as an attempt to establish functional and causal links between genotype, phenotype and natural selection, represents one of the main challenges of the 21st century in evolutionary biology (Schluter 2009).

Materials and methods

Sample preparation

RNA samples were isolated separately from 24 individuals and three different tissue types (white muscle, brain, liver), in order to get a diversified representation of genotypes and expressed genes (Table 1). All RNA samples came from previous gene expression studies and had been kept at −80 °C until thawed for this experiment. As such, fish rearing conditions, euthanasia procedure and RNA extraction protocols are described in details in St-Cyr *et al.* (2008) for pool D and N (liver tissue), Derome *et al.* (2008) (Pool BC: muscle tissue) and Whiteley *et al.* (2008) (Pool BC: brain tissue). Pool D and N respectively represent sympatric dwarf and normal whitefish from Cliff Lake. BC whitefish represent backcross hybrids involving dwarf whitefish from Témiscouata Lake and normal whitefish from Aylmer Lake that were previously used in gene and QTL mapping projects (Rogers & Bernatchez 2007; Rogers *et al.* 2007). In short, total RNA was extracted separately for each individual using the TRIzol Reagent protocol (Invitrogen). Following extraction, all samples were further cleaned by ultra filtration using microcon (Millipore) spin columns. Samples were quantified using Experion™ RNA StdSens Analysis Kit (Bio-Rad). Total RNA was stored in pure water supplemented with SuperaseIn™ RNase Inhibitor (Ambion) and kept at −80 °C.

Enrichment for polyA mRNA was conducted using MicroPoly(A)Purist™ Kit (Ambion). Approximately 100 ng of full-length complementary DNA was synthesized from each polyA mRNA sample following SMART™ PCR cDNA Synthesis Protocol (Clontech). All cDNA samples (3–8 ng) were PCR amplified using Advantage 2 PCR Kit (Invitrogen) and modified SMART™ primers (5′-AAGCAGTGGTATCAACGCAGAGT-3′), which comprised an extra five nucleotide at the 5′ end to serve as an individual specific tag. PCR conditions were as follow: initial denaturation for 1 min at 95 °C, followed by 17–20 cycles depending on sample (1 cycle: 15 s at 95 °C, 30 s at 65 °C, 6 min at 68 °C). Following amplification, all samples were quantified using Quant-iT Picogreen dsDNA Assay Kit (Invitrogen) and three separate pools with equal DNA quantities were prepared; Pool D and N consists of RNA extracted from liver of eight individuals (St-Cyr *et al.* 2008) each whereas Pool BC consisted of four white muscle (Derome *et al.* 2008) and four brain (Whiteley *et al.* 2008) tissue of backcross hybrids. Approximately 5 μg of double-stranded cDNA from each of three cDNA pools was sequenced (0.75 run per pool) on a Roche GS-FLX DNA Sequencer using methods previously described (Margulies *et al.* 2005) at the Genome Quebec Innovation Center (McGill University, Montreal, Canada).

Contig assemblies

Initial quality filtering of whitefish 454 sequences was performed using Roche proprietary analysis software Newbler (Margulies *et al.* 2005). Base calling was performed using PyroBayes, which produces more confident base calls than the native 454 base-calling programme (Quinlan *et al.* 2008). Prior to assembling all sequences, primers and sample specific tags sequences were removed from the data set using a custom made Perl script. CLC Genomics Workbench 3.1 (CLC Bio) was used to assemble sequences *de novo* (similarity 0.97, overlap 0.5). We performed several test

Table 1 Samples used for sequencing and data obtained from 454 GS-FLX pyrosequencing runs

Pool	Lineage	Tissue type	Number of individuals	Quantity sequenced	Number of reads	Length (mean/median)*
D	Cliff Lake Dwarf	Liver†	8	0.75 plate	183365	194/214
N	Cliff Lake Normal	Liver†	8	0.75 plate	210703	191/209
BC	[(Aylmer Lake normal × Témiscouata Lake dwarf) × Aylmer Lake normal]	Muscle‡ Brain§	4 4	0.75 plate	238409	195/216

*Length in nucleotides of read after primers and sample-specific tags were removed.
†N and D samples originally used by St-Cyr *et al.* (2008).
‡Muscle tissue was previously used by Derome *et al.* (2008).
§Brain tissue was previously used by Whiteley *et al.* (2008).

assemblies, based on parameters from recent transcriptome-sequencing studies (Barbazuk *et al.* 2007, >0.95 similarity index; Vera *et al.* 2008, >0.80; Zhao *et al.* 2009, >0.96], and found that using a similarity criterion too low (below 0.9) leads to the assembly of dissimilar sequences, riddled with paralogous sequence variants (PSVs) instead of true SNPs (data not shown). On the other hand, a highly restrictive one (above 0.98) discards too many sequences from the assembly (data not shown). Allowing for 3% mismatch was deemed a reasonable estimate based on relatively low whitefish polymorphism previously observed (1.4 SNPs/kb, Whiteley *et al.* 2008) and average pyrosequencing error (~0.5%, Margulies *et al.* 2005). Note also that our threshold should prevent the assembly of duplicated (paralogous) regions that trace back to an ancient salmonid genome duplication (25–100 Ma, Allendorf *et al.* 1975) as the latter would be expected to have 6–25% sequence divergence, based on a conservative estimate of ~0.25% nuclear sequence divergence per million years.

Consensus sequences were Matched (BLAST, Altschul *et al.* 1997) against a publicly available set of 32 000 salmonids cDNA (cGRASP database, http://web.uvic.ca/grasp/microarray/array.html) in BioEdit (Hall 1999) (BLASTN *e*-value <1e-50). This 32 000 cDNA database had been previously assembled from more than 700 000 EST (expressed sequence tags) sequences obtained from a variety of cDNA libraries. Hence, it should comprise the majority of all cDNA expressed at least in Atlantic salmon, a salmonid closely related to lake whitefish (von Schalburg *et al.* 2008). Mitochondrial genome from the European lake whitefish (*Coregonus lavaretus*) (Miya & Nishida 2000) was also used to verify the mitochondrial origin of candidate genes. Functional categories (gene ontology biological functions) for genes of interest were identified with either the information provided by the cGRASP database or searches on http://amigo.geneontology.org/ and http://www.uniprot.org

SNP discovery and functional characterization of polymorphism

Assembled contigs were screened for SNPs using the software CLC Genomics Workbench 3.1 under the following criteria; minimum coverage of SNP: 6X, and minimum frequency of the least frequent allele: 20%, whereas the remaining parameters were left as default. The analysis of SNP frequencies between normal and dwarf whitefish as well as other statistical tests were calculated in R (v. 2.8.1; The R Foundation for Statistical Computing®, 2009, 3-900051-07-0). Namely, allele frequencies were analysed to identify SNPs that showed significant divergent allelic frequencies between normal and dwarf whitefish (minimum coverage of SNP of 4X for normal and dwarf, Fisher's exact test corrected for multiple hypothesis testing by calculating *Q*-values from *P*-values distribution, Storey 2002). Following this, we arbitrarily defined strongly divergent SNPs as markers for which the frequency of an allele differed by more than 0.5 between populations (this index has a maximum value of 1) and *Q*-value <0.05.

Open reading frames (ORF) for each assembled contig were produced using the program *getorf* in EMBOSS (European Molecular Biology Open Software Suite, Rice *et al.* 2000). The longest open-ended ORF (minimum length of 200 nucleotides) was kept as the most probable translated region of the gene. Lastly, maximum likelihood was used to estimate the ratio of synonymous SNP per synonymous site against non-synonymous SNP per non-synonymous site using PAML 4.2 (runmode = 0, CodonFreq = 2, model = 2; Yang 2007).

Comparison with previous gene expression, QTL an eQTL studies

We used data from previous lake whitefish gene expression (Derome *et al.* 2006; St-Cyr *et al.* 2008; Nolte *et al.* 2009; Renaut *et al.* 2009), QTL and genome scans (Rogers & Bernatchez 2007) as well as eQTL mapping (Derome *et al.* 2008; Whiteley *et al.* 2008) studies to match their gene annotation with genes identified in this study. We provide a legend at the bottom of Table 3 as a summary of the different studies and the rationale for why they were considered as genes of particular interest.

