MOLECULAR ECOLOGY

Molecular Ecology (2010) **19**, 5389–5403 doi: 10.1111/j.1365-294X.2010.04934.x

The transcriptomics of sympatric dwarf and normal lake whitefish (*Coregonus clupeaformis* spp., Salmonidae) divergence as revealed by next-generation sequencing

JULIE JEUKENS,* SÉBASTIEN RENAUT,* JÉRÔME ST-CYR,* ARNE W. NOLTE† and LOUIS BERNATCHEZ*

**Institut de Biologie Intégrative et des Systèmes (IBIS), Québec-Océan, Université Laval, 1030 av. de la médecine, Québec, QC, G1V 0A6, Canada, †Max Planck Institute for Evolutionary Biology, August-Thienemann-Street 2, 24306 Plön, Germany*

Abstract

Gene expression divergence is one of the mechanisms thought to be involved in the emergence of incipient species. Next-generation sequencing has become an extremely valuable tool for the study of this process by allowing whole transcriptome sequencing, or RNA-Seq. We have conducted a 454 GS-FLX pyrosequencing experiment to refine our understanding of adaptive divergence between dwarf and normal lake whitefish species (*Coregonus clupeaformis* spp.). The objectives were to: (i) investigate transcriptomic divergence as measured by liver RNA-Seq; (ii) test the correlation between divergence in expression and sequence polymorphism; and (iii) investigate the extent of allelic imbalance. We also compared the results of RNA-seq with those of a previous microarray study performed on the same fish. Following *de novo* assembly, results showed that normal whitefish overexpressed more contigs associated with protein synthesis while dwarf fish overexpressed more contigs related to energy metabolism, immunity and DNA replication and repair. Moreover, 63 SNPs showed significant allelic imbalance, and this phenomenon prevailed in the recently diverged dwarf whitefish. Results also showed an absence of correlation between gene expression divergence as measured by RNA-Seq and either polymorphism rate or sequence divergence between normal and dwarf whitefish. This study reiterates an important role for gene expression divergence, and provides evidence for allele-specific expression divergence as well as evolutionary decoupling of regulatory and coding sequences in the adaptive divergence of normal and dwarf whitefish. It also demonstrates how next-generation sequencing can lead to a more comprehensive understanding of transcriptomic divergence in a young species pair.

Keywords: 454 sequencing, adaptive divergence, allelic imbalance, Coregonus, gene expression, RNA-Seq, speciation, transcriptome

Received 7 June 2010; revision received 18 August 2010; accepted 29 August 2010

Introduction

Deciphering the molecular bases of population divergence and speciation is one of the main challenges of evolutionary biology (Edmands 2002; Coyne & Orr 2004; de Queiroz 2005). During the past years, the identification of genes and genomic regions associated with adaptive population divergence has become a highly productive research area (e.g. Hoekstra *et al.* 2006; Joron *et al.* 2006; Schemske & Bierzychudek 2007; Barrett *et al.* 2008). Gene expression divergence is one of the mechanisms thought to underlie phenotypic divergence. Indeed, transcription levels have a great potential for evolutionary novelty, which can then be harnessed by natural selection (Oleksiak *et al.* 2002; Wray *et al.* 2003; Fay & Wittkopp 2008).

Molecular tools designed for the study of gene expression have flourished in the past 15–20 years (Kulesh *et al.* 1987; Schena *et al.* 1995). Namely, microarray

Correspondence: Julie Jeukens, Fax: 1 418 656 7176;
E-mail: julie.jeukens.1@ulaval.ca

technologies allow the simultaneous detection of expression for thousands of genes, thus offering a powerful means to investigate population and species divergence on a transcriptome-wide scale (Ranz & Machado 2006). Moreover, recent advances in sequencing technologies now allow highly efficient whole transcriptome shotgun sequencing (RNA-Seq), which holds the promise of a more informative and accurate view of the transcriptome at lower costs compared with high throughput Sanger sequencing (Wang *et al.* 2009). Given the growing amount of sequence data produced as a result of the increasing accessibility of these new tools, the notion of 'nonmodel organism' is likely to change dramatically in the years to come (Gilad *et al.* 2009; Bernatchez *et al.* 2010).

Lake whitefish (*Coregonus clupeaformis*), which comprises multiple pairs of sympatric forms (normal and dwarf) engaged in the process of speciation, is a powerful system to investigate ecological speciation, according to which phenotypic divergence and ultimately speciation are the outcomes of divergent natural selection (Schluter 2000). Previous studies conducted at the genomic, transcriptomic, phenotypic and ecological levels have yielded substantial insights into adaptive divergence and its genetic bases in the whitefish species complex (reviewed in Bernatchez *et al.* 2010). The limnetic dwarf lake whitefish, despite its young origin (15 000 YBP, Pigeon *et al.* 1997), strikingly differs from the benthic normal whitefish in morphology, but more so in life history traits, metabolism and behaviour. Microarray experiments have led to the identification of expression divergence for genes and key gene functions potentially implicated in the adaptive divergence of dwarf and normal lake whitefish (Derome *et al.* 2006; St-Cyr *et al.* 2008; Nolte *et al.* 2009). A genetic basis has also been demonstrated for traits that differ between both species such as swimming behaviour (Rogers *et al.* 2002), growth (Rogers & Bernatchez 2005), morphology and life history (Rogers & Bernatchez 2007). Moreover, strong intrinsic and extrinsic post-zygotic reproductive barriers have been identified between them (Lu & Bernatchez 1998; Rogers & Bernatchez 2006; Rogers *et al.* 2007; Whiteley *et al.* 2008; Renaut *et al.* 2009). Finally, the integrated use of linkage, phenotypic and gene expression mapping has provided insights into the genetic architecture of adaptive traits differentiating dwarf and normal whitefish, with the identification of key genomic regions that appear to have high pleiotropic effects on gene expression (Derome *et al.* 2008; Whiteley *et al.* 2008).

Next-generation sequencing technologies, especially high throughput pyrosequencing (454 Life Sciences, Margulies *et al.* 2005), have quickly become extremely valuable tools for nonmodel species such as whitefish by making genome sequencing, transcriptome sequencing and single nucleotide polymorphism (SNP) discovery efficient and accessible, even in the absence of a reference genome (Quinn *et al.* 2008; Vera *et al.* 2008; Kristiansson *et al.* 2009; Renaut *et al.* 2010). For instance, 454 pyrosequencing of cDNA allows a combined study of both gene expression and polymorphism. While divergence in expression was hypothesized to be correlated with nonsynonymous substitution rate (amino acid replacement rate) because both should co-evolve when under common selective pressures (Nuzhdin *et al.* 2004), both positive (Nuzhdin *et al.* 2004; Khaitovich *et al.* 2005) and nonsignificant correlations (Kohn *et al.* 2008; Tirosh & Barkai 2008) have been found between these two evolutionary modes. Nevertheless, this type of comparison has rarely been attempted on species as evolutionarily young as dwarf and normal whitefish. Furthermore, measuring allele-specific expression rather than total gene expression (as in the case of cDNA microarrays) can offer even greater insight into regulatory variation by providing direct evidence of *cis* and *trans*-regulatory differences (Wittkopp *et al.* 2004; Guo *et al.* 2008; Serre *et al.* 2008; Graze *et al.* 2009).

Renaut *et al.* (2010) recently conducted a 454 pyrosequencing experiment allowing the assembly of over 130 megabases (Mb) of nonnormalized cDNA as well as the identification of SNP markers for documenting the extent of genetic divergence between dwarf and normal whitefish. Using the same data set, the present study focuses on the transcriptomic characterization of gene expression divergence as seen by RNA-Seq. The objectives were to: (i) determine if phenotypic divergence between dwarf and normal whitefish is associated with transcriptomic divergence as measured by RNA-Seq in the liver; (ii) test the hypothesis that divergence in expression is correlated with nonsynonymous substitution rate; and (iii) investigate the extent of allelic imbalance (AI), defined as a difference in expression levels between alleles of a given gene, within and among species of whitefish. We also used RNA-Seq data to validate a previous microarray study which used the exact same biological samples (St-Cyr *et al.* 2008). In doing so, we gained more refined insights towards a better understanding of the role played by the transcriptome in the adaptive divergence of this young species pair.

