SÉBASTIEN RENAUT

Génomique de la spéciation chez le grand corégone (*Coregonus clupeaformis*) : divergence adaptative et isolement reproducteur.

Thèse présentée
à la Faculté des études supérieures de l'Université Laval
dans le cadre du programme de doctorat en biologie
pour l'obtention du grade de Philosophiae Doctor (Ph. D.)

DÉPARTEMENT DE BIOLOGIE
FACULTÉ DES SCIENCES ET GÉNIE
UNIVERSITÉ LAVAL
QUÉBEC

2010

Résumé court

La mise en place de barrières à la reproduction et, par conséquent, la spéciation elle-même, engendrent et maintiennent la biodiversité. Les principaux objectifs de mes travaux étaient premièrement d'identifier l'ampleur des différences d'expression de gènes entre des jeunes espèces de grand corégones nains, normaux, ainsi que leurs hybrides. Par la suite, grâce à une technique révolutionnaire de séquençage, nous avons obtenu de l'information de séquence et de polymorphisme pour plusieurs milliers de gènes. Finalement, le génotypage de marqueurs SNPs en populations naturelles et simultanément dans une famille de poissons hybrides a permis d'évaluer l'effet de la sélection naturelle sur la divergence génétique des corégones. Ainsi, cette thèse apporte une meilleure compréhension de la divergence adaptive et de l'isolement reproducteur chez le corégone, tout en identifiant spécifiquement des gènes candidats impliqués dans le processus de spéciation.

Résumé long

Contrairement à la grande quantité d'information écologique appuyant le rôle de la sélection naturelle comme une cause principale de la divergence des populations, la compréhension des mécanismes génomiques sous-jacents reste nébuleuse. Le premier objectif de mes travaux était d'identifier l'ampleur des différences d'expression de gènes entre des jeunes espèces de grand corégones (*Coregonus clupeaformis*) nains, normaux, ainsi que leurs hybrides. À l'aide de biopuces, nous avons identifié chez les poissons juvéniles 14 fois plus de gènes différentiellement exprimés que chez les embryons. Chez les embryons, l'expression moyenne entre les parents et les hybrides différait pour très peu de gènes, ceci en contraste avec les poissons juvéniles. Cependant, chez ces embryons, certains gènes clés du développement semblaient être fortement dérégulés. Nous avons aussi trouvé des évidences d'une dérégulation accrue de l'expression où la non additivité de l'expression chez les juvéniles expliquait une plus grande fraction des patrons de transmission chez les rétrocroisements. Par la suite, en comparant les embryons rétrocroisés survivants à ceux moribonds, nous avons observé près de 2000 gènes dérégulés, ainsi démontrant que la dérégulation des gènes essentiels du développement était associée à l'isolement reproducteur post-zygotique. Lors du chapitre suivant, grâce à une technique révolutionnaire de séquençage, nous avons pu bâtir une large banque de données génomiques et ainsi démontrer, à partir de plus de 6000 marqueurs SNPs putatifs identifiés, qu'une faible portion du génome montrait une divergence de fréquence d'allèles élevée entre nains et normaux. Finalement, en procédant à des analyses de balayage génomique avec une centaine de SNPs, nous avons pu caractériser l'effet de la sélection naturelle sur la variation génétique, tout en associant ces marqueurs à quatre phénotypes adaptatifs. À ce titre, l'utilisation intégrée de données de balayage génomique, d'association génotype-phénotype et de génomique fonctionnelle a permis d'identifier plus précisément deux gènes candidats impliqués dans la divergence écologique récente des corégones nains et normaux. En conclusion, cette thèse apporte de manière globale une meilleure compréhension de l'effet de la divergence adaptive sur l'isolement reproducteur et sur l'architecture génétique des corégones, tout en identifiant de manière spécifique des gènes candidats impliqués dans le processus de spéciation.

Abstract

The evolution of reproductive isolation and therefore speciation itself, generates and maintains biodiversity. Unlike the vast amount of ecological information supporting the role of natural selection as a principal causative agent of population divergence, an understanding of the genomic basis underlying this divergence remains nebulous. The main objectives of my doctoral thesis were first, to identify the magnitude of gene expression differences among young species of dwarf and normal lake whitefish (*Coregonus clupeaformis*), as well as their hybrids. Using microarrays, we showed that 16-weeks old juvenile fish had 14 times more genes displaying significant regulatory divergence than embryos. In embryos, very few transcripts differed in average expression between parents and hybrids, in contrast to juvenile fish. Nevertheless, in embryos some key developmental genes seemed to be highly deregulated. We also found evidence for increased misexpression in juveniles, whereby non-additivity explained a larger fraction of hybrid inheritance patterns in backcross compared to F1-hybrids. Subsequently, by comparing surviving backcross embryos with moribund ones, we found that a large portion of the transcriptome was deregulated. This demonstrated that deregulation of genes essential for development was associated with post-zygotic reproductive isolation. In the next chapter, using revolutionary 454 sequencing, we built a large database of genomic data and quantified, from more than 6,000 putative SNP markers, that a small portion of the genome showed high differences in allele frequency between dwarf and normal whitefish. Finally, using SNP markers, genome scans were performed to characterize the effect of natural selection for genes of known functions, while at the same time associating these SNPs with four adaptive phenotypes. As such, the integrated use of phenotypic, transcriptomic and functional genomic information provided good evidence for the role of two candidate genes in the recent ecological divergence of lake whitefish. In conclusion, this doctoral thesis aimed to provide a global understanding of the intricate effects of adaptive divergence and reproductive isolation on the genetic architecture of whitefish populations, while concurrently identifying specific candidate genes involved in the speciation process.

Avant-Propos

Remerciements

Je tiens à remercier en premier lieu mon directeur de recherche Louis Bernatchez. C'est lui qui m'a accueilli dans son laboratoire, m'a soutenu positivement et a toujours cru en mes capacités, du tout début à la toute fin. Merci Louis pour tes conseils, ton support, ton enthousiasme et nos nombreuses discussions scientifiques. J'ai appris énormément de toi et il ne fait aucun doute que ces années passées au labo influenceront fortement le futur de ma carrière scientifique.

Je tiens à remercier Arne Nolte avec qui j'ai directement collaboré sur plusieurs aspects du projet. Il m'a beaucoup aidé à développer mon esprit critique, approfondir mes connaissances en biologique moléculaire et finalement à développer mes qualités de pêcheur sportif. Je remercie aussi Nicolas Derome qui s'est officiellement joint tardivement à mon projet en tant que co-superviseur, mais avec qui j'ai beaucoup discuté, roulé et skié depuis mes débuts à Québec. Et tout les autres ayant travaillé ou travaillant sur le projet corégone. Plus particulièrement, Julie Jeukens, avec qui depuis mon arrivée au labo, j'ai souvent discuté, ventilé mes frustrations scientifiques, collaboré et fêté. Sean Rogers, qui, en tant qu'ancien étudiant sur le projet corégone, a fait un travail admirable sur lequel plusieurs idées de mon projet reposent; Jérome St-Cyr, Andrew Whiteley ayant aussi fait partie du projet corégone.

J'ai grandement bénéficié de l'apport scientifique et amical de la cinquantaine de personnes que j'ai côtoyées au cours des quatre dernières années dans cette grande entreprise coopérative qu'est le labo Bernatchez. Plus particulièrement, je tiens encore une fois à remercier certaines personnes: Guillaume Côté pour son aide dans le laboratoire, Éric Normandeau pour nos discussions scientifiques et son aide précieuse en bioinformatique, Vincent Bourret, Jesus Mavàrez, Christopher Sauvage, Scott McCairns, Mélanie Dionne et Christian Roberge pour leurs conseils et discussions scientifiques, Geneviève Ouellet-

Cauchon pour la relecture de ma thèse et la correction du français et finalement Lucie Papillon et Vicky Albert pour leur aide technique.

Les gens du LARSA, particulièrement Serge Higgins et Jean-Christophe Therrien, m'ont beaucoup aidé à élever et garder en vie mes poissons, travail pour lequel je n'avais aucune expérience en arrivant à l'Université Laval. Je remercie Jérome Laroche du service de bio-informatique, Gaétan Légaré au séquençage, Julie Turgeon et Nathalie Isabel, membres de mon comité de soutien de doctorat, et Betty Pelgas en cartographie génétique. Je remercie les gens du cGRASP à Victoria et les gens du centre d'innovation de Génome Québec à l'Université McGill, Alexandre Belisle, Gary Levesque, Alexandre Montpetit et Pierre Lepage qui m'ont conseillé afin d'optimiser le mieux possible le séquençage 454 et le génotypage. Je remercie également tous les organismes qui m'ont soutenu financièrement: le CRSNG, l'Université Laval, Québec-Océan, le département de biologie de l'Université Laval, l'IBIS, la fondation Marc Bourgie, la fondation Richard Bernard, l'Université de Washington et l'ACCEBUL.

Finalement, je remercie chaleureusement mes parents, qui m'ont toujours soutenu et poussé à aller jusqu'au bout de mes capacités.

Organisation de la thèse

Cette thèse aborde différents aspects de la biologie, de la spéciation et de la détection de la sélection au niveau moléculaire chez le grand corégone. Puisqu'aucun processus biologique ne devrait être étudié en vase clos, une des idées de mon projet était de réunir différents niveaux d'information provenant de collaborations avec d'autres chercheurs et d'études antérieures déjà publiées. Le but ultime restant toujours d'élucider les bases génétiques des changements évolutifs ayant pour conséquence l'isolement reproducteur des populations. En pratique, cela a favorisé ma participation, à titre de collaborateur, à plusieurs études portant sur la génomique de la spéciation chez le grand corégone lors de mon projet de doctorat.

Cette thèse est organisée en six chapitres. Les chapitres 2, 3 ,4 ,5 ont tous fait l'objet d'un article de recherche publié ou en voie de l'être. Par ailleurs, j'ai aussi collaboré à 4 autres articles. De ces derniers, je n'ai inclus que l'article de Nolte *et al.* (2009, *BMC evol. biol.*) en annexe, puisque le chapitre 2 de ma thèse découle directement de cette étude et que ma collaboration à ce premier article fut directe tant au niveau de la conception, des manipulations et de l'analyse que de l'écriture. Mon directeur Louis Bernatchez est co-auteur sur chacun des articles puisqu'il a participé au niveau du financement, de la conception et de l'écriture, sans compter les innombrables discussions, commentaires et débats scientifiques. Étant arrivé tardivement comme co-directeur officiel de mon projet, Nicolas Derome n'est directement impliqué que dans la dernière partie de mon projet (chapitre 5), bien que j'ai profité de ses judicieux conseils depuis le tout début de mes travaux.

Le chapitre 2 est publié sous la référence: 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. Puisque cette étude fait suite directe à un premier article auquel j'ai collaboré étroitement avec le Dr Nolte, ce chapitre contient aussi un court résumé de cette première partie et, en annexe, l'article qui en a découlé: 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.

Renaut et al. 2009: SR, AWN et LB ont conçu le projet. SR et AWN ont acquis les données. LB a supervisé le projet. SR a analysé les données et écrit le manuscrit.

Nolte et al. 2009: SR, AWN et LB ont conçu le projet. SR et AWN ont acquis les données. LB a supervisé le projet. AWN a analysé les données et écrit le manuscrit.

Le chapitre 3 est accepté en date du 18 octobre pour la revue *Heredity* sous la référence: Renaut S, Bernatchez L (2010) Transcriptome-wide signature of hybrid breakdown associated with intrinsic reproductive isolation in lake whitefish species pairs (*Coregonus* spp. Salmonidae).

SR et LB ont conçu le projet. SR a acquis les données. LB a supervisé le projet. SR a analysé les données et écrit le manuscrit.

Le chapitre 4 est publié sous la référence: Renaut S, Nolte AW, Bernatchez L (2010) Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. Salmonidae). *Molecular Ecology* **19**(Suppl. 1), 115-131.

SR, AWN et LB ont conçu le projet. SR et AWN ont acquis les données. LB a supervisé le projet. SR a analysé les données et écrit le manuscrit.

Le chapitre 5 est accepté en date du 18 octobre pour la revue *Molecular Ecology* sous la référence: Renaut S, Nolte AW, Rogers SM, Derome N, Bernatchez L. Gradients of ecological speciation, SNP signature of selection on standing genetic variation, and association with adaptive phenotypes in lake whitefish species pairs (*Coregonus* spp.).

SR, AWN et LB ont conçu le projet. SMR a fait les mesures phénotypiques. ND et LB ont supervisé le projet. SR a analysé les données et écrit le manuscrit.

Voici finalement la liste des autres projets portant sur la génomique de la spéciation, publiés ou en voie de l'être, en ordre chronologique:

Whiteley AR, Derome N, Rogers SM, St-Cyr J, Nolte AW, Renaut S, Jeukens J, Laroche J, Labbe A, Bernatchez L (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.

Contribution au niveau de l'acquisition de données de séquençage afin de démontrer qu'il existait peu de divergence nucléotidique entre les corégones nains et normaux.

Bernatchez L, Renaut S, Whiteley AW, AW, Derome N, Jeukens J, Landry L, Lu G, Nolte AW, Østbye K, Rogers SM, St-Cyr J (2010) On the origin of species: insights from the ecological genomics of lake whitefish. *Philosophical Transactions of the Royal Society Biological Sciences*. **365**, 1783-1800.

Contribution au niveau de l'analyse et de la revue critique de la rédaction du manuscrit. En outre, j'ai préparé pour cet article une liste exhaustive, disponible comme matériel supplémentaire, d'environ 500 gènes candidats identifiés au cours des dernières années à travers les différentes études d'expression de gène, de QTL et d'eQTL.

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

Contribution au niveau de l'acquisition des données et de la revue critique de la rédaction et des analyses.

"Nothing is as characteristic of biological processes as interactions at all levels, among genes of the genotype, between genes and tissues, between cells and other components of the organism, between the organism and its inanimate environment, and between different organisms. It is precisely this interaction of parts that gives nature as a whole ... its most pronounced characteristics" (Ernst Mayr at the wise age of 100, 2004a)

"Rien n'est aussi caractéristique des processus biologiques que leurs interactions à tous les niveaux, entre les gènes et le génotype, les gènes et leurs tissu, les cellules et les autres composantes de l'organisme, les organismes et leur environnement inanimé et entre les différents organismes. C'est précisément cette interaction des éléments qui donne à la nature ... ses caractéristiques les plus distinctives."

(Ersnt Mayr à l'âge vénérable de 100 ans, 2004a)

Table des Matières

Liste des tableaux

Liste des figures

Lexique des abréviations utilisées

ADN, ADNc	Acide désoxyribonucléique, ADN complémentaire.
AFLP	*Amplified fragment length polymorphism* (Polymorphisme de longueur de fragments amplifiés).
ARN, ARNm	Acide ribonucléique, ARN messager.
ANOVA	Analysis of variance (Analyse de variance).
BC	Backcross (rétrocroisement).
BLAST	Basic local alignment search tool (Algorithme d'assemblage de séquences).
bp, kb, Mb	Basepair, kilobase, megabase (paire de base, kilobase, megabase).
cis	Ce dit d'un mutation qui affecte directement le gène prêt duquel celle-ci se trouve (voir *trans*).
d_N/d_S	Taux de mutations non-synonymes / taux de mutations synonymes.
eQTL	Expression quantitatif trait loci (Loci de trait quantitatif d'expression, voir *QTL*).
EST	Expressed sequence tag (Étiquette d'une séquence exprimée).
FDR	False discovery rate (Taux de faux positifs pour les hypothèses jugées significatives).
F_{ST}	Fixation Index (indice de divergence génétique).
indel	Insertion-deletion (insertion-délétion).
ORF	Open reading frame (Cadre de lecture, région traduite en protéine).
PCR	Polymerase chain reaction (réaction de polymérase en chaine).
emPCR	Réaction de polymérase en chaine en émulsion.
p_N/p_S	Taux de polymorphismes non-synonymes / taux de polymorphismes synonymes.
***P*-value**	Valeur de probabilité associé à un test statistique.
PSV	Paralogous sequence variant (Variation dû à une séquence paralogue).
QTL	Quantitatif trait loci (Loci de trait quantitatif).
***Q*-value**	Valeur de probabilité ajustée pour tests multiples selon la méthode de Storey (2002) *ou encore* le taux de faux positifs pour les hypothèses jugées significatives d'après une valeur de probabilité *P*.
SNP	Single nucleotide polymorphism (polymorphisme d'un seul nucléotide).
trans	Ce dit d'une mutation qui affecte un gène qui se trouve loin de celle-ci.
UTR	Untranslated region (Région non traduite en protéine).

Chapitre 1 : Introduction générale

1.1 Évolution, biodiversité et spéciation

La nature de toute entité biologique est déterminée par son bagage génétique, son environnement et l'interaction de ces deux paramètres. Cette double causalité régit la diversité biologique et tout organisme évolue constamment par l'effet des forces de la sélection et de la dérive. De plus, la mise en place de barrières à la reproduction et, par conséquent, la spéciation elle-même engendrent et maintiennent cette biodiversité. Selon la théorie écologique de la radiation adaptative, cette divergence des populations est causée par la sélection naturelle divergente et a pour conséquence directe l'isolement reproducteur de celles-ci (Mayr 1942, Schluter 2000, Rundle & Nosil 2005). Ainsi, les traits sous sélection naturelle divergente ou des traits génétiquement corrélés à ceux-ci affecteront aussi la capacité de reproduction entre individus. À ce jour, élucider les mécanismes responsables de la divergence adaptative et de l'isolement reproducteur reste encore l'un des objectifs fondamentaux de la biologie évolutive (Schluter 2000, Coyne & Orr 2004, Schluter 2009, Presgraves 2010). Ces concepts de base feront donc partie intégrante de mon travail de doctorat sur la génomique de la spéciation chez le grand corégone.

Dans le cadre de ce projet, le concept biologique d'espèce développé par Dobzhansky (1937) et Mayr (1942) et stipulant que « les espèces sont des groupes d'individus pour lesquels il existe des opportunités d'échanges génétiques en conditions naturelles et étant isolés reproductivement d'autres groupes semblables » (Mayr 1942) est probablement le plus approprié. Par ailleurs, ce concept doit bien souvent ne pas être utilisé de manière purement restrictive, puisque même pour des espèces reconnues et fortement divergentes, l'isolement reproducteur peut s'avérer incomplet (Mallet 2006). Pour autant que les groupes d'individus parviennent à maintenir une cohésion génétique distincte face à un flux de gène potentiel, ils peuvent être considérés comme des espèces à part entière (Rieseberg & Willis 2007). Avant tout, pour éviter de tomber dans un argument sémantique, cela revient donc à dire que l'étude du processus de spéciation implique directement l'étude de l'isolement reproducteur entre des groupes divergents et des causes de la mise en place de barrières à la reproduction. Plus particulièrement, l'étude de la spéciation chez des jeunes espèces permet d'identifier les premières barrières au flux génique à survenir lors de la divergence des populations. En effet, les changements génétiques identifiés chez une espèce en cours de divergence adaptative sont plus probablement à la base de la divergence elle-même qu'ils ne le seraient si le processus de spéciation avait déjà eu lieu (Schluter 2000, Mallet 2006, Via 2009, Presgraves 2010).

1.2 Génétique de la spéciation

Un des grands défis actuels en biologie évolutive consiste à faire le lien entre un trait adaptatif et les changements génétiques à la base de ce trait (Feder & Mitchell-Olds 2003). Contrairement à la grande quantité d'informations écologiques appuyant le rôle de la sélection naturelle comme une cause principale de la divergence des populations, notre compréhension des bases génomiques sous-jacentes au processus de spéciation reste encore nébuleuse (Coyne & Orr 2004, Schluter 2009, Presgraves 2010). De plus, le nombre de gènes ou régions génétiques influençant les différences entre espèces varie fortement d'une étude à l'autre, si bien qu'il existe relativement peu d'information quant à la nature des gènes responsables des différences entre les espèces et l'ampleur du rôle de la sélection dans le maintien de ces différences génétiques (Orr 2001, Coyne & Orr 2004).

Lors de la spéciation en présence de flux de gènes potentiel (*speciation-with-gene-flow,* Nosil 2008), la sélection divergente contribue à créer une différentiation génomique hétérogène puisque certains loci spécifiques seront la cible directe de la sélection et donc montreront un flux de gène réduit, tandis que la majorité du génome évoluera sous l'effet de forces neutres. Ainsi, ce sont ces régions sous sélection que l'on appelle *îlots de divergence* qui se trouvent dans un océan génomique évoluant sous l'effet des mutations et de la dérive génétique (Turner *et al.* 2005). Par ailleurs, ces îlots expliqueraient aussi les différences adaptatives entre les populations divergentes. Ainsi, à mesure que le processus de spéciation progresse, ces régions tout au début rares, vont s'agrandir en taille et en nombre, jusqu'à ce que, ultimement, les génomes des deux espèces deviennent complètement divergents et l'isolement reproducteur soit complet (Wu 2001, Wu & Ting 2004).

La recherche de ces îlots de divergence a captivé la communauté scientifique et s'est récemment matérialisée par le concept de divergence par auto-stop génétique avancé par Via (Via & West 2008, Via 2009) et par la suite modélisé et testé par simulation (Feder & Nosil 2010). Selon cette hypothèse, des régions de différentiation génétique relativement larges autour du ou des gènes soumis à la sélection divergente se forment, puisque le taux de recombinaison inter-population ainsi que de migration effective sont réduits pour ce(s) gène(s), ce qui amène comme conséquence de perpétuer et d'agrandir ces régions. L'hypothèse alternative stipule que la spéciation est initiée par la sélection divergente agissant plus localement et faiblement mais de manière simultanée sur de nombreux gènes dispersés à travers le génome (Feder & Nosil 2010, Michel *et al.* 2010, Kelleher & Barbash 2010). Ainsi, ces scénarios, non mutuellement exclusifs, représentent les deux situations extrêmes d'un continuum et stipule que l'adaptation d'une population dans un nouvel environnement peut être due à plusieurs gènes ayant chacun un effet mineur, comme elle peut être due à un ou quelques gènes d'effet majeur (Orr 2005, Michel *et al.* 2010).

Un progrès important à été accompli au cours des dernières années et plusieurs "gènes de la spéciation", des gènes impliqués dans l'isolement reproducteur, ont été découverts. Notamment, de nombreux exemples existent chez les mouches du genre *Drosophila* (Presgraves 2010), les levures du genre *Saccharomyces* (Lee *et al.* 2008) ou même chez les vertébrés (*Mus musculus*, Mihola *et al.* 2008). Il demeure cependant difficile de démontrer que ces gènes de la spéciation sont la cause directe de l'isolement reproducteur et ne sont pas des changements qui sont survenus une fois le processus de spéciation complété (Mallet 2006). D'autres études phares en biologie évolutive ont identifié, chez des jeunes espèces, l'implication d'un gène ou région génomique précise dans la variation à un phénotype donné, tout en démontrant le caractère adaptif et par corolaire l'implication dans l'isolement reproducteur de ce trait phénotypique (notamment chez l'épinoche à trois épines Colosimo *et al.* 2005 Miller *et al.* 2007, la souris Hoekstra *et al.* 2006, ou les papillons du genre *Heliconius* Joron *et al.* 2006).

De manière générale, ces cas d'études ne représentent que quelques cas de lien de causalité entre la variation génétique et l'isolement reproducteur ou encore la divergence adaptative. Il existe donc toujours un besoin flagrant d'études empiriques avant de pouvoir généraliser le nombre de gènes et le type de variation génétique à la base de la spéciation. De plus, le paradigme du "gène de la spéciation" devra, selon toute évidence, évoluer afin d'expliquer les causes et conséquences complexes, et se répercutant probablement sur tout le génome, des patrons de divergence observés lors du processus de spéciation (Presgraves 2010, Counterman *et al.* 2010, Michel *et al.* 2010, Kelleher & Barbash 2010).

1.3 Génétique de la spéciation et évolution parallèle

L'évolution parallèle se définit comme l'évolution indépendante d'un même caractère phénotypique dans des populations apparentées (Futuyma 1986). Par ailleurs, lorsqu'elle est associée à un changement similaire d'environnement, l'évolution parallèle est généralement considérée comme une manifestation de l'effet de la sélection naturelle (Schluter *et al.* 2004). Ainsi, l'évolution indépendante de phénotypes similaires a été démontrée dans plusieurs systèmes (Pigeon *et al.* 1997, Nosil *et al.* 2002, Schluter *et al.* 2004). Par exemple, la colonisation indépendante des habitats lacustres par les épinoches à trois épines et l'évolution similaire du nombre de plaques latérales ainsi que de la taille des individus est un des exemples classiques d'évolution phénotypique parallèle (Schluter *et al.* 2004). Ce concept peut donc, en théorie, s'appliquer à différents niveaux de complexité biologique. Ainsi, on peut, par extrapolation, prédire que l'évolution phénotypique parallèle se répercutera autant au niveau transcriptomique (expression de gènes) que génétique. Il existe, dans la littérature, plusieurs cas où cette hypothèse se vérifie : soit au niveau des gènes identiques différentiellement exprimés dans des populations indépendantes (Cooper *et al.* 2003, Roberge *et al.* 2006, Derome *et al.* 2006), ou alors au niveau des gènes eux-mêmes (Yoon & Baum 2004, Colosimo *et al.* 2005, Hoekstra *et al.* 2006, Miller *et al.* 2007) ou encore plus précisément des mêmes mutations ciblées par la sélection (Hoekstra *et al.* 2006). Si de tels exemples existent et prouvent l'implication de la sélection naturelle divergente, il n'en reste pas moins que, dans bien des cas, puisque la sélection n'agit directement que sur le phénotype lui-même, différents mécanismes moléculaires sont recrutés par la sélection pour mener à l'évolution parallèle d'un même phénotype adaptatif (Yoon *et al.* 2004, Hoekstra *et al.* 2006, Arendt & Reznick 2008, Renaut *et al. accepté*).

1.4 Génétique de la spéciation et hybridation

Lorsque deux espèces se rencontrent et s'hybrident, des allèles n'ayant jamais été présents ensembles peuvent interagir, produire de nouveaux phénotypes et engendrer une baisse de la valeur adaptative de cette progéniture hybride soit de manière intrinsèque (ex. mortalité accrue) ou extrinsèque (i.e dépendante de l'environnement). Selon le modèle avancé par Bateson et développé par Dobzhansky et Muller, la manifestation de cet isolement reproducteur post-zygotique est vraisemblablement expliqué par des interactions non-additives (épistatiques) entre allèles à des loci divergents (Bateson 1909, Dobzhansky 1937, Muller 1940). Ainsi, la recombinaison de deux génomes différents brise les complexes de gènes co-adaptés et rompt l'équilibre des forces épistatiques. De surcroît, ce phénomène devrait se manifester tout particulièrement chez les hybrides de seconde génération où la recombinaison homologue (i.e l'enjambement ou *crossover* des chromatides) des deux génomes devient plus propice à engendrer des incompatibilités génétiques (Coyne & Orr 2004, Burton *et al.* 2006). De nombreuses études ont démontré que l'hybridation de deux espèces différentes peut produire des phénotypes extrêmes par rapport à ceux des deux parents; un phénomène que l'on désigne par le terme de ségrégation transgressive (Rieseberg *et al.* 1999). Dans bien des cas, ceci causera une baisse de la valeur adaptative des individus, bien que plusieurs exemples existent où l'hybridation engendre une espèce hybride, adaptée à un nouvel environnement (Mallet 2007, Mavárez & Linares 2008).

1.5 Expression de gènes

Plusieurs niveaux de complexité biologique existent, des gènes encodés dans l'ADN aux protéines qu'ils produisent, en passant par la transcription de l'ARN messager; et chacun de ces niveaux aura une influence sur le phénotype. Par exemple, une mutation alternant la structure ou fonction d'une protéine peut alimenter la sélection naturelle en diversité génétique et ainsi être responsable d'un phénotype adaptatif (Hoekstra & Coyne 2007). D'un autre côté, il a été démontré que des changements au niveau de la régulation d'un gène semblent souvent être la cause directe d'un phénotype adaptatif et représentent donc une source majeure de nouveauté évolutive sur laquelle la sélection peut agir (Whitehead & Crawford 2006, Wray 2007). Ainsi ce fut King et Wilson (1975) qui, en se fondant sur le taux de similarité étonnamment élevé entre les gènes humains et ceux de chimpanzés, furent les premiers à émettre l'hypothèse que des mutations dans des régions régulatrices expliquaient plus vraisemblablement les différences entres espèces évolutivement proches que des mutations structurelles. La base de cette régulation différentielle peut être due à des changements de séquence directement dans une région régulatrice d'un gène (changement en *cis*) ou encore à des changements distaux, par exemple dans un facteur de transcription ou encore un petit ARN non codant. Ces facteurs distaux, ou *trans*, interagissent avec les régions régulatrices et jouent donc un rôle majeur dans l'expression d'un phénotype donné (Ranz & Machado 2007, Chen & Rajewsky 2007, Landry *et al.* 2007a).

1.6 Expression de gènes et hybridation

Les études chez les hybrides offrent une opportunité unique de mettre à jour des incompatibilités régulatrices, habituellement masquées chez les deux espèces séparément par des changements compensatoires (Landry *et al.* 2007a). Plusieurs études ont démontré théoriquement et empiriquement qu'un dérèglement au niveau de la transcription de l'ARN messager pouvait expliquer l'incompatibilité des génomes chez les hybrides (Noor & Feder 2006). La plupart de ces études ont observé, pour une majorité de gènes, des patrons de transmission du niveau de transcription qui suggèrent une base génétique non-additive (Ranz *et al.* 2004, Rockman & Kruglyak 2006, Roberge *et al.* 2008). Par exemple, Ranz et collaborateurs (2004) ont montré que les profils globaux d'expression de *Drosophila melanogaster* et *D. simulans* étaient plus proche l'un de l'autre qu'ils ne l'étaient de leur descendance hybride. D'autres auteurs ont, quant à eux, conclu qu'une majorité des profils de transcriptions observés étaient sous contrôle génétique additif (Swanson-Wagner *et al.*

2006, Cui *et al.* 2006, Rottscheidt & Harr 2007). Par ailleurs, lors du début de mon projet de doctorat, la dynamique évolutive causant une dysfonction de la régulation n'avait été étudiée que chez les espèces modèles en laboratoire, ces études ne s'attardant en général qu'aux hybrides de première génération, chez des espèces ayant souvent divergées depuis plusieurs millions d'années (principalement du genre *Drosophila*, Landry *et al.* 2007a, Ortiz-Barrientos *et al.* 2007). Encore une fois, à partir de telles conclusions, il est difficile d'inférer si les gènes identifiés sont à la base de la divergence elle-même ou bien proviennent de régions génétiques devenant progressivement résistantes à l'introgression (par dérive génétique ou sélection naturelle) une fois le processus de spéciation ayant eu lieu.

1.7 Survol technique et analytique des biopuces

Fort de ces concepts théoriques et grâce, entre autres, à la baisse substantielle du coût d'utilisation de plusieurs technologies de pointe en biologie moléculaire, les biopuces sont devenues un outil de choix en écologie et évolution afin de quantifier le rôle de la divergence d'expression (Kammenga *et al.* 2007). Ainsi, depuis une dizaine d'années, les biopuces ont permis de mettre en évidence la forte variation interindividuelle que présentaient les profils de transcription à l'intérieur des populations naturelles (Townsend *et al.* 2003, Oleksiak *et al.* 2002, Whitfield *et al.* 2003), que ces profils étaient en partie héritables (Brem & Kruglyak 2005, Rockman & Kruglyak 2006) et qu'ils pouvaient répondre rapidement à la sélection (Ferea *et al.* 1999). Toutes ces évidences empiriques corroborent l'hypothèse que la variation de l'expression de gènes, modulée par la sélection, participerait de façon importante à l'évolution de la biodiversité (Townsend *et al.* 2003).

Les biopuces, aussi appelées *puces à ADN*, *microarrays* ou *DNA chips* sont utilisées pour mesurer la quantité d'ARN messager transcrite pour plusieurs milliers de gènes simultanément. La technique repose sur le principe d'hybridation d'un ADN rétro-transcrit et marqué à l'aide d'une sonde fluorescente sur un support (le plus souvent une lame de verre) contenant des ADN complémentaires et spécifiques pour chaque gène. En hybridant deux échantillons marqués avec deux sondes différentes sur une lame, il est possible d'obtenir un signal relatif d'expression pour ces individus, et ce pour un grand nombre de gènes (Brown 2007).

Les biopuces sont très sujettes aux variations de conditions expérimentales, à la fois durant leur impression (dépôt des ADNc amplifiés par PCR sur la lame) et durant les hybridations, d'où la nécessité d'en normaliser les résultats avant de les analyser (Leek & Storey 2007). Les puces à oligonucléotides où les sections d'ADNc sont synthétisées directement sur la lame sont moins sensibles à ces variations, mais plus onéreuses et requièrent plus d'information génomique *a priori*. De manière générale, les données de biopuces peuvent être traitées comme tout autre trait quantitatif si ce n'est de certaines particularités. Ainsi, la forte variation potentielle non modélisée et le faible de taux de réplication représentent bien sûr des limitations inhérentes que certaines procédures statistiques permettent néanmoins d'atténuer (Kerr *et al.* 2000, Leek & Storey 2007). De plus, puisque plusieurs milliers de tests d'associations statistiques sont effectués simultanément, différentes approches statistiques ont été suggérées afin de corriger le taux de faux positifs dus aux tests multiples (Holm 1979, Benjamini & Hochberg 1995, Storey 2004).

Les biopuces utilisées lors de mon de projet de doctorat ont été développées par le cGRASP (consortium for Genomic Research on All Salmon Project, Génome Canada) et contiennent 16 006 ADNc de salmonidés (principalement de saumon atlantique, *Salmo salar*). Les séquences imprimées sur les lames de verre sont en fait des courts fragments d'ADNc (ESTs d'environ 600 paires de bases) provenant de plus de 175 banques d'ADN de saumons de différents tissus et différents stades développementaux (Rise *et al.* 2004). Par ailleurs, des études antérieures ont démontré la validité d'utiliser cette biopuce chez d'autres espèces de salmonidés tel le grand corégone (Derome *et al.* 2006, Whiteley *et al.* 2008), le cisco (Derome & Bernatchez 2006) ou encore l'omble de fontaine (Sauvage *et al.* 2010).

1.8 Le séquençage au 21ième siècle

En 2001, après un travail acharné de près de 10 ans ayant coûté plusieurs milliards de dollars et ayant réuni des centaines de scientifiques de part le monde, le séquençage et l'assemblage du génome humain était annoncé en grande presse (International Human Genome Sequencing Consortium 2001, Venter *et al.* 2001). Aujourd'hui, séquencer ce génome peut se faire en moins d'une semaine et ce pour moins de 5000 dollars (Drmanac *et al.* 2010). Ainsi, ce que l'on appelle dans le langage courant les *nouvelles générations de séquençage* remplacent déjà, pour plusieurs applications, la technique classique de séquençage dite *Sanger* (Sanger *et al.* 1977). Elles révolutionnent ainsi le monde de la génétique en général (Shendure & Ji 2008). Plusieurs compagnies ont développé des approches méthodologiques afin d'offrir une quantité toujours grandissante de données pour un coût toujours moindre (Illumina Genome Analyzer, Bennett 2004; Applied Biosynthesis SOLiD™, Shendure *et al.* 2005; Roche 454 Sequencing™, Margulies *et al.* 2005; Helicos, Milos *et al.* 2008 et Pacific Biosciences SMRT™, Eid *et al.* 2009). Bien que différentes les unes des autres, toutes ces approches réunissent des avancées technologiques provenant de différents domaines scientifiques (photonique, informatique, biochimie) et possèdent certaines similarités au niveau du parallélisme des réactions de séquençage, de la miniaturisation des réactions enzymatiques et de l'usage accru de la bioinformatique.