SNP validation

A subset of polymorphic loci (31) were validated using matrix-assisted laser desorption/ionization time-of-flight mass spectroscopy (MALDI-TOF MS) assays (Sequenom) at Genome Quebec Innovation Center in order to test whether these markers were likely to be true SNPs rather than PSVs. Twenty-nine fish from a lake containing a single panmictic population of Normal whitefish, Lake Aylmer (45°54'N, 71°20'W), were genotyped. Deviation from Hardy–Weinberg equilibrium (chi-squared test corrected for multiple hypothesis testing, *Q*-value; Storey 2002) and expected heterozygosity [$F_{is} = (H_e - H_o)/H_e$] were calculated in R.

Results

Sequencing, contig assembly and annotation

A total of 632 000 sequences with a median length of 212 nucleotides/sequence and totalizing ~130 megabases were obtained from sequencing the D, N and BC

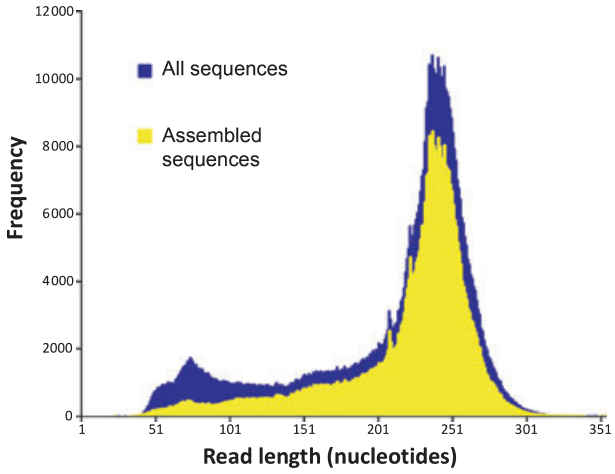

Fig. 1 Frequency distribution of the total number of reads (blue) and assembled ones (yellow).

separate pools of cDNA (0.75 GS-FLX sequencing run per pool; Fig. 1, Table 2 NCBI sequence read archive SRA 009800). By using a similarity criterion of 0.97, we assembled, *de novo*, 428 068 sequences out of 632 000 (68%) into 2674 separate contigs (Table 2), meaning that 32% of all sequences were left as unassembled singletons. Shorter reads were harder to assemble and usually discarded (Fig. 1). Mean contig length was 576 bp, with the smallest contig having a length of 101 and the longest 6116 bp. Coverage was also highly variable due to the fact that the cDNA sequences were not normalized (1.3X–4140X) as another goal of this research will be to document differential gene transcription between dwarf and normal whitefish from this same data set (Jeukens, J., Bernatchez., in prep.) All consensus sequences were matched to the list of 32 000 cDNA from salmonids and good hits (BLASTN *e*-value <1e-50) were obtained for 59% (1577) of them.

Table 2 Summary statistics of assembled contigs

Number of sequences assembled	428 068 (68% of total)
Number of contigs*	2674
Mean length	576
Number of SNPs	6042
Mean SNP/kb (min–max)	3.4 (0–44.8)
Mean coverage (min–max)	8.9X (1.3X–4140X)
Base substitutions	
Transitions	
A–G	1930 (31.7%)
C–T	1867 (30.6%)
Transversions	
A–T	599 (9.8%)
A–C	658 10.8%)
C–G	344 (5.%)
T–G	696 (11.4%)
Number of ORFs†	1904
Mean length of ORF	482
Number of SNPs	3274
p_N/p_S	0.37 (0.0028/0.0075)

*Similarity criterion: 0.97. Minimum overlap: 0.5.
†Mininum length set for accepting open reading frame (ORF): 200 nucleotides.
p_S, number of synonymous SNPs per synonymous sites; p_N, number of non-synonymous SNPs per non-synonymous sites.

SNP discovery and functional characterization

Out of the 6042 putative SNPs, we identified among all 2674 contigs, the proportions of transition substitutions were A/G, 31.7%, and C/T, 30.6%, compared to transversions A/C, 10.8%: G/T, 11.4%: A/T, 9.8% and C/G, 5.6% (Table 2). This corresponds to a transition:transversion ratio of 1.65:1. Mean number of SNP per kilobase was 3.4. A total of 70 contigs out of 2674 (or 2.6%) had a very high polymorphism rate (>20 SNPs/kb). These were involved in several functional classes; mostly mRNA translation and processing (11 hits), DNA transposition (6 hits) and mitotic spindle organization and biogenesis (5 hits), yet only the last two categories were significantly overrepresented compared to observed frequencies of represented functional groups among all contigs assembled (Fishers's exact test, P. value < 0,05, table 3).

A total of 1904 predicted ORFs with a mean length of 482 bp was identified. These contained 3274 polymorphic sites of which 1734 were synonymous and 1540 non-synonymous. There were 2.8 SNPs per 1000 non-synonymous sites and 7.5 SNPs per 1000 synonymous sites, for a genome-wide non-synonymous to synonymous substitution rate ratio of 0.37 ($p_S = 0.0075$, $p_N = 0.0028$; Fig. 2, Table 2). Twenty-nine contigs had a p_S/p_N ratio >1, suggestive of positive selection, and these were involved in several biological functions, most notably, mRNA translation and processing (7 hits). Yet, none of the biological functions was significantly overrepresented compared to all contigs assembled (Fisher's exact test, $P > 0.05$; Table 4).

SNP frequencies between dwarf and normal whitefish

We analysed a subset of 1504 SNPs that met our criterion for inferring allele frequencies (see Materials and methods). Although most SNPs showed little divergence (Fig. 3), 190 SNPs had significant divergent allelic frequencies (*Q*-value <0.05) and 89 of these were strongly divergent between normal and dwarf whitefish (above 0.5 in Fig. 3 & Table 5). These 89 SNPs represented 46 different contigs and several biological functions. Of interest among these, seven mitochondrial genes (*e*-value <1e-50: cytochrome C subunit 1, 2 and 3;

Table 3 Functional annotation (gene ontology biological functions) of ranked contigs with the highest rate of single nucleotide polymorphisms per kilobase (SNPs/kb >20 or 2%)