Materials and methods

Unless specified otherwise, all data manipulations and statistics were computed in R (v. 2.9.1 The R Foundation for Statistical Computing; scripts available upon request).

Sample preparation and sequencing

While all data analyses and interpretations are novel, raw data analysed in this study have been previously published (Renaut *et al.* 2010) and are accessible via the Sequence Read Archive at NCBI [normal whitefish (SRA:SRX010969), dwarf whitefish (SRA:SRX011025)]. The 16 samples used for sequencing were the same as those of a previous study (St-Cyr *et al.* 2008). Briefly, sympatric dwarf and normal whitefish were collected in Cliff Lake (46°23′59″N, 69°15′11″W), located in the Allagash River basin, Maine, USA. Normal and dwarf fish were concurrently caught in the same nets, during the same time period, in June 2003. Nets were pulled every 30 min, ensuring that the fish were still alive prior to tissue collection. Eight adult individuals, half males, half females, were randomly collected from each population. Liver tissue samples were immediately frozen in liquid nitrogen, and later stored at −80 °C. The liver was selected for its role in growth regulation (Rise *et al.* 2006) and food consumption related functions (Trudel *et al.* 2001; St-Cyr *et al.* 2008). Total RNA was extracted using the TRIzol Reagent protocol (Invitrogen, Carlsbad, CA, USA). All RNA samples were cleaned by ultra filtration using microcon spin columns (Millipore, Billerica, MA, USA). Sample quality and concentration were assessed with the Experion™ RNA StdSens Analysis kit (Bio-Rad, Hercules, CA, USA). Samples were stored in RNase free water supplemented with Superase-In™ RNase Inhibitor (Ambion, Austin, TX, USA) at −80 °C.

All samples were subsequently enriched for polyA mRNA using the MicroPoly(A)Purist™ kit (Ambion). Approximately 100 ng of complementary DNA (cDNA) was synthesized from each polyA mRNA sample using the SMART™ PCR cDNA Synthesis kit (Clontech, Mountain View, CA, USA). All cDNA samples (3–8 ng) were PCR amplified using the Advantage 2 PCR kit (Clontech) and modified SMART™ primers (5′-AAGCAGTGGTATCAACGCAGAGT-3′), which comprised an extra five nucleotides at the 5′ end for individual-specific tagging. PCR conditions were as follows: initial denaturation for 1 min at 95 °C; 17–20 cycles depending on the sample (15 s at 95 °C, 30 s at 65 °C, 6 min at 68°C). Following amplification, all samples were quantified using the Quant-iT Picogreen dsDNA Assay kit (Invitrogen) and two separate pools with equal cDNA quantities were prepared: one with the eight normal samples and one with the eight dwarf samples. Nonnormalized cDNA was used to enable gene dosage (i.e. gene expression) in addition to gene discovery. The two cDNA pools were sequenced 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). A first run was performed with a half plate for each pool and a second run was performed with a quarter plate for each pool.

Contig assembly

Initial quality filtering of 454 sequence data was performed using Roche proprietary analysis software Newbler (Margulies *et al.* 2005). Base calling was re-done using PyroBayes, which produces more confident results compared with the 454 program (Quinlan *et al.* 2008). Prior to assembly, primer and sample-specific tag sequences were removed from the data set using a custom made Perl script (available upon request). CLC Genomics Workbench 3.1 (CLC Bio, Aarhus, Denmark) was used to assemble all normal and dwarf sequences into contigs (sets of overlapping DNA segments) *de novo* (minimum match percentage 0.97, overlap 0.33, global alignment). The choice of this specific minimum match percentage was based on a compromise between the risk of assembling many more potential paralogous sequences using a smaller value and that of discarding more sequence data using a higher value (see Renaut *et al.* 2010). Contig consensus sequences were screened for ribosomal RNA by local blast using all salmon putative 18S, 5S and 28S sequences available in GenBank. Reads that did not assemble into contigs were discarded.

To assess transcriptome completeness, a rarefaction curve was computed. This approach, normally used in ecology to determine the expected number of species as a function of the number of individuals sampled, was recently applied to whole transcriptome sequencing (Hale *et al.* 2009). Rarefaction curves generally grow rapidly as the most common 'species' are found and then plateau when only the rarest species remain to be sampled. Here, we computed a curve of the number of contigs detected as a function of the number of mapped reads (i.e. assembled reads) randomly sampled.

Gene expression analyses

Consensus sequences of the contigs were used as a reference for the RNA-Seq analysis, performed with CLC Genomics Workbench 3.1. This analysis was performed separately for normal and dwarf whitefish data by reassembling the raw sequence data onto the reference consensus sequences and transforming sequence counts per contig into reads per kilobase per million mapped reads (RPKM, Mortazavi *et al.* 2008). RPKM values are used to standardize sequence counts as a function of contig length and absolute data set size (i.e. mapped reads: assembled reads) and are calculated as follows: (number

of reads per contig) × (1000/length of contig) × (1×10^6/total number of mapped reads in the assembly). Only for contigs represented by 10 or more reads were differences in expression between dwarf and normal whitefish tested with a chi-square test (Eveland *et al.* 2008) and corrected for multiple hypotheses testing with Qvalue (Storey 2002). GOanna was used for automated contig annotation through blastx of consensus sequences (McCarthy *et al.* 2007). This online tool transfers Gene Ontology (GO) annotations from annotated gene products in other species [including Atlantic salmon (*Salmo salar*) and rainbow trout (*Oncorhynchus mykiss*)] based on blast searches against databases that contain GO annotated sequences and returns both blast results and GO annotations. Through parsing of the GOanna output, the GO 'biological process' term of the best blast hit (*E*-value $<1 \times 10^{-5}$) was selected and classified into one of 16 functional categories according to Table S1 (Supporting information). These categories were subsequently used to identify overrepresentation of gene functions between contigs overexpressed in dwarf fish and contigs overexpressed in normal fish using Fisher's exact test.

Relationship between gene expression and polymorphism

To assess whether polymorphism was related to gene expression, we used stringent criteria to identify SNPs and subsequently parsed them into different polymorphism types. First, all possible open reading frames (ORF, minimum length of 30 codons) for each contig were produced using CLC Genomics Workbench 3.1 and submitted to protein blast (blastp) searches with GOanna. The sequence with the best blast hit (highest score, *E*-value $<1 \times 10^{-5}$) was selected. As for contigs with no blast hit, the longest possible ORF was kept as the most probable translated region. Second, CLC Genomics Workbench was used for SNP identification (as described in Brockman *et al.* 2008) with the following parameters: window of 11 bp, maximum gap + mismatch of 2, minimum average quality of 20, minimum SNP position quality of 25, minimum coverage of 10, minimum SNP frequency of 33%. Third, SNPs were classified as being inside (coding) or outside (noncoding) selected ORFs. Fourth, using all ORF sequences and their coding SNPs, maximum likelihood estimates were used to compute nonsynonymous and synonymous substitution rates with PAML 4.2 (runmode = −2, CodonFreq = 2, model = 2, Yang 2007). These actually corresponded to polymorphism rates, p_n and p_s, because SNPs were generally not fixed between normal and dwarf whitefish. Therefore, we also computed a divergence index equal to absolute [frequency($allele1_{Dwarf}$) −frequency($allele1_{Normal}$)] for each SNP (Renaut *et al.* 2010). Finally, Pearson's coefficient correlation statistics were calculated to test the correlation between gene expression divergence and (i) noncoding SNPs per noncoding site; (ii) p_n; (iii) p_s; and (iv) the sequence divergence index.

Allelic imbalance

Because individuals possessed unique tags in our data set (only at cDNA sequence extremities, see Sample preparation and sequencing), we could also investigate allelic imbalance (AI, e.g. Wittkopp *et al.* 2004; Graze *et al.* 2009), defined as unequal expression levels between alleles of a given gene in a cell (Yan *et al.* 2002). Thus, we first computed per individual allele-specific expression levels for each SNP and discarded SNPs that did not have at least one individual expressing both alleles, which was the only way of ensuring the presence of both alleles in a single genotype. AI was then tested with a paired *t*-test (two-sided) for each SNP and pool (dwarf or normal), which was subsequently corrected for multiple testing with Qvalue (Storey 2002). In other words, the number of reads for allele 1 was compared with the number of reads for allele 2 for each individual of a pool, thereby testing the null hypothesis of per-individual equal expression of alleles for each SNP. Then, because the *t*-test was expected to be highly significant for SNPs with a majority of homozygous individuals, an AI index was calculated such that each individual with imbalance in favour of allele 1 (where allele 2 $\neq$ 0) was given a score of +1 and each individual with imbalance in favour of allele 2 (where allele 1 $\neq$ 0) was given a score of −1. A score of 0 was given for individuals expressing only one allele or equal levels of the two alleles. The sum of these individual scores was calculated separately for dwarf and normal whitefish pools, with the resulting index ranging from −8 to 8. Finally, for a given SNP to be considered a candidate case of AI, it needed a significant paired *t*-test and an AI index $\neq$ 0 either in dwarf, normal or both whitefish species.