Voici, brièvement, la description de l'approche méthodologique dite *séquençage 454* que j'ai utilisée dans mon projet de doctorat (Roche 454 Sequencing™, Margulies *et al.* 2005). Au départ, une source d'ADN à séquencer est fournie. Si l'ADN est de grande taille, celui-ci sera fragmenté mécaniquement en plus petits morceaux (500bp-1kb). Par la suite, des adaptateurs seront rajoutés aux bouts des fragments afin que chaque fragment puisse se lier sur une microbille de polymère. Lors de l'étape suivante, une PCR directement sur chaque bille dans chaque goutte d'eau (*emPCR*) multipliera le nombre de copies présentes sur chaque bille. À l'aide d'un courant électrique, les billes tomberont sur une plaque de plastique contenant plusieurs millions de micro-puits. Survient ensuite l'étape de pyroséquençage en tant que telle. L'un après l'autre, les nucléotides (A,C,T ou G) seront rajoutés. S'il y a lieu, l'ADN polymérase insèrera ce nucléotide à la chaine d'acides nucléiques et libèrera un groupement pyrophosphate. Celui-ci sera utilisé par une autre enzyme qui émettra de la fluorescence. Ainsi, plus il y aura de nucléotides rajoutés par l'ADN polymérase, plus l'intensité lumineuse sera élevée. Une caméra à haute définition (*CCD*) mesurera la quantité de lumière émise dans chaque puits. Par la suite, on nettoiera les puits en retirant tous les nucléotides non-incorporés. Ce cycle sera répété plusieurs centaines de fois, chaque fois avec un nucléotide différent, pour obtenir une chaîne d'acides nucléiques pouvant aller jusqu'à 1000 nucléotides de longueur totale (Margulies *et al.* 2005, Mardis 2008, Roche 2010)

1.9 Les nouvelles générations de séquençage et le polymorphisme nucléotidique

Les nouvelles techniques de séquençage sont en train de rapidement transformer les domaines de l'écologie, de l'évolution et de la génétique (Mardis 2008, Rokas & Abbot 2009, Tautz *et al.* 2010). Le séquençage 454 se démarque encore du lot en écologie et évolution, bien que cette disparité soit probablement vouée à disparaître. A ce jour, il apporte une taille de fragment nettement plus longue que toute autre technique. Ceci s'avère important afin d'améliorer l'assemblage des données, surtout pour des espèces non modèles, souvent utilisées en écologie et évolution, pour lesquelles on ne possède souvent pas de génome de référence. Ainsi, cette avalanche de données permet non seulement de répondre à des questions avec plus de précision, mais aussi d'aborder de nouvelles problématiques impensables il y a seulement cinq ans (Tautz *et al.* 2010).

Un des objectifs principaux de ces projets de séquençage réside dans la découverte de la variation génétique qu'il peut exister au sein d'une population, telle la variation du nombre de copies d'un gène (*CNVs*), les insertions/délétions (*indels*) ou le polymorphisme d'un nucléotide simple (*SNPs*). Plus particulièrement, les SNPs sont devenus rapidement des marqueurs génétiques d'intérêt en écologie et évolution (Schlötterer 2004, Morin *et al.* 2004). Leur attraction principale réside dans le fait que, contrairement aux marqueurs de type AFLP ou microsatellite, ils peuvent être directement liés à des gènes d'intérêt. De plus, l'automatisation relativement simple de leur génotypage a pour conséquence une baisse substantielle des coûts par marqueur (Ehrich *et al.* 2005, Shen *et al.* 2005, van Tassel *et al.* 2008). Néanmoins, bien qu'ils soient en théorie très abondants et facilement génotypables, le développement de marqueurs SNPs peut requérir plusieurs étapes de validation et rester prohibitif même pour des projets de taille modeste.

Grâce à l'identification du polymorphisme nucléotidique, il devient possible d'évaluer les effets de la sélection sur la variation génétique fonctionnelle et l'effet de celle-ci sur la variance phénotypique et transcriptomique. L'identification des polymorphismes de séquences dans les régions transcrites du génome est d'un intérêt primordial pour tenter de caractériser les effets de la sélection, puisque les impacts de cette variation dépendront de sa localisation génomique exacte (intron, exon, région non traduite-*UTR*). Les mutations au sein des régions codantes, qui affecteront la composition en acides aminés d'une protéine et donc sa fonctionnalité, peuvent être facilement évaluées. Ainsi, le taux d'accumulation de mutations non synonymes (d_N) par rapport au taux de mutations synonymes (d_S) donne un aperçu de l'effet de la sélection conduisant à l'évolution d'une séquence codante, puisqu'un gène avec un ratio d_N / d_S élevé (e.g. > 1) est susceptible d'évoluer sous l'influence de la sélection positive (McDonald & Kreitman 1991, Axelsson *et al.* 2008, Ellegren 2008).

Une autre utilité des nouvelles générations de séquençage réside dans le fait qu'elles permettent de quantifier l'expression de gènes en partant du principe que plus le nombre d'ARN messagers séquencés pour un gène donné est élevé, plus ce gène est fortement exprimé. Puisque le coût relativement faible par séquence est une caractéristique de base des nouvelles générations de séquençage, il est donc possible de séquencer les ARNm et ainsi quantifier directement les niveaux de transcription (Wang *et al.* 2009). En isolant et séquençant le transcriptome complet d'une espèce, on peut obtenir de l'information similaire et potentiellement plus précise qu'avec les biopuces (Fu *et al.* 2009). De surcroît, tout au contraire des biopuces, cette approche ne requiert aucune information *a priori* et permet aussi de quantifier l'expression de manière allèle spécifique (Gilad *et al.* 2009, Fontanillas *et al.* 2010, Jeukens *et al.* 2010).

1.10 Association génotype-phénotype

Tel que mentionné en section 1.2, un des grands défis actuels en biologie consiste à faire le lien entre un trait phénotypique et la variations génétiques à la base de ce trait (Feder & Mitchell-Olds 2003). Une des approches utilisées afin d'identifier des gènes à la base de traits adaptatifs importants repose sur des méthodes de génétique quantitative et de détection de la variation génétique sous-jacente à un trait phénotypique (MacKay 2001). À l'aide de ces approches, des associations statistiques sont identifiées entre des marqueurs génétiques variables et des phénotypes prédéterminés. Par ailleurs, si l'on connait la distance entre ces marqueurs, basée sur le taux de recombinaison entre ceux-ci, on peut aller jusqu'à identifier des régions génétiques expliquant la variance pour un trait phénotypique donné (QTLs, MacKay 2001) ou même transcriptomique (eQTLs, Gibson & Weir 2005). Bien entendu, ceci requiert au bas mot, quelques centaines, voir plusieurs millions en génomique humaine, de marqueurs génétiques informatifs. La technique reste cependant limitée par le fait que de tels segments chromosomiques identifiés comme expliquant une fraction de la variabilité pour un phénotype donné peuvent néanmoins contenir un grand nombre de gènes. De plus, elle informe peu sur les mécanismes de régulation en tant que tel. Néanmoins, de telles approches permettent de disséquer la base polygénique de traits quantitatifs complexes. Par exemple, une étude exhaustive par Miller et collaborateurs (2007) a permis d'identifier chez l'épinoche à trois épines un QTL d'effet majeur pour la pigmentation, un trait phénotypique jouant fort probablement un rôle dans l'adaptation à l'eau douce. Par la suite, ces chercheurs ont démontré qu'un gène connu comme impliqué dans la coloration *(Kitlg)* se trouvait effectivement dans l'intervalle défini par ce QTL, tout en montrant l'évolution parallèle d'un allèle régulant l'expression différentielle de *Kitlg*. Ainsi, l'intégration de données phénotypiques, de l'effet de la sélection, d'expression, ainsi que d'association génotype-phénotype permet donc d'accéder à la "boîte noire" qui se trouve encore trop souvent entre le génotype et le phénotype.

1.11 Le grand corégone, *Coregonus clupeaformis*

De nombreux attributs biologiques font du genre *Coregonus* un modèle d'étude fort pertinent dans le but d'élucider les bases génomiques de la divergence adaptive et de l'isolement reproducteur dans un contexte écologique. Premièrement, il est le genre le plus riche en espèces au sein de la famille des salmonidés, avec une estimation minimale de 28 taxons reconnus et répartis à travers l'hémisphère Nord (Reshetnikov 1988), bien que le nombre réel d'espèces biologiques pourrait largement dépasser ce chiffre (Kottelat & Freyhof 2007). Depuis la fin du Pléistocène, les membres de ce genre ont évolué afin d'occuper un vaste éventail d'habitats, ce qui a eu comme conséquence de créer de fortes variations phénotypiques entre les espèces taxonomiquement reconnues ainsi qu'au sein même des espèces. D'intérêt particulier est la présence en Amérique du Nord de deux formes lacustres de grand corégones vivant en sympatrie à l'intérieur des mêmes lacs (taxon *C. clupeaformis*).

Le cadre naturel des populations de grand corégones étudiées lors de ma thèse doctorale se situe dans le bassin du fleuve Saint-Jean (nord-est des États-Unis et sud du Québec, voir figure 5.1), où des formes sympatriques ont été reportées dans six lacs (Lu & Bernatchez 1999). L'isolement géographique à la fin du Pléistocène a causé une certaine divergence génétique entre les populations qui habitaient des refuges glaciaires à l'est et l'ouest de cette zone, mais sans qu'aucune divergence phénotypique distinctive ne se soit formée entre les races glaciaires en allopatrie (Bernatchez & Dodson 1990, 1991). Le contact secondaire de ces lignées évolutives est survenu il y a environ 12 000 ans au sein de lacs nouvellement créés par la fonte des glaciers et ce relèvement isostatique du continent (Castric *et al.* 2001). L'opportunité écologique ainsi que les interactions compétitives ont contribué à l'évolution parallèle d'une forme naine limnétique, qui a divergé en sympatrie de la forme normale benthique (Bernatchez 2004, Landry *et al.* 2007b). Par ailleurs, le niveau d'isolement et de différenciation phénotypique de ces deux formes dépendent fortement des conditions écologiques et environnementales spécifiques de chaque lac (Landry *et al.* 2007b, Landry & Bernatchez 2010). Ainsi, l'accumulation de différences génétiques au cours de l'isolement géographique en allopatrie, le contact secondaire subséquent et la divergence écologique en sympatrie conduisent à la spéciation écologique contemporaine des corégones nains et normaux (Lu & Bernatchez 1998, Rogers & Bernatchez 2006, Bernatchez *et al.* 2010).

Plusieurs études menées par le passé ont démontré des différences adaptatives entre les corégones nains et normaux en rapport aux traits morphologiques (Lu & Bernatchez 1999, Rogers *et al.* 2002, Bernatchez 2004), d'histoire de vie (Bernatchez *et al.* 1999), comportementaux (Rogers *et al.* 2002), physiologiques (Trudel *et al.* 2001, Rogers & Bernatchez 2005, 2007) et transcriptomiques (Derome *et al.* 2006, St Cyr *et al.* 2007, Nolte *et al.* 2009a). De plus, des études sur la génétique de populations des formes sympatriques ont montré que dans plusieurs lacs celles-ci restaient isolées génétiquement malgré un flux de gènes potentiels (Lu & Bernatchez 1999, Lu *et al.* 2001). Ainsi, bien que l'existence de formes hybrides soit possible, celles-ci souffrent d'une importante réduction de leur valeur sélective. La manifestation de cet isolement reproducteur postzygotique entre les écotypes semble être due aussi bien à des forces intrinsèques (performance réduite des spermatozoïdes et mortalité accrue des hybrides, Lu & Bernatchez 1998, Whiteley *et al.* 2009, Rogers & Bernatchez 2006, Renaut & Bernatchez *accepté.*), qu'à des forces extrinsèques dépendantes du contexte environnemental (Rogers & Bernatchez 2006, Nolte *et al.* 2009a).

Au moment où j'ai amorcé mon projet, il existait déjà plusieurs ressources disponibles pour étudier les bases génomiques du processus de spéciation chez les salmonidés. Entres autres, comme mentionné plus haut, une biopuce de 16000 gènes pour les salmonidés (cDNA microarray) a été développée récemment (von Schalburg *et al.* 2005, voir section 1.6), ainsi qu'une importante quantité d'information de séquences provenant d'autres espèces de salmonidés proche du corégone a été collectée (Rise *et al.* 2004, cGRASP 2010). Plus particulièrement pour le corégone, des cartes génétiques basées sur des marqueurs AFLPs et microsatellites ont aussi été développées (Rogers *et al.* 2007) et, par la suite, de nombreux QTLs phénotypiques (Rogers & Bernatchez 2007) et d'expression (eQTLs, Derome *et al.* 2007, Whiteley *et al.* 2007) ont été identifiés. Ainsi, la combinaison des études de cartographie ainsi que de balayage génomique en milieu naturel avait donc déjà permis de mettre en évidence certaines régions génomiques comme potentiellement impliquées dans la divergence adaptative des corégones. Néanmoins, l'utilisation de marqueurs génétiques anonymes restreint la portée réelle de tels résultats à identifier concrètement des gènes candidats impliqués dans la spéciation (Campbell & Bernatchez 2004, Rogers & Bernatchez 2007). De manière générale, toutes ces ressources justifient l'utilisation du grand corégone comme espèce modèle pour des études qui permettront d'approfondir nos connaissances sur les mécanismes génomiques régissant l'évolution des espèces (Bernatchez *et al.* 2010).

1.12 Objectifs de la thèse

Une des idées principales de ma thèse était premièrement de mieux comprendre, de manière globale, l'effet de la divergence adaptive sur l'isolement reproducteur et l'architecture génétique des populations de corégone. Deuxièmement, l'objectif était de bâtir sur les études mentionnées dans la section précédente et ainsi intégrer différents niveaux de complexité biologique afin d'identifier spécifiquement des gènes candidats impliqués dans le processus de spéciation. Ces gènes candidats pourront par la suite faire lieu de confirmations ou l'objet d'études fonctionnelles plus approfondies.

Lors du deuxième chapitre, nous nous sommes attardés, à l'aide de la technique des biopuces, à étudier la divergence d'expression entre les nains et normaux à deux stades développementaux. Étant donné que, tôt dans l'ontogénèse, le développement embryonnaire est fortement conservé au cours de l'évolution tant au niveau phénotypique que transcriptomique (Slack *et al.* 1993, Irie & Sehara-Fujisawa 2007), nous avions donc émis comme hypothèse que les différences d'expression entre nains et normaux devraient se manifester plus tard que tôt dans l'ontogénèse (Nolte *et al.* 2009a). Par la suite, nous avons comparé ces résultats à la divergence d'expression identifiée chez les hybrides de première et de deuxième génération et à ainsi quantifier les patrons d'expression additifs, dominants, non-additifs ou encore transgressifs (Renaut *et al.* 2009). Comme expliqué en section 1.7, l'incompatibilité des génomes devrait engendrer une dérégulation de l'expression chez les hybrides, et, plus particulièrement, chez ceux de deuxième génération.

Ayant lors du deuxième chapitre identifié des patrons d'expression anormaux plus importants chez les hybrides de deuxième génération en conformité avec la théorie, nous avons suivi dans le détail le développement de ceux-ci lors du troisième chapitre. Ainsi, puisqu'un pourcentage relativement important d'hybrides rétrocroisés se développait de manière anormale, l'objectif était donc de quantifier le niveau de différentiation d'expression associé à cette manifestation d'isolement reproducteur post-zygotique et, par la suite, de vérifier si certaines catégories de gènes étaient plus particulièrement ciblées (Renaut & Bernatchez *accepté*).

Lors du quatrième chapitre, nous avons utilisé la technique de pyroséquençage 454 afin de séquencer le transcriptome d'individus nains, normaux et hybrides. Cette approche, permet d'obtenir directement des donnés de séquences et d'expression (Renaut *et al.* 2010, Jeukens *et al. sous presse*). Les objectifs de cette partie étaient en premier lieu 1) de développer une large banque de SNPs potentiels, puis 2) d'évaluer la distribution de ces SNPs à travers les régions codantes, 3) d'évaluer la divergence de fréquence d'allèles entre nains et normaux, 4) d'évaluer les effets de l'hybridation sur l'expression et finalement si possible 5) d'identifier des gènes ou des mutations responsables de la divergence et de l'isolement des nains et normaux.

Lors du cinquième chapitre, nous avons effectué des analyses de balayage génomique en utilisant des marqueurs SNPs que nous avons développés afin de caractériser l'effet de la sélection naturelle sur la variation génétique pour des gènes aux fonctions connues (Renaut *et al. accepté*). Les cinq lacs analysés, caractérisés par une différente intensité de compétition ainsi que de divergence génétique et phénotypique, représentent donc un continuum de spéciation écologique (Bernatchez *et al.* 1999, Landry *et al.* 2007b, Landry & Bernatchez 2010). Nous avions donc comme objectif de vérifier si les lacs avec une faible intensité de compétition avaient peu de régions du génome touchées par la sélection divergente par rapport aux lacs présentant des environnements hautement compétitifs. Par la suite, en utilisant le même ensemble de marqueurs SNPs, nous avions pour but de tester l'association entre la variation génétique et des traits phénotypiques adaptatifs. À ce titre, l'utilisation intégrée de données de balayage génomique, d'association génotype-phénotype et de génomique fonctionnelle permet d'identifier plus précisément des gènes candidats impliqués dans la divergence écologique récente des corégones nains et normaux.

Chapitre 2 : Divergence in gene regulation at young life history stages and hybrid misexpression between lake whitefish species pairs (*Coregonus* spp. Salmonidae).

2.1 Résumé de l'étude de Nolte, Renaut et Bernatchez incluse en annexe (2009) : *Divergence in gene regulation at young life history stages of whitefish (Coregonus sp.) and the emergence of genomic isolation*

Contexte : L'étude de l'évolution des barrières à la reproduction est d'un intérêt crucial dans notre compréhension du processus de spéciation. Cependant, il existe probablement un certain biais pour l'étude de l'isolement reproducteur postzygotique chez des paires d'espèces anciennes par rapport à l'apparition de barrières au flux de gènes agissant de manière extrinsèque, spécifique à l'environnement. Ce travail évalue l'importance relative des deux processus dans l'évolution de l'isolement génomique des espèces naissantes de corégones (*Coregonus clupeaformis*) pour lesquelles des données préliminaires suggèrent que l'isolement reproducteur postzygotique intrinsèque agit au stade embryonnaire, mais aussi de manière extrinsèque à l'âge adulte.

Résultats : Grâce aux biopuces, l'expression de gènes a été quantifiée afin d'identifier à quel stade (tôt dans le développement embryonnaire ou au stade juvénile) les différences d'expression se manifestaient. Ainsi, les poissons juvéniles possédaient 14 fois plus de gènes différentiellement exprimés entre les nains et normaux que les embryons. De plus, les différences d'expression chez les juvéniles corroborent les patrons d'expression observés chez les adultes et suggèrent donc que les divergences d'expression chez les poissons juvéniles persistent tout au long de la phase adulte. Des analyses comparatives montrent qu'au moins 26 facteurs génétiques identifiés chez les juvéniles sont aussi différentiellement exprimés chez les poissons adultes. Huit de ceux-ci montrent des changements d'expression parallèles, indépendamment du type de tissu ou de l'âge du poisson. Ces derniers sont associés au métabolisme énergétique, un caractère complexe, connu comme impliqué dans la divergence adaptative des corégones nains et normaux.

Conclusion : L'analyse présentée ici a identifié peu de gènes candidats chez les embryons, corroborant les études antérieures montrant une absence de divergence écologique entre les nains et les normaux au stade larvaire. A l'opposé, la divergence d'expression liée à des caractères adaptatifs chez les juvéniles semble être dictée par la sélection naturelle divergente. Ces résultats suggèrent donc que l'isolement reproducteur postzygotique extrinsèque, i.e. dépendant de l'environnement, peut être plus important pour expliquer l'apparition de barrières reproductrices que des obstacles intrinsèques.

2.2 Abstract of study by Nolte, Renaut & Bernatchez (2009) included in annex: *Divergence in gene regulation at young life history stages of whitefish (Coregonus sp.) and the emergence of genomic isolation.*

Background: The evolution of barriers to reproduction is of key interest to understand speciation. However, there may be a current bias towards studying intrinsic postzygotic isolation in old species pairs as compared to the emergence of barriers to gene flow through adaptive divergence. This study evaluates the relative importance of both processes in the evolution of genomic isolation in incipient species of whitefish (*Coregonus clupeaformis*) for which preliminary data suggest that postzygotic isolation emerges with intrinsic factors acting at embryo stages but also due to extrinsic factors during adult life.

Results: Gene expression data were screened using cDNA microarrays to identify regulatory changes at embryo and juvenile stages that provide evidence for genomic divergence at the underlying genetic factors. A comparison of different life history stages shows that 16-week old juvenile fish have 14 times more genes displaying significant regulatory divergence than embryos. Furthermore, regulatory changes in juvenile fish match patterns in adult fish suggesting that gene expression divergence is established early in juvenile fish and persists throughout the adult phase. Comparative analyses with results from previous studies on dwarf-normal species pairs show that at least 26 genetic factors identified in juvenile fish are candidate traits for adaptive divergence in adult fish. Eight of these show parallel directions of gene expression divergence independent of tissue type or age of the fish. The latter are associated with energy metabolism, a complex trait known to drive adaptive divergence in dwarf and normal whitefish.

Conclusion: Although experimental evidence suggests the existence of genetic factors that cause intrinsic postzygotic isolation acting in embryos, the analysis presented here provided few candidate genes in embryos, which also corroborates previous studies showing a lack of ecological divergence between sympatric dwarf and normal whitefish at the larval stage.

In contrast, gene expression divergence in juveniles can be linked to adaptive traits and seems to be driven by positive selection. The results support the idea that adaptive differentiation may be more important in explaining the emergence of barriers to gene flow in an early phase of speciation by providing a broad genomic basis for extrinsic postzygotic isolation rather than intrinsic barriers.

2.3 Résumé de l'article de Renaut, Nolte et Bernatchez (2009) : *Gene expression divergence and hybrid misexpression between lake whitefish species pairs (Coregonus spp. Salmonidae).*

Les analyses du transcriptome ont confirmé chez les hybrides interspécifiques que la dérégulation de l'expression de gènes peut être à la base de l'isolement reproducteur. Ici, en utilisant une biopuce à ADNc contenant 16 006 points, nous avons comparé et contrasté la divergence d'expression à deux stades de développement chez des jeunes espèces de grand corégones (*Coregonus clupeaformis*) à celle d'hybrides de première (normal X nain) et deuxième [(normal X nain) X nain] génération. Notre but était d'identifier le mode d'action responsable de l'expression de gènes en plus d'identifier des gènes candidats dérégulés chez les hybrides. Au stade embryonnaire, l'expression moyenne différait pour très peu de gènes (5 sur 4950 exprimés) entre les parents et les hybrides, ce qui contrastait avec les poissons juvéniles, pour lesquels 617 des 5359 transcrits différaient significativement. Nous avons aussi trouvé des évidences d'une dérégulation accrue de l'expression chez les hybrides rétrocroisés, puisque la non additivité de l'expression expliquait une plus grande fraction des patrons de transmission chez les rétrocroisements (54%) que chez les hybrides F1 (9%). L'expression de gènes chez les hybrides montrait plus de variabilité chez les formes pures et les patrons d'expression transgressive étaient ubiquitaires chez ceux-ci. En particulier, chez les rétrocroisements, l'expression de gènes développementaux impliqués dans le repliement des protéines et la traduction de l'ARN messager était sévèrement dérégulée. Ainsi, la dérégulation de l'expression chez les hybrides se rajoute aux autres facteurs précédemment identifiés comme contribuant à l'isolement reproducteur des espèces naissantes de grand corégones.

2.4 Abstract of Renaut, Nolte & Bernatchez (2009): *Gene Expression Divergence and Hybrid Misexpression between Lake Whitefish Species Pairs (Coregonus spp. Salmonidae)*

Genomewide analyses of the transcriptome have confirmed that gene misexpression may underlie reproductive isolation mechanisms in inter-specific hybrids. Here, using a 16,006 features cDNA microarray, we compared and contrasted gene expression divergence at two ontogenetic stages in incipient species of normal and dwarf whitefish (*Coregonus clupeaformis*), to that of first generation (normal X dwarf) and second generation hybrid crosses [backcross: (normal X dwarf) X normal]. Our goal was to identify the main mode of action responsible for gene transcription and to discover key genes misexpressed in hybrids. Very few transcripts (5 out of 4950 expressed) differed in mean expression level between parentals and hybrids at the embryonic stage, in contrast to 16-week old juvenile fish for which 617 out of 5359 transcripts differed significantly. We also found evidence for more misexpression in backcross hybrids whereby non-additivity explained a larger fraction of hybrid inheritance patterns in backcross (54%) compared to F1-hybrids (9%). Gene expression in hybrids was more variable than in pure crosses and transgressive patterns of expression were ubiquitous in hybrids. In backcross embryos in particular, the expression of three key developmental genes involved in protein folding and mRNA translation was severely disrupted. Accordingly, gene misexpression in hybrids adds to other factors previously identified as contributing to the reproductive isolation of incipient species of lake whitefish.

2.5 Introduction

Under the biological species concept, reproductive isolation arises as a consequence of population divergence, itself driven by natural selection, sexual selection or genetic drift (Coyne and Orr 2004; Bell 2008). The evolution of reproductive isolation is often viewed as a gradual process whereas, over time, more and more barriers will tend to separate lineages and reinforce their divergence (de Queiroz 1998). Frequent gene flow between parental lineages is a characteristic of many early divergence events (Bernatchez 2004; Wu and Ting 2004; Nosil 2008). In these situations, postzygotic isolation may result from the interaction of genetic factors in the parental lineages that, while functional in their normal genetic backgrounds, reduce fitness when recombined in hybrids (Rundle et al. 2000; Burton et al. 2006, Rogers and Bernatchez 2006; Gow et al. 2007). These hybrids can be unfit due to intrinsic factors resulting in increased embryonic mortality or external - environmentally driven- factors, for instance the lack of finding a suitable ecological niche (Schluter 2000).

In nature, rare F1-hybrids, encountering few mates of their kind, may backcross to parental species. This progeny often suffers more problems than first generation hybrids (Barton 2001; Coyne and Orr 2004). Namely, recombination is known to release cryptic genetic variation, resulting in phenotypes that are extreme relative to those of either parental line (Endler 1977; Rieseberg et al. 1999, 2003; Mallet 2007). Ultimately, extreme phenotypes can be lethal or sterile (hybrid breakdown), while transgressive segregation refers to phenotypic values in hybrids that extend significantly outside the range defined by the parents (DeVicente and Tanksley, 1993; Rockman and Kruglyak 2006). As such, a particular trait in hybrids may, on average, be similar to the parents, yet be transgressive and thus maladapted due to an increased phenotypic variability. Both hybrid breakdown and transgressive segregation may explain the underlying basis of post-zygotic isolation in early divergent lineages (Burke and Arnold 2001; Rogers and Bernatchez 2006). The manifestation of such extreme traits supposes non-additive, epistatic, gene interactions (Dobzhansky 1937; Muller 1942), although transgressive segregation can also be caused solely by the complementary action of genes with additive effects (Rieseberg et al. 1999).

Based on the premise that transcriptional regulation constitutes a major component of the genetic basis for phenotypic evolution (Wray et al. 2003; Wray 2007; but see Hoekstra and Coyne 2007; Carroll 2008), the analysis of gene expression has allowed the identification of many candidate genes underlying phenotypic divergence (Ranz and Machado 2006; Derome et al. 2006; Landry et al. 2007; St-Cyr et al. 2008; Jeukens et al. 2009). Moreover, hybrids have proven extremely valuable to identify cryptically differentiated genetic factors, whereby the combination of divergent regulatory elements into a common genetic background resulted in gene misexpression (reviewed by Ortiz-Barrientos et al. 2007 & Landry et al. 2007). For example, Ranz and colleagues (2004) have shown that the global expression profile of *Drosophila melanogaster* and *D. simulans* are more closely related to each other than to their hybrid progeny. Undeniably, there is a large body of evidence pointing towards largely non-additive inheritance of gene expression in hybrids (Rockman and Kruglyak 2006), including in fishes (Roberge et al. 2008), which may consequently explain their selective disadvantage. Other studies, however, have reported the predominance of additive patterns of gene expression in F1-hybrids (Hughes et al. 2006, Rottscheidt and Harr 2007), such that the type of inheritance responsible for gene transcription levels in hybrids remains a contentious issue (Rockman and Kruglyak 2006, Moehring et al. 2007). Finally, patterns of gene expression in young species pairs and post-F1 hybrid generations have been little explored and the underlying transcriptomic basis of reproductive isolation mechanisms remains largely unknown.

Recent post-glacial ecological divergence (12 000-15 000 years ago) of the lake whitefish *Coregonus clupeaformis* (salmonidae) has repeatedly led to the formation of two, benthic and limnetic, whitefish species occurring in sympatry and hereafter referred to as *normal* and *dwarf*, respectively (Bernatchez 2004). Extensive experimental work has pointed to differentiation at morphological (Lu and Bernatchez 1999; Rogers et al. 2002; Bernatchez 2004), life history (Bernatchez et al. 1999), physiological (Trudel et al. 2001; Rogers and Bernatchez 2005, 2007) and genetic levels (Bernatchez 2004; Rogers et al.

2007). The recent advent of microarray technology developed for salmonids (von Schalburg et al. 2005) has permitted to identify consistent gene expression divergence between normal and dwarf whitefish in both natural and laboratory settings (Derome et al. 2006; St-Cyr et al. 2008; Derome et al. 2008; Whiteley et al. 2008; Nolte et al. 2009). In particular, these studies, combined with physiological data (Trudel et al. 2001) have shown that, at the juvenile and adult life stages, energy metabolism plays a fundamental role in driving whitefish adaptive divergence. In particular, Nolte et al. (2009) identified a fourteen fold increase of differentially expressed genes later (juvenile stage) rather than sooner (embryonic stage) in ontogeny, which may potentially explain the emergence of reproductive isolation as a by-product of adaptive divergence on adult characters. At the same time, the relative lack of divergence in embryos of pure crosses may imply that individual genes are still evolving, while the net gene expression outcome is not altered. Thus Nolte et al. (2009) hypothesized that gene misexpression could manifest as genetic factors segregate in hybrid crosses, and particularly so in backcrosses (Nolte et al. 2009). In line with this, both intrinsic (increased hybrid mortality) and extrinsic (transgressive segregation in hatching time) post-zygotic isolation mechanisms where shown to be more prevalent and severe in backcross individuals than F1-hybrids (Lu and Bernatchez 1998; Rogers and Bernatchez 2006).

The present study anchors itself on a previous analysis of gene expression data of pure normal and dwarf whitefishes at two developing stages (embryonic and juvenile, Nolte et al. 2009). Here, the goal was to document the main mode of action responsible for gene transcription in hybrids and to identify genes misexpressed relative to dwarf and normal whitefish. More specifically, under the assumption that similar genes involved in the adaptive divergence of these species are also responsible for driving their reproductive isolation, and based on previous comparisons between pure forms (Nolte et al. 2009), we predicted an excess of hybrid misexpression at the juvenile compared to the embryonic stage. Secondly, we also predicted more evidence of hybrid misexpression in second generation (backcross) compared to first generation hybrids.

2.6 Methods

Strains, crosses and fish maintenance

Details regarding strain origin, crosses and fish maintenance are provided in Nolte et al. (2009). Briefly, eggs were obtained from outbred laboratory strains (normal whitefish originating from Aylmer Lake (45° 54'N, 71° 20'W), dwarf whitefish originating from Témiscouata Lake (47° 41'N, 68° 47'W), as detailed in Lu and Bernatchez 1998) at the Laboratoire de Recherche en Sciences Aquatiques (LARSA, Université Laval, Quebec, Canada). We also used wild dwarf whitefish caught in Témiscouata Lake in October 2006. In order to reduce family specific effects, we used crosses that were composed of several parents, depending on the availability of mature fish (table 2.1). The group D-1 was created using one lab strain dwarf female crossed to five different dwarf males, all originating from Témiscouata Lake. D-2 was created by crossing wild caught dwarf whitefish from the same lake using multiple females and multiple males. Two groups of Normal whitefish (N-1 and N-2) were created from 1 and 5 as well as 2 and 3, females and males of the lab strain of normal whitefishes from Aylmer Lake. F1-hybrids (F1-1) were generated between the females of group N-1 and males of Group D-1. Likewise, another group of F1-hybrids (F1-2) were created among the parents of the second pair of parental groups (D-2 and N-2). Finally, an independent group of backcross (BC) was obtained using an F1-hybrid female generated in the laboratory in a previous study (Rogers et al. 2006) crossed to five normal whitefish (lab strain). As such, the backcrosses are composed of a 75% normal and 25% dwarf genetic background.

Table 2.1 Origins of strains and crosses used for gene expression analysis.

Experimental Group	Lineage	Crosses
D (1)	Témiscouata Lake Dwarf	Lab strain: single female crossed with 5 different males.
D (2)	Témiscouata Lake Dwarf	Wild parental fish, several females crossed with several males.
N (1)	Aylmer Lake Normal	Lab strain: 2 females crossed with 3 males.
N (2)	Aylmer Lake Normal	Lab strain: single female crossed with 5 different males.
F1 (1)	Témiscouata Lake Dwarf - Aylmer Lake Normal	Aylmer Lake female (same as in N1) crossed with 5 Témiscouata Lake dwarf males (same as D1).
F1 (2)	Témiscouata Lake Dwarf - Aylmer Lake Normal	Wild caught Dwarf females (multiple) with three males from Aylmer Lake (same as N2).
BC	F1 -Aylmer Lake Normal	F1-hybrid (derived from Aylmer, Témiscouata lab strains) female crossed with five males (Aylmer Lab strain Normal).

Note. Two experimental groups were created for dwarf, normal and F1, and individuals used for gene expression experiments were composed, in equal part, from the (1) and (2) duplicate families. One experimental group was created for backcross (BC). Dwarf and normal families were used in a previous study by Nolte et al. (2009).

Sampling

We sampled embryos during the beginning of the segmentation period. For our experiments we chose embryos that had formed approximately 20 segments in the detached portion of their tail, which was also curved at an angle of approximately 30°. Furthermore, in this stage, the optic primordium begins to hollow, thus initiating the formation of the eye lens. Viable eggs with well-formed embryos were individually selected, preserved in RNA later (Ambion, Austin, TX) and frozen at –20 °C for storage.

All hatched larvae were transferred to tanks and we sampled juvenile fish at an age of 16 weeks (May 2007), when these reached a weight of approximately 860 mg (500–1190 mg). Young immature fish chosen for gene expression analysis were well developed and in good general shape. Sampling was done in the morning following an 18 hour fast. Fish were then killed with a blow, kept on ice (no longer than 20 min.), homogenized in Trizol Reagent (Invitrogen, Carlsbad, CA) using a polytron homogenizer and stored at –80°C prior to RNA extraction.

Choice of samples and analysis of gene expression

Total RNA was extracted using the Trizol Reagent protocol. For the embryo experiment, a pool of five whole embryos preserved in RNA later was homogenized using a bead mill (Qiagen, Germantown, MD). RNA pooling is a common practice when quantity of material is limiting and inference for most genes is not affected by this (Kendziorski et al. 2005). For the juvenile fish experiment, a single whole juvenile fish was used. Crude total RNA was further cleaned by ultra filtration using microcon (Millipore, Billerica, MA) spin columns (embryo experiment) or a combination of a lithium chloride precipitation (1 vol. 5 M LiCl, precipitation at –20 °C for 2 hrs, centrifugation at 16.000 g at 4 °C for 30 min., 70% ethanol wash) and subsequent ultra filtration (juvenile experiment). Total RNA was quantified and quality checked using Experion™ RNA StdSens Analysis Kit (Bio-Rad, Hercules, CA). Total RNA was stored in pure water supplemented with Superase-In™ RNase Inhibitor (Ambion) at –80°C.