Gene product*	Functional groups	SNPs/kb	Match to previous studies†
60S ribosomal protein L22	Translation (GO:0006412)	44.8	
40S ribosomal protein S5	Translation (GO:0006412)	39.4	10
Nucleolar RNA helicase 2	mRNA splicing (GO:0000398)	39.6	10
Sequestosome-1	Regulation of I-kappaB kinase/NF-kappaB cascade (GO:0043122)	38.1	
Ubiquitin	Positive regulation of transcription (GO:0045941)	37.8	1,5,6,10,11
Tubulin alpha chain	Mitotic spindle organization and biogenesis (GO:0007052)	37.7	10,11
60S ribosomal protein L7	Translation (GO:0006412)	35.3	
Vacuolar ATP synthase catalytic subunit A	Proton transport (GO:0015992)	34.9	
Tubulin alpha chain	Mitotic spindle organization and biogenesis (GO:0007052)	33.5	10,11
Transposable element Tc1 transposase	Transposition, DNA-mediated (GO:0006313)	31.5	
Transposable element Tc1 transposase	Transposition, DNA-mediated (GO:0006313)	29.8	
Collagen alpha-2(I) chain precursor	Skin development (GO:0030199)	27.4	
Retinol dehydrogenase 3	Metabolism (GO:0008152)	26.1	
Transcription factor PU.1	Negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	26	
60S ribosomal protein L27a	Translation (GO:0006412)	25.6	
Similar to Calsequestrin	Calcium ion binding (GO:0005509)	25.5	
Transposable element Tcb1 transposase	Transposition, DNA-mediated (GO:0006313)	25.4	
60S ribosomal protein L5	Translation (GO:0006412)	25.2	6,7,10
Proteasome subunit beta type-7 precursor	Ubiquitin-dependent protein catabolic process (GO:0006511)	24.6	
Ubiquitin carboxyl-terminal hydrolase 28	Ubiquitin-dependent protein catabolic process (GO:0006511)	24.5	
Ubiquitin-like protein FUBI	Translation (GO:0006412)	23.9	
60S ribosomal protein L17	Translation (GO:0006412)	23.5	
Thimet oligopeptidase	Proteolysis (GO:0006508)	23.3	
NADH dehydrogenase iron–sulphur protein 2	Response to oxidative stress (GO:0006979)	23.1	
Probable RNA-directed DNA polymerase from transposon BS	Transposition, DNA-mediated (GO:0006313)	22.8	6,8
Zinc finger protein ZIC 2	Cell differentiation (GO:0030154)	22.2	
Transposable element Tcb1 transposase	Transposition, DNA-mediated (GO:0006313)	22.1	
Acetyl-CoA acetyltransferase, cytosolic	Metabolism process (GO:0008152)	22	
Heterogeneous nuclear ribonucleoprotein G	mRNA processing (GO:0006397)	21.9	6
Fibrinogen beta chain precursor	Blood coagulation (GO:0007596)	21.9	
Protein SEC13 homolog	Protein transport (GO:0015031)	21.6	
Oncorhynchus kisutch 5S ribosomal RNA gene	Translation (GO:0006412)	21.5	
Cold-inducible RNA-binding protein	Response to cold (GO:0009409)	21.3	
Histidyl-tRNA synthetase, cytoplasmic	Translation (GO:0006412)	20.9	
Tubulin alpha chain	Mitotic spindle organization and biogenesis (GO:0007052)	20.9	10,11
Transposable element Tcb2 transposase	Transposition, DNA-mediated (GO:0006313)	20.7	
14-3-3 protein beta/alpha	Ras protein signal transduction (GO:0007265)	20.5	
Stathmin	Mitotic spindle organization (GO:0007052)	20.3	6
THO complex subunit 4	mRNA transport (GO:0051028)	20.3	
Tubulin alpha chain	Mitotic spindle organization and biogenesis (GO:0007052)	2	10,11
Unknown	Unknown	43.9	
Schistosoma japonicum SJCHGC04625 protein	Unknown	38.9	
14-3-3 protein zeta	Unknown	38.5	

Table 3 *Continued*

Gene product*	Functional groups	SNPs/kb	Match to previous studies†
Unknown	Unknown	36.9	
Unknown	Unknown	33.2	
Unknown	Unknown	32.3	
Unknown	Unknown	31.4	
Unknown	Unknown	31.1	
Unknown	Unknown	31	
Unknown	Unknown	30.1	
Protein DJ-1	Unknown	29.3	
Unknown	Unknown	29	
Unknown	Unknown	27.9	
Unknown	Unknown	26.2	
Unknown	Unknown	26.2	
Unknown	Unknown	26	
Unknown	Unknown	25.6	
Unknown	Unknown	25.4	
Unknown	Unknown	25.3	
Unknown	Unknown	24.5	
Unknown	Unknown	24.3	
Unknown	Unknown	24.3	
Unknown	Unknown	24	
Unknown	Unknown	23.5	
Unknown	Unknown	21.7	
Unknown	Unknown	21.5	
Unknown	Unknown	20.9	
Unknown	Unknown	20.8	
Unknown	Unknown	20.6	
Unknown	Unknown	20.3	11
Unknown	Unknown	21.9	
Unknown	Unknown	22.5	
Unknown	Unknown	22.4	

*Note that several contigs may correspond to the same gene annotation. These may be either splice variants of the same gene or different paralogues of that gene.
†Match to previous studies that either showed differential expression between dwarf, normal or hybrid whitefish, or mapped to eQTL:
1: Parallel non-directional change in gene expression between dwarf and normal natural whitefish (white muscle, adults; Derome *et al.* 2006).
2: Parallel directional change in gene expression between dwarf and normal natural whitefish (white muscle, adults; Derome *et al.* 2006).
3: Parallel non-directional change in gene expression between dwarf and normal natural and controlled environment populations (liver, adults; St-Cyr *et al.* 2008).
4: Parallel directional change in gene expression between dwarf and normal natural and controlled environment populations (liver, adults; St-Cyr *et al.* 2008).
5: Parallel directional change in gene expression between dwarf and normal natural populations (liver, adults; St-Cyr *et al.* 2008).
6: Change in gene expression between dwarf and normal controlled environment populations (whole fish, juveniles; Nolte *et al.* 2009).
7: Change in gene expression between dwarf and normal controlled environment populations (white muscle, adults; Derome *et al.* 2008).
8: Change in gene expression between dwarf and normal controlled environment populations whitefish (whole fish, embryos; Nolte *et al.* 2009).
9: Highly transgressive gene in hybrid whitefish (whole fish, juveniles, Renaut *et al.* 2009)
10: eQTL (white muscle, adults; Derome *et al.* 2008).
11: eQTL (brain tissue, adults; Whiteley *et al.* 2008).

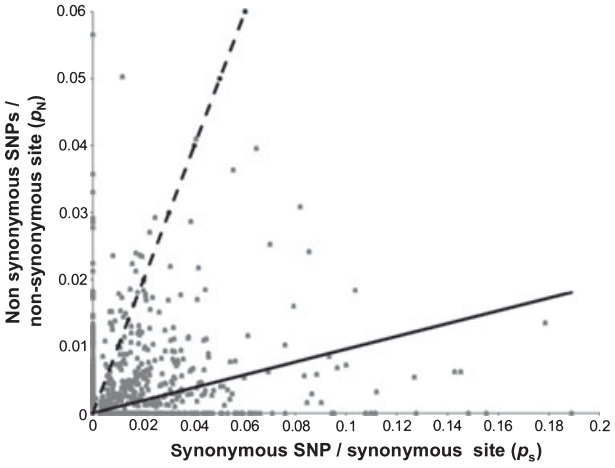

Fig. 2 Non-synonymous mutations per non-synonymous sites compared to synonymous mutations per synonymous sites. Dashed line is the null expectation if mutations were randomly distributed ($p_N = p_S$). Solid line is the slope of experimental data (overall average p_N for all contigs/overall average p_S for all contigs = 0.37).

NADH-dehydrogenase 1, 4 and 5; and cytochrome B) and seven nuclear genes (cytochrome b-c1 complex subunit 6, ATP synthase subunit d, Malate dehydrogenase, glyceraldehyde-3-phosphate dehydrogenase, creatine kinase, Succinyl-CoA ligase and angiopoietin-related protein 3 precursor) were all involved in energy metabolic pathways.

Comparison with previous studies

Twelve contigs identified as highly polymorphic (i.e. above 20 SNPs/kb, Table 3) matched to genes previously identified as candidates in different gene expression studies, and the expression of most of those genes had been previously linked to a specific genomic region (eQTL). Two of these (60S ribosomal protein L5, ubiquitin) were also identified as differentially expressed between normal and dwarf in several independent studies (Table 3). Thirteen contigs with a high p_N/p_S ratio matched to genes previously identified in different gene expression studies, and again the expression of those genes have been lined to an eQTL. Three of those (60S ribosomal protein L5, ornithine decarboxylase antizyme 1 and creatine kinase) were also identified as differentially expressed between normal and dwarf in several independent studies (Table 4).

Eighteen contigs, containing at least one SNP, which showed highly divergent allelic frequencies between normal and dwarf, were annotated to genes previously identified as potential candidates based on expression studies (Table 5). Among these, genes related to energy metabolism (cytochrome C subunit 1, 2 and 3; NADH-dehydrogenase 1, 4 and 5; cytochrome B; cytochrome b-c1 complex subunit 6; ATP synthase subunit d; Malate dehydrogenase; glyceraldehyde-3-phosphate dehydrogenase; creatine kinase; Succinyl-CoA ligase and angiopoietin-related protein 3 precursor) are of particular interest as candidates underlying adaptive divergence between dwarf and normal whitefish since they consistently showed differential expression in independent studies.