Comparison with microarray results

We then compared RNA-Seq results with previous microarray results (St-Cyr *et al.* 2008), which are available via the Gene Expression Omnibus [GEO:GSE21130]. First, sequences from the 16 K GRASP salmonid cDNA microarray (Genomic research on Atlantic salmon project, Rise *et al.* 2004; von Schalburg *et al.* 2005) were used as the reference transcriptome for RNA-Seq analysis, which was performed as described above. As such, this analysis imitated the previous microarray

experiment. Second, only genes that were (i) expressed according to both methods; (ii) expressed in dwarf and normal fish (because this was always the case in the microarray data set); and (iii) represented by at least 10 sequencing reads were selected. Then, selected RPKM values were $\log_2$-transformed and a dwarf/normal ratio of expression was computed for each gene as was done with the microarray data (normalized R/Lowess signal intensity in $\log_2$, St-Cyr *et al.* 2008). The correlation between both data sets was tested using Pearson's coefficient correlation statistic. Finally, 34 candidate expressed sequence tags (EST) that were identified as differentially expressed in liver between dwarf and normal whitefish from two natural lakes and a controlled environment by St-Cyr *et al.* (2008) were validated with RNA-Seq results. This was performed by associating each candidate microarray EST sequence to its corresponding *de novo* contig consensus sequence by local blast. Results for both methods were then confronted.

Results

Contig assembly and differential gene expression

Table 1 presents a summary of the raw sequence data set (Renaut *et al.* 2010) and its assembly into contigs [see Table S2 (Supporting information) for a complete list of *de novo* contigs with annotation and expression data; Dryad Digital Repository: doi:10.5061/dryad.1924]. Three contigs (1318 reads) were identified as ribosomal RNA and approximately 5% of contigs effectively annotated with GOanna were putative ribosomal proteins. According to the rarefaction curve (Fig. 1), approximately half of all mapped reads, i.e. ~100 000, was sufficient to detect 95% of all contigs assembled.

Of 1953 contigs assembled *de novo*, the normal whitefish transcriptome contained 1827 contigs, the dwarf whitefish transcriptome contained 1794 contigs and 948 contigs were significantly differentially expressed (≥10 reads, χ^2 test q-value ≤0.01) between dwarf and normal whitefish. This translated into differential representation

Table 1 Summary of pyrosequencing data and assembly

N pool + D pool*	Count	Average length (bp)
Total reads	395 950	209
Assembled reads *de novo*	225 881	218
Contigs†	1953	459
Contigs with number of reads ≥10†	1198	597

*Normal (N) and dwarf whitefish (D) pools: eight liver cDNA samples for each species, 0.75 plate of 454 GS-FLX sequencing each.
†According to *de novo* assembly.

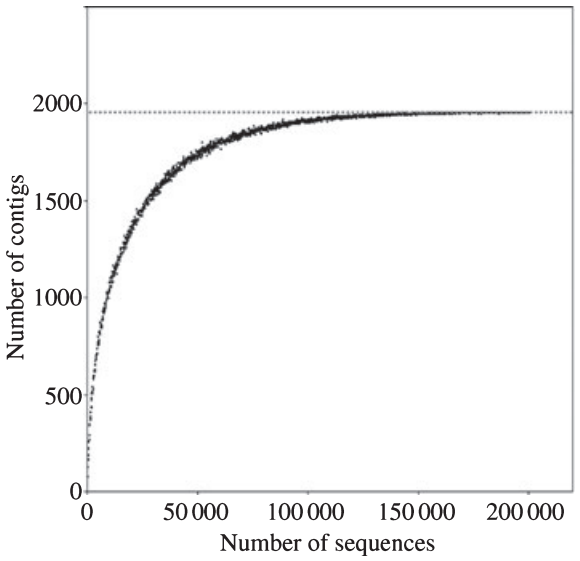

Fig. 1 Rarefaction curve of the whitefish liver transcriptome. Number of contigs detected as a function of the number of reads randomly sampled from the 225 881 reads assembled *de novo*; Grey: 1953 contigs in the complete data set.

of several functional categories of genes (Fig. 2, Table S3 Supporting information). Namely, normal whitefish showed significantly more overexpressed genes associated with protein synthesis (116 contigs) while dwarf fish significantly showed more overexpressed genes related to energy metabolism (51 contigs), immunity (22 contigs), as well as DNA replication and repair (29 contigs). The proportion of unknown contigs (i.e. with no GO biological process term) was not significantly different between contigs overexpressed in normal whitefish and contigs overexpressed in dwarf whitefish (P-value = 0.946, Table S3 Supporting information).

Tables 2 and 3 present contigs with the highest levels of expression divergence between dwarf and normal whitefish. According to expression ratio (D/N, ratio of dwarf and normal RPKM, Table 2), protein synthesis was the most pervasive functional category among the 10 most overexpressed contigs in normal whitefish while candidate contigs overexpressed in dwarf whitefish were associated with a more diverse array of functions, including DNA replication and repair for the three most overexpressed contigs. As for candidate contigs based on absolute expression difference [abs(D-N), absolute difference in RPKM, Table 3], functional categories were also very diverse, with no functional trend.

Relation between gene expression and polymorphism

Figure 3 presents scatter plots of gene expression divergence on the y axis, represented by D/N and abs(D-N),

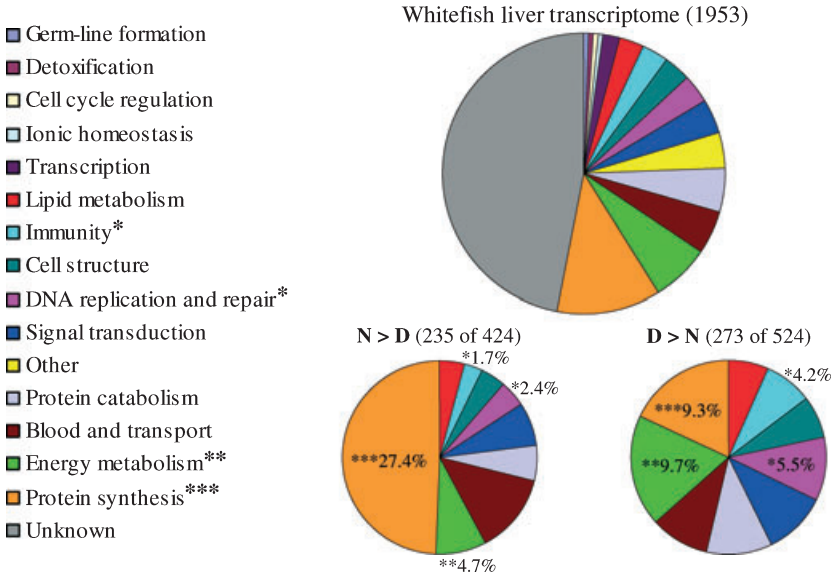

Fig. 2 Whitefish liver transcriptome divided into functional gene categories. Functional categories described in Table S1 (Supporting information) (Unknown: no Gene Ontology (GO) biological process term). Between parentheses: number of contigs represented; N > D: contigs significantly overexpressed in normal whitefish (χ^2 test, ≥10 reads, *q*-value ≤0.01); D > N: contigs significantly overexpressed in dwarf whitefish (χ^2 test, ≥10 reads, *q*-value ≤ 0.01); N > D and D > N: only nine major categories are represented, percentages are relative to the total number of overexpressed contigs; **P*-value ≤0.05 (Fisher's exact test between N > D and D > N); ***P*-value ≤0.01; ****P*-value ≤0.001.