Gene expression analysis was performed using the 16K (v2.0) Salmon cDNA microarray provided by the cGRASP consortium (von Schalburg et al. 2005). Following the vendor's protocol, Genisphere (Hatfield, PA) 3DNA Array Detection Array 350™ Kit (Cy3/Alex647) was used in the embryo experiment since it requires less starting material (we used 3-5ug of RNA). Genisphere 3DNA Array Detection Array 50™ Kit (Cy3/Cy5) was used in the juvenile experiment, where we used 15-20ug of starting RNA. Reverse transcription reactions were performed using Superscript II Kit™ (Invitrogen). Microarrays were scanned using a ScanArray™ Express scanner (Packard Bioscience, Wellesley, MA).

Samples of dwarf, normal, F1-hybrid and backcross were hybridized in a loop design, involving eight biological replicates, and dye swap performed between each replicate (figure 2.1). In this way, technical replication for dwarf and normal samples was three fold while F1-hybrid and backcross samples were each involved in two pair wise comparisons. As a result, we obtained a final set of 40 experiments for each, embryo and juvenile, datasets.

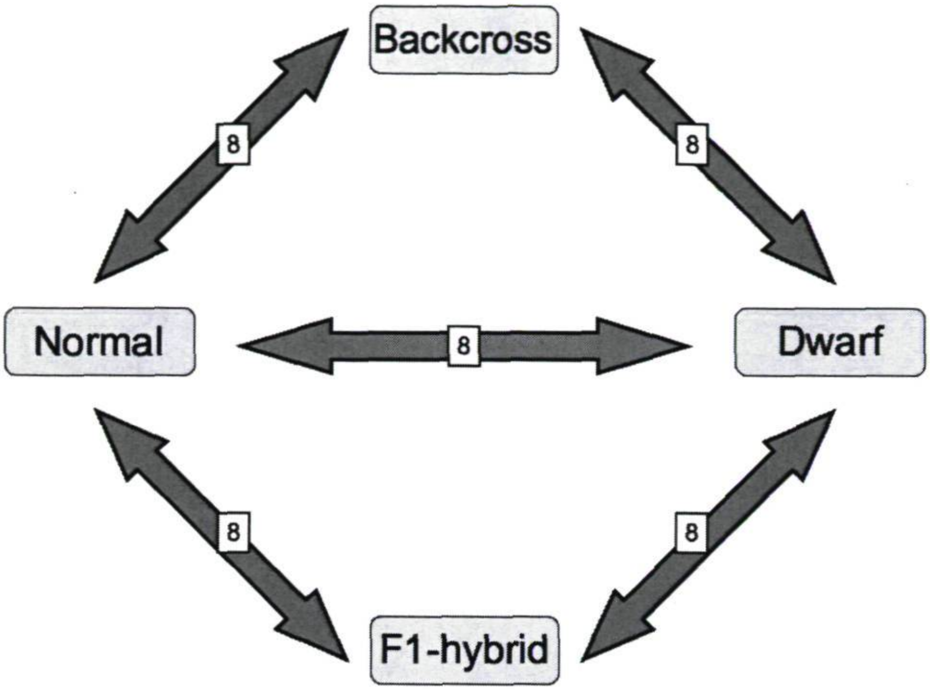

Figure 2.1 Microarray design with the four experimental groups. Each double headed arrow represents one microarray slide such that one complete loop corresponds to 5 slides. Each loop was replicate 8 times for a total of 40 slides for both the embryo and juvenile datasets.

We used an ANOVA based approach using the R package Rmaanova (v1.4.1, Kerr et al. 2000) to identify transcripts differentially expressed. The mixed model used included the following terms as fixed sources of variance: Group (normal, dwarf, F1-hybrid or backcross) and Dye (Fluorescent dye). Sample (biological sample) and Array (individual microarray) were included as random sources of variance. Statistical testing for overall divergence in gene expression is based on an F test (1000 permutations, Fs test option). We corrected for multiple testing using a False Discovery Rate (FDR) cut off value of 0.05, as implemented in Rmaanova. To test for specific pair wise differences between groups (N, D, F1, BC), we used the contrast option (t-test) implemented in Rmaanova (1000 permutations, Fs test option, FDR cut off value of 0.05). In order to remain conservative in interpreting the number of significant features, transcripts with less than 2% sequence divergence where also compressed into a single transcript.

Bayesian analysis of gene expression

Relative level of expression was also estimated for each transcript using a Bayesian approach (BAGEL V.2, Townsend and Hartl 2002). Normalized ratio data were implemented in the software (using the default parameters), and the most probable relative gene expression values were then calculated per transcript, per group (N, D, F1 or BC) and per replicate. The average relative gene expression and variance estimate were then calculated from the eight replicates. Previous work has shown that this Bayesian method strongly corroborates results obtained through analysis of variance-based methods, yet provided a simpler interpretation of relative gene expression and variance (Meiklejohn and Townsend 2005). In our juvenile dataset, we verified that, for all comparisons, fold changes calculated from the relative gene expression values obtained through the Bayesian method and fold changes provided by Rmaanova were highly correlated ($r^2 > 0.97$, data not shown). Similar correlation values were obtained in the embryo dataset, but the low number of transcripts differentially expressed limited the validity of the approach.

Variance estimation in gene expression

Test of homogeneity of variance (Bartlett 1937) between the groups (N, D, F1, BC) was performed to identify general patterns of variability and transgressivity in hybrids compared to parental species, using the relative gene expression values calculated for each replicate. Histograms of the variance estimates were drawn for each group from all the transcripts that showed heterogeneity of variance (at *p*-value <0.05, Bartlett test). Non-parametric Wilcoxon rank sum test where then used to compare the groups (Bauer 1972).

Classification of gene expression inheritance patterns

In order to categorize different types of inheritance, we analyzed the distribution of dominance effects (d/a ratio distribution, Gibson et al. 2004; Hughes et al. 2006, Rottscheidt and Harr 2007) to decipher between additivity, dominance or non-additivity for genes differentially expressed between parentals. For F1-hybrids, $d = \mu_{\text{F1-hybrid}} - ((\mu_{\text{normal}} + \mu_{\text{dwarf}}) / 2)$, $a = (\mu_{\text{normal}} - \mu_{\text{dwarf}}) / 2$, where μ are relative gene expression values. In backcross, formulas were modified to take into account the 75 % normal, 25% dwarf hybrid background, $d = \mu_{backcrosss} - ([\mu_{normal} + (\mu_{dwarf} + \mu_{normal})/2] / 2)$ and $a = [(\mu_{normal} - (\mu_{dwarf} + \mu_{normal})/2] / 2$.

Additivity- Transcript whose hybrid gene expression value corresponds to the mid-value of the parents. A d/a ratio of 0 corresponds to perfect within-locus additivity (*i.e.* $d = 0$). We then set up an arbitrary range of -0.5 to + 0.5 to include transcripts showing patterns of gene expression resembling additivity, rather than dominance.

Dominance- Transcript whose hybrid gene expression value resembles more closely one parent than another. A d/a ratio = -1 or +1 corresponds to complete dominance. In order to also include transcripts showing near complete dominance, we applied a d/a ratio threshold of +0.5 to +1.5 (normal dominance) or -0.5 to -1.5 (dwarf dominance).

Non-additivity- Transcript whose hybrid gene expression is lower or higher than both parents. A d/a ratio greater than +1.5 or smaller than -1.5 corresponds to non-additivity. These genes were further classified as under / overdominant if the hybrid has an expression lower / higher than the mean of both parents. We also included in this category transcripts not significantly differentiated between the parents, but for which hybrid expression was significantly different from the parental values, either by being under or overexpressed.

Transgressivity- As defined by Brem and Kruglyak (2005) and Rockman and Kruglyak (2006), transgressivity corresponds to a transcript whose level of expression in segregating hybrids does not necessarily differ in average from mid-parent values but whose variance extends outside the range of both parents values. Firstly, the maximum range of values defined for the parents was calculated as mean expression plus 2 standard deviation ($\mu_{parent} + 2\sigma_{parent}$) of the highest parent (either normal or dwarf) minus ($\mu_{parent} - 2\sigma_{parent}$) of the lowest parent. Similarly, the range of values for each hybrid group was calculated as ($\mu_{\text{F1-hybrid}} + 2\ \sigma_{\text{F1-hybrid}}$) - ($\mu_{\text{F1-hybrid}} - 2\ \sigma_{\text{F1-hybrid}}$) and ($\mu_{\text{BC}} + 2\ \sigma_{\text{BC}}$) - ($\mu_{\text{BC}} - 2\ \sigma_{\text{BC}}$). Transcripts were qualified as transgressive if the hybrid had a range of values greater than the maximum range of the parents. The total number of transgressive transcripts (as defined above) for each hybrid group was then calculated. We estimated the total number of false positive transgressive transcripts expected by permuting for each transcript separately, the identity of each of the four groups and then calculating a new total number of transgressive transcripts for hybrids as aforementioned (4 groups X 8 replicate loops = 32 columns permuted using the function *sample* in R). *P*-values were then estimated as the number of times, over 1000 permutations, that the total number of false positive transgressive transcripts was greater than the observed number of transgressive transcripts. Lastly, we examined the function and identity of the transcripts that showed the highest transgressivity values in hybrids (range of hybrid, either F1-hybrid or backcross, minus maximum range of parents > 0.5). This cutoff value, although arbitrary, permits to identify the most severely transgressive transcripts, which one would predict to be more likely biologically relevant compared to nearly transgressive transcripts. *P*-values for those transcripts were also estimated as the number of times a randomized dataset would produce a hybrid transcript more transgressive than the real calculated value (1000 permutations).

2.7 Results

This study primarily focuses on hybrid gene expression relative to that reported recently between pure parentals by Nolte et al. (2009). As such, results obtained for pure parental comparisons are not treated in details but reported for the sake of comparison only. In brief, Nolte et al. (2009) found 33 and 502 transcripts differentially expressed in the embryo and juveniles, respectively. Their numbers are slightly different than the ones we obtained for the same comparisons; that is, five and 543 transcripts, respectively. This discrepancy is likely due to the overall different number of technical replicates in both studies. Yet it does not influence any of the earlier conclusions reported by these authors, namely that the number of significant transcripts between normal and dwarf is much larger at the juvenile than the embryonic stage.

Gene expression differentiation among groups.

The number of transcripts (Expressed Sequence Tag clones spotted on the microarray) for which we obtained gene expression data of sufficient quality for subsequent analyses was 4950 and 5359 for the embryos and for the juvenile dataset, respectively. After an FDR correction (0.05) and a compression of replicate spots, we identified five transcripts (0.1 % of all transcripts expressed) and 573 transcripts (12%) as differentially expressed in the embryo and juvenile datasets, respectively among all groups compared. In the embryo dataset, most of that difference was due to normal-dwarf comparison, but the small number of differentiated transcripts hampered the interpretation of this trend (table 2). In the juvenile dataset, most of the observed differences were in the normal-dwarf comparison since 501 of the 573 transcripts (87%) were differentially expressed between the parents. For comparisons involving hybrids, there was much less differentiation between F1 and normal (94 genes or 16% of all transcripts differentially expressed) compared to differences between F1 and dwarf (403 genes or 70%). In contrast, 161 (28%) transcripts differed in the backcross-dwarf comparison compared to 343 (60%) in the backcross – normal comparison. Finally, 177 (31%) transcripts significantly differed between F1 and BC hybrids (see table 2.2 and supplementary table 2.1 for all the transcript IDs, fold changes and relative expression).

Table 2.2 Number of transcripts identified in the ANOVA (FDR corrected permutation p-value < 0.05, Fs test option, 1000 permutations) and the subsequent t-tests between the different experimental groups for embryos and juveniles (FDR corrected permutation p-value < 0.05, Fs test option, 1000 permutations).

Embryos				
		N	D	F1
ANOVA	5			
Backcross		0	5	2
Normal			5	0
Dwarf				1

Juveniles				
		N	D	F1
ANOVA	573			
Backcross		343	161	177
Normal			501	94
Dwarf				403

Note. N=Normal, D=Dwarf, F1=F1-hybrid.

Table 2.3 Number of transcripts identified as additive, dominant, non-additive or transgressive.

	F1-hybrid	Backcross
Additive		
embryos	1 (60%)	2 (40%)
juveniles	217 (44%)	98 (20%)
Dominant		
embryos	4 (40%)	3 (60%)
juveniles	237 (47%)	133 (27%)
Non-additive		
embryos	0 (0%)	0 (0%)
juveniles	47 (9%)	269 (54%)
-overdominant	19 (3%)	100 (20%)
-underdominant	28 (6%)	169 (34%)
Overexpressed	5 (7%)	15 (20%)
Underexpressed	3 (4%)	3 (4%)
Transgressive		
Embryos	1306 (26%)	2622 (53%)
juveniles	2097 (39%)	2316 (43%)

Note. See Material and Methods for criteria defining the categories. Percentage relate to the total number of transcripts differentiated between the parents (5 in embryo dataset, 501 in juvenile dataset), except for over- under-expressed transcripts where percentage refers to the total number of genes specifically differentiated in the juvenile hybrids (75), and for transgressive transcripts, where percentages refer to the total number of transcripts significantly expressed in the embryo (4950) and juvenile (5359) dataset.

Estimates of gene expression variability

In the embryo dataset, 799 transcripts (16% of all transcripts expressed) showed heterogeneity of variance in expression between groups (Bartlett test, p-value < 0.05). Of those, backcross hybrids showed the highest mean and median values of variance, followed by dwarf, F1-hybrids, and normal. All comparisons were significant (Wilcoxon test, p-value < 0.0001, Figure 2.2). In juvenile fish, 656 transcripts (13%) showed significant heterogeneity of variance among groups (Bartlett test, p-value < 0.05). On average, F1-hybrid and backcross were more variable than both normal and dwarf (Wilcoxon test, p-value < 0.0001) but not different from one another (Wilcoxon test, p-value =0.7, Figure 2.3).

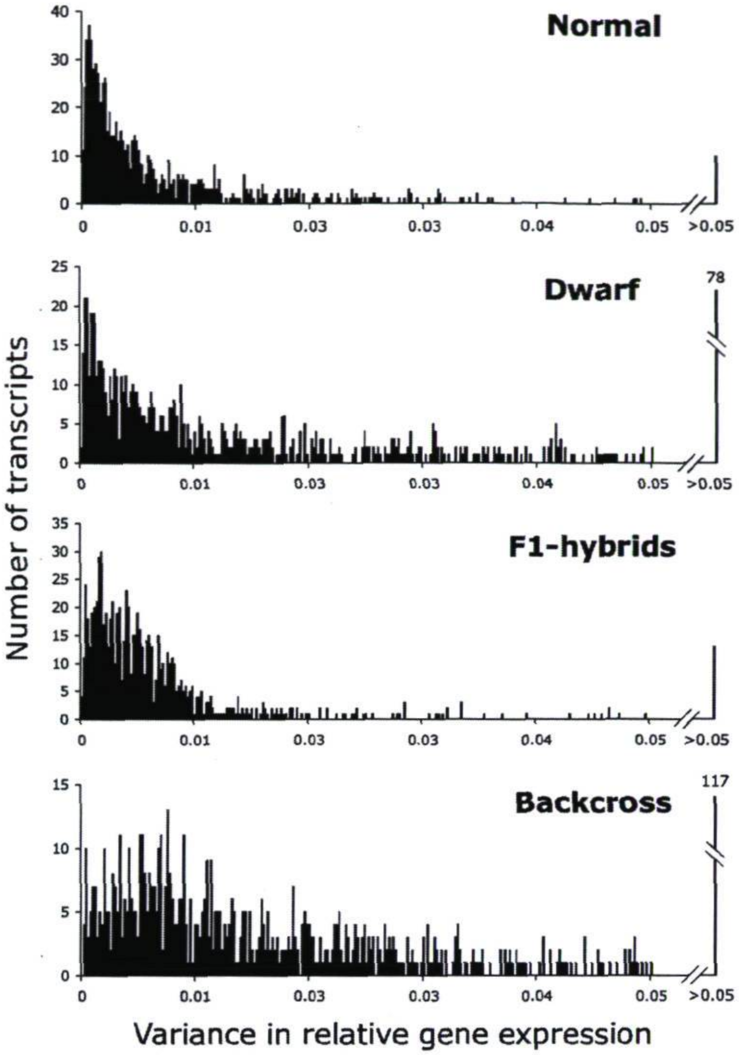

Figure 2.2 Frequency distribution of variance of relative gene expression (embryo dataset) for the 4 groups for genes showing heterogeneity of variance according to a Bartlett test (*p*-value < 0.05). Backcross show the greatest mean and median of variance (mean=0.026, median=0.014), followed by dwarf (mean=0.018, median=0.0089), F1-hybrids (mean=0.0075, median=0.0046), and normal (mean=0.0081, median=0.0035). All pair wise comparisons are significant (Wilcoxon rank test, *p*-value < 0.0001).

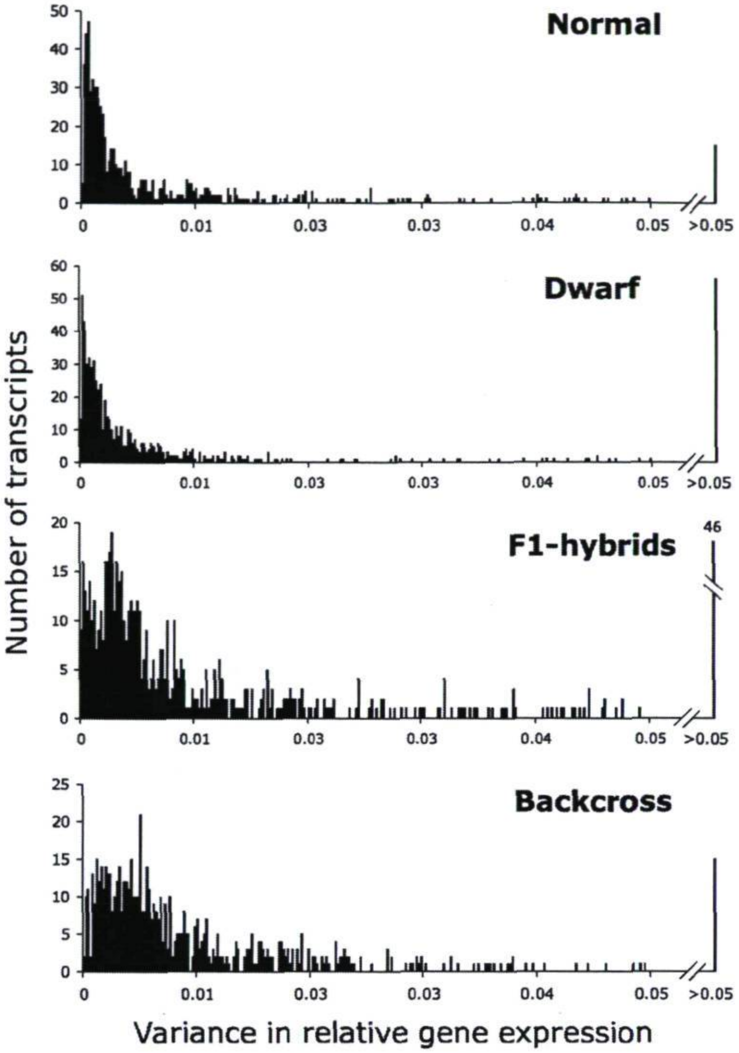

Figure 2.3 Frequency distribution of variance of relative gene expression (juvenile dataset) for the 4 groups for genes showing heterogeneity of variance according to a Bartlett test (*p*-value < 0.05). Backcross (mean=0.012, median=0.0059), and F1-hybrids (mean=0.014, median=0.0053) show comparable and greatest mean and median of variance followed by dwarf (mean=0.017, median=0.0024) and normal (mean=0.008, median=0.025). BC–F1-hybrid comparison and N-D comparison are not significant (Wilcoxon rank test, *p*-value = 0.76 and *p*-value = 0.86, respectively), while all other comparisons are significant (Wilcoxon rank test, *p*-value < 0.0001).

Type of inheritance observed in hybrids

It is noteworthy that the same transcripts generally showed different patterns of expression in F1 and backcross hybrids such that there was actually little correlation between d/a values for specific transcripts in F1 and in backcross hybrids ($r^2 = 0.1$ for juveniles). Secondly, the distribution of dominance effect was skewed towards pure normal whitefish (mean = 0.38, *p*-value < 0.0001, t-test) in the F1-hybrids distribution. This reflects the fact that, as stated above, patterns of expression observed in F1-hybrids were generally more similar to normal than dwarf whitefish. Conversely, backcross d/a ratio distribution displayed skewness towards dwarf (mean = -1.27, *p*-value < 0.0001, t-test, fig 2.4).

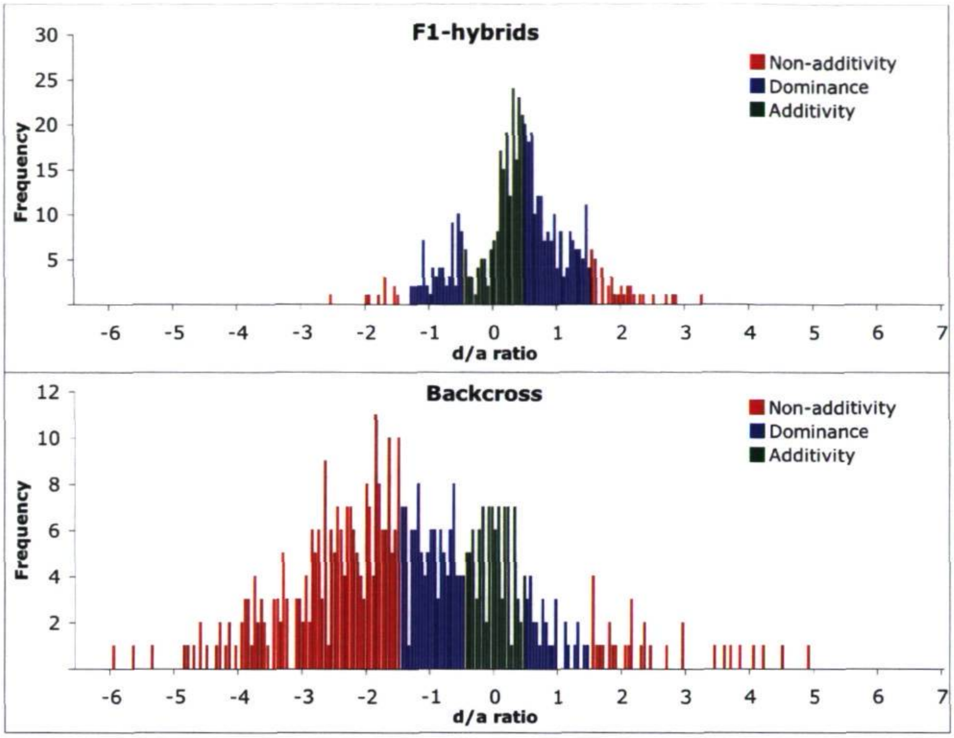

Figure 2.4 Distribution of dominance effects (d/a ratio) for F1-hybrids and backcross for transcripts significantly different between the parents (501 juvenile transcripts). A |d/a| ratio between 0.5 and 1.5 is considered as dominant, >1. 5 non-additive, <0.5 additive (see Material and Methods). Positive values imply that hybrids are more closely related to the normal parent while negative values imply that hybrids are more closely related to the dwarf parent.

Embryos

Since the identification of inheritance patterns relies mostly on differentially expressed transcripts, the very small number of differentiated transcripts hampered the use of this approach in the embryos. On the other hand, we identified a large fraction of transcripts showing evidence of transgressivity both in F1-hybrids (1306, or 26% of all transcripts expressed) and backcross (2622, 53%, table 2.3). The number of false positive transgressive transcripts expected based on 1000 permutation was 1452 (29%) and consequently there were significantly more transgressive transcripts than expected by chance in backcross (*p*-value = 0.01), but not F1-hybrids (*p*-value = 0.55). Nine transcripts showed very high transgressivity in backcross hybrids (range of backcross - maximum range of parents > 0.5). These transcripts, not differentially expressed in any comparisons, are involved in protein folding, mRNA translation, signal transduction, germ-line formation and endocytosis (table 2.4). Moreover, five of those nine transcripts closely match with three different homologs (*Immunoglobulin binding protein* [protein folding], *e*-value: 1e-157; *translation elongation factor alpha 1* [mRNA translation], *e*-value: 1e-15 and *40S ribosomal protein s11* [mRNA translation], E-value: 3e-93) identified as essential for early embryonic development of *Danio rerio* (Amsterdam et al. 2004, table 2.4).

Table 2.4 Range of relative gene expression level (in %) of highly transgressive transcripts for both embryo and juvenile datasets.

Embryos

Transcript ID	Maximum range (parents)	Range (backcross)	Range (F1-hybrids)	Gene product	Functional group
CB485951	[36-181]	[0-233]***	[50-157]	Heat shock cognate 70 kDa protein [a]	Protein folding
CK991158	[46-165]	[17-214]**	[72-136]	Heat shock cognate 70 kDa protein [a]	Protein folding
CB516765	[29-164]	[15-219]**	[39-190]	Fish-egg lectin	Lipopolysaccharide-binding protein
CA060826	[38-179]	[9-226]*	[66-148]	Elongation factor 1 alpha [b]	Translation
CK990889	[48-155]	[28-197]**	[76-139]	Guanine nucleotide-binding protein	Signal transduction
CA051954	[56-150]	[34-198]**	[78-126]	Protein kinase C	Germ line formation
CB502683	[39-174]	[16-224]**	[74-135]	Heat shock cognate 70 kDa protein [a]	Protein folding
CB502825	[69-125]	[45-151]**	[65-156]	Asialoglycoprotein receptor 2	Endocytosis
CN442505	[58-123]	[37-153]***	[72-156]	40S ribosomal protein S11 [c]	Translation

Juveniles

Transcript ID	Maximum range (parents)	Range (backcross)	Range (F1-hybrids)	Gene product	Functional group
CB507670	[78-123] [a]	[74-128]	[49-145]***	Collagen alpha-2(I) chain precursor	Muscle contraction
CA064346	[59-126]	[97-157]	[24-142]**	Proproteinase E precursor	Protein degradation
CA053777	[76-117]	[76-120]	[67-159]***	SJCHGC04882	Unknown
CK990215	[18-178]	[14-156]	[4-216]*	UNKNOWN	Unknown
CA043836	[67-120]	[62-123]	[63-168]**	Phosducin-like protein 3	OTHER
CA037858	[70-135]	[58-144]	[30-147]***	UNKNOWN	Unknown
CB488336	[94-117]	[78-122]	[52-126]***	Collagen alpha-1(I) chain precursor	Muscle contraction

CB510992	[76-138]	[65-126]	[33-148]***	Apolipoprotein A-I precursor	Lipid metabolism
CB504468	[68-132]	[87-136]	[29-147]***	Elastase-1	Muscle contraction
CA038612	[60-146]	[70-148]	[12-151]**	Serotransferrin-2 precursor	Transport
CB500533	[84-115]	[65-141]	[41-137]**	Collagen alpha-1(I) chain precursor	Muscle contraction
CB507066	[80-116]	[74-128]	[41-143]***	Collagen alpha-1(I) chain precursor	Muscle contraction
BU965755	[86-123]	[73-131]	[45-138]***	Coiled-coil domain	OTHER
CA038358	[86-121]	[76-128]	[45-137]**	Proteasome subunit alpha type 2	Protein degradation
CB496771	[58-148]	[65-148]	[7-159]**	Serotransferrin precursor	Transport
CA043815	[67-144]	[67-124]	[21-163]***	UNKNOWN	unknown
CB492384	[82-111]	[55-139]***	[71-147]	Creatine kinase B-type	Energy metabolism
CB497013, CB496702	[47-131]	[47-200]**	[29-175]**	Myosin heavy chain	Muscle contraction

Note. See Material and Methods for criteria defining the categories. Values below 100 % imply underexpression compared to the average of all four groups, while values over 100% imply overexpression compared to the average of all four groups. Transcripts for which the gene expression range of hybrid minus the maximum range of the parents > 50: *** *p*-value <0.01, ** *p*-value < 0.05, * *p*-value < 0.1 (based on 1000 permutations as described in Material and Methods). Gene product and functional groups are based on the latest annotation file provided by cGRASP (May 2008).

Corresponding homologs and knockdown phenotypes in *Danio rerio* study (Amsterdam et al. 2004):

[a] Immunoglobulin binding protein (BLASTn, *e*-value: 1e-157)

-Knockdown phenotype: Day 1: pinched midbrain/hindbrain boundary. Day 2: small head and eyes, inflated hindbrain ventricle, thin body. Day 5: very small head and eyes, thin body with underdeveloped liver/gut.

[b] Eukaryotic translation elongation factor 1 alpha (BLASTn, e value: 1e-15).

-Knockdown phenotype: Day 3: small head and eyes, rounder yolk. Day 5: increasingly necrotic.

[c] 40S ribosomal protein s11 (BLASTn, *e*-value: 3e-93)

-Day 1: pinched midbrain/hindbrain boundary. Day 2: small head and eyes, inflated hindbrain ventricle, thin body. Day 5: small head and eyes, thin or necrotic body, round grey yolk, underdeveloped liver/gut

Juvenile stage

For the 501 transcripts that were differently expressed between the pure forms at the juvenile stage, 217 (44%) and 98 (20%) had a relative expression resembling additivity for F1 and backcross hybrids, respectively. A total of 237 transcripts (47%) had an inheritance pattern close to dominance in F1-hybrids compared to 133 (27%) in backcross. A total of 47 (9%) and 269 (54%) of all transcripts showed level of expression that fell outside the mean of the parents (non-additive) for F1-hybrid and backcross, respectively ($|d/a|$ ratio > 1.5, table 2.3). Twenty-eight and 169 transcripts were under-dominant compared with 19 and 100 over-dominant in F1-hybrid and backcross, respectively. Seventy-five transcripts were not differentiated between the parents (suppl. table 2.1), yet still identified in the ANOVA; most (49) of those significantly differentiated from only one parent yet never from both. In juvenile F1-hybrid, three (underexpressed) and five (over-expressed) transcripts had a mean level of expression falling significantly outside the parental range. Under the same criteria, three and 15 transcripts were respectively under and overexpressed in the backcross hybrids.

We detected a large fraction of transcripts showing evidence of transgressivity both in F1-hybrids (2097, or 39% of all transcripts expressed) and backcross (2316, 43%). The number of false positive transgressive transcripts expected based on a randomized dataset was 1605 (33%) and thus, there were more transgressive transcripts than expected by chance, yet this was non-significant in F1 (*p*-value = 0.16, based on 1000 permutations as described in Material and Methods) and backcross hybrids (*p*-value = 0.09) respectively. Eighteen and three transcripts showed highly significant transgressivity in F1 and backcross hybrids, respectively. These transcripts, not differentially expressed in any comparisons, belong to muscle contraction, energy metabolism, lipid metabolism, protein degradation and transport functional categories (table 2.4).

2.8 Discussion

The main objective of this study was to document patterns of gene expression divergence in first (*normal* X *dwarf*) and second generation hybrid crosses [backcross: (*normal* X *dwarf*) X *normal*], and compare them to pure normal and dwarf parental forms, at both embryonic and juvenile ontogenetic stages. More specifically, under the assumption that similar genes involved in the adaptive divergence of these species are also responsible for driving their reproductive isolation (Nolte et al. 2009), we predicted more evidences of misexpression at the juvenile compared with the embryonic stage. In general, our results supported this prediction as we observed that few genes differed in average expression in hybrids compared to parentals at the embryonic stage, while many more did so at the juvenile stage. Secondly, we predicted more evidence of hybrid misexpression in backcross hybrids and the fact that non-additivity was more prevalent in backcross compared to F1-hybrids supported this prediction. Lastly, extreme transgressivity of several key developmental genes was observed in backcross embryos. This emphasizes that, at the transcriptomic level, intrinsic hybrid misexpression may also play a role in explaining reproductive isolation of dwarf and normal whitefish. Below, we discuss the potential implications of those results, also considering the limitations of the data.

Patterns of inheritance in hybrids

Strikingly, patterns of inheritance were quite distinct between backcross and F1-hybrids, and d/a values were not correlated between them. Under an additive model of inheritance, we would also have expected F1-hybrid to be the mid-value of their parents and backcross to be closer to the normal phenotype (with which they share 75% of their genome). This was not the case, as more genes differentiated F1-hybrids to dwarf while, and to a lesser extent, more genes differentiated backcross to normal, a result also exemplified by the asymmetry in the direction of dominance in both F1-hybrids (normal dominance) and backcross (dwarf dominance). This idiosyncratic result cannot be ignored, yet is difficult to interpret beyond the fact that, as it has been emphasized many times before, gene expression is a complex phenotype, whose behavior is hard to predict and whose inheritance often does not follow simple Mendelian rules (Rockman and Kruglyak

2006). Indeed, this asymmetry of gene expression divergence towards one parent is apparently quite common in F1-hybrids event though the underlying mechanistic reasons responsible for this trend are poorly understood (*Drosophila*: Ranz et al. 2004; Gibson et al. 2004; *Mus*: Rottscheidt and Harr 2007; *Salvelinus*: Mavàrez et al. 2009).

The transcriptomic basis of ecological (extrinsic) reproductive isolation factors

Previous gene expression studies (Derome et al. 2006, St-Cyr et al. 2008; Nolte et al. 2009) combined with physiological data (Trudel et al. 2001) have shown that changes in the expression of metabolic genes are largely responsible for the physiological adaptation to distinct whitefish benthic (normal) and limnetic (dwarf) niches. Notably, a suite of six key metabolic genes (glyceraldehyde-3-phosphate dehydrogenase, Fructose-bisphosphate aldolase A, Beta-enolase, Trypsin-1 precursor, Cytochrome c oxidase polypeptide VIa and Nucleoside diphosphate kinase) was identified as consistently divergent between normal and dwarf whitefish (Nolte et al. 2009). We may then hypothesize that misexpression for those metabolic genes could contribute, to an atypical physiological phenotype and to an inferior, ecologically maladapted, individual. Here, we found that, in F1-hybrids juveniles, those genes mostly showed an intermediate pattern of expression (suppl. table 2.1) and no transgressivity compared to parents. In backcross hybrids, two of those genes (G3PDH, FBPA A) showed additivity of expression, the rest being slightly non-additive, while none revealed transgressivity.

According to the ecological theory of adaptive radiation, intermediate hybrid phenotypes may be selected against if no suitable ecological niche for them exists in nature (Schluter 2000). Recent work in sticklebacks (Gow et al. 2007) and cichlids (van der Sluijs et al. 2008) has shown such environment driven natural selection may be key in explaining incipient population divergence. As such, reproductive isolation of lake whitefish could be at least partly seen as a by-product of divergent selection acting on metabolic genes. Admittedly, a clear demonstration of the association between the expression of such genes and phenotypic variation between dwarf and normal whitefish is lacking and we are currently conducting quantitative trait loci (QTL), expression QTL and gene mapping studies towards this end (Renaut S, Nolte AW, Bernatchez L, unpublished data).

A possible role of transgressivity in reproductive isolation

Gene expression in hybrids was generally more variable than parental, both at embryonic and juvenile stages. One might argue that the patterns of variance observed are confounded by a different number of families used in each treatment. While this cannot be entirely ruled out, it is noteworthy that backcross individuals, who showed the highest level of variance in gene expression, consisted of a single female crossed to five males, relative to all other treatments, which consisted of many half-sib families. This family effect probably also explains the relatively large variance of dwarf whitefish, which comprised both half-sib families and many wild-caught natural families. It then seems unlikely that this factor would explain the general increased variance observed in the backcross group. Alternatively, recombination can release hidden variation and generate transgressive phenotypes (Rieseberg et al. 1999, 2003; Mallet 2007). Populations are known to accumulate cryptic variation only revealed under certain genotypic or environmental conditions (Le Rouzic and Carlborg 2007). Recently, Landry et al. (2007) have illustrated how in hybrids, the regulation of coevolved *cis* regulatory regions and *trans* transcription factors could be disrupted and lead to increased phenotypic novelties in hybrids. This may explain why many whitefish genes showed increased variance in expression and transgressivity in hybrids, and yet were not differentially expressed between parentals. In fact, all the highly transgressive transcripts presented in table 4 were not differentially expressed in any comparisons.