High rate of transposition in hybrids

Given that we identified many highly polymorphic contigs annotated to DNA transposition (Table 3), these were further investigated. Forty-four contigs matching to six different DNA transposons and retrotransposons elements (BLASTN *e*-value <1e-50, Table 6) were detected. These contigs were also, on average, four times more polymorphic than the rest of the assembly (10.8 SNPs/kb compared to 3.4 overall, *t*-test, $P < 0.0001$). As sequencing was performed on non-normalized cDNA, the number of reads per population may be used as a proxy for gene expression (Torres *et al.* 2007; Ledford 2008). A total of 4600 sequences assembled into these 44 contigs and, invariably, there was a strong bias such that 70% of the sequences matching these came from backcross hybrids, whereas the data set was composed of only 38% backcross sequences (chi-squared test, $P < 1e\text{-}16$).

SNP validation

Twenty-nine individual fish were genotyped for a subset of 31 polymorphic SNPs within a single lake (Lake Aylmer). Six markers deviated significantly from expected Hardy–Weinberg frequencies due to heterozygous excess ($Q < 0.05$, Fig. 4). SNPs genotyped came from contigs with polymorphism ranging from 1.4 to 38 SNPs/kb and there was no apparent correlation between amount of polymorphism and F_{is} estimates (Pearson's correlation coefficient = −0.08, $P = 0.69$).

Discussion

By sequencing a total of two and one-quarter runs on the 454 GS-FLX system, 632 000 reads with a mean length, once primers and sample-specific tags were removed, of 193 nucleotides were obtained. The fact that we obtained ~30% fewer sequencing reads than what would be theoretically expected (400 000 sequences/run) is, at least in part, due to the nature on the DNA itself. First, cDNA sizes, which are quite variable, render the shearing process prior to sequencing more difficult. Second, mature cDNA usually contains large polyA stretches that are harder to sequence and cause many reads to be rejected due to poor quality or very short lengths (Gary Levesque,

Table 4 Contigs with the highest ratio of non-synonymous SNPs per non-synonymous site (p_N)/synonymous SNPs per synonymous site (p_S)

Gene product*	Functional groups	p_N/p_S	Match to previous studies†
40S ribosomal protein S5	Translation (GO:0006412)	4.35	10
Glutamine synthetase	Response to glucose stimulus (GO:0009749)	2.98	
14 kDa apolipoprotein	G-protein coupled receptor protein signalling pathway (GO:0007186)	2.28	
Basement membrane-specific heparan sulphate proteoglycan	Cell adhesion (GO:0007155)	2.24	
Betaine–homocysteine S-methyltransferase 1	Methionine biosynthetic process (GO:0009086)	2.17	3,6
60S ribosomal protein L5	Translation (GO:0006412)	2.06	6,7,10
Aldehyde dehydrogenase, mitochondrial precursor	Carbohydrate metabolic process (GO:0005975)	2	
Complement C3-1	G-protein coupled receptor protein signalling pathway (GO:0007186)	1.81	
Keratin, type I cytoskeletal 13	Epidermis development (GO:0008544)	1.75	6
Tubulin alpha chain	Mitotic spindle organization (GO:0007052)	1.73	10,11
40S ribosomal protein S16	Translation (GO:0006412)	1.62	11
40S ribosomal protein S13	Translation (GO:0006412)	1.59	
Ornithine decarboxylase antizyme 1	Polyamine metabolic process (GO:0006595)	1.52	6,11
40S ribosomal protein S8	Translation (GO:0006412)	1.44	11
Stathmin	Mitotic spindle organization (GO:0007052)	1.28	11
Creatine kinase M-type	Phosphocreatine biosynthetic process (GO:0046314)	1.27	1,6,9,11
Transposable element Tc1 transposase	Transposition, DNA-mediated (GO:0006313)	1.21	
Heterogeneous nuclear ribonucleoprotein G	mRNA processing (GO:0006397)	1.21	6
Beta-2-glycoprotein 1 precursor	Heparin binding (GO:0008201)	1.21	
Unknown	Protein amino acid phosphorylation (GO:0006468)	1.2	
ATP-binding cassette sub-family F member 1	Translation (GO:0006412)	1.15	
Nucleolar RNA helicase 2	RNA processing (GO:0006396)	1.01	10
Unknown	Unknown	1.45	
Unknown	Unknown	1.42	
Similar to fatty acid desaturase domain family, member 6	Unknown	1.41	
Unknown	Unknown	1.26	
Unknown	Unknown	1.23	
Unknown	Unknown	1.03	
Unknown	Unknown	1.01	

*Note that several contigs may correspond to the same gene annotation.
†See legend in Table 3.

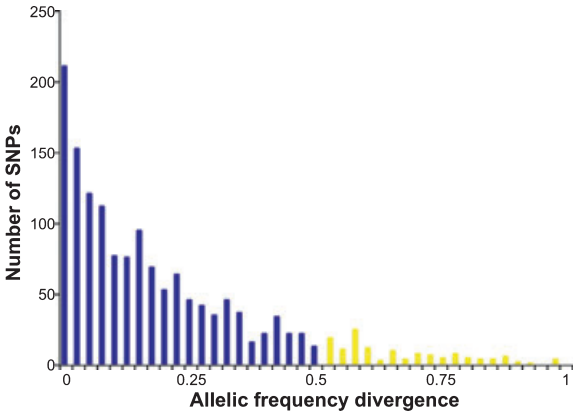

Fig. 3 Frequency distribution of allelic frequency differences between normal and dwarf whitefish. Allele divergence value above one (yellow) and with a *Q*-value <0.05 were considered as highly divergent single nucleotide polymorphism (SNP) markers. Allele divergence value = absolute value of [frequency($allele_{1Dwarf}$) − frequency($allele_{1Normal}$)]. Note that 1504 SNPs from 387 different contigs were used to draw this distribution.

McGill University, pers. comm.). Nevertheless, we obtained over 130 mb of sequencing reads, which, compared to Sanger sequencing technology, required several orders of magnitude less time and money. As expected, the amount of sequences assembled is strongly dependent not only on the read length (i.e. shorter reads are harder to assemble, Fig. 1) but also on the stringency of the assembly performed. Here, by using a similarity criterion of 0.97 (see Materials and methods for rationale behind using 0.97), 68% of all reads were assembled into 2674 different contigs.

SNP discovery, validation and functional characterization

We identified over 6000 putative SNPs. If all substitutions were equally likely, a 1:2 transition (ts) to transversion (tv) ratio would be expected, as there are twice as many possible transversions than transitions. In reality, a biased ts:tv ratio is thought to be a universal characteristic of the nucleotide composition landscape (Lynch 2007). At the same time, some authors (e.g. Keller *et al.* 2007) have recently suggested that biased ts:tv ratio may be a sampling artefact as conclusions are based upon experimental data from a few model species (Lynch 2007). Here, in lake whitefish, a strongly biased ratio towards transitions (1.65:1) was identified, supporting the view that this trend is ubiquitous at least among vertebrates.

Determining an exact number of sequence polymorphisms largely depends on the stringency of the assembly and the criteria used to define a true SNP (i.e. coverage and minimum frequency of SNP). Using fairly stringent criteria (minimum similarity: 0.97; minimum coverage of SNP: 6X; and minimum frequency of the least frequent allele: 20%) reduced the amount of false positives. Nevertheless, as salmonids underwent an ancient whole genome duplication event and given that over 50% of their genome is still considered duplicated (Allendorf *et al.* 1975), we cannot refute the possibility that a significant proportion of putative SNPs may actually be PSVs. For example, in Atlantic salmon, 19% of polymorphic SNPs predicted to be of high quality showed heterozygous excess most likely due to genome duplication (Hayes *et al.* 2007). This is a problem inherent to SNP markers even in well-characterized species, including humans. Fredman *et al.* (2004) showed, using fully homozygous cell masses, that only 50% of sequence variants (i.e. putative SNPs) in duplicated regions of the human genome are true SNPs. In fact, our own SNP validation assay revealed that 19% (6/31) of the genotyped loci significantly deviated from expected Hardy–Weinberg frequencies because of heterozygous excess (Fig. 4), a tell-tale sign that these SNPs may be variants between duplicated regions of the genome (Fredman *et al.* 2004). Although this may be true, several alternative explanations may also be responsible for this pattern: small sample size, heterozygote advantage, frequency-dependent selection or presence of null alleles. Finally, as we address in the last section of the Discussion and Conclusion, although a single SNP only provides circumstantial evidence of its importance in the adaptive divergence of lake whitefish, we strongly emphasize (as suggested by others; cf. Vasemagi & Primmer 2005; Stinchcombe & Hoekstra 2008) that combining experimental evidence targeting different biological levels (e.g. variation at the DNA, gene expression and phenotypic levels) represents the best strategy towards deciphering the genetic basis of evolutionary change. Nonetheless, we recognize that a large data set of SNP markers identified using high-throughput methods probably needs to be validated by alternative methods before being used in further studies as true, experimentally confirmed, genetic markers. Until fully homozygous lines or haploid individuals can be produced, it will be difficult to truly disentangle the effect of gene duplication and genomic divergence.