Table 2 Top 20 candidate contigs according to RNA-Seq (criterion 1)

Rank*	Contig no.†	D/N ratio‡	Putative gene name	Functional category§
D1	552	4.75	Retrotransposable element Tf2 type 3	DNA replication and repair
D2	1217	3.90	Uricase	DNA replication and repair
D3	824	3.24	Reverse transcriptase	DNA replication and repair
D4	1744	3.04	Haemoglobin subunit alpha-4	Blood and transport
D5	1736	3.04	Diablo homolog, mitochondrial	Cell cycle regulation
D6	102	2.78	Retrotransposon-like protein 1	Protein catabolism
D7	1642	2.72	UPF0362 protein C20orf149 homolog 2	Unknown
D8	742	2.68	Myosin heavy chain, fast skeletal muscle	Other
D9	12	2.60	Growth arrest and DNA-damage-inducible protein	Signal transduction
D10	1310	2.56	Apolipoprotein A-IV	Lipid metabolism
N1	1777	0.16	60S ribosomal protein L36	Protein synthesis
N2	1849	0.21	Ribosomal protein S29	Protein synthesis
N3	1242	0.23	C-type lectin domain family four member E	Unknown
N4	237	0.23	60S ribosomal protein L30	Protein synthesis
N5	1600	0.26	60S ribosomal protein L36	Protein synthesis
N6	9	0.26	Inositol oxygenase	Signal transduction
N7	1859	0.26	Ribosomal protein, large P1, like 2	Protein synthesis
N8	1453	0.29	Ribosomal protein S16	Protein synthesis
N9	1521	0.29	Heat shock 90kDa protein 1 beta isoform a	Protein synthesis
N10	1434	0.35	40S ribosomal protein S28	Protein synthesis

*Top 10 of contigs overexpressed in dwarf (D) whitefish liver and top 10 of contigs overexpressed in normal (N) whitefish liver. Both D and N had to have at least 10 reads according to the RNA-Seq analysis. Unannotated contigs were excluded.
†Numerical tag assigned by CLC Genomics Workbench during *de novo* assembly.
‡Criterion: Ratio of standardized sequence counts (RPKM).
§Described in Table S1 (Supporting information).

Table 3 Top 20 candidate contigs according to RNA-Seq (criterion 2)

Rank*	Contig no.†	abs(D-N)‡	Putative gene name	Functional category§
D1	1624	7353	Ribosomal protein S26	Protein synthesis
D2	1230	6365	Cystein proteinase inhibitor protein	Unknown
D3	1647	4435	Ribosomal protein S6-2	Protein synthesis
D4	1593	3551	Apolipoprotein C-I	Lipid metabolism
D5	1845	3219	Cystein proteinase inhibitor protein	Unknown
D6	1620	3077	RAB39B, member RAS oncogene family	Signal transduction
D7	1285	2955	Mid1-interacting protein 1	Unknown
D8	43	2724	Complement C4	Immunity
D9	1934	1998	Complement factor H1 protein	Immunity
D10	1819	1812	Retinol dehydrogenase 3	Other
N1	1211	23840	Proteasome 26S subunit	Protein catabolism
N2	1391	13513	Apolipoprotein A-I	Lipid metabolism
N3	1093	6894	Serum albumin 2	Blood and transport
N4	1486	6125	Vitellogenin	Lipid metabolism
N5	1910	5165	Serum albumin 2	Blood and transport
N6	1760	4988	Ribosomal protein L37a	Protein synthesis
N7	1810	4378	Vitellogenin	Lipid metabolism
N8	1890	3880	Ribosomal protein S5-2	Protein synthesis
N9	1335	3860	Vitellogenin	Lipid metabolism
N10	1630	3505	Ribosomal protein S2	Protein synthesis

*Top 10 of contigs overexpressed in dwarf (D) whitefish liver and top 10 of contigs overexpressed in normal (N) whitefish liver. Both D and N had to have at least 10 reads according to the RNA-Seq analysis. Unannotated contigs were excluded.
†Numerical tag assigned by CLC Genomics Workbench during *de novo* assembly.
‡Criterion: Absolute difference in standardized sequence counts (RPKM).
§Described in Table S1 (Supporting information).

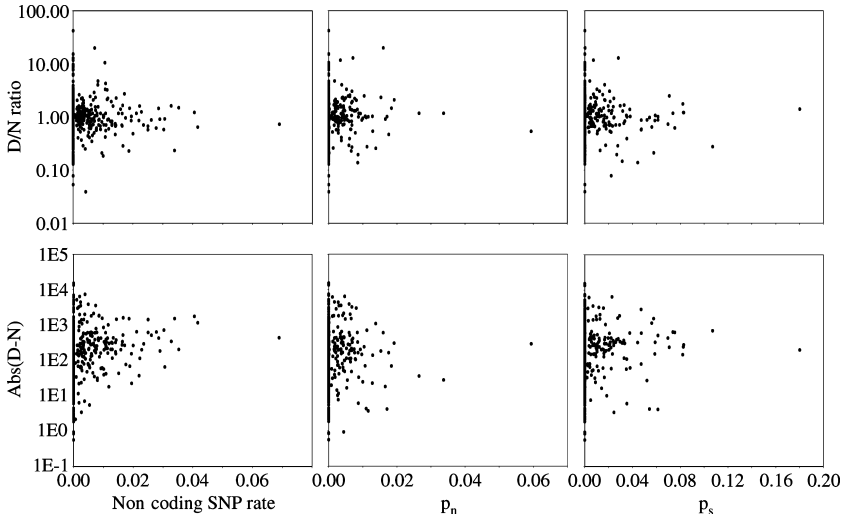

Fig. 3 Scatter plots of gene expression divergence measured with RNA-Seq and polymorphism rate. All contigs represented have a mean coverage ≥10. D/N: ratio of dwarf (D) and normal (N) standardized (RPKM) sequence counts, log-scale; Abs(D-N): absolute difference in standardized (RPKM) sequence counts, log-scale; Noncoding SNP rate: number of SNPs per noncoding site [not inside an open reading frame (ORF)]; p_n: number of nonsynonymous SNPs per nonsynonymous site; p_s: number of synonymous SNPs per synonymous site.

and three measures of polymorphism on the *x* axis: noncoding SNPs per noncoding site, nonsynonymous SNPs per nonsynonymous site (p_n) and synonymous SNPs per synonymous site (p_s). Only contigs with a mean coverage ≥10 were plotted, as most contigs under this threshold did not meet SNP detection criteria. Results showed neither positive nor negative trends but instead a cone-shaped pattern, wherein genes that are the most extreme in terms of gene expression divergence show no or little polymorphism. This pattern did not seem to be influenced by polymorphism type and was even observed on scatter plots of gene expression divergence as measured by microarray and polymorphism rates, either calculated from SNPs among whitefish samples or between whitefish and other salmonid sequences spotted on the array (results not shown).

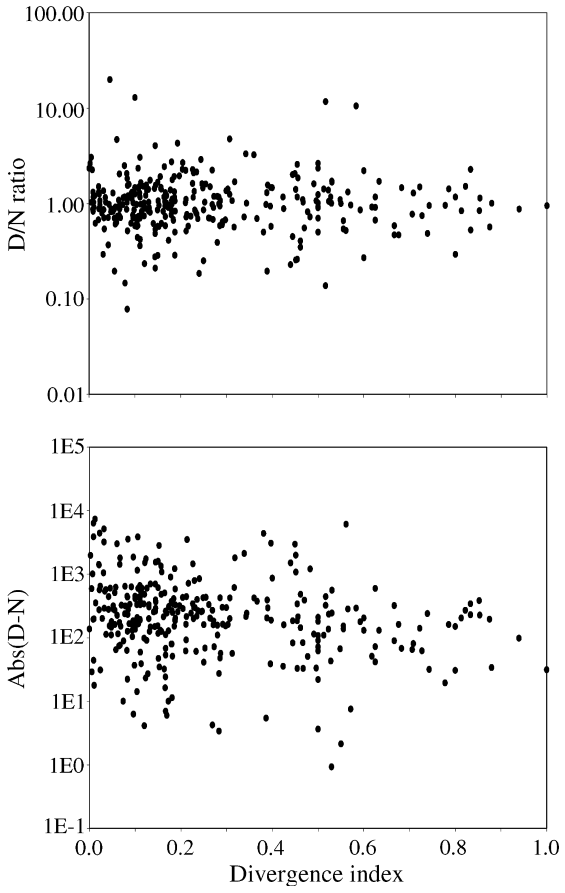

Fig. 4 Scatter plots of gene expression divergence measured with RNA-Seq and sequence divergence. All contigs represented have a mean coverage ≥10 and at least one SNP. D/N: ratio of dwarf (D) and normal (N) standardized (RPKM) sequence counts, log-scale; Abs(D-N): absolute difference in standardized (RPKM) sequence counts, log-scale; Divergence index: mean of absolute[frequency(allele1$_{Dwarf}$) − frequency(allele1$_{Normal}$)] for all SNPs in a contig.