Transgressive segregation may in some cases create fitter phenotypes (for example, hybrid species resulting from selected "hopeful monsters", Barton 2001; Mallet 2007). Conversely, it also underlies post-zygotic isolation mechanisms such that transgressive hybrids often suffer a highly reduced survival (Barton 2001; Coyne and Orr 2004). We propose that the overall patterns of transgressivity we observed, including misexpression of several key developmental genes, may contribute to abnormal hybrid development and increased embryonic mortality identified by Lu and Bernatchez (1998) and also Rogers et al. (2006) as a plausible post-zygotic reproductive isolation mechanism. Namely, five of the nine transcripts identified as highly transgressive in embryos and involved in protein folding and mRNA translation are especially good candidates since knockdown mutants in *Danio rerio* for those genes are known to show visible embryonic defect and almost invariably die prior to, or early after, hatching (Amsterdam et al. 2004). This proportion (five out of nine or 54%) was also significantly higher (*p*-value < 0.001, one tailed Fisher's Exact test) than the actual proportion of transcripts in the whole embryo dataset that matched to essential genes identified in *D. rerio* (257 out of 4950 or 5%). In fact, the abnormal phenotypes described in the *D. rerio* study closely match our own observation that a large fraction of backcross eggs (35%) started to show visible defects (asymmetric axial body plan, small eyes, heart not beating, deformed tail) 15 days after our sampling and eventually die prior to hatching (Renaut S, unpublished data). A previous study on reproductive isolation in lake whitefish also showed that a large fraction of the backcross progeny died around the same developmental time (Rogers and Bernatchez 2006). Of course, there is a leap between linking a knockdown mutation completely obliterating a gene product (as it is the case in the *D. rerio* study) and a simple increase in biological variation. Yet, it is noteworthy that these key developmental genes are the most transgressive out of 4950 surveyed in the embryo dataset and were generally significantly more transgressive than expected by chance. Moreover, it is plausible that hybrid genetic combinations creating even greater misexpression for those genes may have caused early lethality (prior to our sampling) and thus may have reduced our ability to pick out such abnormal phenotypes. Consequently, the increased patterns of variance observed in hybrid embryos are likely to be conservative estimates.

Gene expression studies of speciation

Most gene expression differentiation we observed was between normal and dwarf parental forms rather than between hybrids. These results contrast with many recent gene expression-speciation studies that have identified pervasive non-additive patterns of gene expression in first generation hybrids (see recent reviews by Landry et al. 2007, Ortiz-Barrientos et al. 2007). Namely, our study brings new lights into gene expression studies of speciation for two main reasons. Firstly, the bulk of the work, done mostly in *Drosophila*, particularly in the *melanogaster* group (Michalak and Noor 2003; Ranz et al. 2004; Landry et al. 2005; Haerty and Singh 2006; Moehring et al. 2007), and to a lesser extent in *Xenopus* (Malone et al. 2007) and *Mus* (Rottscheidt and Harr 2007) involves biological species that have diverged millions of years ago [e.g. *D. simulans-D. mauritiana*: 0.93 MYA, *D. simulans-D. melanogaster*: 5.1 MYA (Tamura et al. 2004), *X. laevis-X. muelleri:* >20 MYA (Evans et al. 2004), *M. musculus* subspecies: 0.3-1.0 MYA (Boursot et al. 1996)]. Since genetic incompatibilities continue to accumulate over time even after complete reproductive isolation has been established, this complicates the identification of the loci that initially led to the divergence event (Mallet 2006). In these studies, most hybrids are known to be poorly fit, sterile, or simply inviable (Ranz et al. 2004) rendering it difficult to disentangle whether gene expression misexpression is the cause rather than the consequence of hybrid inviability. In contrast, in young diverging lineages such as whitefish species pairs that diverged 12 000-15 000 years ago (Bernatchez 2004), much less genetic divergence is expected due to frequent gene flow or recent common ancestry. The effects of hybridization may then be subtler and the genetic changes identified more likely to be involved in the very early steps of reproductive isolation. Secondly, even though the effect of early reproductive barriers may be more important in later hybrid generations (Barton 2001), most recent studies have focused on first generation hybrids. Clearly, we identified different genes and patterns of inheritance in first and second-generation hybrids. Moreover, backcross hybrids have also revealed increased non-additivity as well as transgressive expression of essential developmental genes. To our knowledge, no gene expression studies of speciation in natural systems have previously used genome-wide gene expression data to compare the expression profile of second-generation hybrids to that of parental lineages.

2.9 Conclusion

We have attempted to determine the role of gene expression divergence in the development of reproductive isolation of recently evolved lineages of lake whitefish. We demonstrated that F1 and backcross hybrids showed different patterns of expression and that gene misexpression at both embryonic and juvenile stages might take different forms (intermediacy, non-additivity and transgressivity of expression). We then identified candidate genes whose role in driving reproductive isolation will have to be further confirmed. As pointed out in a recent review, the identification of candidate genes does not constitute an end in itself, but rather the beginning of a new set of evolutionary relevant questions (Stinchcombe and Hoekstra 2007). Ultimately, by combining information obtained through gene expression studies such as this one, to QTL mapping, gene sequencing, mapping and genome scan data, we will be able to better answer questions regarding the underlying genetic architecture of adaptive traits and expression phenotypes, the role of standing genetic variation and *de novo* mutations in driving the emergence of those traits, and the role of natural selection and/or drift in maintaining this divergence.

2.10 Acknowledgments

We would like to thank the Ministère des Ressources Naturelles et de la Faune du Québec (MRNF) for issuing scientific collecting permits; J.-C. Therrien and S. Higgins (LARSA) for their assistance in spawning and rearing whitefish; J. Mavárez, E. Normandeau and C. Roberge for many insightful suggestions and help with the analysis. This project was funded by a Natural Science and Engineering Research Council of Canada (NSERC) postgraduate scholarship to SR, NSERC Discovery grant and Canadian Research Chair to LB, and a postdoctoral research stipend from the German Research Foundation to AN. This paper is a contribution to the research program of Québec Océan.

Chapitre 3 : Transcriptome-wide signature of hybrid breakdown associated with intrinsic reproductive isolation in lake whitefish species pairs (*Coregonus* spp. Salmonidae).

3.1 Résumé

L'étude de la génétique de la spéciation a porté presque exclusivement sur des analyses rétrospectives de l'isolement reproducteur entre des espèces hautement divergentes. Cependant, une compréhension du processus de spéciation dans son ensemble devra englober une analyse des conséquences de la divergence génomique dans des jeunes lignées évolutives encore capables d'échanger des gènes en conditions naturelles. L'accumulation de variation génétique conditionnellement neutre peut conduire à l'évolution de réseaux de gènes divergents. Dans un contexte hybride, l'équilibre épistatique entre ces mutations peut être perturbé et ainsi entraîner l'apparition de traits désavantageux, y compris la dérégulation de l'expression des gènes. Ici, nous avons documenté la divergence d'expression de gènes entres des jeunes espèces de grand corégones normaux, nains et leurs hybrides au début du développement embryonnaire. Une proportion importante (33%) des hybrides issus de rétrocroisements a montré des anomalies du développement, non observées chez les formes parentales, qui menaient ultimement à la mort des larves. Alors que le transcriptome des formes parentales était pour ainsi dire identique, suggérant un rôle de la sélection stabilisatrice, tous les hybrides ont montré une forte divergence d'expression. En comparant les hybrides survivants aux hybrides moribonds, nous avons observé plus de 2000 gènes dérégulés. En particulier, la dérégulation était significativement biaisée pour les gènes essentiels du développement. De plus, les gènes précédemment documentés comme étant fortement transgressifs (variance inter-individuelle exagérée) furent presque toujours sous-exprimés chez les hybrides. Nos résultats montrent de manière convaincante une dérégulation totale du transcriptome et un lien explicite entre la dérégulation des gènes essentiels du développement et l'isolement reproducteur post-zygotique.

3.2 Abstract

Genetic analyses of speciation have focused nearly exclusively on retrospective analyses of reproductive isolation between highly divergent species. Yet, a full understanding of the speciation process must encompass analysis of the consequences of genomic divergence in young lineages still capable of exchanging genes under natural conditions. The accumulation of conditionally neutral genetic variation may lead to the evolution of divergent gene networks. In a hybrid background, such mutations may no longer compensate one another, resulting in the appearance of selectively disadvantageous traits, including disruption of gene expression regulation. Here, we documented genome-wide patterns of gene expression divergence between young lineages of normal and dwarf lake whitefish and their backcross hybrids for which strong, yet incomplete post-zygotic isolation barriers exist. A significant proportion (33%) of backcross hybrids showed developmental abnormalities not seen in parental forms and eventually leading to death. While the transcriptome of parental forms was nearly identical during embryonic development, suggesting a role for stabilizing selection, all hybrids displayed strongly divergent patterns of gene expression. By comparing healthy, surviving hybrids against moribund ones showing abnormal development, we observed that over 2000 genes were misregulated in these abnormal embryos. In particular, misregulation was significantly biased towards essential developmental genes which were severely underexpressed. Furthermore, genes previously documented to be highly transgressive (exaggerated inter-individual variance) were almost invariably underexpressed in hybrids. Our results thus clearly showed a transcriptome-wide signature of hybrid breakdown in young, incipient species and demonstrated a persuasive link between misexpression of essential developmental genes and post zygotic isolation.

3.3 Introduction

Despite ongoing efforts, the nature of the genomic changes underlying reproductive isolation and ultimately leading to speciation remain elusive (Coyne and Orr 2004; Schluter 2009; Presgraves 2010). Based on our current comprehension of post-zygotic isolation, hybrid sterility and/or inviability arises as a consequence of incompatible allele combinations in hybrids (Dobzhansky-Muller incompatibilities) that have diverged between pure species or populations (Dobzhansky 1937; Muller 1942). These alleles, whether neutral or positively selected within their own genetic background interact negatively when brought together in inter-specific hybrid genomes. The break up of co-adapted gene complexes will therefore generate mosaic chromosomes composed of the two diverging genomes and is expected to be more deleterious in later generation hybrids compared to F1 generations (i.e. hybrid breakdown; Rieseberg et al. 1999; Coyne and Orr 2004; Burton et al. 2006). It is believed that the most significant point in the speciation process is the initial development of reproductive isolation (Dettman et al. 2007; Via 2009). By studying diverging lineages which are only partially reproductively isolated, early barriers to gene flow and their underlying genetic basis can be identified before they become confounded with other differences that accumulate after speciation. Yet, studies of the effect of genomic incompatibilities on gene regulation have focused nearly exclusively on retrospective analyses of reproductive isolation between highly divergent species where gene flow does not occur in natural conditions (Michalak and Noor 2003; Ranz et al 2004; Noor and Feder 2006; Rottscheidt and Harr 2007; Landry et al. 2007).

At the gene transcription level, genomic incompatibilities can lead to the disruption of the transcriptional machinery and the appearance of novel, unforeseen patterns of gene expression in hybrid lineages (Landry et al. 2007). This is often hypothesized to be a major factor underlying hybrid breakdown. Analyses of gene expression profiles using microarrays have shown that the combination of divergent regulatory elements within a common genetic background often results in the disruption of gene expression (reviewed by Ortiz-Barrientos et al. 2007; Landry et al. 2007). For example, the global expression profile of *Drosophila melanogaster* and *D. simulans* are more closely related to each other than to their hybrid progeny (Ranz et al. 2004). Similarly, in sympatric but ecologically divergent populations of brook charr *(Salvelinus fontinalis)*, dramatic breakdown of gene expression patterns in hybrids compared with their parental relatives was observed (Mavarez et al. 2009).

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; Rogers and Bernatchez 2007; Bernatchez et al. 2010). Following the last glacier retreat (less than 15 000 years ago), secondary contact of lake whitefish (*Coregonus clupeaformis*) evolutionary lineages isolated during the Pleistocene (100 000-200 000 years ago) has led to the parallel evolution of two morphologically and ecologically divergent sympatric species in several lakes of northeastern North America: benthic normal and limnetic dwarf whitefish (Bernatchez 2004). Furthermore, the ecological specificity of each lake seems to be the main driving isolation factor since, in certain lakes, secondary contact of these same lineages resulted in a hybrid swarm (Lu et al. 2001; Landry et al. 2007). As expected from a recent divergence event, the overall level of genetic differentiation between species pairs is relatively low and hybrids are still present at low frequency in natural populations (Falush et al. 2007). Moreover, early in ontogeny, normal and dwarf fish are phenotypically and ecologically indistinguishable (Chouinard and Bernatchez 1998; Bernatchez 2004). At the same time, hybrid whitefish experience striking fitness consequences of genomic incompatibilities and ongoing reproductive isolation (increased mortality, disruption of hatching time and reduced sperm performance, Rogers and Bernatchez 2006; Whiteley et al. 2009; Bernatchez et al. 2010)

Based on this knowledge, we followed the ontogeny of pure dwarf and normal whitefish as well as their hybrids to identify the precise moment at which embryos started to show early signs of hybrid breakdown. By quantifying genome-wide levels of gene expression when developmental defects were the most extreme, we could identify transcriptome changes potentially associated with post zygotic isolation. Most specifically, we tracked patterns of regulation for genes known to be essential for early fish development (Amsterdam et al. 2004). We predicted that these essential genes would be the most affected in hybrids since their expression is especially critical and has a severe cascading effect on embryonic development and fitness in general.

3.4 Methods

Strains, crosses and fish maintenance

Eggs were obtained from outbred laboratory strains [normal whitefish originating from Aylmer Lake (45° 54'N, 71° 20'W), dwarf whitefish originating from Témiscouata Lake (47° 41'N, 68° 47'W), as detailed in Lu and Bernatchez 1998] at the Laboratoire de Recherche en Sciences Aquatiques (LARSA, Université Laval, Quebec, Canada). We created half-sib families as follows. The dwarf group was created using one lab strain dwarf female crossed to five different dwarf males. The normal whitefish group was created by crossing one lab strain normal female to three separates normal males. Finally, an independent group of backcross individuals was obtained using an F1-hybrid female generated in the laboratory in a previous study (Rogers et al. 2006) crossed to five normal males. As such, the backcross individuals are composed of a 75% normal and 25% dwarf genetic background. Fish were anesthetized with 0.001% eugenol solution whereupon eggs and semen were stripped and immediately fertilized *in vitro*. Fertilized eggs were disinfected (0.0001 % iodine solution) and incubated on submerged grids in the same flow-through system (4.5 - 5.5°C). To avoid contamination by fungi, dead eggs were removed on a daily basis and all samples were treated weekly with malachite green oxalate.

Sampling

We sampled embryos approximately 60 days post fertilization (between 280-288 degree-days) in normal, dwarf and backcross families. Individually chosen eggs were preserved in RNA later (Ambion, Austin, TX) and frozen at –20 °C for storage. Embryos sampled in the backcross family were separated into normally healthy looking ones, hereafter referred as *backcross-healthy*, and ones that showed developmental problems, hereafter referred as *backcross-moribund* (see results for the description of the developmental phenotype of each group). To avoid sampling dead eggs where RNA degradation would affect gene expression measurements, 40 independent moribund embryos were observed daily to confirm that all were still alive at least several days following sampling (yolk sac of dead embryos changes from translucent to opaque). At the same time, these 40 moribund embryos as well as nearly all eggs classified as backcross-moribund died in the following weeks, prior to or just after hatching (at ~100 days).

Experimental design

Gene expression analysis was performed using the 16,000 (v2.0) Salmon cDNA microarray provided by the cGRASP consortium (von Schalburg et al. 2005). More than 175 cDNA libraries constructed from a wide variety of tissues and different developmental stages (Atlantic salmon and rainbow trout) were used to produce the array. Furthermore, the validity of this array has been amply demonstrated in lake whitefish, whereas hybridization signal is the same as for Atlantic salmon (von Schalburg et al. 2005). Sequence divergence between normal and dwarf is also very low and therefore did not affect hybridization kinetics (Derome et al. 2006, Whiteley et al. 2008). Samples of dwarf, normal, backcross-healthy and backcross-moribund were hybridized in a loop design, involving eight biological replicates for the backcross-healthy and backcross-moribund comparison and six for the others. Dye swap was performed between each replicate. As a result, we obtained a final set of 32 microarray slides (fig. 3.1).

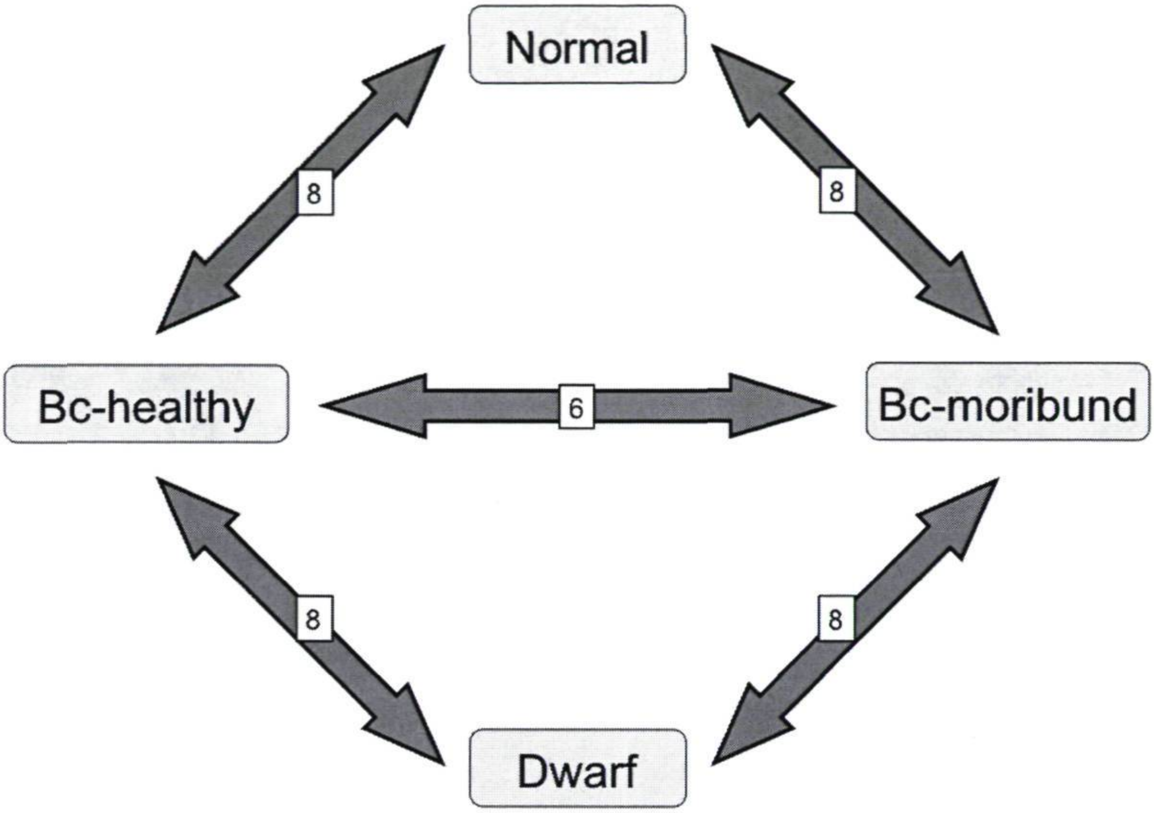

Figure 3.1 Microarray design with the four experimental groups. Each double-headed arrow represents one microarray slide. Numbers written on arrows refer to the number of comparisons done between each group for a total of 32 comparison and 28 separate biological samples.

Total RNA was extracted using the Trizol Reagent protocol (Invitrogen, Carlsbad, CA). A pool of five whole embryos preserved in RNA later was homogenized using a bead mill (Qiagen, Germantown, MD) for each sample hybridized on the array. RNA pooling is a common practice when quantity of material is limiting and inference on gene expression patterns for most genes is not affected by this (Kendziorski et al. 2005). Crude total RNA was treated with DNase I for 15 min at room temperature (1 unit/ug, Invitrogen) and further cleaned by ultra filtration using microcon YM30 spin columns (Millipore, Billerica, MA). Total RNA was eluted in pure water supplemented with Superase-In™ RNase Inhibitor

(Ambion), quantified and quality checked using Experion™ RNA StdSens Analysis Kit (Bio-Rad, Hercules, CA), then stored at -80°C. Reverse transcription reactions were performed using Superscript II Kit (Invitrogen). Following this, Genisphere (Hatfield, PA) 50Array Expression Array Detection Kit™ (Cy3/cy5) was used following the vendor's protocol (~20ug of cDNA material hybridized). Microarrays were scanned using a ScanArray™ Express scanner (Packard Bioscience, Wellesley, MA) and spots were located and quantified using the histogram method in QuantArray 3.0 (Perkin Elmer, Waltham, USA). Gene expression data were deposited at Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo, Series accession GSE23095).

Data analysis

Unless stated otherwise, all analyses were performed in R (v.2.9.0. The R Foundation for Statistical Computing®. 2009. 3-900051-07-0). First, local background was subtracted from spots. Only spots above the mean of empty (blank) spots plus two standard deviations, in at least one channel, were kept for further analysis. Missing data were imputed using the K-nearest neighbors (20 neighbors) and data were $\log_2$ transformed and normalized using Lowess algorithm. Data were fitted into a mixed model ANOVA (R/maanova package v1.14, Kerr et al. 2000) to identify differential expression. The mixed model included the following terms as fixed sources of variance: Group (dwarf, normal, backcross-healthy and backcross-moribund) and Dye (two fluorescent dyes). Sample (28 biological samples) and Array (32 individual microarrays) were included as random sources of variance. Finally, a surrogate variable analysis (Leek and Storey 2007) enabled the identification of an eigengene explaining 68% of the residual variance and corresponding to a strong block effect (eight slides of the backcross-healthy / backcross moribund comparisons done in August 2007, the rest of the slides done in June 2008). As a result, this block effect was also included in the mixed model as a random term. A permutation based F test was performed to test for statistically significant divergence in gene expression and restricted maximum likelihood was used to solve the mixed model equations (p-value < 0.05, 1000 permutations, Fs test). Following this, p-values were corrected for multiple hypotheses testing using q-value correction (q-value < 0.05, Storey

2002). To test for specific pairwise differences between groups (N, D, BC-healthy, BC-moribund), permutation based *t*-tests between the four groups for all genes identified as differentially expressed in the ANOVA (*q*-value < 0.05, 1000 permutations, Fs test) were performed. Finally, fold changes for the group effect ($\log_2$) were also extracted from the analysis for all significant transcripts. In order to remain conservative in interpreting the number of significant features, transcripts with less than 1% sequence divergence over 95 % of the sequence printed on the array were also compressed into a single gene.

Best linear unbiased estimates (BLUE) of the dye, array and block effects (i.e. technical variation) were subtracted from the normalized fitted values of gene expression data for each gene. Following this, each gene expression value was divided by the mean of its channel such that values should only represent the group and sample effects (i.e. biological variation). Lastly, since the same biological sample is sometimes represented on several arrays, we calculated mean gene expression for each of the 28 different samples used in the study. These values were then used to perform a hierarchical clustering analysis to construct a heatmap, with pairwise gene and sample distance matrices estimated from Pearson correlation coefficients.

Functional classification and assessment of significant differential representation of functional classes were performed using DAVID bioinformatics resources (v6.7, Huang et al. 2009). First, Unigen clusters were obtained from cGRASP for all annotated transcripts printed on the microarray. Then, overrepresented gene ontology classes (molecular function or biological process) amongst differentially expressed genes were identified [modified Fisher's test (*Ease score*) corrected for multiple hypothesis testing (*Benjamini correction*) < 0.05]. Lastly, previously published data (Amsterdam et al. 2004) were used to identify patterns of gene expression for essential embryonic development genes in fish. Briefly, Amsterdam and colleagues (2004) identified a suite of 315 genes in early embryonic development in which knockdown mutations are lethal in zebrafish (*Danio rerio*). Furthermore, their study also confirmed that the propensity of being an essential gene is highly conserved throughout evolution. We then used BLASTn algorithm (Altschul et al. 1997) to match these essential genes to our own sequences printed on the array (*e*-value < 1e-15).

3.5 Results

Development

Approximately 60 days post fertilization (between 280-288 degree-days) embryos were sampled in normal, dwarf and backcross families. At this stage, embryos have developed into the pharyngula period, which corresponds to the phylotypic stage (cf. *Danio rerio* developmental staging in Kimmel et al. 1995, Slack et al. 1993). Nearly all normal and dwarf embryos showed the same developmental phenotype, and only six embryos out of 104 (5.8%) assessed for developmental characteristics were abnormally developing in these pure crosses. It is noteworthy that these six embryos had major developmental defects that were substantially different (no axial body plan, "cyclops") from the backcross moribund group described below. Normally developing pure and backcross embryos were phenotypically similar and characterized by: differentiated pigment cells, a beating heart pumping blood throughout the circulatory system, eyes pigmented and a fully mobile tail, detached from the yolk sac (Fig. 3.2a-b). In the backcross family, a significant proportion of embryos [33%, 52 out of 156 embryos visually inspected (Fisher's exact test, *p*-value<0.0001)] showed either evident lag in development or atypical phenotype (small head, small eyes, deformed eye lens, heart not beating, reduced cell pigmentation, deformities of the tail) and where hereafter referred to as backcross-moribund embryos (Fig. 3.2c-d). Thus, while not all backcross moribund possessed exactly the same phenotype, they undoubtedly represented a discrete phenotype different from all other pure and healthy backcross embryos.

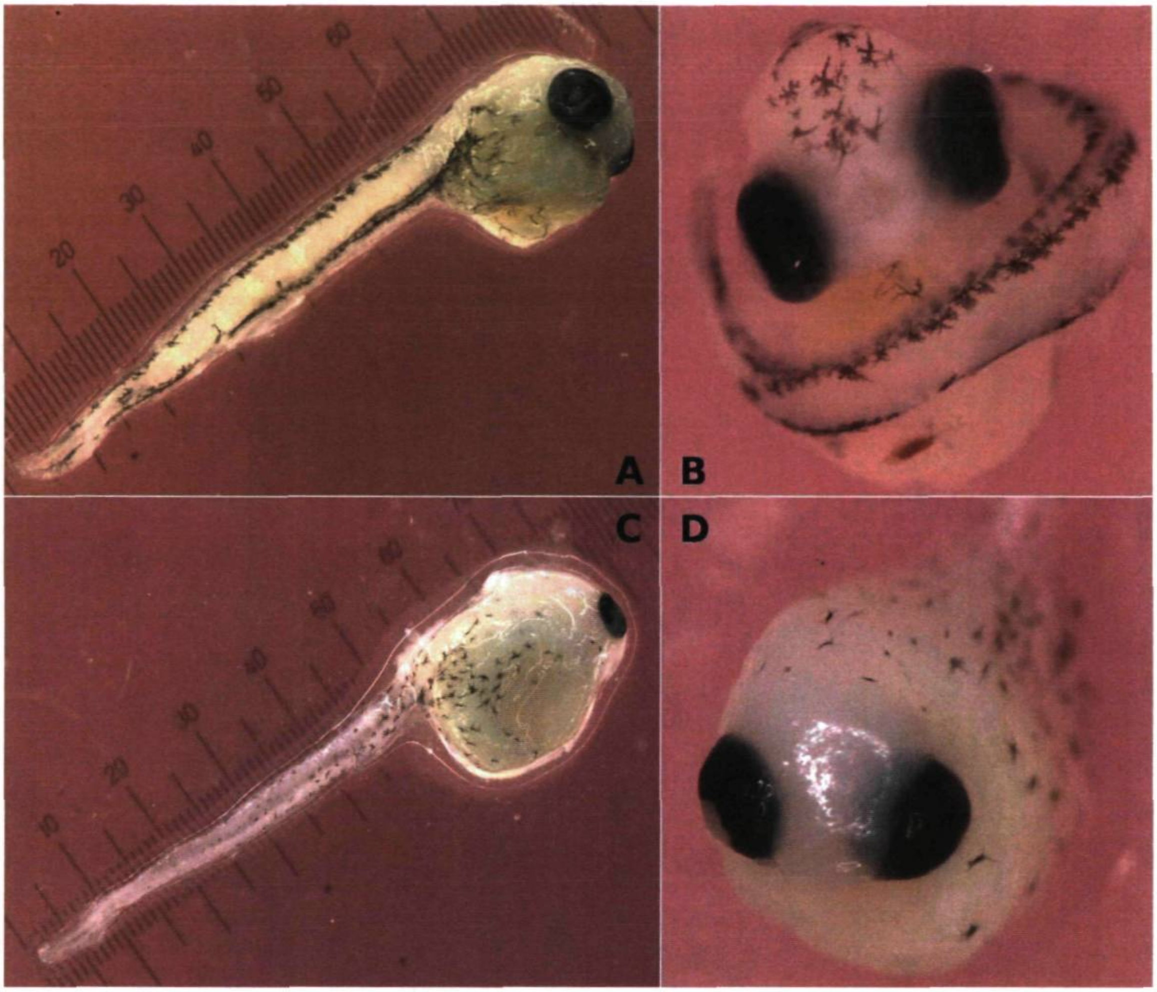

Figure 3.2 A-B: General developmental characteristics of normal, dwarf and backcross-healthy embryos (differentiated pigment cells, beating heart pumping blood throughout the circulatory system, eyes pigmented, eye lens completely formed and a fully mobile tail, detached from the yolk sac). **C-D**: General phenotype of backcross-moribund embryos (small head, small eyes, deformed eye lens, reduced cell pigmentation, deformities of the tail). Note that egg chorion was removed (**A-D**) and embryos were manually unfolded with dissecting needles in **A** and **C**. Units on scale are 0.1mm.

Gene expression regulation

Very little difference in gene expression was found between normal and dwarf whitefish, with only two transcripts differentially expressed (60s ribosomal protein L23a, Hemoglobin subunit beta) after correction for multiple hypotheses testing and replicate spotting on the array, 0.1 % of all transcripts expressed (FDR *q*-value < 0.05). This was in stark contrast with backcross-healthy hybrids for which 162 and 77 (3.0 and 1.5 %) genes were differentially expressed compared to normal and dwarf, respectively. Even more strikingly, 2214 (39%) transcripts were differentially expressed between backcross-healthy and backcross-moribund, 1993 (35%) between normal and backcross-moribund and 1964 (35%) between dwarf and backcross-moribund (Fig. 3.3, suppl. table 3.1).

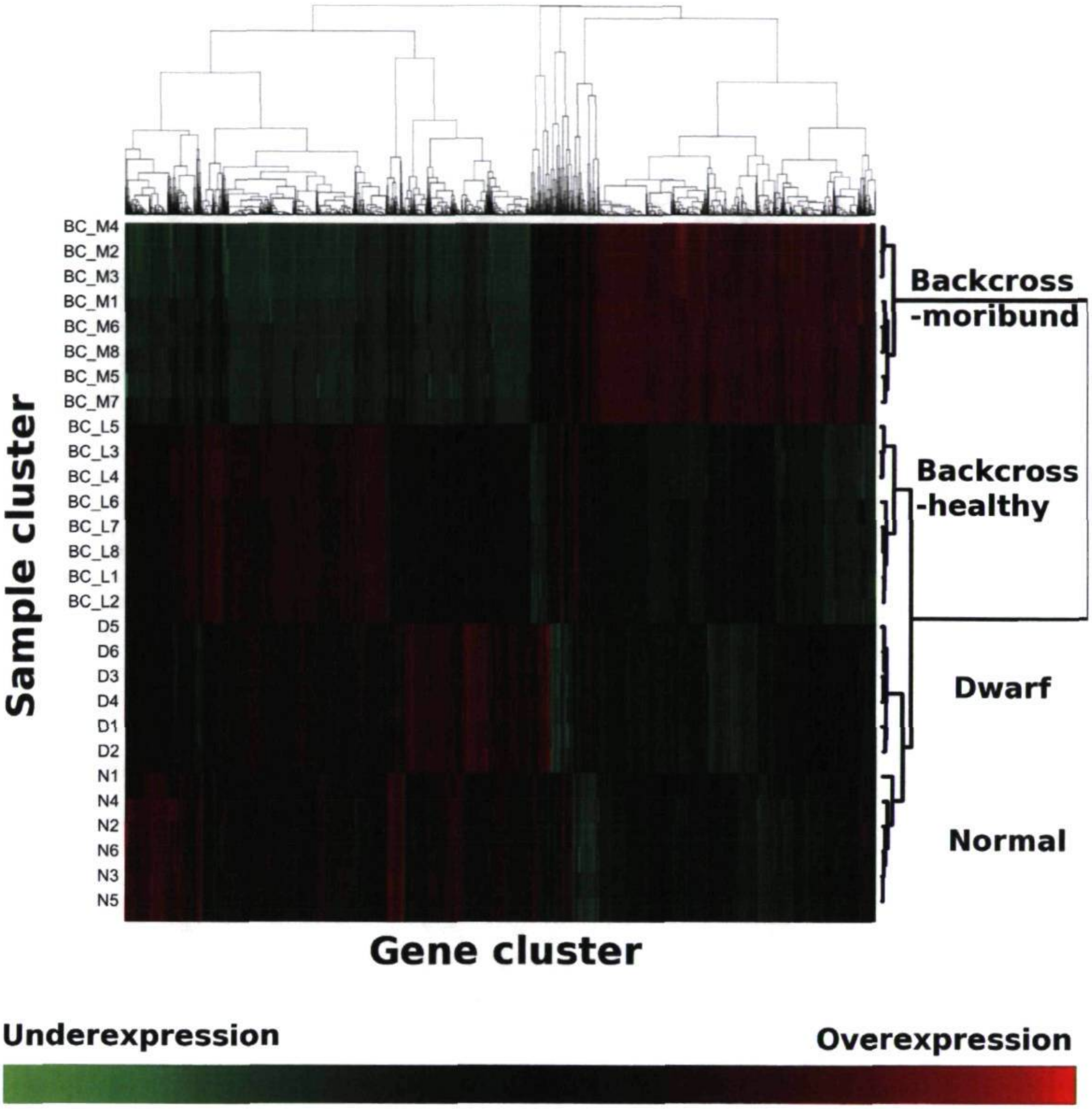

Figure 3.3 Heatmap representing hierarchical clustering based on gene [all genes differentially expressed between groups as identified by ANOVA (q-value < 0.05)] distance matrix and sample distance matrix (28 biological samples).

Functional classification

Functional analyses using DAVID bioinformatic resources (v6.7) revealed that several biological processes and molecular functions were overrepresented amongst the lists of differentially expressed genes. Firstly, translation and generation of precursor metabolites and energy gene ontology (GO) terms were significantly overrepresented in all comparisons involving backcross-moribund individuals (Table 3.1). Comparatively, glycolysis, alcohol and carbohydrate catabolic processes were overrepresented in the list of genes differentially expressed between backcross-healthy and dwarf, while no GO terms were overrepresented in the backcross-healthy/normal or normal/dwarf categories. Transcripts differentially expressed in backcross-moribund were further divided into two categories: transcripts underexpressed and overexpressed compared to the average of all groups. In overexpressed genes, a total of 15 different GO terms, which were further grouped into seven general categories, were overrepresented. These were mostly related to transport and energy metabolism functions. In underexpressed genes, 11 different GO terms, which were further grouped into six general categories, were overrepresented. These were related to completely distinct functional categories; mostly macromolecule metabolism and regulation of mRNA translation.

Table 3.1 Significant over-representation of gene ontology (GO) categories (biological processes: BP and molecular function: MF) among genes which showed different transcription levels between all genes expressed on the microarray and the different experimental groups.