Several functional categories were identified among the list of highly polymorphic contigs. Namely, ribosomal proteins (mRNA translation), tubulin (mitotic spindle organization) and transposable elements (DNA transposition) are all part of multigenic families found in numerous copies throughout the genome. Such genes are probably particularly prone to biases due to PSVs and therefore putative SNPs for these should be used with vigilance. At the same time, based on our genotyp-

Table 5 Single nucleotide polymorphism markers with significant divergent allelic frequencies between sympatric normal and dwarf whitefish

Description†	Functional category‡	Allele 1 (D)§	Allele 1 (N)§	Total number of sequences (D, N)¶	abs[$f(a_{1D}) - f(a_{1N})$]††	Match to previous studies‡‡
Angiopoietin-related protein 3 precursor	Fatty acid metabolic process (GO:0006631)	1	0.2	9,5	0.8*	
ATP synthase subunit d, mitochondrial	ATP synthesis (GO:0015986)	0.71	0.06	7,16	0.65*	
Creatine kinase M-type	(2) Phosphocreatine biosynthetic process (GO:0046314)	0.77	0.21	43,67	0.56**	1,6,9,10
Creatine kinase M-type	(13) Phosphocreatine biosynthetic process (GO:0046314)	0.73	0.11	391,383	0.62**	1,6,9,10
Cytochrome b	Electron transport chain (GO:0022900)	1	0.06	25,31	0.94***	
Cytochrome b-c1 complex subunit 6, mitochondrial precursor	Electron transport chain (GO:0022900)	0.6	0.09	43,35	0.51***	
Cytochrome c oxidase subunit 1	Oxidation reduction (GO:0055114)	0.98	0.18	184,186	0.8***	6,7
Cytochrome c oxidase subunit 2	Oxidation reduction (GO:0055114)	0.99	0.1	102,97	0.89***	6,7
Cytochrome c oxidase subunit 3	(3) Oxidation reduction (GO:0055114)	0.99	0.22	159,157	0.77***	1,3,6,7,10
Glyceraldehyde-3-phosphate dehydrogenase	Glycolysis (GO:0006094)	0.8	0	20,6	0.8**	1,2,4,5,7,10
Malate dehydrogenase, cytoplasmic	Tricarboxylic acid cycle (GO:0006099)	0.6	0	15,13	0.6**	4
Malate dehydrogenase, cytoplasmic	Tricarboxylic acid cycle (GO:0006099)	0.67	0.14	24,14	0.53*	4
NADH dehydrogenase subunit 1	Electron transport chain (GO:0022900)	1	0.17	37,35	0.83***	6
NADH dehydrogenase subunit 1	Electron transport chain (GO:0022900)	0.99	0.16	97,153	0.83***	6
NADH dehydrogenase subunit 4	Electron transport chain (GO:0022900)	1	0.18	24,17	0.82***	6,7
NADH dehydrogenase subunit 5	Electron transport chain (GO:0022900)	1	0	6,7	1**	
Succinyl-CoA ligase, mitochondrial precursor	Tricarboxylic acid cycle (GO:0006099)	0.68	0.15	28,27	0.53**	
40S ribosomal protein S9	Translation (GO:0006412)	0.9	0.1	29,20	0.8***	6,10
60S acidic ribosomal protein P2	Translation (GO:0006412)	0.95	0.23	19,13	0.72***	11
60S ribosomal protein L27a	(2) Translation (GO:0006412)	1	0.23	26,26	0.77**	
60S ribosomal protein L39	Translation (GO:0006412)	0.75	0.07	8,28	0.68**	6
Actin, cytoplasmic 1	(2) Cytoskeleton (GO:0005856)	0.76	0.09	85,125	0.67**	6
Alpha-1-antitrypsin precursor	Blood coagulation (GO:0007596)	0.97	0.45	29,33	0.52**	
C-type lectin domain family 4 member E	Immune response (GO:0006955)	0.78	0.25	9,63	0.53*	
Coagulation factor × precursor	(3) Unknown	0.76	0.19	81,90	0.57*	
Coagulation factor × precursor	Unknown	0.83	0.29	23,24	0.54**	
Complement C5 precursor	Complement activation, alternative pathway (GO:0006957)	1	0.17	12,6	0.83**	
Complement factor H precursor	(5) Complement activation, alternative pathway (GO:0006957)	0.82	0.21	553,516	0.61**	
Ferritin, heavy subunit	(2) Regulation of transcription (GO:0045892)	0.82	0.28	61,53	0.54**	5,11
Fibrinogen beta chain precursor	(14) Platelet activation (GO:0030168)	0.82	0.03	170,163	0.79**	
Fibrinogen beta chain precursor	Platelet activation (GO:0030168)	0.7	0.05	23,39	0.65***	
Fibrinogen beta chain precursor	(5) Platelet activation (GO:0030168)	0.67	0.1	139,192	0.57**	
Fibronectin precursor	Cell adhesion (GO:0007155)	1	0	8,4	1*	
Heat shock protein HSP 90-beta	Regulation of nitric oxide biosynthetic process (GO:0045429)	0.8	0.19	10,31	0.61*	
Haemopexin precursor	Cellular iron homeostasis (GO:0006879)	0.98	0.46	206,371	0.52***	
Metallothionein mRNA	Unknown	0.64	0	11,24	0.64***	6
Nucleolar RNA helicase 2	Nuclear mRNA splicing (GO:0000398)	0.92	0.4	12,48	0.52*	10
Selenoprotein Pa precursor	Response to oxidative stress (GO:0006979)	0.69	0.02	29,50	0.67***	

Table 5 *Continued*

Description†	Functional category‡	Allele 1 (D)§	Allele 1 (N)§	Total number of sequences (D, N)¶	abs[$f(a_{1D})$ – $f(a_{1N})$]††	Match to previous studies‡‡
Subunit of Ca^{2+}-dependent complex	(2) Unknown	0.87	0.12	52,19	0.75**	
Unknown	Unknown	1	0	7,5	1*	
Unknown	(2) Unknown	0.99	0.1	149,80	0.89***	
Unknown	Unknown	0.78	0.04	9,52	0.74***	
Unknown	Unknown	1	0.3	6,17	0.7*	
Unknown	(2) Unknown	0.61	0	38,59	0.61***	
Unknown	(2) Unknown	0.77	0.19	43,33	0.58*	

†Note that several contigs may correspond to the same gene annotation.
‡Numbers in parentheses indicate that several SNPs within contig were divergent and the subsequent allelic frequencies and *Q*-value are an average for these SNPs.
§Frequency of the most common dwarf allele and frequency of its corresponding normal allele.
¶Number of sequences from dwarf and normal fish used to calculate allele frequencies.
††abs[$f(a_{1D})$ – $f(a_{1N})$] = absolute value of (frequency($allele_{1Dwarf}$) – frequency($allele_{1Normal}$)]. *Q < 0.05, **Q < 0.01, ***Q < 0.001: Probability value of Fisher's exact test corrected for multiple hypothesis testing (Q-value) calculated for each SNP as the total number of sites identified to each allele in normal and dwarf.
‡‡See legend in Table 3.

ing results, we did not find any significant correlation between F_{is} (as a potential indication of PSVs) and polymorphism rate (P = 0.69).