In contrast, Fig. 4 presents scatter plots of gene expression divergence [represented by D/N and abs (D-N)] and mean sequence divergence between dwarf and normal whitefish. It also differs from Fig. 3 in that only contigs with at least one SNP are represented. These results clearly show the absence of correlation between the amount of sequence divergence at polymorphic sites and gene expression divergence.

Allelic imbalance

Following *de novo* assembly, 46% of mapped reads were tagged (cDNA sequence extremities, see Sample preparation and sequencing). Tagged reads parsing revealed that the proportion of tagged over total mapped reads was the same for normal and dwarf pools and that the variance of the number of reads per individual fish was also the same between normal and dwarf pools (CV = 0.25).

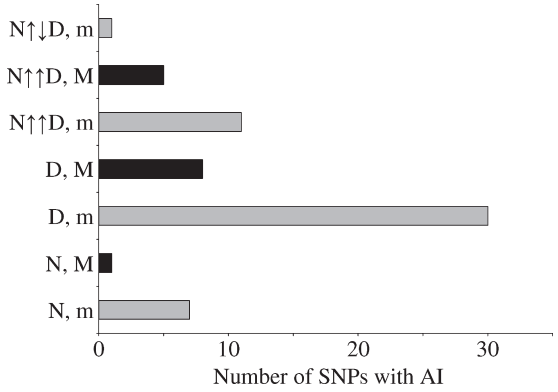

Fig. 5 Candidate SNPs for allelic imbalance. All SNPs represented have a coverage ≥10. AI: allelic imbalance, i.e. unequal expression levels between alleles of a SNP; N: normal whitefish; D: dwarf whitefish; N↑↑D: overrepresentation of the same allele in N and D; N↑↓D: overrepresentation of different alleles in N and D; overrepresentation of an allele: m (grey): minor candidate, ≤4 individuals with AI, *q*-value ≤0.1 (paired *t*-test); M (black): major candidate, >4 individuals with AI, *q*-value ≤0.05 (paired *t*-test).

Tagged reads were used to investigate per individual allele-specific expression levels. Of 685 SNPs that were polymorphic in the transcriptome of at least one individual, 63 (9.2%) showed significant allelic imbalance (AI, Fig. 5). These were classified according to the presence of AI in a minority (≤4, *q*-value ≤0.1) or a majority (>4, *q*-value ≤0.05) of individuals and if this was in dwarf, normal or both whitefish species. Dwarf whitefish clearly had the highest number of AI cases, with 38 SNPs (30 minor, eight major) showing significant AI in dwarf whitefish only, as opposed to 8 SNPs for normal whitefish (seven minor and one major). The second most common type of AI was overrepresentation of the same allele in both normal and dwarf whitefish, with a total of 16 SNPs (11 minor and five major). The 63 candidate SNPs corresponded to 43 contigs (Table S4 Supporting information), almost half of which were implicated in protein synthesis (19 contigs). Among other contigs with candidates SNPs for AI, five were implicated in transport functions, three had immunity related functions, two were implicated in energy metabolism, two were related to signal transduction, one was implicated in DNA replication and repair and one was implicated in transcription. While the abovementioned functional categories were evenly distributed across contigs showing AI in dwarf, normal and both forms of whitefish, the overall proportions of contigs related to protein synthesis and immunity were significantly higher among these 43 contigs than in all differentially expressed contigs

(Fisher's exact test, P-value = 0.0001 and P-value = 0.04 respectively).

Comparison with microarray results

The correlation between dwarf to normal (D/N) ratios of gene expression for common genes between RNA-Seq and a previous microarray data set (St-Cyr *et al.* 2008) was positive and highly significant, but modest (r = 0.38, P-value = 2.2 e−16, Fig. S1 supporting information). Table 4 shows 34 candidate ESTs for differential expression between dwarf and normal whitefish according to the previous microarray study. Differential expression for half of the candidates was confirmed when using a criterion of q-value <0.01, while the use of a more stringent criterion of q-value <0.0001 (and D/N > 1.25) led to the confirmation of about 25% of candidate ESTs.

Discussion

Contig assembly and differential gene expression

RNA-Seq is assumed to provide a reliable estimate of absolute gene expression levels (Fu *et al.* 2009). However, while sources of bias in gene expression estimates based on next-generation sequencing data are not yet fully understood, it is becoming clear that base composition, library preparation and transcript length all have an impact on RNA-Seq data (Marioni *et al.* 2008; Gilad *et al.* 2009; Oshlack & Wakefield 2009). Recent technological advances already offer promising solutions to overcome at least some of these issues, namely through direct single molecule RNA sequencing (Ozsolak *et al.* 2009). Here, samples were prepared and sequenced in parallel and our data suggest that efforts to reduce sources of bias were successful. Namely, analysis of tagged reads revealed that normal and dwarf pools had equal proportions of tagged reads and equal relative variance in per-individual sequence number. The RPKM method was used to normalize for the difference in data set size between normal and dwarf pools (Mortazavi *et al.* 2008). This method is appropriate only if the overall composition of the RNA population is equivalent between samples (Robinson & Oshlack 2010). Indeed, if one of two conditions has a large number of uniquely or highly expressed genes, the expression of the other genes, i.e. common genes between both conditions, is expected to be low for a given sequencing effort. Here, all samples came from whitefish adult livers and among contigs with 10 or more reads, the dwarf and normal pools only had four and 11 uniquely expressed contigs respectively. Moreover, all uniquely expressed contigs were lowly expressed (≤20 reads).

Although the starting material was nonnormalized cDNA, the rarefaction analysis clearly demonstrated that our sampling of the whitefish liver transcriptome was sufficient to reliably represent contig diversity. This is consistent with a previous RNA-Seq study in lake sturgeon gonads where comparative rarefaction computation for normalized and nonnormalized cDNA demonstrated that gene discovery was equivalent with 5000 or more reads (Hale *et al.* 2009). In total, we identified 1953 contigs, with a mean length of 459 bp. A comparable study of lake trout (*Salvelinus namaycush)* liver cDNA using the same sequencing technology, similar sequencing effort and CLC Genomics Workbench for contig assembly (Goetz *et al.* 2010) yielded highly similar results, i.e. 2276 contigs from 425 821 reads. However, it is difficult to compare our results with other studies that used programs that were not specifically created for 454 data, CAP3 for instance, as they typically lead to the assembly of at least 10 times as many contigs with very few reads per contig on average (e.g. Kristiansson *et al.* 2009).