Comparisons	Category	Description[a]	Fold[b]	count (%)[c]	*p*-value[d]	Number of GO terms[a]
Backcross-moribund / Backcross-healthy	BP	Generation of precursor metabolites and energy	1.74	67 (10.2%)	0.000	1
	BP	Translational elongation	1.91	28 (4.2%)	0.066	1
	MF	Structural molecule activity	1.49	65 (9.8%)	0.022	1
	MF	Hydrogen ion transporter activity	1.86	28 (4.2%)	0.047	1
Backcross-moribund / dwarf	BP	Generation of precursor metabolites and energy	1.72	60 (10.1%)	0.000	1
	BP	Translation	1.84	70 (11.7%)	0.022	2
	MF	Structural molecule activity	1.58	62 (10.4%)	0.006	1
Backcross-moribund / normal	BP	Generation of precursor metabolites and energy	1.78	63 (10.3%)	0.000	1
	BP	Translational elongation	2.07	28 (4.6%)	0.010	1
	MF	Cytoskeletal protein binding	1.75	38 (6.2%)	0.021	1
Backcross-healthy / dwarf	BP	Glycolysis	21.08	6 (22.2%)	0.001	1
	BP	Alcohol catabolic process	15.17	6 (22.2%)	0.009	1
	BP	Carbohydrate catabolic process	14.59	6 (22.2%)	0.006	4

Backcross-healthy / normal		-------				
Dwarf / normal		-------				
All genes differentially expressed / Backcross-moribund overexpressed genes	BP	Generation of precursor metabolites and energy	2.09	62 (18.5%)	0.00	1
	BP	Oxidation reduction	1.76	50 (14.9%)	0.00	1
	BP	Electron transport chain	2.12	30 (8.9%)	0.00	1
	BP	Phosphorylation	1.82	29 (8.6%)	0.05	1
	MF	Catalytic activity	1.24	148 (44%)	0.00	
	MF	Magnesium ion binding	2.05	21 (6.3%)	0.01	
	MF	Ion transporter activity	1.80	28 (8.2%)	0.02	9
All genes differentially expressed / Backcross-moribund underexpressed genes /	BP	Translation	1.41	69 (15.4%)	0.001	1
	BP	Gene expression	1.29	136 (30.3%)	0.000	1
	BP	Macromolecule metabolic process	1.26	182 (40.4%)	0.00	4
	BP	RNA metabolic process	1.45	48 (10.7%)	0.033	2
	MF	Structural constituent of ribosome	1.48	8.9 (40%)	0.029	1
	MF	Nucleic acid binding	1.32	86 (19.2%)	0.016	2
All genes differentially expressed / Essential genes differentially expressed	BP	Translation	3.91	22 (48.9%)	0.000	1
	BP	gene expression	2.64	32 (71.1%)	0.000	1
	BP	Macromolecule biosynthetic process	2.08	30 (66.7%)	0.002	6
	BP	Protein metabolic process	1.97	32 (71.9%)	0.001	3
	MF	Structural constituent of ribosome	5.20	16 (35.6%)	0.000	1

	MF	Structural molecule activity	3.25	16 (35.6%)	0.002	1

Note. [a]Similar GO terms were grouped together as one category using the most inclusive GO term provided (e.g. *Macromolecule metabolic process* grouped within *Macromolecule biosynthetic process*). Number of original GO terms in each category was written in *Number of GO terms* column. [b]Fold enrichment values given by DAVID as compared to the background group (either genes expressed or all genes differentially expressed). [c] Number of unigenes in GO category and percentage of total. [d]Note that when more than one GO term was included in the same category, the *p*-value is the mean *p*-values of all the GO terms. * *p*-value [Ease score (Bonferroni correction implemented by Huang et al. 2009)] < 0.05, ** *p*-value < 0.01, *** *p*-value < 0.001.

Of particular interest were 450 (3.2 % of all transcripts printed on the array) transcripts, which matched (BLASTn *e*-value < 1e-15) to genes essential for early fish development (Fig. 3.4). Essential genes refer to those genes for which knockdown mutations are embryonic lethal in *Danio rerio* (Amsterdam et al. 2004). Among these essential transcripts, 336 were expressed and 204 differentially expressed. Both of these values were significantly higher compared to the proportion of all transcripts expressed and differentially expressed, respectively (Chi square test, *p*-value < 0.01). Moreover, 81% (166 transcripts) of these essential transcripts were underexpressed in backcross-moribund compared to the average of the four groups and this was highly significant compared to the proportion of all transcripts underexpressed in these embryos (56% or 1445) (Chi square test, *p*-value < 0.0001, Fig. 3.4). We also identified several GO terms that were overrepresented amongst the lists of essential genes differentially expressed against all genes differentially expressed. These were related to different functional categories but mostly, macromolecule metabolism and regulation of mRNA translation (table 3.1).

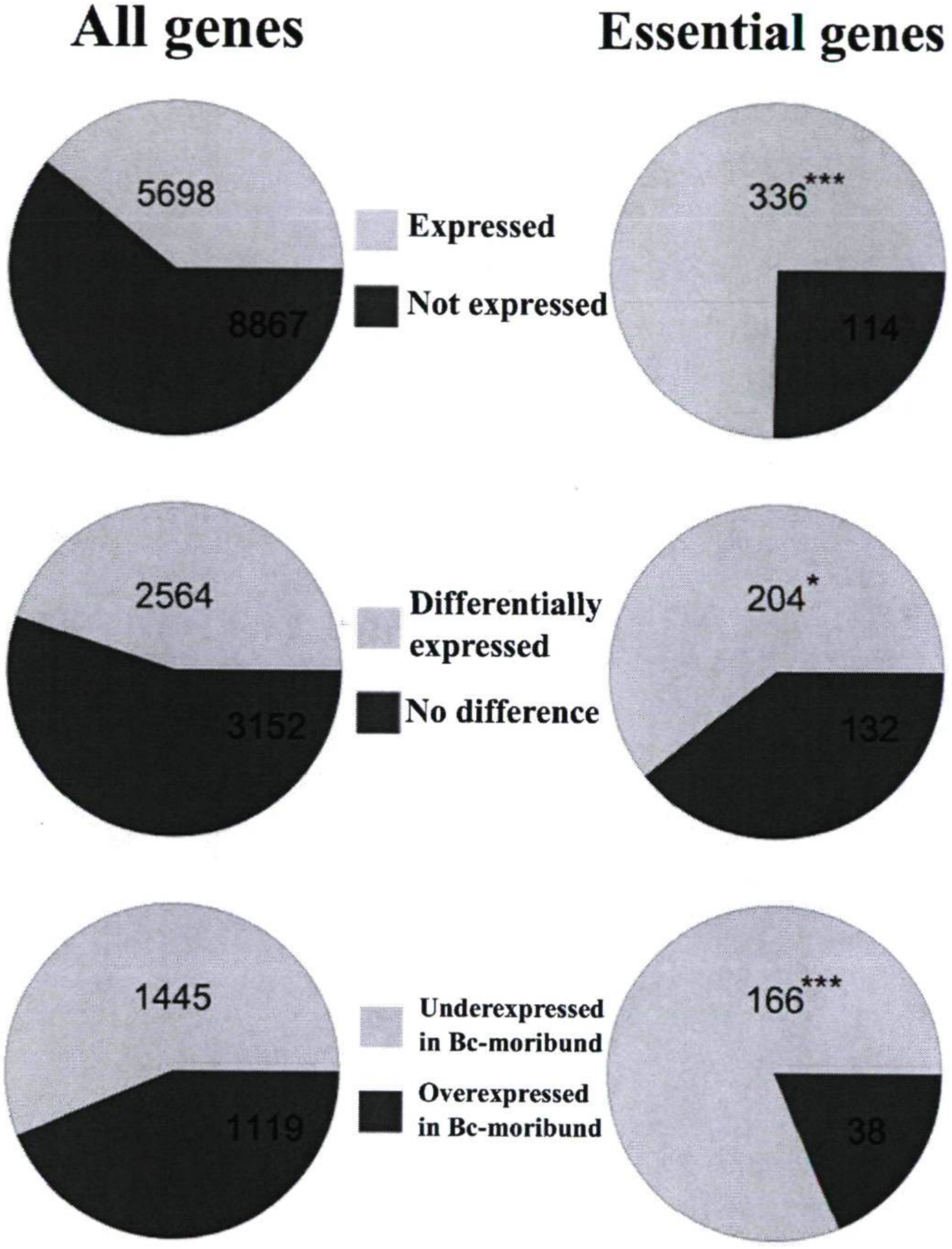

Figure 3.4 Number of transcripts matching to essential early developmental genes. Note that transcripts were defined as essential if they matched (BLASTn *e*-value < 1e-15) to the list of 315 genes identified as essential early developmental genes in zebrafish (Amsterdam et al. 2004) (Chi square test: * p-value < 0.05, *** p-value < 0.001)

Finally, in a previous study looking at gene expression differences in 30 day old whitefish embryos, a suite of seven genes that showed highly transgressive patterns of expression in backcross hybrids and which also comprised three essential developmental genes was identified (Renaut et al. 2009). These genes did not differ in mean expression in 30 days old backcross embryos, but showed exaggerated inter-individual variance extending outside the range of both parents. In the present study, 18 transcripts matched to these seven transgressive genes and all except one, were underexpressed in both backcross groups compared to the parents. Moreover, this effect was even more pronounced for backcross-moribund (table 3.2).

Table 3.2 Relative gene expression for genes previously identified as highly transgressive in 30 days old backcross whitefish embryos (Renaut et al. 2009). Genes products in bold are also essential developmental genes in fish embryos according to Amsterdam et al. (2004). Note that relative gene expression ($\log_2$ values) is expressed relative to the normal group. When several transcripts match to the same gene product, relative gene expression was calculated as the mean of all transcripts matching to that gene product.

Description	Fold change (N)	Fold change (D)	Fold change (Bc-h)	Fold change (Bc-m)	*q*-value (Bc-m /Bc-h)	*q*-value (Bc-m /D)	*q*-value (Bc-m/N)	*q*-value (Bc-h/D)	*q*-value (Bc-h/N)	*q*-value (N/D)
Asialoglycoprotein receptor 2	0.00	0.03	0.22	0.21	n.s.	**	**	n.s.	*	n.s.
Protein kinase C	0.00	-0.23	-0.27	-0.23	n.s.	n.s.	**	n.s.	*	n.s.
Guanine nucleotide-binding protein	0.00	-0.29	-0.42	-0.39	n.s.	n.s.	***	n.s.	***	n.s.
40S ribosomal protein S11 (3 transcripts)	0.00	0.06	-0.03	-0.22	**	***	*	n.s.	n.s.	n.s.
Heat shock 70 kDa protein (10 transcripts)	0.00	-0.21	-0.30	-0.42	n.s.	n.s.	***	n.s.	n.s.	n.s.
Elongation factor 1 alpha	0.00	-0.36	-0.39	-0.56	**	**	***	n.s.	**	n.s.
Fish-egg lectin	0.00	-0.07	-0.27	-0.95	***	***	***	n.s.	n.s.	n.s.

(N: Normal, D: Dwarf, Bc-h: Backcross-healthy, Bc-m: Backcross moribund. *q*-value for the respective t-tests. n.s.: non significant, * *q*-value < 0.05, * *q*-value < 0.01, *** *q*-value < 0.001).

3.6 Discussion

Our main objective was to quantify genome-wide levels of gene expression when developmental defects were the most extreme and thus potentially associate changes in expression to post zygotic isolation. Here, we discuss the potential implications and inherent limitations of these results in light of our understanding of post-zygotic isolation in incipient species of lake whitefish and of the genetics of speciation in general. A significant proportion of backcross embryos showed strong developmental defects. These developmental problems leading to a mortality rate of at least 33% clearly represent a severe fitness cost of producing hybrids between dwarf and normal whitefish in natural conditions, and therefore undeniably act as a strong post-zygotic isolation mechanism. Admittedly, since these observations are drawn from five half-sib backcross families derived from a single hybrid female crossed to five different normal males, they should be interpreted cautiously. However, the fertilization success of those backcross half-sib families eggs was similar to the one observed for pure crosses, as also previously reported (Lu and Bernatchez 1998, Rogers and Bernatchez 2006). In addition, the distinct abnormal embryonic development phenotype observed in backcross embryos was not observed in any of the pure families. Moreover, our observations indicating that strong post zygotic isolation barriers exist between normal and dwarf lake whitefish is corroborated by several lines of evidence previously documented. An elevated mortality rate around the same developmental time (Rogers and Bernatchez 2006) as well as a reduced sperm performance (Whiteley et al. 2009) has been identified in independent backcross families. Elevated, albeit lower than for the backcross, mortality was also observed in F1 hybrids at the same developmental time (Lu and Bernatchez 1998). Thus, increased mortality in hybrids has now been observed in three independent studies. Furthermore, strong segregation distortion for over 30 % of mapped genetic markers reflected differential survival rates among backcross hybrids genotypes (Rogers and Bernatchez 2007). For all these reasons, we propose that the developmental and expression differences observed here are direct consequences of hybridization between diverging lineages.

The specifics of hybrid breakdown on gene expression regulation

Using cDNA microarrays, we observed very little difference in gene expression between pure normal and dwarf whitefish. This lack of divergence between the transcriptome of the parental forms early in ontogeny is in line with the lack of differentiation in developmental time until emergence (Rogers and Bernatchez 2006) and also corroborates previous findings whereas 30 days old embryos had 14 times less genes (n = 5 genes) displaying significant regulatory divergence than 16 weeks old juvenile fish (Nolte et al. 2009, Renaut et al. 2009). In fact, dwarf and normal larval whitefish are nearly indistinguishable and their phenotypic and ecological divergence is expected to take place at the juvenile stage (Chouinard and Bernatchez 1998). Embryos sampled here were in the phylotypic developmental stage corresponding to the pharyngula period in *Danio rerio* development (Kimmel et al. 1995). During this highly conserved stage of development, individuals are expected be more similar to each other than during any earlier or later developmental periods (Slack 1993). This observation has frequently been made even for highly divergent taxa and should also apply for gene expression differences (Irie and Sehara-Fujisawa 2007).

In stark contrast with parental forms, hybrids showed developmental problems that had far-reaching repercussions on gene expression regulation. Backcross hybrids all had strong divergent patterns of gene expression compared to parental crosses and this was especially true for moribund hybrids. The nature of breakdown in gene regulation observed in whitefish backcross hybrids is more likely a consequence of genomic incompatibilities rather than the mere result of a stochastic RNA degradation process. Genes were affected in a precise way depending on their functionality. This is also confirmed by functional analysis, whereby markedly different gene ontology terms (whether biological process or molecular function) were overrepresented whether a gene was under or over expressed in the backcross-moribund group (table 3.1). Therefore, developmental problems seem to be associated with this breakdown in gene regulation in hybrids. Nevertheless, the challenge will remain to establish whether this association is through causality and this remains an unavoidable limit of all gene expression studies of speciation (Noor and Feder 2006).

As predicted, essential fish developmental genes were the most affected in hybrids. Their expression is especially critical and has a severe cascading effect on embryonic development and fitness in general. Theses genes, whose essentiality is highly conserved throughout evolution (Amsterdam et al. 2004), were not only more differentially expressed than expected but also strongly underexpressed in backcross compared to the average of the four groups (81% of those are underexpressed in backcross embryos, Fig. 3.4). The functional analysis of GO terms also revealed that similar categories were overrepresented in essential genes differentially expressed as in backcross-moribund underexpressed genes and these were mostly genes involved in macromolecule metabolism and regulation of mRNA translation (table 3.1).

Furthermore, the early signs of gene misexpression previously documented in 30 day old embryos (Nolte et al. 2009, Renaut et al. 2009) culminated in dramatic differences in development and gene expression regulation in the present study in 60 days old hybrid embryos, particularly so for genes known to be essential in early embryonic development (table 3.2). For example, in zebrafish, knockdown mutation of Heat Shock Cognate 70 leads after five days to: pericardial edema, a necrotic head and a general degeneration of the body; and knockdown of elongation factor 1 alpha leads to a small head and eyes, rounder yolk and increased necrosis (Amsterdam et al. 2004). In lake whitefish backcross hybrids, both of these genes were among the most transgressive of all at 30 days (Renaut et al. 2009), were severely underexpressed in the present study (at 60 days, table 3.2), and embryos showed similar general phenotype to zebrafish mutants (Fig. 3.1c-d).

A remaining challenge is now to determine how many genomic regions actually contribute to reproductive barriers. One possibility is that this is attributed to the accumulation of conditionally neutral mutations throughout the genome. This is an attractive hypothesis as mutations accumulate throughout the genome at all times in all populations. As such, they provide different architectural starting points, which may subsequently be recruited by adaptive processes and potentially lead to different evolutionary trajectories, whether under stabilizing or divergent selection pressures (Lynch 2007). Here, as very little difference was observed between normal and dwarf whitefish, postzygotic reproductive isolation in these young species appears to involve mostly genes under stabilizing selection for gene expression, as it has been previously shown in drosophilids (Haerty and Singh 2006) and brook charr (Mavarez et al. 2009).

The fact that in whitefish, over 30 % of mapped genetic markers show locus-specific deviations from expected Mendelian segregation (segregation distortion) tends to support the idea that many "ordinary loci" are associated with reproductive isolation (Rogers and Bernatchez 2007). On the other hand, we cannot refute that one or a few genetic loci may underlie large scale disruption of expression, especially so if those are associated with essential developmental genes, where knockdown mutations are known to be lethal. Another explanation worth mentioning pertains to the fact that large scale hybrid gene misexpression can also result from a variety of mechanisms related to the disrupting the chromatin integrity or the expression of non coding RNAs (Michalak 2009). For example, miRNA networks, which have been shown to regulate gene expression in early development in many vertebrate species, can have a large cascading effect if disrupted and thus provide a likely "major gene" effect (Lee et al. 2007, Michalak 2009). As such, all other differences in gene expression and phenotype would be a downstream effect of this major factor. Other minor incompatibilities could still exist, which would account for the differential expression of genes observed in the healthy backcross fish versus parental types. Along this line, we previously observed that a least one locus associated with embryonic mortality is also linked to a regulation (eQTL) hotspot and may harbor a major regulatory gene with strong pleiotropic effect on the expression of numerous genes (Bernatchez et al. 2010), thus representing a possible mechanism explaining the genome-wide disruption of regulation in hybrids.

3.7 Conclusion

By comparing healthy, surviving hybrids against moribund ones, our results identified a transcriptome wide signature of hybrid breakdown in young, incipient species and revealed a persuasive link between misexpression of essential developmental genes and post zygotic isolation. Our analysis of the genomic basis of ongoing speciation in young evolutionary fish lineages helps to bridge the gap between ecological studies of reproductive isolation with limited knowledge of their genetic basis, and genetic studies of speciation with an incomplete ecological perspective. Quite clearly, both are needed towards a truly general theory of the genetics of speciation.

3.8 Acknowledgements

We thank JC. Therrien and S. Higgins for their assistance in spawning and rearing whitefish, AW Nolte for suggestions on experimental design and analysis, J. Jeukens, CR Landry, E. Normandeau and J. St-Cyr for comments on an earlier version of the manuscript. This research 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 L.B. and a NSERC postgraduate scholarship to S.R. This paper is a contribution to the research program of Québec-Océan.

Chapitre 4 : Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. Salmonidae).

4.1 Résumé

Les nouvelles générations de séquençage permettent de découvrir un nombre élevé de marqueurs de type SNP pour des espèces pour lesquelles peu d'information génomique était au préalable disponible. Ici, nous avons assemblé *de novo* plus de 130Mb d'ADNc non-normalisé en utilisant des données de pyroséquençage 454 provenant de grand corégones normaux, nains et hybrides rétrocroisés. Le but principal de l'étude était d'assembler un large jeu de données de marqueurs SNPs, de documenter leur distribution dans les régions codantes, d'évaluer les effets de la divergence des espèces sur les fréquences d'allèles et, finalement, de combiner ces informations avec des études génomiques précédentes afin d'identifier des gènes candidats sous-jacents à la divergence adaptive du grand corégone. Nous avons identifié 6094 SNPs potentiels dans 2674 contigs (taille moyenne: 576 paires de bases, gamme [101-6116]), 1540 mutations synonymes et 1734 mutations non synonymes pour un ratio global de taux de polymorphisme non synonymes / synonymes (p_{N} / p_{S}) de 0.37. Comme attendu, basé sur le jeune âge (< 15 000 ans) des paires d'espèces de corégone, la divergence moyenne entre les deux était relativement faible. Néanmoins, 89 SNPs montraient des divergences de fréquences d'allèles prononcées entre nains et normaux. De ceux-ci, les SNPs dans des gènes annotés pour des fonctions métaboliques étaient les plus abondants. En combinaison avec des données expérimentales précédentes, ceci apporte la preuve que les gènes impliqués dans le métabolisme énergétique font d'excellents candidats expliquant la divergence des paires d'espèces de corégones nains et normaux. Finalement, de manière imprévue, nous avons identifié 44 séquences contiguës (contigs), annotées à des éléments transposables. Ceux-ci étaient composés, de manière prédominante, de séquences d'hybrides rétrocroisés; ce qui indiquerait une augmentation de l'activité des éléments transposables pouvant expliquer la baisse de la valeur adaptative des hybrides, documentée par le passé.

4.2 Abstract

Next generation sequencing allows the discovery of large numbers of single nucleotide polymorphisms 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 dataset 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 nonsynonymous mutations for a genome-wide nonsynonymous 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.

4.3 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 (Shendure & Ji 2008; Branton *et al.* 2008; Metzker 2009) has led some researchers to believe that many technical approaches (e.g.: Sanger sequencing, DNA microarrays), which where themselves revolutionary 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 (CNVs), Insertion-Deletions (indels) or Single Nucleotide Polymorphisms (SNPs) by sequencing pools of genetically heterogeneous individuals (Vera *et al.* 2008; Barbazuk *et al.* 2007; 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 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 Tassel *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 labor 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 since their effect on amino acid composition and therefore protein functionality can be easily assessed. Similarly to *dn/ds* ratios, the rate of accumulation of nonsynonymous 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 *et al.* 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 before present 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 QTLs suggest that a small proportion of the whitefish genome (approx. 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 program 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 dataset 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).

4.4 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 4.1). 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 *et al.* 2007, Rogers and Bernatchez 2007). In short, total RNA was extracted separately for each individual using the TRIzol Reagent protocol (Invitrogen, Carlsbad, CA). Following extraction, all samples were further cleaned by ultra filtration using microcon (Millipore, Billerica, MA) spin columns. Samples were quantified using Experion™ RNA StdSens Analysis Kit (Bio-Rad, Hercules, CA). Total RNA was stored in pure water supplemented with Superase-In™ RNase Inhibitor (Ambion, Austin, TX) and kept at –80°C.

Table 4.1 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]
D	Cliff Lake Dwarf	Liver [a]	8	0.75 plate	183365	194 / 214
N	Cliff Lake Normal	Liver [a]	8	0.75 plate	210703	191 / 209
BC	[(Aylmer Lake normal X Témiscouata Lake dwarf) X Aylmer Lake normal]	Muscle [b]	4	0.75 plate	238409	195 / 216
		Brain [c]	4			

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

Enrichment for polyA mRNA was conducted using MicroPoly(A)Purist™ Kit (Ambion). Approximately 100ng of full length complementary DNA was synthesized from each polyA mRNA sample following SMART™ PCR cDNA Synthesis Protocol (Clontech, Mountain View, CA). All cDNA samples (3-8ng) 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^{0}C, followed by 17-20 cycles depending on sample [1 cycle : 15 sec. at 95^{0}C, 30 s at 65^{0}C, 6min. at 68^{0}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 5ug 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 done using PyroBayes which produces more confident base calls than the native 454 base calling program (Quinlan *et al.* 2008). Prior to assembling all sequences, primers and sample specific tags sequences were removed from the dataset using a custom made Perl script. CLC Genomics Workbench 3.1 (CLC Bio, Aarhus, Denmark) was used to assemble sequences de novo (similarity 0.97, overlap 0.5). We performed several test 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 MYA, 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 / MY.

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 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 www.uniprot.org.

SNP discovery and functional characterization of polymorphism

Assembled contigs were screened for single nucleotide polymorphisms 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%, while 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 analyzed 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 nonsynonymous SNP per nonsynonymous 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, Renaut *et al.* 2009, Nolte *et al.* 2009), QTL and genome scans (Rogers & Bernatchez 2007) as well as eQTL mapping (Whiteley *et al.* 2008, Derome *et al.* 2008) studies to match their gene annotation with genes identified in this study. We provide a legend at the bottom of table 4.3 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, San Diego, USA) 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 square test corrected for multiple hypothesis testing, *q*-value, Storey 2002) and expected heterozygosity ($F_{ST} = (H_e - H_o) / H_e$) were calculated in R.

4.5 Results

Sequencing, contig assembly and annotation

A total of 632 000 sequences with a median length of 212 nucleotides/sequence and totalizing approximately 130 megabases were obtained from sequencing the D, N and BC separate pools of cDNA (0.75 GS-FLX sequencing run per pool; Fig. 4.1, Table 4.2 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 4.2), meaning that 32% of all sequences were left as unassembled singletons. Shorter reads were harder to assemble and usually discarded (Fig. 4.1). 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 (1.3 X - 4140 X) as another goal of this research will be to document differential gene transcription between dwarf and normal whitefish from this same data set (Jeukens et al. *accepted*). 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 4.2 Summary statistics of assembled contigs.

Number of sequences assembled		428068 (68% of total)
Number of contigs [a]		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 Open Reading Frames [b]		1904
Mean length of ORF		482
Number of SNPs		3274
p_N / p_S [c]		0.37 (0.0028 / 0.0075)

[a] Similarity criterion: 0.97. Minimum overlap: 0.5.

[b] Mininum length set for accepting Open Reading Frame: 200 nucleotides.

[c] p_S: Number of synonymous SNPs per synonymous sites, p_N: Number of nonsynonymous SNPs per nonsynonymous sites.

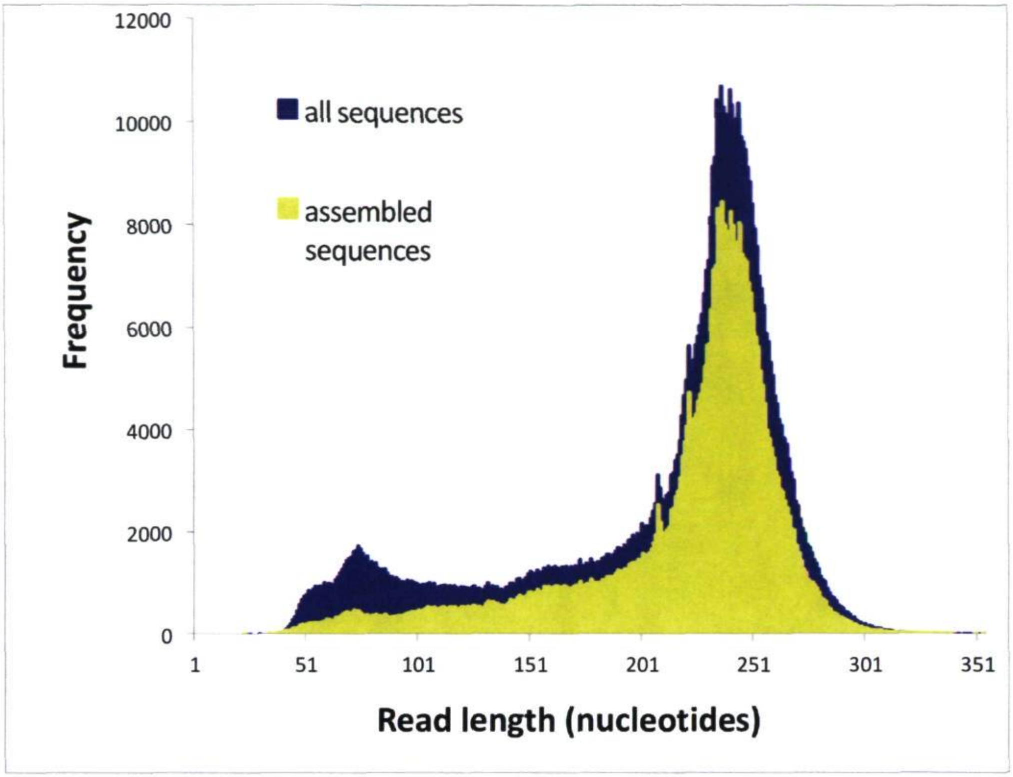

Figure 4.1 Frequency distribution of the total number of reads (blue) and assembled ones (yellow).

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 4.2). 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 (Fisher's exact test, *p*-value < 0.01, table 4.3).

Table 4.3 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 [a]	Functional groups	SNPs /kb	Match to previous studies [b]
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 organisation 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 organisation 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-sulfur 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 organisation 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 organisation 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	
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	

[a] Note that several contigs may correspond to the same gene annotation. These may be either splice variants of the same gene or different paralogs of that gene.

[b] **Match to previous studies that either showed differential expression between dwarf, normal or hybrid whitefish, or mapped to eQTL.**

1: Parallel nondirectional 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 nondirectional 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)

A total of 1904 predicted Open Reading Frames (ORF) with a mean length of 482 bp was identified. These contained 3274 polymorphic sites of which 1734 were synonymous and 1540 nonsynonymous. There were 2.8 SNPs per 1,000 nonsynonymous sites and 7.5 SNPs per 1,000 synonymous sites, for a genome-wide nonsynonymous to synonymous substitution rate ratio of 0.37 (p_S=0.0075, p_N=0.0028, Fig. 4.2, table 4.2). Twenty-nine contigs had a p_N / p_S 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 were significantly overrepresented compared to all contigs assembled (Fisher's exact test, *p*-value> 0.05; table 4.4).

Table 4.4 Contigs with the highest ratio of nonsynonymous SNP per nonsynonymous site (p_N) / synonymous SNP per synonymous site (p_S).

Gene product [a]	Functional groups	p_N / p_S	Match to previous studies [b]
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 signaling pathway (GO:0007186)	2.28	
Basement membrane-specific heparan sulfate 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 signaling 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

[a] Note that several contigs may correspond to the same gene annotation.

[b] See legend in table 3

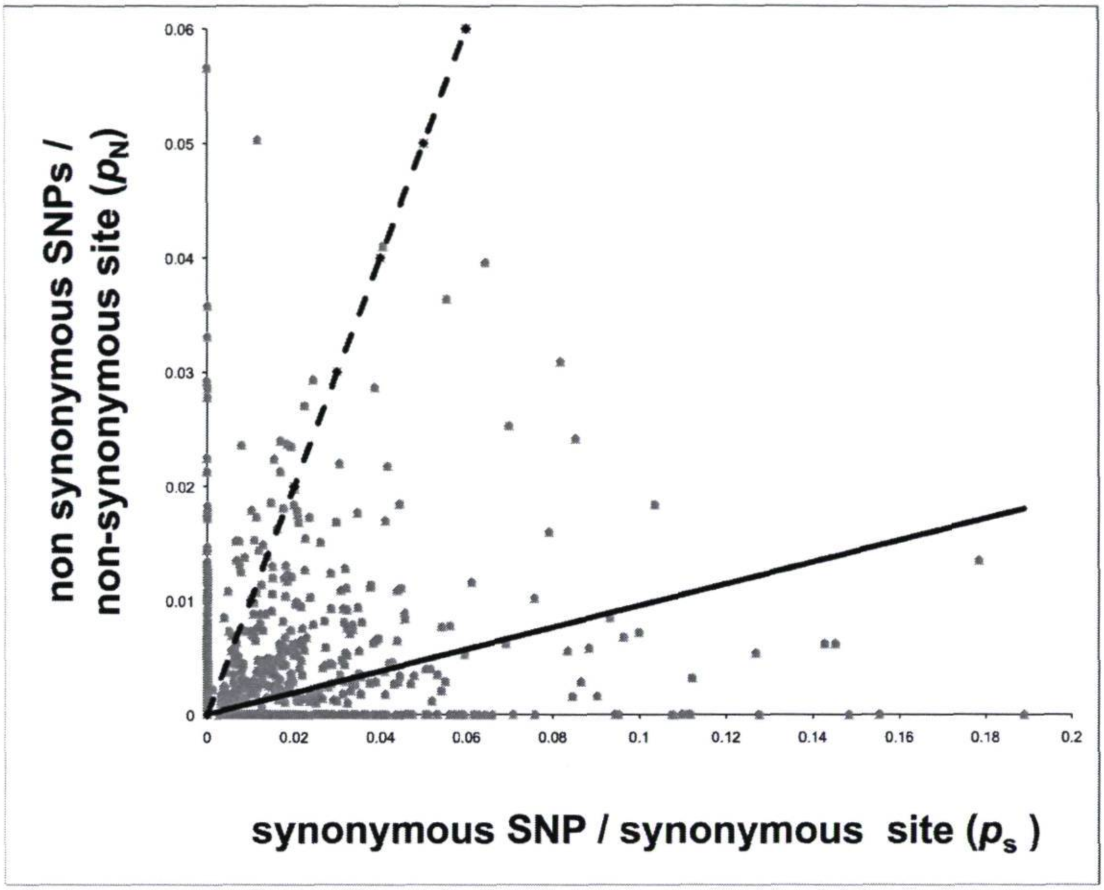

Figure 4.2 Nonsynonymous mutations per nonsynonymous sites compared to synonymous mutations per synonymous sites. Dashed line is the null expectation if mutations were randomly distributed ($p_N = p_S$). Solid line is the slope of experimental data (overall average p_N for all contigs / overall average p_S for all contigs = 0.37).

SNP frequencies between dwarf and normal whitefish

We analyzed a subset of 1504 SNPs that met our criterion for inferring allele frequencies (see Materials and Methods). While most SNPs showed little divergence (Fig. 4.3), 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. 4.3 & table 4.5). These 89 SNPs represented 46 different contigs and several biological functions. Of interest among these, seven mitochondrial genes (*q*-value < 1e-50: Cytochrome C subunit 1, 2 & 3, NADH-dehydrogenase 1, 4 & 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.

Table 4.5 SNP markers with significant divergent allelic frequencies between sympatric normal and dwarf whitefish.

Description [a]	Functional category [b]	Allele 1 (D) [c]	Allele 1 (N) [c]	Total number of sequences (D, N) [d]	abs[$f(a_{1D}) - f(a_{1N})$] [e]	Match to previous studies [f]
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 X precursor	(3) unknown	0.76	0.19	81,90	0.57*	
Coagulation factor X precursor	unknown	0.83	0.29	23,24	0.54**	
Complement C5 precursor	Complement activation,	1	0.17	12,6	0.83**	

	alternative pathway (GO:0006957)					
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*	
Hemopexin 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***	
subunit of Ca2+-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*

[a] Note that several contigs may correspond to the same gene annotation.

[b] Numbers in parentheses indicate that several SNPs within contig were divergent and the subsequent allelic frequencies and *q*-values are an average for these SNPs.

[c] Frequency of the most common dwarf allele and frequency of its corresponding normal allele.

[d] Number of sequences from dwarf and normal fish used to calculate allele frequencies.

[e] abs[$f(a_{1D}) - f(a_{1N})$] = absolute value of [frequency(allele$_{1Dwarf}$) - frequency(allele$_{1Normal}$)]. * *q*-value < 0.05, ** *q*-value < 0.01, *** *q*-value < 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.

[f] See legend in table 4.3.