Nucleotide substitution effect on predicted ORFs

By identifying 1904 predicted ORFs, this permitted to estimate a transcriptome-wide non-synonymous to synonymous substitution rate ratio (p_N/p_S) of 0.37. As such, on average, the p_N/p_S ratio per gene is much lower than a ratio of one expected if mutations were randomly distributed (Fig. 2). This is generally interpreted as indicative of the effect of purifying selection against deleterious amino acid altering changes. Alternatively, ORFs with an elevated p_N/p_S ratio (e.g. above 1) may indicate genes evolving under the effect of positive selection. Here, 29 contigs had a p_N/p_S ratio above 1 and were involved in several biological functions. These may constitute candidates under the effect of natural selection responsible for the adaptive divergence of lake whitefish. However, three caveats must be mentioned from such an analytical approach. First, by definition, ORFs represent 'potential' region of the genome translated into a protein and therefore do not necessarily code for the actual polypeptide chain. Second, as the number of polymorphic site per base pair is relatively low, only 13% of all contigs detected had an ORF and a p_N and p_S value above zero. Lastly, with few mutations per gene, ratios can vary drastically if one or a few polymorphic sites are misidentified. As such, although this type of information may be useful to look at general transcriptome-wide trends or in combination with other experimental evidence, inferring the effect of selection on single candidate genes, solely looking at the distribution of synonymous and non-synonymous mutations, must be done with caution.

Differences between normal and dwarf whitefish

As expected based on the young age (<15 000 years) of whitefish species pair, the overall level of divergence between them was relatively weak. In fact, out of 1504 SNPs, only 89, coming from a maximum of 45 different genes (Table 5), had significant highly divergent allelic frequencies between normal and dwarf populations. This represents 6% of all SNPs for which we had enough sequence information to perform this analysis and good candidates for genomic islands of early divergence. In fact, 6% is comparable to what genome scan studies of young species pairs have found looking for genetic loci with divergent allele frequency (5–10%, reviewed in Nosil *et al.* 2009). For example, Turner *et al.* (2005) have identified only three genomic regions, encompassing a maximum of 67 genes, showing evi-

Table 6 Expression (total number of sequences) annotated to transposon elements in normal and dwarf whitefish as well as backcross hybrids

Gene product	No. contigs†	Total number of sequences		
		Normal	Backcross	Dwarf
Transposable element Tc1 transposase	16	133	584**	157
Transposable element Tcb1 transposase	12	231	1142**	298
Transposable element Tcb2 transposase	6	96	740**	151
Non-LTR retrotransposon	4	52	320**	104
PREDICTED: similar to transposase (*Strongylocentrotus purpuratus*)	1	1	9*	2
Probable RNA-directed DNA polymerase from transposon BS	6	68	394**	90

†Several assembled contigs were annotated (*e*-value <1e-50) to the same gene product.
*P = 0.08, **P < 1e-16. Chi-squared tests based on the expected proportion of sequences (in whole assembly, 62% of all sequences are either normal or dwarf, 38% are backcross).

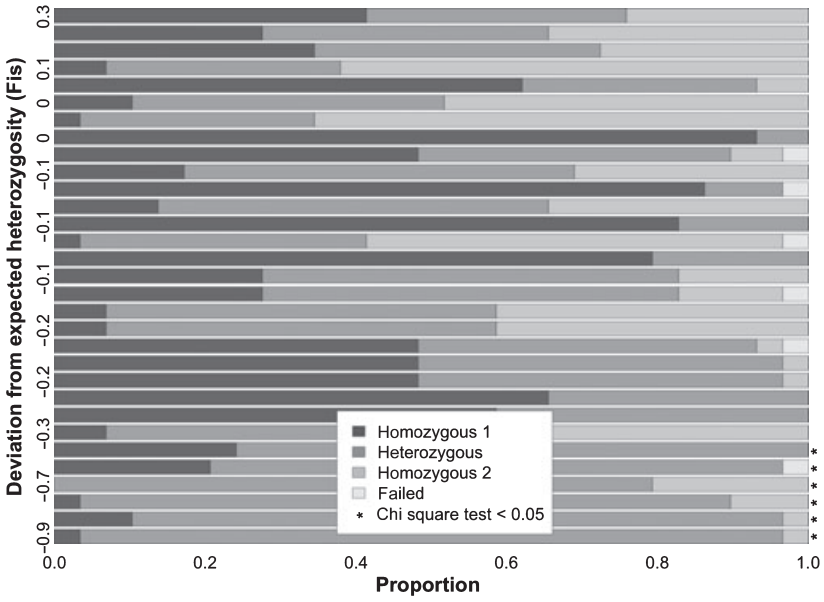

Fig. 4 Single nucleotide polymorphism (SNP) validation for 29 individuals originating from a single lake (Lake Aylmer) and genotyped for 31 polymorphic markers. SNPs were ranked according to F_{is} values (*y*-axis, left side). Deviation from expected Hardy–Weinberg frequencies (chi-square test, 1 d.f., *Q*-value <0.05) were included on the *y*-axis (right side).

dence of reduced gene flow in African malaria mosquitoes (*Anopheles gambiae*), a system characterized by strong assortative mating. In lake whitefish, using anonymous AFLP markers, previous genome scan studies have suggested that as little as 1.2% of the genome (which may still represent several hundred genes) might be under the effect of directional selection during the adaptive divergence of lake whitefish (Campbell & Bernatchez 2004; Rogers & Bernatchez 2005).

Furthermore, the proportion of divergent SNPs identified in this study may represent an overestimate due to several factors. First, SNP frequencies were estimated from sequences from a maximum of eight dwarf and eight normal individuals that were available. Given the relatively small number of individuals and allele copies, depicted differences should thus be interpreted with caution. Nevertheless, this analytical approach represents a necessary preliminary step towards identifying potential candidate SNPs. Second, as transcribed cDNA was sequenced, it is conceivable that normal and dwarf heterozygous individuals may overexpress a different allele and thus show divergent cDNA allelic patterns despite sharing a common genotype. At this point, it is difficult to clearly distinguish the two alternatives. Yet, both mechanisms point out relevant genetic differences between populations (i.e. differential allele specific expression or true genotypic differences) and we are currently conducting experiments to investigate how these transcriptome allelic frequencies are correlated to genotypic frequencies (Renaut S., Bernatchez L. unpublished).

Increased rate of transposition in hybrids

Aside from sequence or gene expression divergence, a broad variety of mechanisms related to the maintenance of chromatin integrity may be involved in causing hybrid

dysfunctions and possibly reproductive isolation (Fontdevila 2005; Michalak 2009). In fact, during her pioneer work on transposable elements, Barbara McClintock was the first to suggest that hybridization in plants might activate dormant transposons and result in genome restructuring (McClintock 1984). Since then, several studies have shown that transposition rates in plant hybrids can increase by several orders of magnitudes (Shan *et al.* 2005; Ungerer *et al.* 2006). In animals however, contrasting results and limited direct evidence have casted doubts on the role of transposable elements in speciation processes (Coyne 1989; Labrador *et al.* 1999; Coyne & Orr 2004). Here, extensive sequencing data provide compelling evidence of an important increase in transposon activity in hybrids, which may be a consequence of partial incompatibility of normal and dwarf genomes reported in previous studies (Rogers & Bernatchez 2006, 2007). Contigs annotated to transposable elements were also, on average, four times more polymorphic than the rest of the assembly. Transposons are, by nature, highly duplicated, and therefore the high polymorphism rate probably reflects the fact that several duplicated copies are activated. Lastly, as cDNA from liver tissue in normal and dwarf and white muscle and brain in the backcross was sequenced, elevated activity of transposon could also be a tissue-specific effect. Nonetheless, it would be peculiar and unheard of in the literature that transposable elements would be more active in muscle and brain than in liver tissues.