Figure 2 clearly demonstrated transcriptomic divergence between dwarf and normal whitefish. Like the previous microarray experiment (St-Cyr *et al.* 2008), these results are consistent with the observed trade-off in life history traits among whitefish species pairs, wherein dwarfs have a higher metabolic rate, necessary for increased foraging and predator avoidance in the limnetic niche, while normal whitefish allocate a much larger fraction of their energy budget to growth (Trudel *et al.* 2001). Dwarf whitefish also overexpressed more genes related to immunity, which is widely recognized as a potent feature of local adaptation that can underlie population divergence (e.g. Dionne *et al.* 2007) and speciation (Turelli *et al.* 2001; Buckling & Rainey 2002; Eizaguirre *et al.* 2009). Goetz *et al.* (2010) observed significant overexpression of genes associated with the immune function in limnetic (vs. benthic) lake trout and suggested that it could be related the increased temperature variation and pathogen exposure experienced by these fish. DNA replication and repair related functions were also overrepresented in dwarf compared with normal whitefish, and it could be argued that the higher metabolic rate of the former (Trudel *et al.* 2001) induces more DNA damage and therefore more active DNA repair pathways. This idea has received a reasonable amount of support in mammals (Adelman *et al.* 1988; Foksinski *et al.* 2004), but it remains hypothetical for whitefish species pairs until further investigation is performed. Nevertheless, as the number of differentially expressed genes between normal and dwarf whitefish was relatively large, mainly because RPKM normalization automatically increased all numerical values, it should be noted that results at the level of functional

Table 4 RNA-Seq validation of previous microarray results

Functional group	EST clone number and putative gene name[1]	Microarray		RNA-Seq[4]				
		Ratio[2]	*P*-value[3]	Contig[5]	Ratio[6]	Abs(D-N)[7]	*q*-value[8]	Validation[9]
Cell cycle regulation	CB517934 ARF GTPase-activating protein GIT2	1.37	0.0038	na	na	na	na	
	CB492176 Polyposis locus protein 1 homolog	0.86	0.0220	na	na	na	na	
Detoxification	CA057214 Liver carboxylesterase 22 precursor	1.35	0.0019	194	1.12	106.0	0.0026	√*
	CB496876 Liver carboxylesterase 22 precursor	1.33	0.0030	194	1.12	106.0	0.0026	√*
	CB496493 Glutathione S-transferase Mu 5	1.13	0.0258	1507	1.17	24.5	0.0291	
	CB497579 Glutathione S-transferase Mu 5	1.10	0.0347	1507	1.17	24.5	0.0291	
Energy metabolism	CB516178 Anionic trypsin II precursor	1.55	0.0096	na	na	na	na	*
	CB515463 Elastase 2 precursor	1.23	0.0208	na	na	na	na	
	CA045033 Anionic trypsin II precursor	1.24	0.0271	na	na	na	na	*
	CA062911 Fructose-bisphosphate aldolase A	1.21	0.0151	na	na	na	na	
	CB502483 Fructose-bisphosphate aldolase B	1.19	0.0094	**1208**	**1.41**	**139.8**	**<0.0001**	√
	CB497681 Glyceraldehyde-3-phosphate dehydrogenase	1.13	0.0459	1661	1.02	550.3	0.0050	√
	CA768062 Glyceraldehyde-3-phosphate dehydrogenase	1.16	0.0080	1661	1.02	550.3	0.0050	√
	CB493574 Glyceraldehyde-3-phosphate dehydrogenase	1.23	0.0081	1661	1.02	550.3	0.0050	√
	BU965756 Glyceraldehyde-3-phosphate dehydrogenase	1.31	0.0076	1322	1.16	135.7	0.0003	√
	CB498361 Glyceraldehyde-3-phosphate dehydrogenase	1.23	0.0072	1322	1.16	135.7	0.0003	√
	CB514460 Glyceraldehyde-3-phosphate dehydrogenase	1.26	0.0073	1322	1.16	135.7	0.0003	√
	CB491826 Glyceraldehyde-3-phosphate dehydrogenase	1.18	0.0345	1322	1.16	135.7	0.0003	√
	CA055883 Homogentisate 1,2-dioxygenase	1.28	0.0051	**1284**	**1.86**	**156.8**	**<0.0001**	√
	CB493498 Malate dehydrogenase, cytoplasmic	1.34	0.0013	**1928**	**2.37**	**418.1**	**<0.0001**	√
	CB518115 Malate dehydrogenase, cytoplasmic	1.25	0.0074	**1928**	**2.37**	**418.1**	**<0.0001**	√
	CA062141 Cytochrome c oxidase polypeptide VIa-heart	0.87	0.0339	na	na	na	na	
Germ-line formation	CA062348 Estradiol 17-beta-dehydrogenase 2	1.12	0.0256	**1819**	**1.69**	**1811.7**	**<0.0001**	√
	CB496948 Prostaglandin-H2 D-isomerase precursor	1.20	0.0132	**681**	**1.46**	**449.2**	**<0.0001**	√
Immunity	CA037686 C1q-like adipose specific protein	1.40	0.0011	na	na	na	na	
Iron homeostasis	CB515893 Heme oxygenase	1.83	0.0001	na	na	na	na	
Lipid metabolism	CK990220 Adipocyte Fatty acid-binding protein	1.28	0.0032	na	na	na	na	
Protein synthesis	CA063352 Peptidyl-prolyl cis-trans isomerase B precursor	0.84	0.0244	67	0.92	14.4	0.0690	
	CA048973 Protein disulfide-isomerase A4 precursor	0.75	0.0036	na	na	na	na	
Unknown	CA063623 Unknown	1.83	<0.0001	na	na	na	na	
	CB509889 Unknown	1.23	0.0190	**681**	**1.46**	**449.2**	**<0.0001**	√
	CB498458 Unknown	1.28	0.0072	**1593**	**1.54**	**3551.0**	**<0.0001**	√
	CK990521 Unknown	0.79	0.0047	na	na	na	na	
	CA045102 Unknown	0.80	0.0073	na	na	na	na	

All expressed sequence tags (EST) represented showed parallel changes in expression between dwarf and normal whitefish from two natural lakes and controlled environmental conditions according to a previous microarray study (modified from St-Cyr *et al*. 2008). Microarray and RNA-Seq data represented are for the same eight normal and eight dwarf whitefish from Cliff Lake, MA, USA.

[1]EST clone number represents a single expressed sequence tags (EST) sequence on the microarray.

[2]Mean dwarf expression level divided by mean normal expression level (D/N; normalised R/Lowess signal intensity in $\log_2$).

[3]Permutated *P*-values (ANOVA, F3 test, 1000 permutations).

[4]na: insufficient data (<10 reads) or no data; bold characters: *q*-value <0.0001 (and ratio >1.25).

[5]Identified by local blast.

[6]Ratio of dwarf (D) and normal (N) standardized (RPKM) sequence counts.

[7]Absolute difference in standardized (RPKM) sequence counts.

[8]χ^2 test (N, D) and correction with Qvalue software (Storey 2002).

[9]A check mark (√) corresponds to a *q*-value <0.01, *the expression difference was validated by real-time PCR in a previous study (Jeukens *et al*. 2009).

categories were the same when read counts were only normalized to 200 000 total reads for each pool and only 332 contigs were significant for differential expression using the same statistical approach.

Another way of extracting ecologically relevant information from transcriptome-wide expression data is through a candidate gene approach, as was shown in Tables 2 and 3. In addition to the functional categories identified in Fig. 2, these results revealed candidate genes associated with blood and transport, cell cycle regulation, protein catabolism, signal transduction and lipid metabolism. A noteworthy example is putative haemoglobin subunit alpha-4 (rank D4, Table 2), as haemoglobin subunits are the object of ongoing research in whitefish, both at the transcriptional and sequence levels (Evans ML & Bernatchez L, unpublished data). Tetrameric haemoglobin mediates oxygen transport in the blood and the study of its genes in vertebrates has lead to the elucidation of molecular and structural bases for physiological adaptation to temperature and altitude variations (e.g. Storz *et al.* 2009; Campbell *et al.* 2010). Likewise, elevated haemoglobin transcription in dwarf compared with normal whitefish could be related to their divergent metabolic rates thought to be directly associated with the bioenergetic cost incurred by habitat selection in the limnetic environment (Trudel *et al.* 2001; Rogers *et al.* 2002; Rise *et al.* 2006).

Relation between gene expression and polymorphism

The interpretation of our SNP data should always consider potential paralogy (Renaut *et al.* 2010), especially given that salmonids have pseudo-tetraploid genomes due to recent whole genome duplication (Allendorf & Thorgaard 1984). Indeed, according to direct validation approaches and model fitting in salmon, 8–20% of putative polymorphic contigs were expected to be paralogous or multiple sequence variants (PSVs or MSVs, Hayes *et al.* 2007; Moen *et al.* 2008). Moreover, according to model fitting, the average SNP density (SNPs/bp) in duplicated regions was about three times that of unduplicated regions. By applying these results to our data, it could be reasonable to assume that about 80% of polymorphic contigs are true variants of a single gene and that their frequency distribution is skewed towards the lower range of observed polymorphism rates.