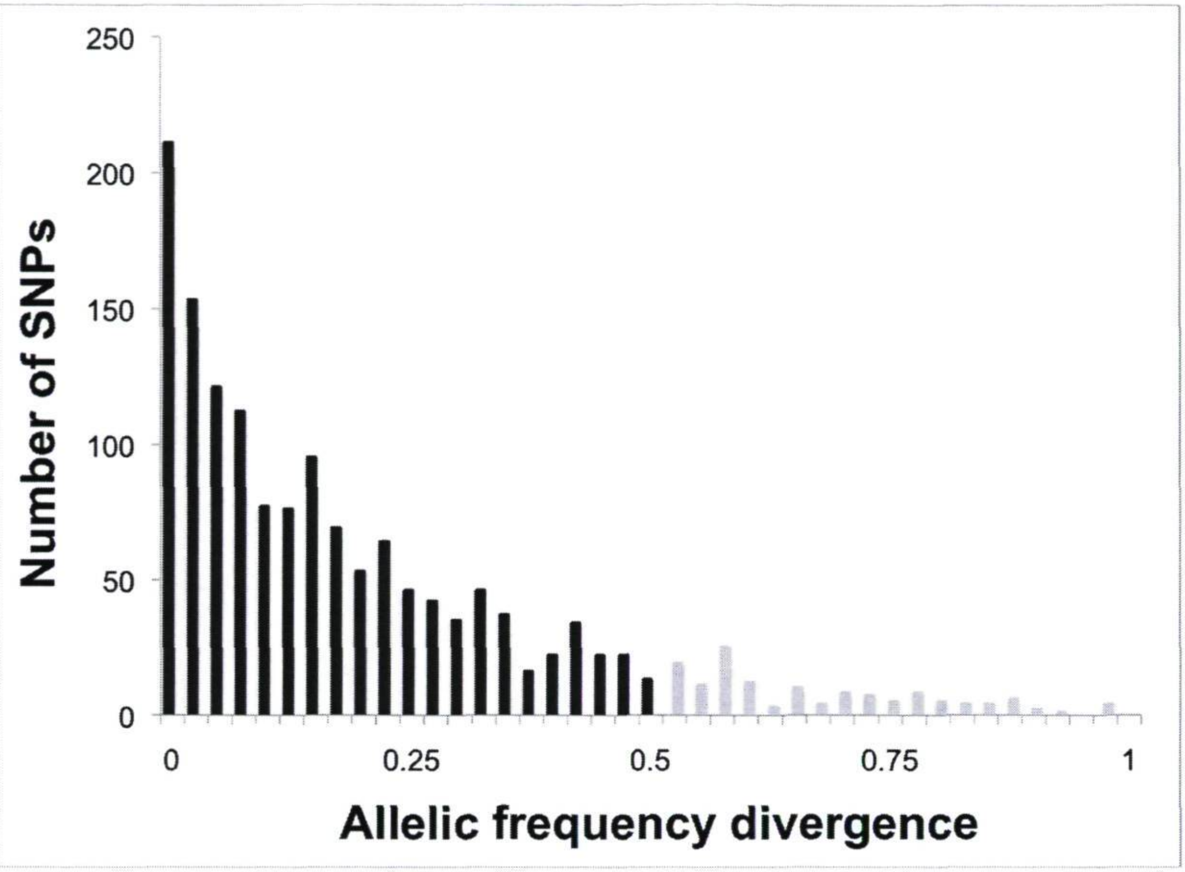

Figure 4.3 Frequency distribution of allelic frequency differences between normal and dwarf whitefish. Allele divergence value above one (yellow) and with a q-value < 0.05 were considered as highly divergent SNP markers. Allele divergence value = absolute value of [frequency(allele$_{1\text{Dwarf}}$) - frequency(allele$_{1\text{Normal}}$)]. Note that 1504 SNPs from 387 different contigs were used to draw this distribution.

Comparison with previous studies

Twelve contigs identified as highly polymorphic (i.e. above 20 SNPs / kb, table 4.3) 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 4.3). 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 4.4).

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 4.5). Among these, genes related to energy metabolism (Cytochrome C subunit 1, 2 & 3, NADH-dehydrogenase 1, 4 & 5, and 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 (see table 4.3), these were further investigated. Fourty-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*-value < 0.0001). Since sequencing was done on non-normalized cDNA, 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, while the dataset was composed of only 38% backcross sequences (chi square test, *p*-value < 1e-16).

Table 4.6 Expression (total number of sequences) annotated to transposon elements in normal and dwarf whitefish as well as backcross hybrids.

		Total number of sequences		
Gene product	**No. contigs** [a]	**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 (*S. purpuratus*)	1	1	9*	2
Probable RNA-directed DNA polymerase from transposon BS	6	68	394**	90

[a] Several assembled contigs were annotated (*e*-value < 1e-50) to the same gene product. **p*-value = 0.08, ** *p*-value < 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).

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*-value < 0.05, Fig. 4.4). 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*-value = 0.69).

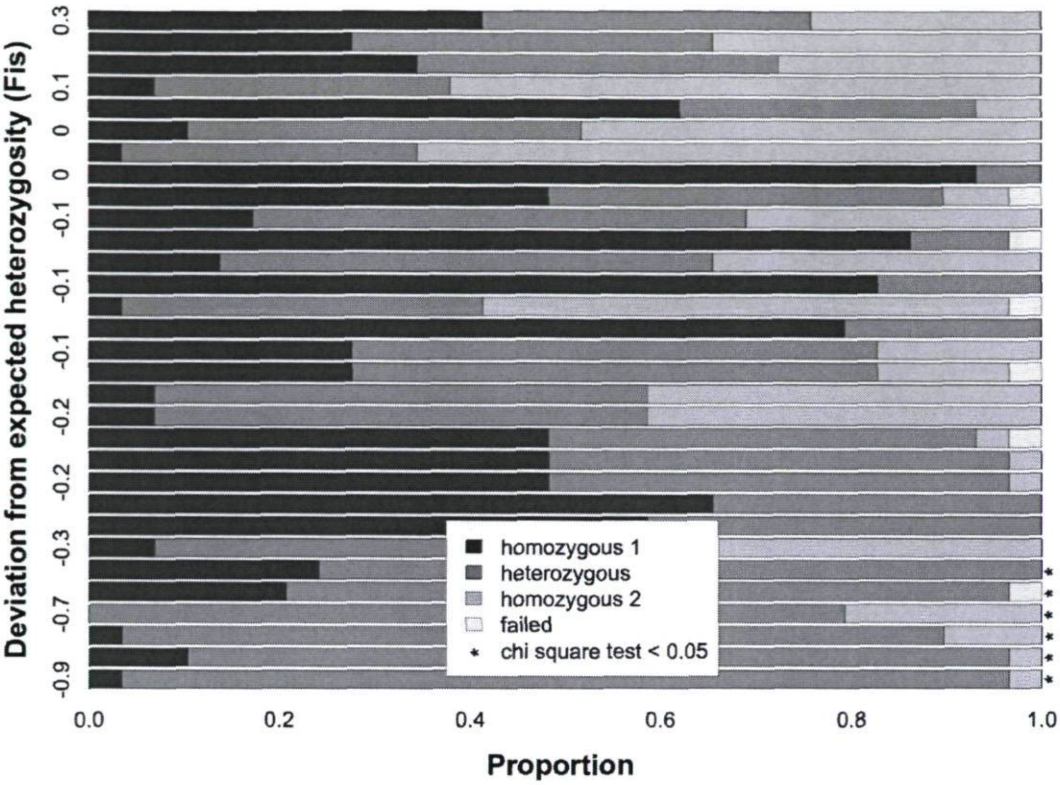

Figure 4.4 SNP validation for 29 individuals originating from a single lake (Lake Aylmer) and genotyped for 31 polymorphic markers. SNPs were ranked according to F_{IS} values (y-axis, left side). Deviation from expected Hardy-Weinberg frequencies (chi square test, 1 df, q-value < 0.05) were included on the y-axis (right side).

4.6 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 about 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. Firstly, cDNA sizes, which are quite variable, render the shearing process prior to sequencing more difficult. Secondly, mature cDNA usually contains large polyA stretches which are harder to sequence and cause many reads to be rejected due to poor quality or very short lengths (Gary Levesque, 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 on the read length (i.e. shorter reads are harder to assemble, Fig. 4.1) 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, since 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 artifact since 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, since 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 Paralogous Sequence Variants. 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 and colleagues (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 single nucleotide polymorphisms. 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. 4.4), a tell-tale sign that these SNPs may be variants between duplicated regions of the genome (Fredman *et al.* 2004). While 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, while 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.* Vasemägi & 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 dataset of SNPs 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 multi-genic 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 genotyping results, we did not find any significant correlation between F_{IS} (as a potential indication of PSVs) and polymorphism rate (p-value = 0.69).

Nucleotide substitution effect on predicted open reading frames

By identifying 1904 predicted Open Reading Frames, this permitted to estimate a transcriptome wide nonsynonymous 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. 4.2). 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 (eg. 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. Firstly, by definition, ORFs represent "potential" region of the genome translated into a protein and therefore do not necessarily code for the actual polypeptide chain. Secondly, since the number of polymorphic sites 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, while 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 nonsynonymous 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 4.5), had significantly 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 and colleagues (2005) have identified only three genomic regions, encompassing a maximum of 67 genes, showing evidence 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. Firstly, 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. Secondly, since transcribed cDNA was sequenced, it is conceivable that normal and dwarf heterozygous individuals may over-express 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. unpubl.).

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; Rogers *et al.* 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, since 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. Firstly, previous gene expression studies (Derome *et al.* 2006; St-Cyr *et al.* 2008; Jeukens *et al.* 2009; Renaut *et al.* 2009; Nolte *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 and Bernatchez 2004; Rogers and Bernatchez 2007) to identify genes under the influence of divergent selection. However, they provided only indirect evidence since 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 artifact 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 and Dodson 1990; Lu *et al.* 2001). Consequently, while 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 less than ten 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 and colleagues (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* versus *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, that could be either physically linked to other candidate genes or be selected due to strong epistatic interactions with metabolic genes.

4.7 Conclusion

Next generation sequencing technologies is 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 dataset of putative SNP markers, analyze 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 program 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.

4.8 Acknowledgments

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 paper is a contribution to the research program of Québec Océan.

4.9 Research Interest of the authors

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.

Chapitre 5 : Gradients of ecological speciation, SNP signature of selection on standing genetic variation, and association with adaptive phenotypes in lake whitefish species pairs (*Coregonus* spp.).

5.1 Résumé

Lorsqu'une population s'adapte à un nouvel environnement, la sélection divergente contribuera à promouvoir une différenciation génomique hétérogène, en provoquant une réduction du flux de gènes pour les loci associés aux caractères adaptatifs. Ici, en utilisant plus de 100 marqueurs SNPs récemment développés pour des jeunes espèces de grand corégones nains et normaux (*Coregonus clupeaformis*), nous avons effectué des analyses de balayage génomique afin de caractériser l'effet de la sélection naturelle dans cinq lacs distincts contenant des paires d'espèces sympatriques. Une proportion différente de SNPs (entre 0-12%) a été identifiée comme aberrante (*outliers*) et ceci en relation avec l'intensité prévue des interactions compétitives entre les corégones nains et normaux propre à chaque lac. De plus, la forte réduction de l'hétérozygotie pour les loci *outliers* chez les nains, mais non les normaux indiquerait que la sélection directionnelle a agit sur la variation génétique présente *ab initio* de manière plus forte sur le phénotype nain que normal. Par la suite, en utilisant le même ensemble de marqueurs SNPs, nous avons testé leur association avec neuf phénotypes adaptatifs chez les poissons hybrides issus de rétrocroisements. Quatre caractères adaptatifs (croissance, activité, nombre de branchiténies et facteur de condition) étaient associés à 16 gènes différents. Nous n'avons observé aucune relation simple entre le niveau de différenciation génétique (F_{ST}) et l'association avec des phénotypes adaptatifs. De même, nous n'avons pas observé de signatures génétiques de divergence adaptative parallèle. Au contraire, les résultats spécifiques à chaque lacs sous-tendent l'évolution indépendante de ceux-ci. Finalement, l'utilisation intégrée de données phénotypiques, transcriptomiques et fonctionnelles a mené à la découverte de certains gènes candidats (ex. Sodium Potassium ATPase et Triosephosphate isomerase) impliqués dans la divergence écologique du grand corégone. En conclusion, l'identification de plusieurs marqueurs sous sélection divergente suggère que de nombreux gènes, d'une manière spécifique à chaque lac, sont recrutés par sélection qui agit sur la variation génétique présente avant la spéciation écologique des corégones et ultimement, contribue à l'évolution du phénotype nain. À ce titre, nos résultats représentent un cas convaincant du rôle prédominant de la variation génétique présente *ab initio* (plutôt que des mutations *de novo*) dans les premières étapes de la spéciation écologique.

5.2 Abstract

As populations adapt to novel environments, divergent selection will promote heterogeneous genomic differentiation via reductions in gene flow for loci underlying adaptive traits. Using a dataset of over 100 SNPs markers, genome scans were performed to investigate the effect of natural selection maintaining differentiation in five lakes harbouring sympatric pairs of normal and dwarf lake whitefish (*Coregonus clupeaformis*). A variable proportion of SNPs (between 0-12%) were identified as outliers, which corroborated the predicted intensity of competitive interactions unique to each lake. Moreover, strong reduction in heterozygosity was typically observed for outlier loci in dwarf but not in normal whitefish indicating that directional selection has been acting on standing genetic variation more intensively in dwarf whitefish. SNP associations in backcross hybrid progeny identified 16 genes exhibiting genotype-phenotype associations for four adaptive traits (growth, swimming activity, gill-rakers and condition factor). However, neither simple relationship between elevated levels of genetic differentiation with adaptive phenotype, nor conspicuous genetic signatures for parallelism at outlier loci were detected, which underscore the importance of independent evolution among lakes. The integration of phenotypic, transcriptomic and functional genomic information identified two candidate genes (Sodium Potassium ATPase and Triosephosphate isomerase) involved in the recent ecological divergence of lake whitefish. Finally, the identification of several markers under divergent selection suggests that many genes, in an environment specific manner, are recruited by selection and ultimately contributed to the repeated ecological speciation of a dwarf phenotype.

5.3 Introduction

One of the main objectives of evolutionary biology is to elucidate the genetic basis of adaptive phenotypic traits. As natural selection acts to shape phenotypic diversity, it modulates the underlying genomic architecture in intricate ways. Despite tremendous advances in genetic studies, a link between adaptive phenotypes and genotype has been made for only a small number of traits in an even smaller number of organisms (e.g. Colosimo *et al.* 2005; Hoekstra *et al.* 2006; Miller *et al.* 2007; Counterman *et al.* 2010). These demonstrations require the combination of different research approaches targeting various functional and biological levels (e.g. variation at the DNA, gene expression and phenotypic levels) and represent the best strategy for deciphering the genetic basis of evolutionary change and diversification driven by natural selection (Vasemägi & Primmer 2005; Stinchcombe & Hoekstra 2008: Storz & Wheat 2010). Moreover, progress towards this goal will be best accomplished by the comparative study of evolutionarily young and ecologically distinct lineages where genetic conflicts are not fully resolved and natural introgressive hybridization, albeit limited, is still possible (Mayr 1963; Schluter 2000; Via 2009; Presgraves 2010). As such, the genetic changes contributing to the early steps of adaptive divergence and reproductive isolation can be studied before they become confounded and erased by additional genetic differences that accumulate after speciation is complete.

According to the ecological theory of adaptive radiation, shifts of organisms into novel habitats are hypothesized to be adaptive, whereby populations should diverge for specific phenotypes and genotypes that influence survival and reproduction when exposed to different environments (Mayr 1963; Schluter 2000). As a consequence, divergent selection will create heterogeneous genomic differentiation by causing specific loci (and those physically linked to them) to flow between populations less readily than others. This will result in accentuated genetic divergence of regions affected by selection while, on the contrary the homogenizing effects of gene flow will preclude divergence in other regions (Lewontin & Krakauer 1973; Wu 2001; Nosil *et al.* 2009). As speciation takes its course, regions under the effect of divergent selection, expected to be originally rare, will tend to grow in size and number until eventually the whole genome becomes fully incompatible and true *sensus stricto* biological species are formed (Wu 2001; Wu & Ting 2004).

Several methods have been developed to identify regions of genetic divergence (F_{ST} outliers genome scan methods, Lewontin & Krakauer 1973; Beaumont & Nichols 1996; Beaumont & Balding 2004; Foll & Gaggiotti 2008; Excoffier *et al.* 2009). Nevertheless, any of these approaches has its limitations and may be biased towards identifying only markers under particularly strong selective pressure (Michel *et al.* 2010; Storz & Wheat 2010). Thus, genome scan methods should be complemented by other approaches towards linking the effect of selection with genetic and ultimately, adaptive phenotypic divergence (Butlin 2008; Michel *et al.* 2010). Accordingly, demonstrating the effect of divergent selection for specific loci while simultaneously associating these same loci to adaptive characters known to influence assortative mating brings compelling evidence of the genetic basis of ecological speciation. Nevertheless, such demonstrations remain few and difficult to document (Noor & Feder 2006; Schluter 2009; Presgraves 2010). Towards this goal, Via and West (2008) showed that quantitative trait loci (QTL) for adaptive traits between pea aphid populations were, albeit weakly, linked to regions of higher genetic differentiation. Via (2009) further highlighted a similar scenario in lake whitefish (Rogers & Bernatchez 2007) and hypothesized that natural selection should create relatively large region of genetic differentiation as among populations effective rate of recombination around these markers become highly reduced. Conversely, speciation can also be initiated by selection acting simultaneously on many physically unlinked loci. These alternative scenarios are not mutually exclusive as divergent selection may act strongly on individual genes, creating large regions of differentiation through the process of divergence hitchhiking, while concurrently acting in a global, intricate manner (Feder & Nosil 2010; Michel *et al.* 2010).

Lake whitefish from the St-John River basin (southeastern Quebec, Canada and northeastern USA) are characterized by the occurrence of several lakes harbouring dwarf and normal sympatric whitefish (Bernatchez *et al.* 2010). They represent a rare illustration of a continuum of both morphological and genetic differentiation within a given taxon, spanning from complete introgression to near complete reproductive isolation, depending on the history (Lu *et al.* 2001) or the unique ecological characteristics of each lake (Lu & Bernatchez 1999; Landry *et al.* 2007, Landry & Bernatchez 2010). Furthermore, mounting evidence has indicated that dwarf is the derived phenotype, evolved from a normal whitefish ancestor (Rogers & Bernatchez 2007; Landry *et al.* 2007; Landry & Bernatchez 2010). For instance, dwarf whitefish exclusively occur in sympatry with normal whitefish. In addition, lakes inhabited by sympatric pairs and isolated since the last glacial retreat about 12 000 years ago indicate that dwarf whitefish have evolved in parallel more than once (Pigeon *et al.* 1997). At the genetic level, genome scans have provided evidence of markers under divergent selection while concurrently identifying limited parallel patterns of genetic differentiation between independent lakes (Campbell & Bernatchez 2004; Rogers & Bernatchez 2007). In addition, a genetic basis has been demonstrated through common garden experiments and QTL mapping for adaptive traits known to differ between both forms: namely swimming behaviour, growth, morphology and gene expression variation (Rogers *et al.* 2002; Rogers & Bernatchez 2007; Derome *et al.* 2008; Whiteley *et al.* 2008). These comprehensive studies using anonymous AFLP markers nevertheless beg the question as to the nature and functional identity of the genes underlying adaptive traits under selection.

Here, through the use of a set of over 100 informative single nucleotide polymorphisms (SNPs) markers developed from lake whitefish coding regions, we aimed to complement this largely anonymous genetic basis of adaptive divergence with a more functional ecogenomics approach by conducting both genome scans in five distinct lakes containing sympatric normal dwarf species pairs as well as genotype-phenotype associations. These five lakes, differentiated in their potential for competitive interactions, phenotypic and genetic divergence between normal and dwarf represent a continuum of ongoing ecological speciation (Bernatchez *et al.* 1999; Lu *et al.* 2001; Campbell & Bernatchez 2004; Landry *et al.* 2007, Landry & Bernatchez 2010). As such, we predicted that lakes with a lower potential for competition (and associated weaker genetic and phenotypic differentiation) should exhibit fewer SNP markers affected by natural selection compared to lakes with potentially higher competitive environments. Following this, using the same set of SNP markers, we tested the statistical association between genetic variation and phenotypic traits known to underlie the differential adaptation of normal and dwarf lake whitefish. Then, we investigated the hypothesis that genetic markers showing elevated levels of genetic differentiation should also be more strongly associated with adaptive phenotypes than other neutral markers. Finally, through the integrated use of F_{ST} outliers genome scan, genotype-phenotype association and functional genomics, we identified candidate genes involved in the recent ecological divergence of lake whitefish normal and dwarf species pairs.

5.4 Methods

Samples and study system

We used DNA samples previously collected (Campbell & Bernatchez 2004) from lakes harbouring sympatric populations of normal (N) and dwarf (D, Fig. 5.1): Cliff Lake (27 N, 30 D), Webster Lake (26 N, 22 D), Indian Pond (13 N, 28 D) and East Lake (24 N, 24 D) as well as material collected in 2007 from Témiscouata Lake (47°41'N, 68°47'W; 24 N, 24 D). The colonization history of all lakes, except East involved a secondary contact between two independent evolutionary lineages isolated for 100 000 - 200 000 YBP (Acadian and Atlantic, Bernatchez & Dodson 1990). In addition, in these lakes, the dwarf phenotype most likely evolved from a normal phenotype of the Acadian lineage ancestry (Lu *et al*. 2001). In contrast, East Lake has been colonized only by the Acadian lineage (Pigeon *et al.* 1997). We also included samples collected from an allopatric normal population from Pohénégamook Lake, providing information regarding ancestral standing genetic variation that existed in the pure Acadian lineage prior to secondary contact. DNA for genotyping and sequencing was extracted using a standard Proteinase K digest of tissues in SDS buffer and consecutive chloroform and high NaCl isolation.

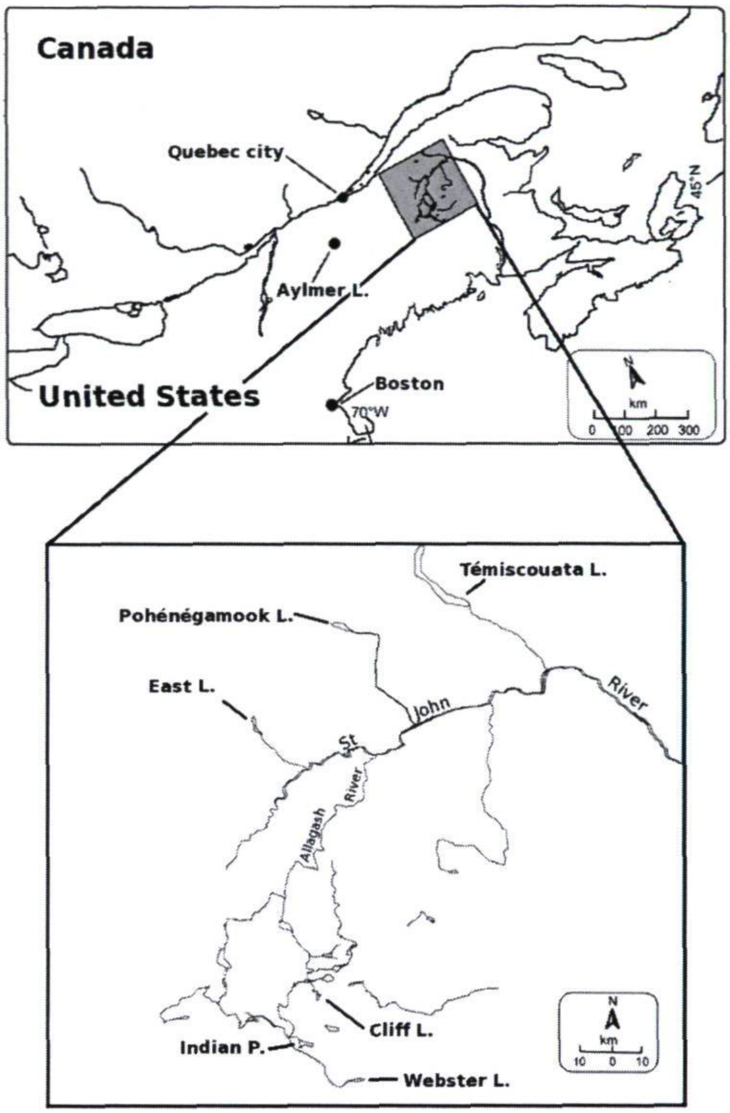

Figure 5.1 Map of the study area, with locations of population samples in the St-John River basin.

For the association study, we used DNA samples previously extracted from 196 backcross ((Normal X Dwarf) X Dwarf) individuals (see Rogers *et al.* 2007 for details). These backcross individuals trace their origins back to Témiscouata Lake (dwarf, Acadian lineage) and Aylmer Lake (normal, Atlantic-Mississippian lineages). When necessary, we re-extracted DNA using standard Proteinase K digest of tissues in SDS buffer and consecutive chloroform and high NaCl isolation from fin clips preserved in ethanol. For these individuals, nine different phenotypes previously measured and found to differentiate normal and dwarf lake whitefish were used (behavioral traits - depth selection, burst swimming, directional change, activity; physiological traits - growth rate, condition factor; morphological traits - gill-rakers; life history traits - onset of sexual maturation, gonado-somatic index, see Rogers & Bernatchez 2007 for details about phenotypic measurements).

SNP identification and genotyping

Two approaches were used for SNP discovery. First, we designed primers based on salmon ESTs that were used in microarray studies to detect differences in gene expression (Nolte *et al.* 2009a; Renaut *et al.* 2009). These primers were used to PCR amplify fragments from the genomic DNA from pools of dwarf whitefish from Témiscouata Lake and normal whitefish from Aylmer Lake. PCR amplicons were Sanger sequenced and then visually screened for putative SNPs. Additional SNPs were chosen from a 454 sequencing experiment of cDNA libraries derived from dwarf and normal whitefish from Cliff Lake as well as from the same backcross individuals used in the association study. Putative SNPs previously identified (see Renaut *et al.* 2010 for all details on SNP identification criteria) were visually inspected in an attempt to discard erroneous assemblies or low quality SNPs, which may cause errors in primer design and amplification during genotyping. Briefly, regions 200 bp upstream and downstream of a SNP of interest and which contained two or more SNP or indel in full linkage disequilibrium were discarded as they are likely to represent paralogous sequence variants. SNPs closer than 100 bp from the contig end were also discarded. All sequences were matched (BLASTn) against NCBI nr/nt database and only the best hit for each amplicon was retained for annotation purposes.

Genotyping assays were designed and performed using Matrix-assisted laser desorption / ionization time-of-flight mass spectrometry (MALDI-TOF) developed by Sequenom (San Diego, CA, USA) at the Genome Quebec Innovation Center (McGill University, Montreal, Canada). Genotyping assays were developed for 470 putative SNPs. Replicate genotyping of positive controls indicated a maximum error calling rate of 4.10% (24 inconsistencies out of 586 genotypes).

We calculated pairwise F_{ST}, observed and expected heterozygosities for each whitefish population independently using MICROSATELLITE ANALYZER (Dierenger & Schlötterer 2003, suppl. table 3). For the purpose of identifying loci subject to selection, we used F_{ST} estimates from the five lakes harboring sympatric populations. BAYESFST (Beaumont & Balding 2004) was used to test for the significance of outlier loci for all polymorphic SNPs within each lake (two populations defined, normal and dwarf). We interpreted evidence of positive selection as suggested by Beaumont and Balding (2004). For each locus, a positive value of the locus parameter (α_i) suggests that locus *i* is subject to divergent selection, whereas a negative value suggests that balancing selection tends to homogenize allele frequencies over populations. We calculated 10 000 values of α_i and defined α_i to be 'significant at level *p*' if *p* percent of the values were positive (evidence of divergent selection). We ran simulations five times with different seed values for the algorithm to ensure reproducibility of probability values (Pearson's correlation coefficient > 0.99). Note that BAYESFST deals with the problem of multiple hypotheses testing through the prior distribution of the regression parameter for the locus parameter α_i. Therefore *p*-values calculated in BAYESFST are very conservative compared to frequentist method based on summary statistics (e.g. FDIST, Beaumont & Nichols 1996; Beaumont & Balding 2004).

A chi-square test was performed to verify whether parallel trends of genetic divergence between lakes were observed more often than expected at random. Expected numbers of loci showing parallel trends between two lakes were calculated from the product of the percentage of outliers in the first and second lake by the total number of markers surveyed (Campbell & Bernatchez 2004).

Normalized phenotypic data for the nine adaptive phenotypes (Rogers & Bernatchez 2007) were used to perform an association analysis in the R environment (v.2.10.1. The R Foundation for Statistical Computing®, 2009) with the package SNPassoc (v1.6, Gonzalez *et al.* 2007) using a codominant genetic model. We applied, for each SNP, a general linear model and a likelihood ratio test to obtain probability values. For each test, *p*-values were then corrected for multiple hypotheses testing using *q*-value correction (*q*-value package, Storey 2002).

5.5 Results

SNP identification and genotyping

After exclusion of failed marker assays, monomorphic markers, and those markers exhibiting excess heterozygosity, a set 112 (16 identified through Sanger sequencing, 96 through 454 sequencing) were retained for further analyses (see supplementary table 1 (genome scan) and 2 (association) for lists of informative SNPs). Ninety-six of those SNPs were informative in the genome scan of natural populations (mean missing data rate of 7.5% across individuals, suppl. table 5.1) and 87 for genotype-phenotype associations (mean missing data rate of 16.1% of across individuals, suppl. table 5.2), such that 63 % of all markers (70/112) were informative in both datasets.

Genome-scan of sympatric normal and dwarf species pairs

Cliff Lake had the highest mean pairwise F_{ST} value (0.28), followed by Webster (0.11), Indian (0.06), East (0.02) and Témiscouata (0.01, table 5.1). These values were highly concordant with F_{ST} values calculated previously from both microsatellite (Pearson's correlation coefficient = 0.99) and AFLP (Pearson's correlation coefficient = 0.83) markers as well as with the extent of phenotypic differentiation (Pearson's correlation coefficient = 0.86) previously observed between normal and dwarf whitefish (Lu & Bernatchez 1999, Campbell & Bernatchez 2004).

Table 5.1 Estimates of genetic differentiation between dwarf and normal lake whitefish.

Normal - dwarf pairwise F_{ST} (mean)	Cliff	Webster	Indian	East	Témiscouata
SNP[a]	**0.28**	**0.11**	**0.06**	**0.02**	**0.01**
AFLP[b]	0.22	0.17	0.04	0.11	NA
Microsatellite[c]	0.26	0.14	0.08	0.06	0.04

[a]Estimates for SNP are based on 94 polymorphic nuclear loci, [b]440 AFLP loci (reported from Campbell & Bernatchez 2004) and [c]six microsatellite loci (reported from Lu & Bernatchez 1999). NA: Témiscouata Lake was not analyzed using AFLP markers by Campbell & Bernatchez (2004).

We identified outlier loci showing accentuated patterns of genetic differentiation (Fig. 5.2 and table 5.2). Outliers were detected in every pairwise comparison, except for Témiscouata Lake. At a *p*-value of 0.2 calculated from the 10 000 iterations of the locus parameter (α_i) in BAYESFST, the number of outliers was 12, 7, 5, 3, 0 for Cliff, Webster, Indian, East and Témiscouata, which represented 15, 8.4, 6.1, 3.7 and 0 percent of the markers tested respectively. Furthermore, the number and percentage of outliers was positively correlated with the extent of genetic divergence between normal and dwarf (Pearson's correlation coefficient = 0.96). The two mitochondrial markers (Cytochrome c oxidase subunit 3 and NADH ubiquinone oxidoreductase chain 5) were completely fixed between normal and dwarf in Cliff Lake ($F_{ST} = 1$), as previously reported for the whole mitochondrial genome (Bernatchez & Dodson 1990). Glucose-6-phosphatase ($F_{ST} = 0.94$, Cliff Lake), a gene playing a key role in regulating glucose levels in the blood, was the nuclear gene showing the greatest F_{ST} value. Several other genes also had a high outlier F_{ST} value depending on lakes (e.g.: Probable ubiquitin carboxyl terminal hydrolase with $F_{ST} = 0.93$ in Cliff Lake and $F_{ST} = 0.41$ in Indian Lake; Heat shock protein HSP 90 beta with $F_{ST} = 0.71$ in Webster Lake, Cyclin I with $F_{ST} = 0.21$ in East Lake). Four genes (Cyclin I, Heat Shock 27kDa protein, Probable ubiquitin carboxyl terminal hydrolase and Sodium potassium transporting ATPase subunit alpha) were also identified as parallel F_{ST} outliers in more than one lake, although in only one comparison (Cliff - East Lake), was the number of outliers greater than expected by chance alone (chi square test, *p*-value = 0.04, table 5.3).

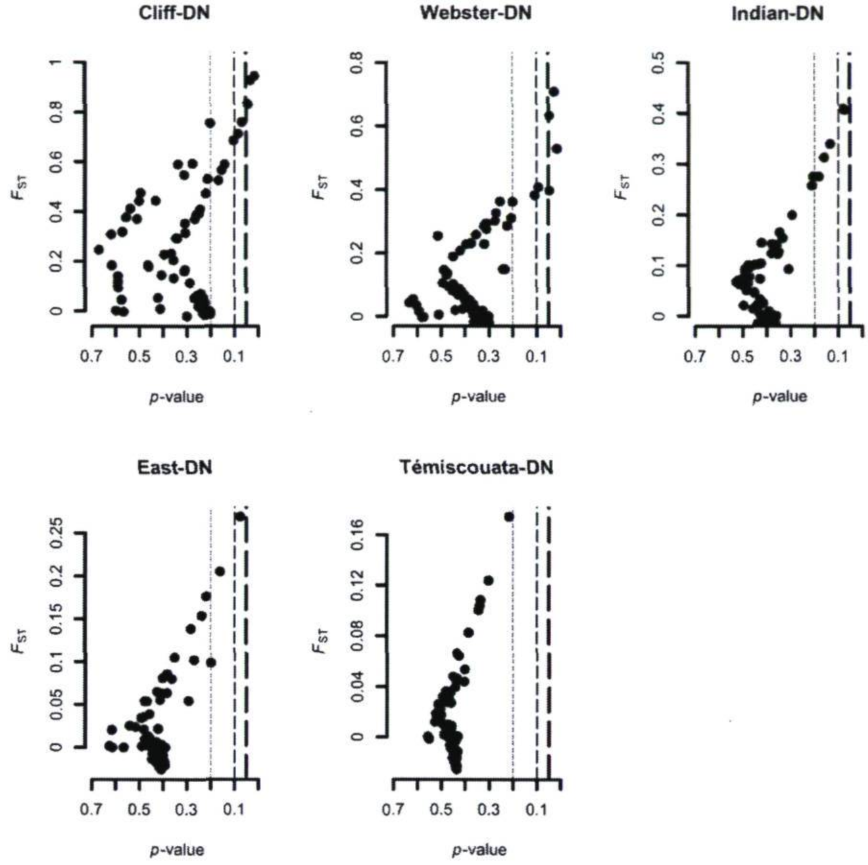

Figure 5.2 Pairwise F_{ST} as a function of probability between sympatric dwarf and normal whitefish in five lakes. Dashed lines represent 0.2, 0.1 and 0.05 *p*-values for all five genome scans performed independently.

Table 5.2 Summary of all F_{ST} outliers identified from genome scans between sympatric dwarf and normal whitefish in five lakes.

SNP functional annotation	Cliff	Webster	Indian	East	Témiscouata
26S protease regulatory subunit 4	**0.83*****	0.04	0.10	0.00	-0.02
Antithrombin III precursor	0.39	0.02	**0.41****	0.08	-0.01
ATP binding cassette sub family E member 1	0.37	0.15	**0.28***	-0.02	-0.01
Glucose 6 phosphatase	**0.94*****	0.29	0.20	-0.02	0.00
Triosephosphate isomerase	**0.69****	0.06	-0.03	0.00	-0.02
Proteasome subunit beta type 8 precursor	**0.57***	0.36	0.14	-0.01	-0.02
Fibrinogen beta chain precursor	0.47	**0.38***	0.14	0.02	-0.02
T complex protein 1 subunit epsilon	0.16	**0.40****	-0.03	0.01	0.08
ATP synthase subunit e	0.14	**0.63****	0.17	-0.01	0.01
Multifunctional protein ADE2	0.10	0.13	**0.31***	0.01	-0.02
Probable ubiquitin carboxyl terminal hydrolase	**0.93*****	-0.01	**0.41****	0.04	0.00
Heat Shock 27kDa protein	0.39	**0.41****	**0.28***	0.18	0.01
NADH ubiquinone oxidoreductase chain 5 (mitochondrial)	**1.00*****	-0.04	0.15	0.00	0.00
Cyclin I	**0.53***	**0.36***	0.08	**0.21***	0.02
Cytochrome c oxidase subunit 3 (mitochondrial)	**1.00*****	-0.02	0.15	0.00	0.00
Heat shock protein HSP 90 beta	0.44	**0.71*****	0.05	-0.01	0.07
complement component C9	**0.59****	-0.02	0.13	0.02	0.03
Angiotensinogen	0.41	**0.53*****	0.34	-0.03	-0.02
Salmo salar RED protein	**0.76****	0.31	0.14	0.09	-0.02
Sodium potassium transporting ATPase subunit alpha	**0.71****	0.00	NA	**0.10***	-0.02

Sodium potassium transporting ATPase subunit beta	**0.75***	0.32	0.09	0.06	0.05
Ribulose phosphate 3 epimerase	-0.01	0.07	-0.03	**0.27****	-0.01

Significant outliers are in bold. *P*-values calculated from BAYESFST outputs as described in MM: * *p*-value < 0.2, ** *p*-value < 0.1, *** *p*-value < 0.05. Genes underlined were also associated with adaptive phenotypes (see table 3)

Table 5.3 Chi-Square test to assess whether parallel trends for F_{ST} outlier loci were observed more often than expected by chance. Genes included in this table were those that were outliers in more than one lake. Below the diagonal are expected values; above the diagonal are the outlier loci. * *P*-value = 0.04 in East - Cliff comparison (two genes), all other comparisons non significant.