Comparison with previous studies

The integration of results from this study with previous analyses of gene expression, QTL and genome scans in whitefish significantly adds to our understanding of the genetic basis of the adaptive divergence of sympatric dwarf and normal whitefish in several ways. First, previous gene expression studies (Derome *et al.* 2006; St-Cyr *et al.* 2008; Jeukens *et al.* 2009; Nolte *et al.* 2009; Renaut *et al.* 2009) combined with physiological data (Trudel *et al.* 2001) have provided ample evidence that changes in the expression of genes involved in energetic metabolism pathways are largely responsible for the adaptation to distinct whitefish benthic (normal) and limnetic (dwarf) niches. Nevertheless, these studies lacked the empirical evidence linking expression differences to actual genotypic divergence for the same genes. Whiteley *et al.* (2008) addressed this question by combining eQTL information with F_{st} outlier loci obtained from genome scan studies (Campbell & Bernatchez 2004; Rogers & Bernatchez 2007) to identify genes under the influence of divergent selection. However, they provided only indirect evidence as eQTLs may correspond to the location of the gene itself (*cis*), or the location of another gene regulating its expression (*trans*). Our study brings a more direct link between genetic divergence (reduced gene flow) and gene expression divergence. The most salient finding is that 14 genes involved in energy metabolism (both mitochondrial and nuclear) showed pronounced allele frequency differences in this study and were also identified in several previous gene expression studies as differentially expressed in parallel between normal and dwarf whitefish. Namely, very similar allele frequencies observed for mitochondrial SNPs provide confidence that this signal is not a sampling or statistical artefact given that all mitochondrial genes are in full linkage disequilibrium. In addition, previous studies investigating mitochondrial divergence between lake whitefish populations showed that normal and dwarf from the same lake (Cliff Lake) are predominantly associated with distinct mitochondrial lineages from independent glacial refuge origins (Bernatchez & Dodson 1990; Lu *et al.* 2001). Consequently, although genetic variation and differentiation may have arisen in allopatry during the Pleistocene glaciation, its sorting and maintenance in sympatry during the last 15 000 years appears to be promoted by natural selection. Corroborating this claim is the fact that, in the absence of selection against hybrids, gene flow has been shown to homogenize recently diverged limnetic and benthic three-spined stickleback species pairs in <10 years (Taylor *et al.* 2006). Therefore, the whole mitochondrial genome, due to its non-recombining nature, is probably under strong selective constraints and we hypothesize that, in conjunction with the maintenance of pronounced allelic divergence at nuclear genes also involved in energy metabolism, it may confer different metabolic efficiencies involved in the adaptive divergence of dwarf and normal whitefish. Consequently, breakdown or mis-regulation of mitochondrial bioenergetics functions in hybrids could play an important role in the speciation process of dwarf and normal whitefish, as revealed recently in other systems (Ellison & Burton 2008; Gershoni *et al.* 2009).

That metabolic genes associated with the mitochondrion machinery are the underlying targets of selection leading to the adaptive divergence of lake whitefish is further supported by one of the main findings from Whiteley *et al.* (2008). Namely, their combined eQTL–F_{st} outlier approach indicated that an eQTL for cytochrome c oxidase (subunit VI) was linked to an F_{st} outlier locus in three independent lakes inhabited by sympatric normal and dwarf whitefish populations. Hopefully, through ongoing candidate gene mapping efforts, SNP markers will also permit to elucidate the genomic architecture of expression regulation (*cis* vs. *trans* regulation) for such candidate genes and

strengthen the association between genotype (SNPs from candidate genes) and phenotype (QTLs).

In addition, several contigs with functions unrelated to energy metabolism were matched to previous findings. For example, the 60S ribosomal L5 gene involved in mRNA translation, which was identified as highly polymorphic and potentially evolving under the effect of positive selection (p_N/p_S ratio = 2.06), had been previously linked to parallel gene expression differences between both wild normal and dwarf adult (Derome *et al.* 2008) and juvenile whitefish reared in the laboratory (Nolte *et al.* 2009). Also, ubiquitin, a conserved regulatory protein, was highly polymorphic, previously showed parallel gene expression differences between normal and dwarf in wild adult whitefish (Derome *et al.* 2006; St-Cyr *et al.* 2008), laboratory-reared juveniles (Nolte *et al.* 2009) and associated with an eQTL in white muscle (Derome *et al.* 2008) and brain tissue (Whiteley *et al.* 2008). These genes represent examples of additional candidates for divergent selection, which could be either physically linked to other candidate genes or be selected due to strong epistatic interactions with metabolic genes.

Conclusion

Next-generation sequencing technologies are already revolutionizing the way science is done in ecology and evolution. Here, sequencing the transcriptome of two incipient species of lake whitefish and backcross hybrids allowed to gather a large data set of putative SNP markers, analyse their distribution among genes, highlight an apparent increased activity of transposons in hybrids and identify potential targets of divergent selection. Mitochondrial and nuclear genes involved in energy metabolism emerge as prime candidates underlying the adaptive divergence of sympatric species of lake whitefish. Thorough investigations using genome scan in natural population as well as candidate gene mapping will permit to confirm this hypothesis. The rationale of our research programme on the adaptive divergence of lake whitefish is that integrating results targeting different functional and biological levels (e.g. variation at the DNA, gene expression and phenotypic levels) represents the best strategy towards deciphering the genetic basis of evolutionary change and diversification driven by natural selection.

Acknowledgements

We would like to thank J. Laroche, E. Normandeau and C. Sauvage for help with the bioinformatics, as well as J. Jeukens and N. Derome for insightful suggestions on earlier versions of the manuscript. This project was funded by a Natural Science and Engineering Research Council of Canada (NSERC) and Canadian Research Chair in Genomics and Conservation of Aquatic Resources to LB, a NSERC postgraduate scholarship to SR and a postdoctoral research stipend from the German Research Foundation to AN. This study is a contribution to the research programme of Québec Océan.

Conflicts of interest

The authors have no conflict of interest to declare and note that the sponsors of the issue had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Allendorf FW, Utter FM, May BP (1975) Gene duplication within the family Salmonidae: II. Detection and determination of the genetic control of duplicate loci through inheritance studies and the examination of populations. In: *Isozymes*, Vol. IV (ed. Marken CL), pp. 415–432. Academic Press, London.

Altschul SF, Madden TL, Schaffer AA *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Axelsson E, Hultin-Rosenberg L, Brandstrom M, Zwahlen M, Clayton DF, Ellegren H (2008) Natural selection in avian protein-coding genes expressed in brain. *Molecular Ecology*, **17**, 3008–3017.

Barbazuk WD, Emrich S, Chen HD, Li L, Schanble PS (2007) SNP discovery via 454 transcriptome sequencing. *The Plant Journal*, **51**, 910–918.

Bernatchez L (2004) Ecological theory of adaptive radiation: an empirical assessment from Corigonine fishes (Salmoniformes). In: *Evolution Illuminated: Salmon and Their Relative* (eds Hendry AP, Stearns S), pp. 176–207. Oxford University Press, Oxford, U.K.

Bernatchez L, Chouinard, Lu G (1999) Integrating molecular genetics and ecology in studies of adaptive radiation: whitefish, Coregonus, as a case study. *Biological Journal of the Linnean Society*, **68**, 173–194.

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

Bernatchez L, Dodson JJ (1991) Phylogeographic Structure in Mitochondrial DNA of the Lake Whitefish (*Coregonus clupeaformis*) and Its Relation to Pleistocene Glaciations. *Evolution*, **45**, 1016–1035.

Branton D, Deamer DW, Marziali A *et al.* (2008) The potential and challenges of nanopore sequencing. *Nature Biotechnology*, **26**, 1146–1153.

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.

Coyne JA (1989) Mutation rates in hybrids between sibling species of Drosophila. *Heredity*, **63**, 155–162.

Coyne JA, Orr HA (2004) *Speciation*. Sinauer Associates, Sunderland, MA.

Derome N, Duchesne P, Bernatchez L (2006) Parallelism in gene transcription among sympatric lake whitefish (*Coregonus clupeaformis*, Mitchill) ecotypes. *Molecular Ecology*, **15**, 1239–1249.

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

Dinsdale EA, Pantos O, Smriga S (2008) Microbial ecology of four coral atolls in the Northern Line Islands. *PLoS ONE*, **3**, e1584.

Ehrich M, Bocker S, van den Boom D (2005) Multiplexed discovery of sequence polymorphisms using base-specific cleavage and MALDI-TOF MS. *Nucleic Acids Research*, **33**, e38.