Phenotypes evolve through changes in both protein structure and gene expression. However, the relationship and relative importance of both mechanisms is poorly understood (Hoekstra & Coyne 2007). Although some studies have reported significant positive correlations between both modes of evolution (e.g. Nuzhdin *et al.* 2004; Khaitovich *et al.* 2005), Tirosh & Barkai (2008), after observing no such correlation in yeast, suggested that the strength of negative selection can vary between them, such that certain genes are more sensitive to changes in coding sequences, whereas others are more sensitive to changes in expression. Comparing results obtained here with those of previous studies is not straightforward, as in comparison with other systems investigated thus far (but see below), dwarf and normal whitefish divergence is still in progress and nearly no fixed genetic differences exist between them (only two fixed SNPs identified in this study). Moreover, allele frequencies are actually allele-specific expression frequencies and therefore, an imperfect representation of true genomic frequencies in the population. However, our results showed a complete absence of correlation between gene expression divergence and any polymorphism rate. They rather showed a seemingly cone-shaped trend, where the genes that are the most extreme in gene expression divergence showed no or little polymorphism. While few data points exceeded 25% of the range of any polymorphism rate, this idiosyncratic pattern remained pervasive across panels of Fig. 3. Gene expression divergence and the sequence divergence index were also uncorrelated. Such results could be consistent with the aforementioned idea of Tirosh & Barkai (2008): as a whole, coding-sequence divergence and gene expression divergence are uncorrelated because each one is correlated with different properties of genes, that is, protein structure vs. expression regulation.

Our results are also in line with what was observed among evolutionarily young strains of *Drosophila melanogaster*, where differentiation of gene expression and of coding sequences were also uncorrelated (Kohn *et al.* 2008). Moreover, divergence in the five prime sequences of genes appeared to be correlated with expression divergence, hence supporting evolutionary decoupling of *cis*-regulatory and coding regions of genes (Kohn *et al.* 2008). Ongoing BAC library screening and sequencing for specific candidate genes should allow the study of five prime sequence divergence in relation to gene expression in whitefish (Jeukens *et al.* in prep.). Of course, *trans*-regulation of genes is likely implicated as well, especially considering the previous identification in whitefish of key genomic regions with apparent pleiotropic effects on gene expression (Derome *et al.* 2008; Whiteley *et al.* 2008).

Allelic imbalance

While *cis*-regulatory changes affect transcription in an allele-specific manner, *trans*-regulatory changes modify the expression of factors that interact with *cis*-regulatory sequences of both alleles in a diploid cell (Davidson

2001). Accordingly, AI in a single individual or species provides evidence of *cis*-regulatory differences, as both alleles are assumed to be in the same *trans*-regulatory background. The prevalence of AI in dwarf whitefish could be consistent with harnessing of ancestral alleles for rapid adaptive change (e.g. Colosimo *et al.* 2005). Indeed, in most cases, both alleles were also present in normal whitefish. Moreover, this trend was observed despite the fact that we had 15% more data for the normal pool compared with the dwarf pool. It should also be noted that existing data on both neutral and coding polymorphism do not reveal any significant difference in overall homozygosity levels between normal and dwarf whitefish (Lu & Bernatchez 1999; Campbell & Bernatchez 2004; Renaut *et al.* 2010), thus bias in genotypic composition alone cannot explain the observed prevalence of AI in dwarf whitefish. Protein synthesis and immunity functional annotations were significantly overrepresented among candidate contigs for AI, hence further supporting previous identifications of these functions as potential targets of divergent selection in whitefish species pairs (St-Cyr *et al.* 2008). Indeed, evolutionary changes in the expression of key gene functions could occur at both the gene-specific and the allele-specific levels (von Korff *et al.* 2009). For instance, allele-specific underexpression of protein synthesis genes in dwarf fish could hypothetically be related to their reduced growth rate.

Comparison between RNA-Seq and microarray results

Previous studies comparing independent microarray and RNA-Seq experiments in model organisms have reported correlations varying between 46% and 75% ('t Hoen *et al.* 2008; Marioni *et al.* 2008). Similar results were obtained when the microarray was constructed with consensus sequences from a prior RNA-Seq experiment in a nonmodel organism (62%, Kristiansson *et al.* 2009). Here, RNA-Seq and microarray data were significantly correlated but only modestly (38%). This is not surprising given that a previous quantitative real-time PCR study has shed reasonable doubt on the hybridization specificity of the salmonid cDNA microarray used by St-Cyr *et al.* (2008), especially in the face of multigene family expression (Jeukens *et al.* 2009). Nevertheless, the comparison between RNA-Seq and microarray results with a sequence-specific validation approach confirmed differential expression for 25–50% of ESTs, depending on the stringency of the criterion. Moreover, in accordance with EST annotation, the association between multiple ESTs and a single RNA-Seq contig clearly suggests that they represent a single gene. These results also revealed that all microarray candidates had relatively small absolute differences in expression between dwarf and normal whitefish (14-3500 RPKM). For candidate gene selection among RNA-Seq results, Goetz *et al.* (2010) found that selecting contigs with the highest absolute differences in expression (approximately 3000–20 000 in our data, see Table 3) increased the success rate of real-time PCR validation compared with selecting contigs only according to fold differences in expression levels.

Real-time PCR was previously performed for two candidate genes on the same samples that where used for the experiment presented here and the microarray experiment (Jeukens *et al.* 2009). The first gene, *carboxylesterase*, was significantly overexpressed in dwarf whitefish according to all three methods. However, the second gene, *anionic trypsin*, was significantly overexpressed in dwarf whitefish according to microarray and real-time PCR, but it was absent from the RNA-Seq *de novo* contigs. Upon closer inspection, a few reads actually corresponded to this gene, but where not assembled *de novo* because of a lack of overlapping sequences. This highlights the limitations of *de novo* assembly and the fact that corroborating results from any one of these methods may improve the search for candidate genes based on their level of expression.

To summarize, the main goal of this study was to characterize transcriptomic divergence between incipient species of lake whitefish by means of RNA-Seq. *De novo* contig assembly and gene expression analysis led to the identification of 948 differentially expressed contigs between dwarf and normal whitefish. The former showed more overexpressed genes related to energy metabolism, immunity as well as DNA replication and repair, whereas the latter showed more overexpressed genes associated with growth (protein synthesis). RNA-Seq allowed SNP discovery, which combined with gene expression data permitted to uncover the relationship between expression and sequence evolution in a young species pair. As a result, we observed no correlation between gene expression divergence and any measure of polymorphism, regardless of it being a general measure of the complete data set or an index of divergence between dwarf and normal whitefish. However, 9.2% of tested SNPs showed significant AI, and this phenomenon was significantly more common in the recently diverged dwarf form. Furthermore, protein synthesis and immunity functions were overrepresented among candidate contigs for AI. These results support evolutionary decoupling of regulatory and coding regions of genes, at least for very young species pairs (Kohn *et al.* 2008). They also emphasize the role of gene expression divergence and the potential of allele-specific expression divergence in the process of speciation. This study, together with the companion study of Renaut *et al.* (2010), shows how

next-generation sequencing technologies can be harnessed to reach a much more comprehensive understanding of genomic and transcriptomic divergence in a young species pair.

Acknowledgements

We are grateful to E. Normandeau, C. Sauvage, MM Hansen and two anonymous referees for their constructive comments on earlier versions of the manuscript. We also thank C. Sauvage and J. Laroche for their help with analysis software. This research was financially supported by a Natural Sciences and Engineering research Council of Canada (NSERC) and a Fonds québécois de la recherche sur la nature et les technologies (FQRNT) postgraduate scholarships to JJ; a NSERC postgraduate scholarship to SR; a postdoctoral research stipend from the German Research Foundation to AWN and a NSERC Discovery grant and Canadian Research Chair to LB.

References

Adelman R, Saul RL, Ames BN (1988) Oxidative damage to DNA: relation to species metabolic rate and life span. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 2706–2708.

Allendorf FW, Thorgaard GH (1984) Tetraploidy and the evolution of salmonid fishes. In: *Evolutionary Genetics of Fishes* (ed Turner BJ). pp. 1–53, Plenum Press, New York.

Barrett RDH, Rogers SM, Schluter D (2008) Natural selection on a major armor gene in threespine stickleback. *Science*, **322**, 255–257.

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

Brockman W, Alvarez P, Young S *et al.* (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Research*, **18**, 763–770.

Buckling A, Rainey PB (2002) The role of parasites in sympatric and allopatric host diversification. *Nature*, **420**, 496–499.