	Cliff	**Webster**	**Indian**	**East**	**Témiscouata**
Cliff	X	Cyclin I	ubiquitin carboxyl terminal hydrolase	Na-K transporting ATPase subunit alpha & Cyclin I *	0
Webster	1.21	X	Heat Shock 27kDa protein	0	0
Indian	0.88	0.49	X	0	0
East	0.53	0.30	0.22	X	0
Témiscouata	0	0	0	0	X

Selection acting on standing genetic variation in dwarf whitefish.

In Cliff and Webster lakes where dwarf and normal whitefish are the most genetically and phenotypically differentiated, observed heterozygosity was significantly reduced for outlier loci in dwarf whitefish ($H_{o\ (dwarf)}$ = 0.03 and 0.32 for outliers compared to 0.28 and 0.44 for all other markers in Cliff and Webster respectively, t-test, p-value < 0.05), but not in normal whitefish (Fig. 5.3). In Cliff Lake, ten of the twelve outliers that were polymorphic in normal were fixed in dwarf whitefish. This trend of reduced diversity at outlier loci in dwarf but not in normal whitefish was similar, but not significant in Indian and East lakes, possibly due to large variance estimates associated with a smaller number of outliers. In addition, heterozygosity values in Pohénégamook Lake (H_o = 0.28, pure Acadian lineage origin) and East Lake normal whitefish (H_o = 0.33, pure Acadian lineage origin) for loci identified as outliers in Cliff Lake confirmed that these SNPs were polymorphic in the ancestral Acadian lineage (p-value < 0.01 & p-value < 0.001 against dwarf H_o for outliers in Cliff lake).

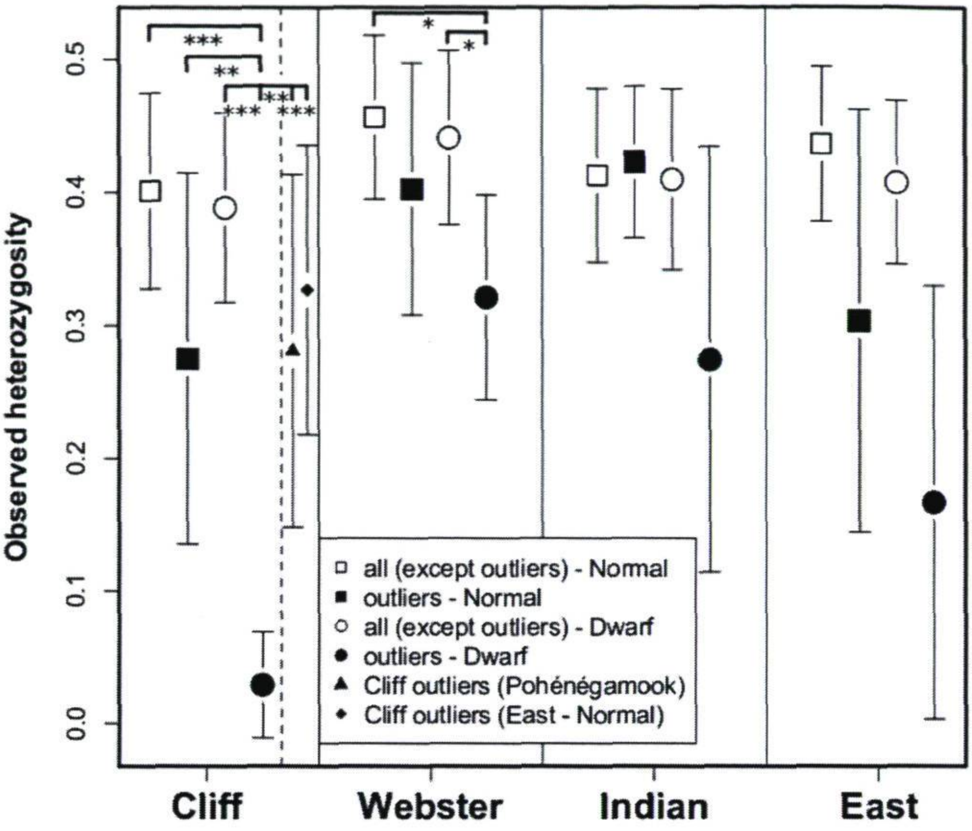

Figure 5.3 Observed heterozygosity for outlier and non-outliers loci, separately for normal and dwarf populations. Heterozygosity was also compared for Cliff Lake outlier loci in Pohénégamook, as well as East Lake normal whitefish, since they represent the ancestral Acadian lineage from which Cliff dwarf whitefish evolved. T-tests comparing mean heterozygosity between each of the groups. * *p*-value < 0.05, ** *p*-value < 0.01, *** *p*-value < 0.001.

Association study

As the fish used in the association study came from a backcross-like family ((HxD)xD), most informative polymorphic markers segregated in a 1:1 homozygous:heterozygous fashion (61 markers), while 26 informative markers segregated in a 1:2:1 fashion (26 markers). Thirty-one out of 87 markers (35%) showed evidence of segregation distortion (*q*-value < 0.2, suppl. table 5.2), which is consistent with microsatellite and AFLP data (Rogers *et al.* 2007). One SNP genotype (Myosin regulatory light chain 2 skeletal, *q*-value = 0.03) was associated with swimming activity (standard deviation of the depth preference of the individual divided by the mean depth observed, as defined by Rogers & Bernatchez 2007), three with condition factor (weight/length^3), and nine with number of gill-rakers. The strongest association was observed for growth (grams of weight gained per day) where ten SNPs were associated with this phenotype (table 5.4). In total, seven SNPs were associated with two phenotypes (aryl hydrocarbon receptor 2 alpha, EAP30 subunit of ELL complex a (Eap30a), Ferritin middle subunit, *Gasterosteus aculeatus* clone VMRC26-150D01, PREDICTED: *Danio rerio* zinc finger protein 638-like, TY3 GYPSY like LTR retrotransposon, Zebrafish DNA from clone DKEY 16P21) and nine with one phenotype (ATP binding cassette sub family E, Inosine monophosphate dehydrogenase 2, Myosin regulatory light chain 2 skeletal, NA-K transporting ATPase subunit alpha 1, Putative ISG12(3) protein, T complex protein 1 subunit epsilon, Tetraspanin 4, Triosephosphaste isomerase, Uncharacterized protein C21orf51).

Table 5.4 SNP genotypes associated with adaptive phenotypes.

Trait	SNP functional annotation	*N*	G1	G2	G3	mean (95% CI) - *G1*		mean (95% CI) -*G2*		mean (95% CI) - *G3*		*p*-val	*q*-val
Behavioral													
activity	Myosin regulatory light chain 2 skeletal	103	AA	AG	-	50.28	(43.61-56.94)	36.38	(31.71-41.04)	-	-	0.000	0.032
Physiological										-	-		
growth	T complex protein 1 subunit epsilon	178	CC	CT	TT	540.33	(318.97-761.69)	527.3	(431-623.59)	232.37	(79.71-385)	0.004	0.057
growth	Gasterosteus aculeatus clone VMRC26-150D01	178	GG	AG	-	319.15	(218.18-420.12)	549	(430.3-668.2)	-	-	0.004	0.057
growth	TY3 GYPSY like LTR retrotransposon	178	CC	CT	-	323.18	(219.71-426.65)	543.1	(426.46-659.74)	-	-	0.006	0.057
growth	aryl hydrocarbon receptor 2 alpha	178	GG	AG	-	543.1	(426.46-659.74)	323.18	(219.71-426.65)	-	-	0.006	0.057
growth	Tetraspanin 4	178	GG	TG	-	323.24	(216.7-429.76)	546.05	(427.2-664.91)	-	-	0.007	0.057
growth	PREDICTED: Danio rerio zinc finger protein 638-like	178	CC	CT	-	330.4	(233.25-427.56)	549.35	(426.22-672.48)	-	-	0.006	0.057
growth	Uncharacterized protein C21orf51	178	AA	AG	GG	-116.7	(-1190.6-957.3)	532.63	(430.36-634.9)	328.15	(260.16-396.14)	0.001	0.057
growth	EAP30 subunit of ELL complex a (Eap30a)	178	AA	AC	-	543.1	(426.46-659.74)	323.18	(219.71-426.65)	-	-	0.006	0.057
growth	Ferritin middle subunit	178	TT	CT	-	319.38	(214.7-424.06)	543.1	(426.46-659.74)	-	-	0.006	0.057
growth	Zebrafish DNA from clone DKEY 16P21	178	GG	AG	-	543.1	(426.46-659.74)	323.18	(219.71-426.65)	-	-	0.006	0.057
Morphological													
gill-rakers	ATP binding cassette sub family E	138	CC	AC	-	23.35	(23.05-23.65)	22.84	(22.49-23.2)	-	-	0.030	0.195
gill-rakers	Gasterosteus aculeatus clone VMRC26-	138	GG	AG	-	23.3	(23-23.64)	22.84	(22.5-	-	-	0.031	0.195

	150D01								23.17)				
gill-rakers	Putative ISG12(3) protein	138	GG	GT	TT	23.56	(23.06-24.05)	23	(22.75-23.24)	22.69	(22.06-23.32)	0.033	0.195
gill-rakers	TY3 GYPSY like LTR retrotransposon	138	CC	CT	-	23.34	(23.03-23.65)	22.84	(22.52-23.16)	-	-	0.030	0.195
gill-rakers	aryl hydrocarbon receptor 2 alpha	138	GG	AG	-	22.84	(22.52-23.16)	23.34	(23.03-23.65)	-	-	0.030	0.195
gill-rakers	PREDICTED: Danio rerio zinc finger protein 638-like	138	CC	CT	-	23.34	(23.04-23.64)	22.82	(22.48-23.15)	-	-	0.023	0.195
gill-rakers	EAP30 subunit of ELL complex a (Eap30a)	138	AA	CA	-	22.84	(22.52-23.16)	23.34	(23.03-23.65)	-	-	0.030	0.195
gill-rakers	Ferritin middle subunit	138	GG	AG	-	23.34	(23.03-23.65)	22.84	(22.52-23.16)	-	-	0.030	0.195
gill-rakers	Zebrafish DNA from clone DKEY 16P21	138	TT	CT	-	22.84	(22.52-23.16)	23.34	(23.03-23.65)	-	-	0.030	0.195
condition factor	NA-K transporting ATPase subunit alpha 1 precursor	182	AA	AC	CC	1.18	(1.13-1.22)	1.18	(1.167-1.2)	1.11	(1.06-1.16)	0.008	0.199
condition factor	Triosephosphaste isomerase	182	CC	AC	-	1.19	(1.17-1.21)	1.14	(1.11-1.17)	-	-	0.005	0.199
condition factor	Inosine monophosphate dehydrogenase 2	182	CC	AC	-	1.14	(1.12-1.16)	1.19	(1.16-1.23)	-	-	0.005	0.199

N refers to the sample size. G1, G2, G3 are the genotypes of the SNP with the mean and 95 % confidence interval for the phenotypic values corresponding to G1, G2 and, when applicable, G3 genotypes.

Link between genome scan and association

There was no significant trend for SNP markers associated with adaptive phenotypes to show greater evidence of reduced gene flow (higher F_{ST}) compared to all other markers (p-value > 0.05 for all five lakes, Wilcoxon rank sum test between candidate loci and all other markers, Fig. 5.4). Three genes (Triosephosphate isomerase, T complex protein 1 subunit epsilon, NA-K transporting ATPase subunit alpha) associated with an adaptive phenotype also exhibited outlier levels of divergence in Cliff, Webster and Cliff, and East Lake respectively, but this was not more than expected at random (p-value > 0.05, chi-square test for all five lakes, whereby the expected numbers of outlier loci associated with phenotypes were calculated by multiplying the percentage of outlier loci for a lake times the percentage of markers associated with phenotypes times the total number of markers surveyed).

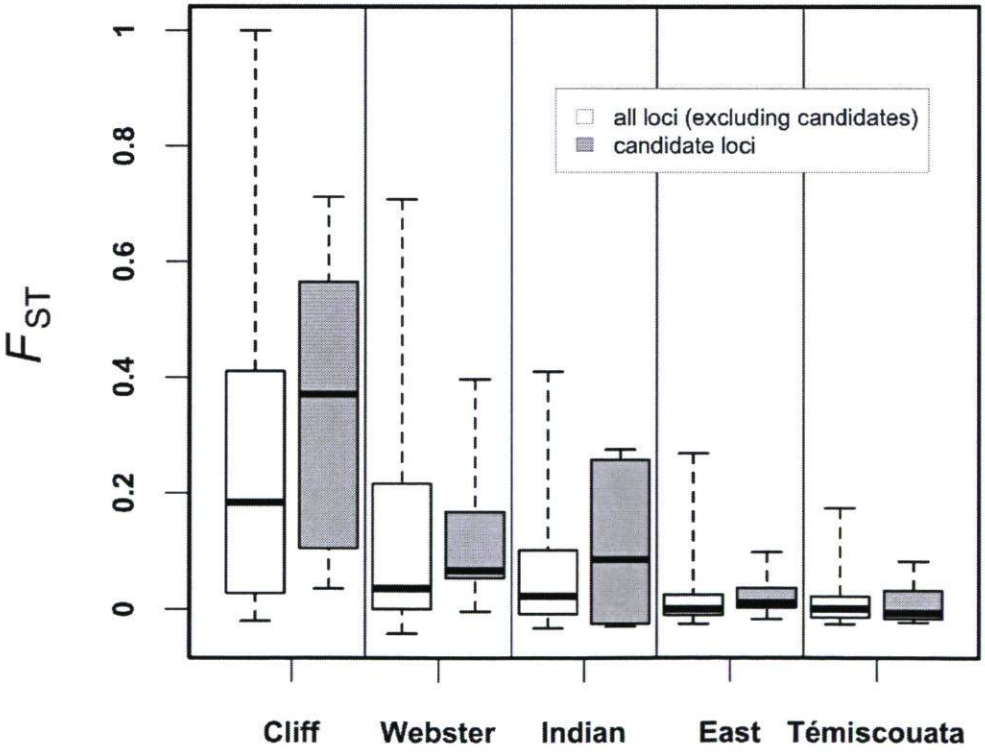

Figure 5.4 Boxplot of F_{ST} in each of the five sympatric lakes for loci associated with adaptive phenotypes (candidate loci) and all other loci. Boxes represent 95% confidence intervals of the median (dark line) and whiskers extend to data extremes.

5.6 Discussion

Genome-scan of sympatric normal and dwarf species pairs

In a recent review, Nosil and colleagues (2009) reported that the proportion of outliers in the literature varied greatly between genome scan studies, ranging from 0.4% to 25 %. According to the authors, this discrepancy between studies is best explained by variation in the number of populations and individuals examined, molecular markers employed, methods for estimating baseline neutral differentiation and criteria determining outlier status. While this evidently explains part of the variation, we believe that this may also reflect the specific intensity of divergent selection towards a new adaptive optimum and/or time since divergence varying between model systems. For instance, genome-wide SNP data for *Mus musculus musculus* and *M. m. domesticus* which have diverged for nearly 1 million years, revealed that genomic islands of differentiation represented 7.5% of autosomal regions and 90% of the X chromosome (Harr 2006). In comparison, thorough genome scan between M and S forms of *Anopheles gambiae*, where divergence is very recent and speciation ongoing, identified three islands of divergence representing less than 3 % of the genome (White *et al.* 2010). While between species comparisons will remain difficult due to the different analytical methods and markers used, cases of closely related species such as dwarf and normal whitefish permit evaluations of the effect of divergent selection according to a gradient of ongoing ecological speciation. Here, by comparing species pairs that evolved independently, we identified that the proportion of markers under divergent selection was associated with the previously recorded lake-specific potential for competition and phenotypic differentiation, as discussed below.

All five lakes studied belong to the same river basin and are situated at relatively the same altitude (from 201m for Témiscouata Lake to 322m for East Lake). Yet, none has direct connection and are therefore physically isolated one from the other. Based on the hydrological history of the area, they most likely have been colonized and isolated from one another at around the same time period, during the isostatic rebound following the removal of the 2.2km thick ice sheet cover (12 000 BP, Castric *et al.* 2001). In contrast to similar colonization times, differences in abiotic and biotic characteristics translated into different ecological landscapes (Landry *et al.* 2007; Landry & Bernatchez 2010). For example, Landry *et al.* (2007) showed that the three lakes harboring the most divergent sympatric populations (Cliff, Webster and Indian lakes) were characterized by: less habitat available due to shallower mean depth and oxygen depletion below the thermocline during the growing season, less zooplanktonic prey biomass and smaller prey size range compared to the least divergent populations (East and Témiscouata lake), which had more habitat available and more prey density. Such a prey structure has also previously been interpreted as evidence for increased potential for competition in other systems (Magnan 1988; Svanbäck & Persson 2004). Presented with this evidence, the authors concluded that resource limitation resulted in increased potential for competition and selective pressure towards optimal normal and dwarf adaptive peaks (Landry *et al.* 2007). Here, by means of genome scans, we showed that this increased potential for competition and intensity of selection at the phenotypic level is also reflected at the genetic level since more markers were identified as being potentially under the effect of divergent selection in Cliff, Webster and Indian (20 different SNPs) compared to East and Témiscouata lakes (two different SNPs). As such, we experimentally confirmed one of the premises of the genic view of speciation (Wu 2001), whereas populations that are more representative of the early steps of ecological speciation in whitefish (Témiscouata and East lakes) have fewer genetic markers under the effect of natural selection compared to more diverged species (Cliff, Webster and Indian lakes). At this point linkage information from the association family regarding the number or size of these islands of divergence is limited given the predicted haploid number of chromosomes (40). Nevertheless, given that these outliers are not in greater linkage disequilibrium than the rest of the genome (suppl. table 5.4), they should not represent a single large block in LD, but more likely, it suggests that outlier markers are situated on distinct linkage groups. In addition, the exact number of significant outliers will depend on the stringency of the multiple hypotheses testing correction and the conclusions still hold with more stringent criteria (at p-value < 0.1, number of outliers is 6, 5, 2, 1, 0 and again strongly correlated with the extent of genetic differentiation, see table 5.2).

Parallel patterns of genetic differentiation

Previous studies have provided evidence for the role of parallel phenotypic evolution in lake whitefish speciation. For example, Lu & Bernatchez (1999) and Rogers & Bernatchez (2005) have documented strong parallelism for phenotypic traits varying between dwarf and normal whitefish in independent lakes. Parallelism was also observed at the transcriptome level whereas genes were differentially expressed between normal and dwarf whitefish in independent populations more often than expected (Derome *et al.* 2006, St-Cyr *et al.* 2008). In contrast, less evidence for parallelism has been observed at the genetic level. For instance, modest, yet significant parallelism between at least two lakes out of four was identified for only six out of 48 anonymous outlier AFLP markers (Campbell & Bernatchez 2004, see also Rogers & Bernatchez 2007). Similarly, here we observed little congruence among lakes (Table 5.2 and 5.3). Only four parallel SNPs out of 96 markers displayed parallel genetic differentiation and this was not more than expected at random except in one lake pair comparison. Although these findings remain to be rigorously confirmed using a greater number of markers, they follow the same trend as previous AFLP genome scans. Consequently, in whitefish at least, parallelism at the phenotypic level (including gene expression) is not mirrored by mutations in coding regions.

Given that selection acts at the level of the phenotype, it is plausible that alternate evolutionary trajectories will be taken as selection recruits different mutations while ultimately, leading whitefish to the same ecological normal (benthic) and dwarf (limnetic) niche space in the adaptive landscape. In beach mouse for example, similar fur coloration evolved independently through alternative mutations (Steiner *et al.* 2009). Recent analyses of the factors that shape parallel hybrid zones in sculpins (*Cottus* spp*)* have also provided evidence that the genetic factors that underlie adaptive differentiation differ between populations (Nolte *et al.* 2009b). Another unequivocal case comes from Stanek and colleagues (2009). Here, the authors set up a simple selection experiment in *E. coli* as a mean to assess the genetic basis of adaptation. While they identified a beneficial mutation rising to fixation and conferring a strong fitness advantage in one population, contrary to expectation they could not find any evidence of parallel adaptation in any of the other 11 replicate populations. As such, they concluded that even for simplistic evolutionary scenarios, the genetic basis of adaptation is highly unpredictable. In our current study, different sets of outlier genes were detected in each species pairs of whitefish, which supports our current working hypothesis that many genes associated with numerous biological functions are involved in the adaptive divergence of lake whitefish.

Selection acting on standing genetic variation in dwarf whitefish.

Mounting evidence has revealed that, in the context of this adaptive radiation, directional selection acted more strongly on dwarf rather than normal whitefish (Bernatchez 2004). Namely, dwarf whitefish appear to be at an "ecological disadvantage" relative to normal, both in terms of growth, fecundity (Rogers & Bernatchez 2005) and survival (Fenderson 1964). Higher mortality rate in dwarf whitefish could be related to higher predation pressure (Kahilainen & Lehtonen 2002), while stunted dwarf growth is probably due energy trade-offs at the profit of higher swimming and metabolic activity (Rogers *et al.* 2002, Trudel *et al.* 2001). In addition, diversity of prey utilized by dwarf whitefish is also less than for normal, translating into a more specialised diet (Bernatchez *et al.* 1999) and more pronounced selection in dwarf acted on the number of gill-rakers, a trait involved in prey selection (Bernatchez 2004).

Here, we identified strong reduction in heterozygosity for outlier loci under selection in dwarf whitefish only. This was especially true in Cliff, the lake harbouring the most divergent dwarf - normal pair, where the fixation of a single allele was observed for ten out of twelve outlier markers (Fig. 5.3). Given the strong effect of selection and low level of gene flow between dwarf and normal whitefish in Cliff Lake, we were able to assess the level of ancestral polymorphism in dwarf whitefish for these outlier SNPs. These loci were polymorphic and did not show any reduction in heterozygozity in Pohénégamook Lake as well as the normal whitefish from East Lake, two populations of pure Acadian ancestry closely related to the ancestors of the extant dwarf whitefish in Cliff Lake (Lu *et al.* 2001). The pattern observed in Cliff dwarf whitefish cannot either be explained by a population bottleneck that occurred during lake colonization since heterozygosity was not reduced for the genome as a whole. Similarly, a general reduction in heterozygosity at outliers in all dwarf populations cannot be explained by a single deterministic event prior to the colonization since dwarf have evolved multiple times in independent lakes (Pigeon *et al.* 1997). Conversely, locus specific reduction in genetic variability is often a telltale sign of selective sweep and positive selection (Nielsen 2005). Accordingly, our results imply that natural selection, by differentially sorting out standing genetic variation present prior to the recent ecological speciation of whitefish, has ultimately contributed to the independent evolution of a dwarf phenotype in each lake. As such, our results represent a clear case of the predominant role of selection acting on standing genetic variation, rather than *de novo* mutations, in driving adaptive divergence in the early steps of ecological speciation.

Association

Even for a relatively small set of 87 markers, 16 SNP genotypes were significantly associated with adaptive phenotypes. Seemingly, the exact number of significant association depends on the stringency of the multiple hypotheses correction. We also believe, as discussed in the previous section, that in order to identify specific targets of speciation, combining several lines of evidence is more informative than relying on a single rigid analysis, which may in any case still present biases (Stinchcombe & Hoekstra 2008; Butlin 2008). The strongest evidence for genotype-phenotype association was with growth as ten SNPs were associated with this phenotype. This corroborates results of Rogers and Bernatchez (2007) showing that growth QTLs were the most common and therefore, the slower growth of dwarf versus normal (Trudel *et al.* 2001; Rogers and Bernatchez 2005) was likely to be under polygenic control. The gill-raker apparatus is another common adaptive trait known to differentiate benthic (few gill-rakers) and limnetic (many gill-rakers) species pairs (McPhail 1993; Bernatchez 2004). Here, this trait also appears to be under polygenic control since it was associated with nine SNPs, although the higher *q*-values (0.19) imply a greater false discovery rate. Finally, we also found a genetic basis for the differences in swimming activity and condition factor, whereas dwarf whitefish are known to be more active swimmers and have a more slender body (smaller condition factor) compared to normal fish (Trudel *et al.* 2001).

Link between genome scan and association

There was no significant overall link between elevated rates of genetic divergence and association with adaptive phenotypes (Fig. 5.4). As suggested, confounding demographic, spatial, or local effects on adaptive divergence may affect F_{ST} among environments (Beaumont & Balding 2004; Storz 2005). This appears to be corroborated by other studies looking at the relationship between selection at the genetic level and adaptive QTLs, which found either weak or no correlation between both (whitefish, Rogers & Bernatchez 2007; sunflowers, Yatabe *et al.* 2007; pea aphids, Via & West 2008; sticklebacks, Makinen *et al.* 2008). Here, variation due to the lineage origin may influence the genetic architecture sparking ecological divergence and sculpted by selection, thus generating unique evolutionary scenarios in each lake. This, in turn, could dampen our power to detect adaptive traits under selection unless we had generated independent hybrid families in each lake, which, at this point at least, is not technically feasible. Lastly, because our study probably examined only a subset of adaptive phenotypic traits, truly outlier loci may nonetheless be associated with adaptive phenotypes not yet examined (enzyme production, parasite avoidance mechanisms, mating behaviors, etc).

Moreover, relationships between adaptive phenotypes and regions of elevated genetic divergence may always be, at best, weak. Given the few recombination events in a hybrid backcross, large chromosomal regions are expected to be in linkage disequilibrium. Therefore, this can explain why using few markers, we were able to find a relatively large number of associations in the backcross family. On the other hand, in natural populations, linkage disequilibrium around a selected locus can, in theory, be much smaller thus distorting the association between divergent selection and adaptive phenotypes. This can explain why we did not find a significant link between divergent selection and adaptive phenotypes, while using a greater number of markers in the same study system, Rogers *et al.* (2007) identified a small, yet significant, connection.

Integrating data towards the identity of candidate genes

Finding outlier loci also responsible for adaptive traits and being differentially expressed allows stronger inferences than the sole use of genome scans about the underlying genes associated with ecological divergence (Stinchcombe & Hoekstra 2008; Nosil *et al.* 2009). Here, we discuss this integrated approach for the two strongest candidates, while being conscious that other candidates could also merit from a more detailed analysis (e.g.: T complex protein 1 subunit epsilon or cyclin I and Heat Shock 27kDa protein as previously explained). Admittedly, we still do not have evidence that these SNPs are the mutations responsible for an adaptive phenotype or the direct target of selection. Nevertheless, they must at least be in strong linkage with a causative mutation nearby. Our ongoing work, involving screening and sequencing BAC libraries, may provide a more in depth appreciation of the relative causative importance of regulatory and/or structural mutations for these two candidate genes.

The first case involves the Sodium/Potassium ATPase gene (Fig. 5.5a). This highly conserved protein complex is composed of two subunits in teleost fish (alpha and beta) and actively transports sodium and potassium ions in opposite directions across the plasma membrane (Lodish *et al.* 2008). Due to its essential role, it is one of the single major users of ATP, responsible for 5-40% steady-state cellular energy consumption (Ewart & Klip 1995). Furthermore, it has frequently been identified as involved in local adaptation in fish (e.g. McCairns & Bernatchez 2010). Therefore, as tradeoffs in energy allocation between high metabolic rate (dwarf) and increased growth (normal) are one of the main factors explaining the differentiation between normal and dwarf phenotypes (Trudel *et al.* 2001; St-Cyr *et al.* 2008; Rogers & Bernatchez 2007), these high ATP consumers' genes emerge as plausible candidates. Previous microarray experiments showed that the NaK-ATPase (alpha) gene is upregulated in normal whitefish compared to dwarf at the juvenile stage

(Nolte *et al.* 2009a). Here, one SNP located in the 3' UTR of the alpha subunit (Fig. 5.5a) was associated with a condition factor phenotype (CC genotype associated with smaller condition factor, Fig. 5.5b). In the genome scan, this SNP was an outlier fixed for the same allele (C) in both Cliff and East lakes (Fig. 5.5c). In fact, the C allele was statistically associated with dwarf whitefish in the association family while the A allele with normal fish (Fig. 5c) and this corroborates the fact that dwarf whitefish are more slender (smaller condition factor relative to normal whitefish, Fig. 5.5b). Finally, another SNP coding for a synonymous mutation in the subunit *beta* was outlier in Cliff Lake and in all other lakes, had an F_{ST} value above the mean F_{ST} for that lake (table 5.2).

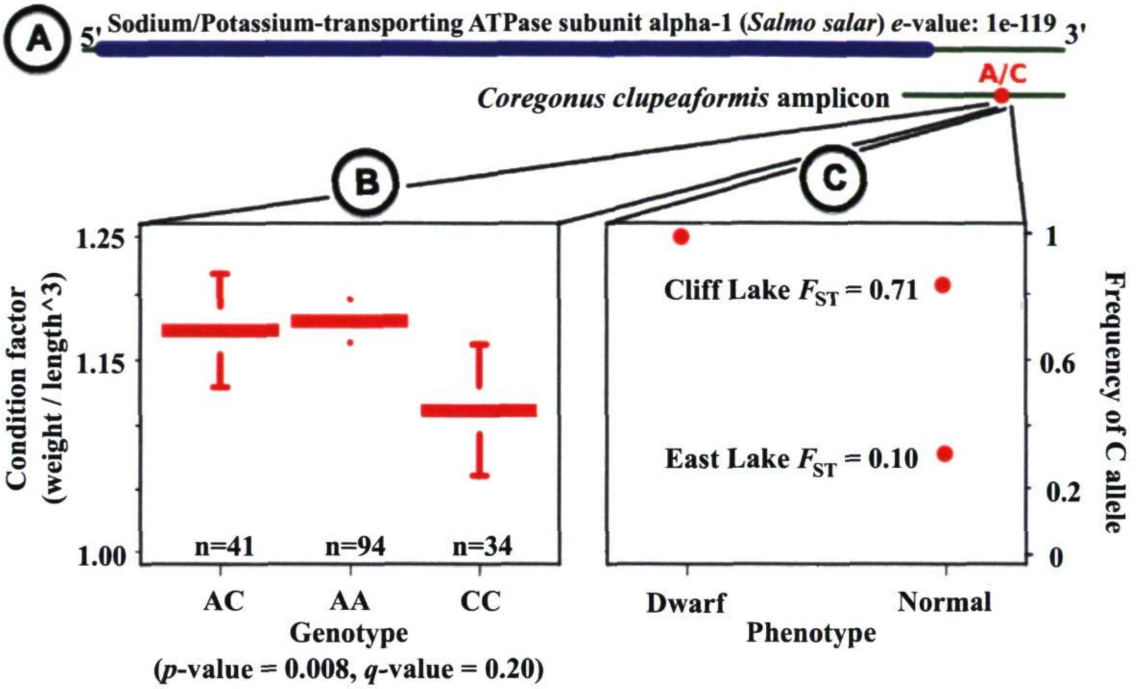

Figure 5.5 A: Genotypic characteristics of a SNP found in Sodium Potassium ATPase subunit alpha gene of whitefish (BLAST *e*-value < 1e-119). The SNP is located in the 3' UTR (in green) according to the *Salmo salar* open reading frame (in blue). B: SNP is associated with condition factor. C: In natural populations, the SNP is an outlier in both Cliff and East lakes.

The second case involves the Triosephosphate isomerase gene (Fig. 5.6a). TPI regulates the fifth step of glycolysis and is essential for efficient energy production (Lodish *et al.* 2008). Again, given the previously identified tradeoffs in energy allocation between dwarf and normal, functional or regulatory changes in genes directly involved in energy production, either through glycolysis or oxidative phosphorylation are predicted (Gershoni *et al.* 2009). Furthermore, TPI expression is upregulated in dwarf compared to normal whitefish at the juvenile stage (Nolte *et al.* 2009a) and down regulated at the adult stage (Derome *et al.* 2006). In this study, one SNP located in an intron of TPI (Fig. 5.6a) was associated with a condition factor phenotype (Fig. 5.6b). This SNP was also identified as an outlier (F_{ST} = 0.69) in Cliff Lake (Fig. 5.6c). Here, the genotype (CC) associated with a robust phenotype was more common in dwarf whitefish compared to the genotype (AC) associated with a slender phenotype and more common in normal whitefish (Fig. 5.6b-c). This counterintuitive result demonstrates again the complex relationship between selection and adaptation. Finally, previous studies identified an outlier AFLP marker with an F_{ST} value of 0.87, linked to an eQTL for TPI (Rogers & Bernatchez 2007; Derome *et al.* 2008). This eQTL for TPI was located within a regulatory hotspot comprising several genes involved in various functions. Therefore, the authors concluded that either the expression of TPI itself (*cis*-regulation) or a gene regulating TPI expression (*trans*-regulation) was under divergent selection (Bernatchez *et al.* 2010).

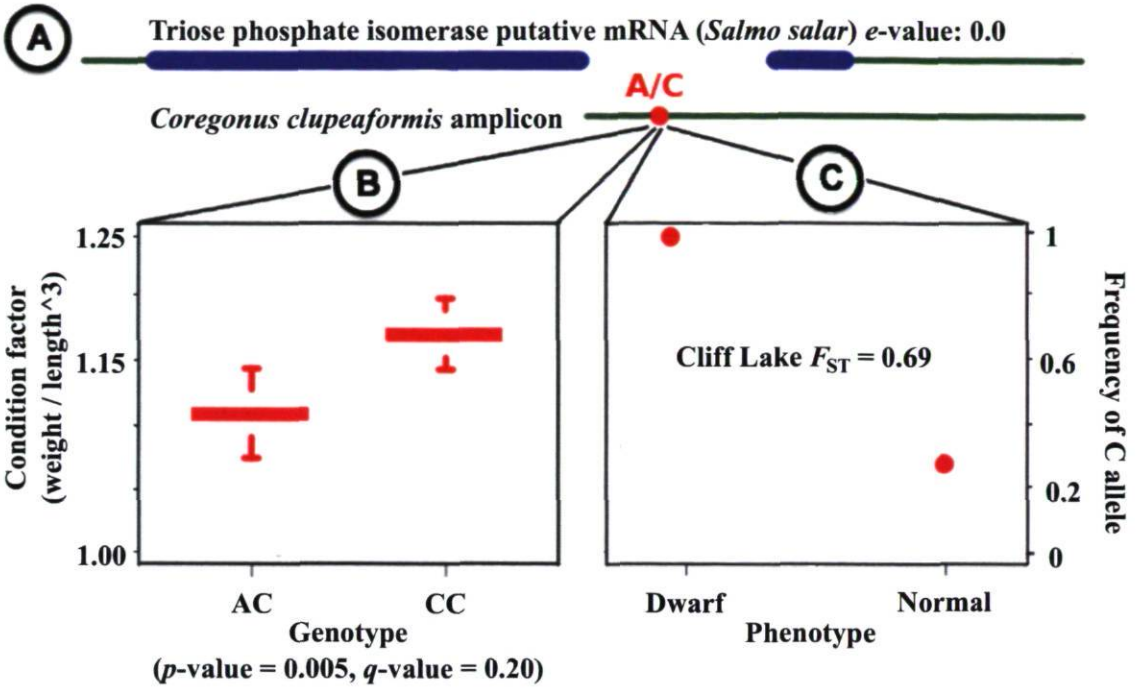

Figure 5.6 A: Genotypic characteristics of a SNP found in Triosephosphate isomerase gene of whitefish (BLAST *e*-value = 0.0). The SNP is found in an intron (in green) according to *Salmo salar* open reading frame (in blue). B: the SNP is associated with condition factor. C: In natural populations, the SNP is an outlier in Cliff Lake.