Ellegren H (2008) Comparative genomics and the study of evolution by natural selection. *Molecular Ecology*, **17**, 4586–4596.

Ellison CK, Burton RS (2008) Genotype-dependent variation of mitochondrial transcriptional profiles in interpopulation hybrids. *Proceedings of the National Academy of Science, USA*, **105**, 15831–15836.

Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, **7**, 574–578.

Fontdevila A (2005) Hybrid genome evolution by transposition. *Cytogenetics and Genome Research*, **110**, 49–55.

Fredman D, White SJ, Potter S *et al.* (2004) Complex SNP-related sequence variation in segmental genome duplications. *Nature Genetics*, **36**, 861–866.

Gershoni M, Templeton AR, Mishmar D (2009) Mitochondrial bioenergetics as a major motive force of speciation. *Bioessays*, **31**, 642–650.

Hall TA (1999) Bioedit: a user-f riendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, **41**, 95–98.

Hayes B, Laerdahl JK, Lien S *et al.* (2007) An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed sequences. *Aquaculture*, **265**, 82–90.

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.

Keller I, Bensasson D, Nichols RA (2007) Transition-Transversion bias is not universal: a counter example from grasshopper pseudogenes. *PloS Genetics*, **3**, e22.

Labrador M, Farre M, Utzet F, Fontdevilla A (1999) Interspecific hybridization increases transposition rates of Osvaldo. *Molecular Biology and Evolution*, **16**, 931–937.

Ledford H (2008) The death of microarrays? *Nature*, **455**, 847.

Lipson D, Razl T, Kieu A *et al.* (2009) Quantification of the yeast transcriptome by single-molecule sequencing. *Nature Biotechnology*, **27**, 652–659.

Lu G, Bernatchez L (1998) Experimental evidence for reduced hybrid viability between dwarf and normal ecotypes of Lake Whitefish (*Coregonus clupeaformis* Mitchill). *Proceedings of the Royal Society of London B: Biological Science*, **265**, 1025–1030.

Lu G, Basley DJ, Bernatchez L (2001) Contrasting patterns of mitochondrial DNA and microsatellite introgressive hybridization between lineages of lake whitefish (*Coregonus clupeaformis*); relevance for speciation. *Molecular Ecology*, **10**, 965–985.

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

Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

McClintock B (1984) The significance of responses of the genome to challenge. *Science*, **226**, 792–810.

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

Metzker ML (2009) Sequencing in real time. *Nature Biotechnology*, **27**, 150–151.

Michalak P (2009) Epigenetic, transposon and small RNA determinants of hybrid dysfuntions. *Heredity*, **102**, 45–50.

Miller W, Drautz DI, Ratan A *et al.* (2008) Sequencing the nuclear genome of the extinct woolly mammoth. *Nature*, **456**, 387–392.

Miya M, Nishida M (2000) Use of mitogenomic information in teleostean molecular phylogenetics: a tree-based exploration under the maximum-parsimony. *Molecular Phylogenetics and Evolution*, **17**, 437–455.

Moen T, Hayes B, Baranski M *et al.* (2008) A linkage map of the Atlantic salmon (*Salmo salar*) based on EST-derived SNP markers. *BMC genomics*, **9**, 223.

Namroud MC, Beaulieu J, Juge N, Laroche J, Bousquet J (2008) Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. *Molecular Ecology*, **17**, 3599–3613.

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**, 925–936.

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

Pigeon D, Chouinard A, Bernatchez L (1997) Multiple modes of speciation involved in the parallel evolution of sympatric morphotypes of Lake Whitefish (*Coregonus clupeaformis*, Salmonidae). *Evolution*, **51**, 196–205.

Quinlan AR, Stewart DA, Stromberg MP *et al.* (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nature Methods*, **5**, 179–181.

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.

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

Rogers SM, Bernatchez L (2005) Integrating QTL mapping and genome scans towards the characterization of candidate loci under parallel selection in the lake whitefish (*Coregonus clupeaformis*). *Molecular Ecology*, **14**, 351–361.

Rogers SM, Bernatchez L (2006) The genetic basis of intrinsic and extrinsic post-zygotic reproductive isolation jointly promoting speciation in the lake whitefish species complex (*Coregonus clupeaformis*). *Journal of Evolutionary Biology*, **19**, 1979–1994.

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

Rogers SM, Isabel N, Bernatchez L (2007) Linkage maps of the dwarf and normal lake whitefish (*coregonus clupeaformis*) species complex and their hybrids reveal the genetic architecture of population divergence. *Genetics*, **175**, 1–24.

Rokas A, Abbot P (2009) Harnessing genomics for evolutionary insights. *Trends in Ecology and Evolution*, **24**, 192–200.

von Schalburg KR, Cooper GA, Leong J *et al.* (2008) Expansion of the genomics research on Atlantic salmon *Salmo salar* L. project (GRASP) microarray tools. *Journal of Fish Biology*, **72**, 2051–2070.

Schlötterer C (2004) The evolution of molecular markers: just a matter of fashion. *Nature Reviews Genetics*, **5**, 63–69.

Schluter D (2009) Evidence for ecological speciation and its alternative. *Science*, **323**, 737–741.

Shan X, Liu Z, Dong Z *et al.* (2005) Mobilization of the active MITE transposons mPing and Pong in rice by introgression from wild rice (*Zizania latifolia* Griseb.). *Molecular Biology and Evolution*, **22**, 976–990.

Shen R, Fan JB, Campbell D *et al.* (2005) High-throughput SNP genotyping on universal bead arrays. *Mutation Research*, **573**, 70–82.

Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nature Biotechnology*, **26**, 1135–1145.

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

Stinchcombe JR, Hoekstra HE (2008) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*, **100**, 158–170.

Storey JD (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, **64**, 479–498.

Taylor EB, Boughman JW, Groenenboom M, Sniatynski M, Schluter D, Gow JL (2006) Speciation in reverse: morphological and genetic evidence of the collapse of a three-spined stickleback (*Gasterosteus aculeatus*) species pair. *Molecular Ecology*, **15**, 343–355.

Torres T, Metta M, Ottenwalder B, Schlötterer C (2007) Gene expression profiling by massively parallel sequencing. *Genome Research*, **18**, 172–177.

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

Turner L, Hahn MW, Nuzhdin S (2005) Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology*, **9**, 1572–1578.

Ungerer MC, Strakosh SC, Zhen Y (2006) Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. *Current Biology*, **16**, R872–R873.

Van Tassell CP, Smith TPL, Matukumalli LK *et al.* (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods*, **5**, 247–252.

Vasemagi A, Primmer CR (2005) Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Molecular Ecology*, **14**, 3623–3642.

Vera JC, Wheat CW, Fescemyer HW *et al.* (2008) Rapid transcriptome characterization for a non-model organism using 454 pyrosequencing. *Molecular Ecology*, **17**, 1636–1647.

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.

Whiteley AR, Persaud KN, Derome N, Montgomerie R, Bernatchez L (2009) Reduced sperm performance in backcross hybrids whitefish species-pairs (*Coregonus* sp.). *Canadian Journal of Zoology*, **87**, 566–572.

Wiedmann RT, Smith TPL, Nonneman DJ (2008) SNP discovery in swine by reduced representation and high throughput pyrosequencing. *BMC Genetics*, **9**, 81–88.

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

Wu CI, Ting CT (2004) Genes and Speciation. *Nature Review Genetics*, **5**, 114–122.

Yang ZH (2007) PAML 4: a program package for phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591.

Zhao Q, Caballero OL, Levy S *et al.* (2009) Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. *Proceedings of the National Academy of Sciences, USA*, **106**, 1886–1891.

The authors are broadly interested in the nature of genetic changes that are associated with speciation. This study is part of Sébastien Renaut's doctoral research, which aims to study the genomic bases of adaptive divergence in the context of a recent ongoing speciation event in lake whitefish. Arne Nolte is interested in the diversity of fishes and understanding the role that environmental and intrinsic factors play in evolution. Louis Bernatchez's research focuses on understanding the patterns and processes of molecular and organismal evolution as well as their significance to conservation.