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

Campbell KL, Roberts JEE, Watson LN *et al.* (2010) Substitutions in woolly mammoth hemoglobin confer biochemical properties adaptive for cold tolerance. *Nature Genetics*, **42**, 536–540.

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

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

Davidson EH (2001) *Genomic Regulatory Systems: Development and Evolution*. Academic Press, San Diego.

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.

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

Edmands S (2002) Does parental divergence predict reproductive compatibility? *Trends in Ecology & Evolution*, **17**, 520–527.

Eizaguirre C, Lenz TL, Traulsen A, Milinski M (2009) Speciation accelerated and stabilized by pleiotropic major histocompatibility complex immunogenes. *Ecology Letters*, **12**, 5–12.

Eveland AL, McCarty DR, Koch KE (2008) Transcript profiling by 3′-untranslated region sequencing resolves expression of gene families. *Plant Physiology*, **146**, 32–44.

Fay JC, Wittkopp PJ (2008) Evaluating the role of natural selection in the evolution of gene regulation. *Heredity*, **100**, 191–199.

Foksinski M, Rozalski R, Guz J *et al.* (2004) Urinary excretion of dna repair products correlates with metabolic rates as well as with maximum life spans of different mammalian species. *Free Radical Biology and Medicine*, **37**, 1449–1454.

Fu X, Fu N, Guo S *et al.* (2009) Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics*, **10**, 161.

Gilad Y, Pritchard JK, Thornton K (2009) Characterizing natural variation using next-generation sequencing technologies. *Trends in Genetics*, **25**, 463–471.

Goetz F, Rosauer D, Sitar S *et al.* (2010) A genetic basis for the phenotypic differentiation between siscowet and lean lake trout (*Salvelinus namaycush*). *Molecular Ecology*, **19**, 176–196.

Graze RM, McIntyre LM, Main BJ, Wayne ML, Nuzhdin SV (2009) Regulatory divergence in *Drosophila melanogaster* and *D. simulans*,a genomewide analysis of allele-specific expression. *Genetics*, **183**, 547–561.

Guo M, Yang S, Rupe M *et al.* (2008) Genome-wide allele-specific expression analysis using Massively Parallel Signature Sequencing (MPSS (TM)) Reveals cis- and trans-effects on gene expression in maize hybrid meristem tissue. *Plant Molecular Biology*, **66**, 551–563.

Hale M, McCormick C, Jackson J, DeWoody JA (2009) Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (*Acipenser fulvescens*): the relative merits of normalization and rarefaction in gene discovery. *BMC Genomics*, **10**, 203.

Hayes B, Laerdahl JK, Lien S *et al.* (2007) An extensive resource of single nucleotide 614 polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed sequences. *Aquaculture*, **265**, 82–90.

Hoekstra HE, Coyne JA (2007) The locus of evolution: evo devo and the genetics of adaptation. *Evolution*, **61**, 995–1016.

Hoekstra HE, Hirschmann RJ, Bundey RA, Insel PA, Crossland JP (2006) A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science*, **313**, 101–104.

't Hoen PAC, Ariyurek Y, Thygesen HH *et al.* (2008) Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Research*, **36**, e141.

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.

Joron M, Papa R, Beltran M *et al.* (2006) A conserved supergene locus controls colour pattern diversity in Heliconius butterflies. *PLOS Biology*, **4**, 1831–1840.

Khaitovich P, Hellmann I, Enard W *et al.* (2005) Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science*, **309**, 1850–1854.

Kohn MH, Shapiro J, Wu CI (2008) Decoupled differentiation of gene expression and coding sequence among Drosophila populations. *Genes & Genetic Systems*, **83**, 265–273.

von Korff M, Radovic S, Choumane W *et al.* (2009) Asymmetric allele-specific expression in relation to developmental variation and drought stress in barley hybrids. *The Plant Journal*, **59**, 14–26.

Kristiansson E, Asker N, Forlin L, Larsson DJ (2009) Characterization of the Zoarces viviparus liver transcriptome using massively parallel pyrosequencing. *BMC Genomics*, **10**, 345.

Kulesh DA, Clive DR, Zarlenga DS, Greene JJ (1987) Identification of interferon-modulated proliferation-related cDNA sequences. *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 8453–8457.

Lu G, Bernatchez L (1998) Experimental evidence of reduced hybrid viability between dwarf and normal ecotypes of lake whitefish (*Coregonus clupeaformis* Mitchill). *Proceedings: Biological Sciences*, **265**, 1025–1030.

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.

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

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, **18**, 1509–1517.

McCarthy FM, Bridges SM, Wang N *et al.* (2007) AgBase: a unified resource for functional analysis in agriculture. *Nucleic Acids Research*, **35**, D599–D603.

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.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, **5**, 621–628.

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.

Nuzhdin SV, Wayne ML, Harmon KL, McIntyre LM (2004) Common pattern of evolution of gene expression level and protein sequence in Drosophila. *Molecular Biology and Evolution*, **21**, 1308–1317.

Oleksiak MF, Churchill GA, Crawford DL (2002) Variation in gene expression within and among natural populations. *Nature Genetics*, **32**, 261–266.

Oshlack A, Wakefield M (2009) Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, **4**, 14.

Ozsolak F, Platt AR, Jones DR *et al.* (2009) Direct RNA sequencing. *Nature*, **461**, 814–818.

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.

de Queiroz K (2005) Different species problems and their resolution. *Bioessays*, **27**, 1263–1269.

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

Quinn NL, Levenkova N, Chow W *et al.* (2008) Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome. *BMC Genomics*, **9**, 404.

Ranz JM, Machado CA (2006) Uncovering evolutionary patterns of gene expression using microarrays. *Trends in Ecology & Evolution*, **21**, 29–37.

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**, 115–131.

Rise ML, von Schalburg KR, Brown GD *et al.* (2004) Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics. *Genome Research*, **14**, 478–490.

Rise ML, Douglas SE, Sakhrani D *et al.* (2006) Multiple microarray platforms utilized for hepatic gene expression profiling of GH transgenic coho salmon with and without ration restriction. *Journal of Molecular Endocrinology*, **37**, 259–282.

Robinson M, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, **11**, R25.

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

Rogers SM, Isabel N, Bernatchez L (2007) Linkage maps of the dwarf and Normal lake whitefish (*Coregonus clupeaformis*) species and their hybrids reveal the genetic architecture of population divergence. *Genetics*, **175**, 375–398.

von Schalburg K, Rise M, Cooper G *et al.* (2005) Fish and chips: various methodologies demonstrate utility of a 16,006-gene salmonid microarray. *BMC Genomics*, **6**, 126.

Schemske DW, Bierzychudek P (2007) Spatial differentiation for flower color in the desert annual Linanthus parryae: was Wright right? *Evolution*, **61**, 2528–2543.

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Schluter D (2000) *The Ecology of Adaptive Radiation*. Oxford University Press, Oxford.

Serre D, Gurd S, Ge B *et al.* (2008) Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic Cis-acting mechanisms regulating gene expression. *Plos Genetics*, **4**, e1000006.

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.

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

Storz JF, Runck AM, Sabatino SJ *et al.* (2009) Evolutionary and functional insights into the mechanism underlying high-altitude adaptation of deer mouse hemoglobin. *Proceedings of the National Academy of Sciences*, **106**, 14450–14455.

Tirosh I, Barkai N (2008) Evolution of gene sequence and gene expression are not correlated in yeast. *Trends in Genetics*, **24**, 109–113.

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.

Turelli M, Barton NH, Coyne JA (2001) Theory and speciation. *Trends in Ecology & Evolution*, **16**, 330–343.

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

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**, 57–63.

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.

Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary changes in cis and trans gene regulation. *Nature*, **430**, 85–88.

Wray GA, Hahn MW, Abouheif E *et al.* (2003) The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution*, **20**, 1377–1419.

Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW (2002) Allelic variation in human gene expression. *Science*, **297**, 1143.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591.

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

Supporting information

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

Table S1 Functional gene categories

Table S2 *De novo* contigs from RNA-Seq of the whitefish liver transcriptome

Table S3 Representation of functional gene categories in the whitefish liver transcriptome

Table S4 Candidate genes for allelic imbalance (AI)

Fig. S1 Scatter plot of gene expression divergence measured by RNA-Seq and microarray.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.