In conclusion, the identification of several markers under divergent selection or linked to adaptive phenotypes suggests that many genes, in a lake specific manner, are recruited by selection acting on standing genetic variation, during the adaptive divergence of lake whitefish. While we are accustomed to thinking of adaptive divergence and reproductive isolation being linked to single causal mutations (Colosimo *et al.* 2005; Hoekstra *et al.* 2006; Miller *et al.* 2007), this paradigm will probably soon shift towards a more intricate, yet more complete view of the genomics of speciation (Stern & Orgogozo 2008; Baxter *et al.* 2010; Counterman *et al.* 2010; Michel *et al.* 2010).

5.7 Acknowledgments

We thank V. Bourret, N. Isabel, B. Pelgas and C. Sauvage for insightful suggestions about the analysis, J. Laroche for help with bioinformatics and G. Côté with help in the laboratory. Special thanks to A. Belisle, A. Montpetit and P. Lepage at the Genome Quebec Innovation Center for genotyping troubleshooting. This research 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, a postdoctoral research stipend from the DFG to AN and NSERC discovery grant to SMR. This paper is a contribution to the research programme of Québec-Océan.

5.8 Research interest of the authors

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 in LB's laboratory, which aims to study the genomic bases of adaptive divergence in the context of a recent ongoing speciation event in lake whitefish. AN is interested in the diversity of fishes and understanding the genetic basis of adaptation. SMR studies the evolutionary mechanisms for coping with environmental change by integrating ecological genomics and quantitative genetics with field studies of natural selection. ND and LB research focuses on understanding the patterns and processes of molecular and organismal evolution as well as their significance to conservation.

Chapitre 6 : Conclusion

6.1 Conclusions générales

L'un des principaux objectifs des travaux présentés dans cette thèse fut premièrement d'apporter une contribution novatrice dans notre compréhension de la base génomique de la divergence adaptative et de l'isolement reproducteur des populations naturelles. Deuxièmement, nous avions comme objectif d'intégrer plusieurs résultats d'études précédentes et ainsi d'identifier spécifiquement des gènes candidats impliqués dans le processus de spéciation. En effet, l'inter-connectivité des processus biologiques, de l'ADN à l'ARN, la cellule, l'individu, les populations et leur environnement, définit de manière idiosyncratique le monde du vivant. Par conséquent, la meilleure stratégie vers l'identification des bases génétiques des changements évolutifs reste encore de combiner des données expérimentales ciblant différents niveaux de complexité biologique. Par exemple, comme il a déjà été suggéré (Vasemägi & Primmer 2005; Stinchcombe & Hoekstra 2008), cela est possible en intégrant des résultats au niveau de la variation de l'ADN, de l'expression des gènes et de la variation phénotypique, tout en analysant les évidences de sélection à chacun de ces niveaux. Ces gènes candidats peuvent par la suite faire lieu de confirmations ou d'études fonctionnelles plus approfondies.

Les populations de corégone représentent un système d'étude particulièrement attrayant dans l'étude du processus de spéciation, mais aussi de manière générale en écologie, évolution, et génomique. Ainsi, la spéciation écologique récente des formes naines et normales, son évolution indépendante dans plusieurs lacs distincts et l'écologie relativement bien documentée de ces formes en nature représentent tous des avantages de ce système biologique. De plus, le développement de ressources génomiques pour le corégone lui-même ainsi que pour le saumon atlantique, une espèce proche, permet déjà d'entrer dans l'ère de l'écogénomique fonctionnelle et en permettra encore davantage dans un futur très proche.

Lors du premier chapitre expérimental, nous nous sommes intéressés à étudier la divergence d'expression entre les corégones nains et normaux à deux stades développementaux. Grâce aux biopuces, l'analyse de l'expression de gènes n'a identifié que peu de gènes candidats chez les embryons, corroborant les études antérieures montrant une absence de divergence écologique entre corégones nains et normaux au stade larvaire (Chouinard & Bernatchez 1998). A l'opposé, la divergence d'expression liée à des caractères adaptatifs chez les juvéniles semble être dictée par la sélection naturelle divergente. Ces résultats suggèrent donc que l'isolement reproducteur postzygotique extrinsèque, dépendant de l'environnement, peut être plus important que des obstacles intrinsèques.

Par la suite, nous avons montré que très peu de gènes chez les embryons différaient quant à leur expression moyenne entre les parents et les hybrides, ce qui contrastait avec les poissons juvéniles. Nous avons aussi trouvé que la non-additivité de l'expression expliquait une plus grande fraction des patrons d'expression chez les rétrocroisements que chez les hybrides F1, ce qui implique une dérégulation accrue de l'expression pour ces hybrides rétrocroisés. De manière plus spécifique, chez les rétrocroisements, l'expression de gènes développementaux impliqués dans le repliement des protéines et la traduction de l'ARN messager était sévèrement dérégulée. Ainsi, cette dérégulation de l'expression chez les hybrides s'ajoute aux autres facteurs précédemment identifiés comme contribuant à l'isolement reproducteur des jeunes espèces de grand corégones.

Fort de ces premiers résultats démontrant des patrons d'expression anormaux importants chez les hybrides de deuxième génération tels qu'attendus selon la théorie, nous avons suivi plus en détails le développement de ceux-ci lors du troisième chapitre. Ainsi, puisqu'un pourcentage relativement important d'hybrides rétrocroisés se développait de manière anormale, nous avons donc quantifié le niveau de différentiation d'expression associé à cette manifestation d'isolement reproducteur post-zygotique. Nous avons observé les patrons d'expression pour tous les gènes clés du développement embryonnaire, toujours identifiés par homologie avec des données fonctionnelles obtenues pour le poisson zèbre (Amsterdam *et al.* 2004). Nos résultats montrent de manière convaincante une dérégulation totale du transcriptome, ainsi qu'un lien explicite entre la dérégulation des gènes essentiels du développement et l'isolement reproducteur post-zygotique. Bien entendu, le défi reste comme toujours d'établir un lien de causalité entre la divergence des populations et la divergence d'expression, et ceci continue d'être une limite inévitable des études d'expression génique en spéciation (Noor & Feder 2006).

Ce fut grâce aux travaux pionniers de Derome et collaborateurs (2006) que l'on a pu réellement commencer à identifier des gènes candidats impliqué dans la divergence et l'isolement reproducteur chez le grand corégone. Similairement, nos expériences de biopuces ont permis de rechercher, à l'échelle du génome entier, des différences du niveau de transcription et donc des gènes candidats. Comme nous en avons discuté dans les chapitres un et quatre, il faut réaliser que cette technique est probablement vouée à disparaître, remplacée par les nouvelles générations de séquençage (Ledford 2008). De plus, il est important de noter que la variation du niveau de transcription n'est qu'un des nombreux niveaux de contrôle existant entre l'ADN et le phénotype. Ainsi, dans une approche globale, il faudra ultimement quantifier de manière plus fonctionnelle les variations d'expression de gènes tant au niveau des modifications épigénétiques et post-transcriptionnelles qu'au niveau protéique.

Lors du quatrième chapitre, nous avons utilisé la technique de pyroséquencage 454 afin de séquencer le transcriptome d'individus nains, normaux et hybrides. Cette approche permet d'obtenir directement des donnés de séquences et d'expression (Renaut *et al.* 2010, Jeukens *et al.* 2010). Ainsi, des résultats forts prometteurs sont ressortis de cette étude. Comme démontré dans l'éditorial du numéro spécial de la revue *Molecular Ecology*, les nouvelles générations de séquençage sont en voie de révolutionner l'étude de la génomique de la spéciation chez les espèces non modèles (Tautz *et al.* 2010). Lors de notre étude, plusieurs milliers de gènes et SNPs ont été séquencés, formant ainsi une banque de données génomiques formidable pour le grand corégone. De plus, en mesurant la divergence de fréquences d'allèles entre les groupes directement à partir des données de séquences, ce type analyse, bien que naïve, constitue une démarche prometteuse dans les études de balayage génomique afin d'identifier les gènes ciblés par la sélection. De la sorte, cela nous a permis d'observer des différences de fréquences d'allèles beaucoup plus élevées que la normale pour les gènes du métabolisme énergétique. Par ailleurs, cette approche, qui combine directement séquençage et génotypage, est, selon moi, voué à un avenir florissant. Puisqu'il est maintenant facile et peu coûteux de développer une étiquette génétique spécifique (*DNA barcode*) pour chaque individu (Roche 2010), il ne reste donc qu'un pas à franchir avant de calculer directement, à partir de données de séquençage, un indice de fixation (ie. F_{ST}) et ainsi d'inférer l'effet de la sélection sur plusieurs milliers de marqueurs de type SNP, indel, CNV ou même sur le génome en entier (ex. Yi *et al.* 2010). Similairement, ce type d'approche de séquençage/génotypage, appliqué dans une population de recombinants hybrides permet de cartographier un grand nombre de marqueurs en une seule expérience (ex. Baird *et al.* 2008).

Également lors de ce chapitre, l'intégration de données, cette fois-ci de séquençage 454 à celles d'expression de gènes, d'études QTL et eQTL, nous a apporté des réponses concrètes sur les bases génomiques de la spéciation. En conséquence, tout porte à croire que les gènes impliqués dans le métabolisme et plus particulièrement la glycolyse et la phosphorylation oxydative (gènes mitochondriaux et nucléaires) sont impliqués directement dans le processus de spéciation. Plusieurs évidences récentes, tant empiriques que théoriques, démontrent le bien fondé de cette hypothèse qui veut que l'émergence d'incompatibilités cytonucléaires en rapport avec le métabolisme énergétique soit la cause de l'isolement reproducteur (Burton *et al.* 2006, Gershoni *et al.* 2009, Lane 2009, Ballard & Melvin 2010).

Finalement, de manière imprévue, nous avons identifié 44 séquences contiguës (contigs) annotées à des éléments transposables qui étaient composés de manière prédominante de séquences hybrides. Ceci indiquerait donc une augmentation de l'activité des éléments transposables pouvant expliquer la baisse de la valeur adaptative des hybrides documentée par le passé. Le rôle de ce phénomène dans la divergence des populations et la spéciation est relativement bien connu chez les plantes (McClintock 1984), mais il est peu documenté et probablement sous-évalué dans le monde animal (Coyne & Orr 2004). Il me permet aussi de mettre en lumière deux idées fondamentales qui régissent l'origine de l'architecture génomique et qui ne devraient jamais être négligées : 1) les génomes sont composés, du moins en partie, d'une mosaïque d'éléments égoïstes ayant co-évolué entre eux (Dawkins 1976) et probablement de par ce fait 2) les génomes sont intrinsèquement instables (Lynch 2007b, Presgraves 2010).

Lors du dernier chapitre expérimental de ma thèse, nous avons effectué de manière rigoureuse des analyses de balayage génomique en utilisant des marqueurs SNPs que nous avions préalablement développés afin de caractériser l'effet de la sélection naturelle sur la variation génétique pour des gènes aux fonctions connues. Ici, en utilisant plus de 100 marqueurs SNPs récemment développés pour le grand corégone (*Coregonus clupeaformis*), nous avons caractérisé l'effet de la sélection naturelle dans cinq lacs distincts. Une proportion différente de SNPs aberrants (*outliers*) fut identifiée (entre 0 et 12%) et ceci en relation avec l'intensité prévue des interactions compétitives entre les corégones nains et normaux propre à chaque lac. Ainsi, nous avons confirmé expérimentalement l'une des prémisses de la spéciation génique (*sensu* Wu 2001) où, au début du processus de spéciation, le flux de gènes entre populations étant encore important, peu de marqueurs devraient évoluer sous l'effet de la sélection par rapport à un stade plus avancé dans le processus de spéciation. De plus, la forte réduction de l'hétérozygotie chez les nains pour les loci *outliers* permet de démontrer que la sélection semble agir en triant les allèles présents dans la population ancestrale afin de produire un phénotype adapté à la niche écologique limnétique (corégones nains) des lacs. Conséquemment, ces résultats amènent une évidence de plus dans le rôle de la variation génétique présente *ab initio* par rapport aux mutations *de novo* dans la divergence adaptative des populations (Barrett & Schluter 2008).

Par la suite lors de cette étude, en utilisant le même ensemble de marqueurs SNPs, nous avons testé leur association avec neuf phénotypes adaptatifs chez les poissons hybrides issus de rétrocroisements. Quatre caractères adaptatifs (croissance, activité, nombre de branchiténies et facteur de condition) étaient associés à 16 gènes différents. Nous n'avons observé aucune tendance significative entre le niveau élevé de différenciation génétique (F_{ST}) et l'association avec des phénotypes adaptatifs. De même, nous n'avons pas observé de signatures génétiques de divergence adaptative parallèle. Au contraire, les résultats spécifiques à chaque lac sous-tendent l'évolution indépendante de ceux-ci. Ainsi, puisque la sélection n'agit directement que sur le phénotype lui-même, différents mécanismes moléculaires peuvent être recrutés et aboutiront ultimement aux mêmes pics phénotypiques nain et normal dans le paysage adaptif des corégones. Finalement, lors de ce chapitre, l'utilisation intégrée de données phénotypiques, transcriptomiques et fonctionnelles mena à la découverte de certains gènes candidats (ex. Sodium Potassium ATPase et Triosephosphate isomerase) impliqués dans la divergence écologique du grand corégone.

En conclusion, nous avons pu documenter l'effet de la sélection agissant de manière plus prononcée sur le phénotype nain et aussi de manière spécifique et indépendante dans chaque lac. Ceci vient donc appuyer notre hypothèse actuelle qui stipule que de nombreux gènes sont impliqués dans la spéciation écologique du grand corégone.

6.2 Perspectives

L'avènement de la synthèse évolutive moderne (Huxley 1942) a réconcilié et intégré le travail de différentes spécialités de la biologie afin d'accomplir un progrès phénoménal dans l'étude de l'origine et de l'évolution des espèces (Coyne & Orr 2004). Ainsi récemment, plusieurs études de cas ont réussi à faire le lien explicite entre la variation à un gène précis responsable de l'isolement reproducteur et l'effet de la sélection (Noor & Feder 2006, Schluter 2009, Presgraves 2010). Néanmoins, ces études ne représentent que quelques cas de lien de causalité et il est possible qu'elles soient biaisées vers certains changements génétiques plus simples et donc plus faciles à identifier. Il existe donc toujours un besoin incontestable d'études empiriques avant de pouvoir généraliser le nombre de gènes et le type de variation génétique à la base de la spéciation.

Par ailleurs, le paradigme du "gène de la spéciation" devra probablement évoluer afin d'expliquer les causes et les conséquences des patrons de divergence observés lors du processus de spéciation et se répercutant probablement sur tout le génome (Presgraves 2010, Counterman *et al.* 2010, Michel *et al.* 2010). Plusieurs études récentes montrent que le disfonctionnement hybride, et donc l'isolement reproducteur, peuvent aussi être la conséquence de facteurs qui ne sont pas relié directement à l'écologie de l'espèce tel qu'on l'entend dans un sens classique, mais plutôt à son environnement génomique. Par exemple, la course à l'armement entre les éléments égoïstes ou les pathogènes intracellulaires et le génome de l'hôte ou encore la duplication et le mouvement stochastique des gènes dans le génome ont aussi leur rôle à jouer dans la mise en place de barrière à la reproduction (Presgraves 2010). De plus, des changements dans les portions non traduites du génome et liés à la préservation de l'intégrité de la chromatine, comme les changements épigénétiques ou l'expression de petits ARN non codants, pourraient avoir un rôle fort important dans l'adaptation et la divergence des populations, mais sont encore très peu documentés (Michalak 2009). Dans toutes ces situations, l'isolement reproducteur pourrait découler du fait que les génomes sont intrinsèquement instables et donc propices à diverger advenant un changement environnemental. Les organismes sont donc, tout comme la reine rouge de Lewis Carroll, voués à une course évolutive perpétuelle, parfois simplement afin de rester à leur place respective dans le paysage adaptif.

D'un côté plus méthodologique, tant l'analyse des biopuces, que des données de séquençage 454 illustre le rôle de l'exploration de données (*data mining*) dans notre vision de l'avenir de la science et des technologies. Ainsi, l'augmentation exponentielle de la quantité de données informatiques et biologiques disponibles, des capacités de stockage et de la puissance des ordinateurs permet non seulement de tester nos hypothèses de base, mais de plus, d'en ressortir *a posteriori*, des patrons inattendus ou des hypothèses nouvelles. Ce type d'approche qui consiste à collecter des quantités massives de données afin par la suite, d'en ressortir des tendances, transforme notre conception de la méthode scientifique et est certainement voué à un avenir prometteur.

Finalement, selon plusieurs, les prochaines années constituent une période cruciale et excitante de recherche en génomique de la spéciation puisque les modèles d'étude ainsi que la quantité d'information disponible se multiplie constamment (Mayr 2004b, Schluter 2009, Presgraves 2010, Bernatchez *et al.* 2010). Ceci devrait donc permettre de faire le lien entre les études où l'on connait bien le rôle de la sélection à promouvoir la divergence et l'isolement reproducteur, mais peu sur les changements génétiques, et d'autre part, les études où l'on possède une bonne connaissance des gènes impliqués dans l'isolement reproducteur, mais peu sur l'histoire évolutive de ces populations. Le grand corégone constitue encore et toujours un modèle important où la quantité d'information génomique continue à croître très rapidement, permettant ainsi d'identifier les changements génomiques à la base de la spéciation dans un contexte écologique et évolutif bien défini.

Chapitre 7 : Bibliographie

(Notez que pour les références ayant plus de dix auteurs, seul les trois premiers sont inscrits, suivis de l'abréviation "et al.")

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* (eds Marken CL), pp. 415-432, Academic Press, London, UK.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* **25**, 3389-3402.

Amsterdam A, Nissen RM, Sun Z, Swindell EC, Farrington S, Hopkins N (2004) Identification of 315 Genes Essential for Early Zebrafish Development. *Proceedings of the National Academy of Science USA*, **101**, 12792-12797.

Anderson E (1949) *Introgressive Hybridization*. J. Wiley, New York, USA.

Arbeitman MN, Furlong EEM, Imam F, Johnson E, Null BH, Baker BS, Krasnow MA, Scott MP, Davis RW, White KP (2002) Gene Expression During the Life Cycle of *Drosophila melanogaster*. *Science*, **297**, 2270-2275.

Arendt J, Reznick D (2008) Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? *Trends in Ecology & Evolution*, **2**, 26-32.

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.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, **3**, e3376.

Ballard JWO, Melvin RG (2010) Linking the mitochondrial genotype to the organismal phenotype. *Molecular Ecology*, **19**, 1523–1539.

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

Barrett R, Schluter D (2008) Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, **23**, 38-44.

Bartlett MS (1937) Properties of sufficiency and statistical tests. *Proceedings of the Royal Society B-Biological Sciences*, **160**, 268-282.

Barton NH (2001) The role of hybridization in evolution. *Molecular Ecology*, **10**, 551-568.

Bateson W (1909) Heredity and variation in modern lights. In: *Darwin and modern science* (ed. A. C. Seward), pp. 85-101. Cambridge University Press, Cambridge, UK.

Bauer DF (1972) Constructing confidence sets using rank statistics. *Journal of the American Statistical Association*, **67**, 687-690.

Baxter SW, Nadeau N, Maroja L, Wilkinson P, Counterman BA, *et al.* (2010) Genomic Hotspots for adaptation: population genetics of Mullerian mimicry in the *Heliconius melpomene* clade. *PLoS Genetics,* **6**, e794.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969-980.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society B: Biological Sciences*, **263**, 1619-1626.

Bell G (2008) *Selection: The mechanisms of evolution*, 2nd ed. Oxford University Press, New York, USA.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological* **57**, 289-300.

Bennett S (2004) Solexa Ltd. *Pharmacogenomics,* **5**, 433-438.

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

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.

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

Bolnick DI, Near TJ (2005) Tempo of hybrid inviability in centrarchid fishes (Teleostei: Centrachidae). *Evolution*, **59**, 1754-1767.

Boursot P, Din W, Anand R, Darviche D, Dod B, VonDeimling F, Talwar GP, Bonhomme F (1996) Origin and radiation of the house mouse: Mitochondrial DNA phylogeny. *Journal of Evolutionary Biology*, **9**, 391-415.

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

Brem RB, Kruglyak L (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Science USA*, **102**, 1572-1577.

Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science,* **296**, 752-755.

Brown TA (2007) *Genome 3*. Garland Science Publishing, New York, USA.

Burke JM, Arnold ML (2001) Genetics and the Fitness of Hybrids. *Annual Review of Genetics*, **35**, 31-52.

Burton RS, Ellison CK, Harrison JS (2006) The Sorry State of F2 Hybrids: Consequences of Rapid Mitochondrial DNA Evolution in Allopatric Populations. *American Naturalist*, **168**, S14-S24.

Butlin RK (2008) Population genomics and speciation. *Genetica,* **138**, 409-418.

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.

Carroll SB (1995) Homeotic genes and the evolution of arthropods and chordates. *Nature*, **376**, 479-485.

Carroll SB (2008) Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell*, **134**, 25-36.

Castric V, Bonney F, Bernatchez L (2001) Landscape Structure and Hierarchical Genetic Diversity in the Brook Charr, *Salvelinus fontinalis*. *Evolution*, **55**, 1016-1028.

cGRASP (Consortium for Genomic Research on All Salmonids Programs) *http://www.cgrasp.org/* Accédé en août 2010.

Chen K, Rajewsky N (2007) The evolution of gene regulation by transcription factors and microRNAs. *Nature Reviews Genetics,* **8**, 93-103.

Chouinard A, Bernatchez L (1998) A study of trophic niche partitioning between larval populations of reproductively isolated white- fish (*Coregonus* sp.) ecotypes. *Journal of Fish Biology*, **53**, 1231-1242.

Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G Jr, Dickson M, Grimwood J, Schmutz J, Myers RM, Schluter D, Kingsley DM (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science*, **307**, 1928-1933.

Cooper TF, Roze DE, Lenski RE (2003) Parallel changes in gene expression after 20,000 generations of evolution in *Escherichia coli. Proceedings of the National Academy of Science USA*, **100**, 1072-1777.

Counterman BA, Araujo-Perez F, Hines HM *et al.* (2010) Genomic Hotspots for Adaptation: The Population Genetics of Mullerian Mimicry in *Heliconius erato*. *PLoS Genetics,* **6**, e1000796.

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

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

Cui XQ, Affourtit J, Shockley KR, Woo Y, Churchill GA (2006) Inheritance patterns of transcript levels in F-1 hybrid mice. *Genetics*, **174**, 627-637.

Darwin C (1859*) The Origin of Species*. John Murray, London.

Dawkins R (1976) *The Selfish Gene.* Oxford University Press, New York City, USA.

De Queiroz K (1998) The General Lineage Concept of Species, Species Criteria, and the Process of Speciation. *Endless Forms Species and Speciation* (eds Howard DJ, Berlocher SH, eds), pp. 57-75. Oxford University Press, New York, USA

Derome N, Bernatchez L (2006) The transcriptomics of ecological convergence between 2 limnetic coregonine fishes (Salmonidae). *Molecular Biology and Evolution,* **23**, 2370-2378.

Derome N, Bougas B, Rogers SM, Whiteley AR, Labbe A, Laroche J, Bernatchez L (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.

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

Dettman JR, Sirfusingh C, Kohn LM, Anderson JB (2007) Incipient speciation by divergent adaptation and antagonistic epistastis in yeast. *Nature*, **447**, 585-588.

DeVicente MC, Tanksley SD (1993) QTL Analysis of Transgressive Segregation in an Interspecific Tomato Cross. *Genetics*, **134**, 585-596.

Dieckmann U, Doebeli M (1999) On the origin of species by sympatric speciation. *Nature*, **400**, 354-357.

Dieckmann U, Doebeli M, Metz JAJ, Tautz D (2004) *Adaptive speciation*. Cambridge University Press, Cambridge, UK.

Dierenger D, Schlötterer C (2003) MICROSATELLITE ANALYSER (MSA): a platform independent analysis tool for large microsatellite data sets. *Molecular Ecology Notes*, **3**, 167-169.

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

Dobzhansky T (1937) *Genetics and the Origin of Species*. Columbia University Press, New York, USA.

Drmanac R, Sparks AB, Callow MJ, *et al.* (2010) Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science*, **327**, 78-81.

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.

Eid J, Fehr A, Gray J (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133-8.

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.

Endler JA (1977) *Geographic variation, speciation, and clines*. Princeton University Press, Princeton, USA.

Evans BJ, Kelley DB, Tinsley RC, Melnick DJ, Cannatella DC (2004) A mitochondrial DNA phylogeny of African clawed frogs: phylogeography and implications for polyploid evolution. *Molecular Phylogenetics and Evolution*, **33**, 197-213.

Ewart HS, Klip A (1995) Hormonal regulation of the Na+-K+-ATPase: mechanisms underlying rapid and sustained changes in pump activity. *American Journal of Physiology - Cell Physiology,* **269**, C295.

Excoffier L, Foll M (2009) Detecting loci under selection in a hierarchically structured populationHierarchical test of selection. *Heredity*, **103**, 285-298

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.

Feder ME, Mitchell-Olds T (2003) Evolutionary and ecological functional genomics. *Nature Reviews Genetics,* **4**, 649–655.

Fenderson O (1964) Evidence of subpopulations of lake whitefish, *Coregonus clupeaformis*, involving a dwarfed form. *Transactions of the American Fishery Society*, **93**, 77-94.

Ferea TL, Botstein D, Brown PO, Rosenzweig RF (1999) Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 9721-9726.

Foll M, Gaggiotti O (2008) Identifying the Environmental Factors That Determine the Genetic Structure of Populations. *Genetics*, **174**, 875-891.

Fontanillas P, Landry CR, Wittkopp PJ, Russ C, Gruber JD, Nusbaum C, Hartl DL (2010) Key considerations for measuring allelic expression on a genomic scale using high-throughput sequencing. *Molecular Ecology*, **19** (Suppl 1), 212-227.

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

Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ (2004) Complex SNP-related sequence variation in segmental genome duplications. *Nature Genetics*, **36**, 861-866.

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

Futuyma DJ (1986) *Evolutionary biology*. Sinauer Associates, Sunderland, USA

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

Gibson G, Riley-Berger R, Harshman L, Kopp A, Vacha S, Nuzhdin S, Wayne M (2004) Extensive sex-specific nonadditivity of gene expression in *Drosophila melanogaster*. *Genetics,* **167**, 1791-1799.

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

Gonzalez JR, Armengol L, Sole X, Guino E, Mercader JM, Estivill X, Moreno V (2007) SNPassoc: an R package to perform whole genome association studies. *Bioinformatics*, **23**, 644-645.

Gow JL, Peichel CL, Taylor EB (2007) Ecological selection against hybrids in natural populations of sympatric threespine sticklebacks. *Journal of Evolutionary Biology,* **20**, 2173-2180.

Haeckel E (1866) *Generelle Morphologie der Organismen.* Bd. 2: Allgemeine Entwickelungsgeschichte der Organismen. Georg Reimer, Berlin, Dl.

Haerty W, Singh RS (2006) Gene regulation divergence is a major contributor to the evolution of Dobzhansky-Muller incompatibilities between species of *Drosophila. Molecular Biology and Evolution*, **23**, 1707-1714.

Hall BK (1997) Phylotypic stage or phantom: is there a highly con- served embryonic stage in vertebrates? *Trends in Ecology and Evolution*, **12**, 461-463.

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

Harr B (2006) Genomic islands of differentiation between house mouse subspecies. *Genome Research*, **16**, 730-737.

Hayes B, Laerdahl JK, Lien S, Moen T, Berg P, Hindar K, Davidson WS, Koop BF, Adzhubei A, Hoyheim B (2007) An extensive resource of single nucleotide 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.

Holm S (1979) A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics,* **6**, 65-70.

Hosack DA, Dennis JrG, Sherman BT, Lane HC, Lempicki RA (2003) Identifying Biological Themes within Lists of Genes with EASE. *Genome Biology*, **4**, R70.

Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protocols,* **4**, 44-57.

Hughes KA, Ayroles JF, Reedy MM, Drnevich JM, Rowe KC, Ruedi EA, Cáceres CE, Paige KN (2006) Segregating Variation in the Transcriptome: Cis Regulation and Additivity of Effects. *Genetics*, **173**, 1347-1355.

Huxley JS (1942) *Evolution: the modern synthesis.* Allen & Unwin, London, UK

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature,* **409**, 860-921.

Irie N, Sehara-Fujisawa A (2007) The vertebrate phylotypic stage and an early bilaterian-related stage in mouse embryogenesis defined by genomic information. *BMC Biology,* **5**, 1.

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

Jeukens J, Renaut S, St-Cyr J, Nolte AW, Bernatchez L (2010) The transcriptomics of sympatric dwarf and normal lake whitefish (*Coregonus clupeaformis* spp., Salmonidae) divergence as revealed by next-generation sequencing. *Mol. Ecol. In Press*

Joron M, Papa R, Beltrán M, *et al.* (2006) A Conserved Supergene Locus Controls Colour Pattern Diversity in *Heliconius* Butterflies, *PLoS Biology*, **4**, e303.

Kahilainen K, Lehtonen (2002) Brown trout (*Salmo trutta* L.) and Arctic charr (*Salvelinus alpinus* (L.)) as predators on three sympatric whitefish (*Coregonus lavaretus* (L.)) forms in the subarctic Lake Muddusjärvi. *Ecology of Freshwater Fish*, **11**, 158-167.

Kelleher ES, Barbash DA (2010) Expanding islands of speciation. *Nature,* **465**, 1019-1020.

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

Kendziorski C, Irizarry RA, Chen KS, Haag JD, Gould MN (2005) On the utility of pooling biological samples in microarray experiments. *Proceedings of the National Academy of Science USA*, **102**, 4252-4257.

Kerr MK, Afshari CA, Bennett L, Bushel P, Martinez J, Walker NJ, Churchill GA (2002) Statistical analysis of a gene expression micro- array experiment with replication. *Statistica Sinica*, **12**, 203-217.

Kerr MK, Martin M, Churchill GA (2000) Analysis of variance for gene expression microarray data. *Journal of Computational Biology,* **7**, 819-837.

Kimmel CB, Ballard WW, Kimmel SE, Ullmann B, Schilling TF (1995) Stages of Embryonic Development of the Zebrafish. *Developmental Dynamics*, **203**, 253-310.

King MC, Wilson AC (1975) Evolution at two levels: molecular similarities and biological differences between humans and chimpanzees. *Science*, **188**, 107-116.

Kottelat M, Freyhof J (2007) *Handbook of European freshwater fishes*. Publications Kottelat, Cornol, CH.

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

Landry L, Bernatchez L (2010) Role of epibenthic resource opportunities in the parallel evolution of lake whitefish species pairs (*Coregonus* sp.). *Journal of Evolutionary Biology,* **23**, *in press.*

Landry CR, Hartl DL, Ranz J (2007a) Genome clashes in hybrids: insights from gene expression. *Heredity,* **99**, 483-493.

Landry L, Vincent WF, Bernatchez L (2007b) Parallel evolution of lake white fish dwarf ecotypes in association with limnological features of their adaptive landscape. *Journal of Evolutionary Biology,* **20**, 971-984.

Landry CR, Wittkopp PJ, Taubes CH, Ranz JM, Clark AG, Hartl DL (2005) Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of *Drosophila. Genetics*, **171**, 1813-1822.

Landry L, Bernatchez L (2010) Role of epibenthic resource opportunities in the parallel evolution of lake whitefish species pairs (*Coregonus* sp.). *Journal of Evolutionary Biology,* **23**, *in press.*

Lane N (2009) On the origins of bar codes. *Nature*, **462**, 272-274.

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

Lee C-T, Risom T, Strauss WM. (2007) Evolutionary Conservation of MicroRNA Regulatory Circuits: An Examination of MicroRNA Gene Complexity and Conserved MicroRNA-Target Interactions through Metazoan Phylogeny. *DNA and Cell Biology,* **26**, 209-218.

Lee HY, Chou JY, Cheong L, Chang NH, Yang SY, Leu JY (2008) Incompatibility of nuclear and mitochondrial genomes causes hybrid sterility between two yeast species. *Cell,* **135**, 1065-1073.

Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics,* **3**, e161.

Le Rouzic A, Carlborg O (2007) Evolutionary potential of hidden genetic variation. *Trends in Ecology and Evolution*, **23**, 33-37.

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

Lipson D, Raz T, Kieu A, Jones DR, Giladi E, Thayer E, Thompson JF, Letovsky S, Milos P, Causey M (2009) Quantification of the yeast transcriptome by single-molecule sequencing. *Nature Biotechnology*, **27**, 652-659.

Lodish H, Berk A, Kaiser CA, Krieger M, Scott MA, Bretscher A, Ploegh H, Matsudaira P (2008). *Molecular and Cell Biology,* 6th edn. W.H. Freeman and Company, New York, USA.

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, 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.

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 (2007a) The evolution of genetic networks by non-adaptive processes. *Nature Reviews Genetics*, **8**, 803-813.

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

Mackay TFC (2001) Quantitative trait loci in Drosophila. *Nature Reviews Genetics,* **2**, 11-20.

Magnan P (1988) Interactions between Brook charr, *Salvelinus fontinalis*, and non salmonid species-ecological shift, morphological shift, and their impact on zooplankton communities. *Canadian Journal of Fisheries and Aquatic Sciences*, **45**, 999-1099.

Mäkinen HS, Cano JM, Merilä J (2008) Identifying footprints of directional and balancing selection in marine and freshwater threespine stickleback (*Gasterosteus aculeatus*) populations. *Molecular Ecology,* **17**, 3565-3582.

Mallet J (2006) What does *Drosophila* genetics tell us about speciation? *Trends in Ecology and Evolution*, **21**, 386-393.

Mallet J (2007) Hybrid speciation. *Nature*, **446**, 279-283.

Malone JH, Chrzanowski TH, Michalak P (2007) Sterility and gene expression in hybrid males of *Xenopus laevis* and *X. muelleri*. *PLoS ONE*, **2**, e781.

Mardis E (2008) The impact of next-generation sequencing technology on genetics, *Trends in Ecology and Evolution,* **24**, 133-141

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

Mavárez J, Audet C, Bernatchez L (2009) Major disruption of gene expression in hybrids between young sympatric anadromous and resident populations of brook charr (*Salvelinus fontinalis* Mitchill). *Journal of Evolutionary Biology*, **22**, 1708-1720.

Mavarez J, Linares M (2008) Homoploid hybrid speciation in animals. *Molecular Ecology*, **17**, 4181-4185.

Mayr E (1942) *Systematics and the Origin of Species*. Dover Publications, New York, USA

Mayr E (1963) *Animal Species and Evolution*. Harvard University Press, Cambridge, UK.

Mayr E (2004a) *What makes biology unique?* Cambridge University Press, Cambridge, UK.

Mayr E (2004b) Happy birthday: 80 years of watching the evolutionary scenery. *Science,* **305**, 46-47

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

McCairns RJ, Bernatchez L (2010) Adaptive divergence between freshwater and marine sticklebacks: insights into the role of phenotypic plasticity from an integrated analysis of candidate gene expression. *Evolution,* **64**, 1029-1047.

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

McPhail JD (1993) Ecology and evolution of sympatric sticklebacks (*Gasterosteus*): Origin of the species pairs. *Canadian Journal of Zoology,* **71**, 515-523.

Meiklejohn CD, Townsend JP (2005) A Bayesian method for analysing spotted microarray data. *Briefings in Bioinformatics*, **6**, 318-330.

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

Michalak P, Noor MA (2003) Genome-wide patterns of expression in *Drosophila* pure species and hybrid males. *Molecular Biology and Evolution*, **20**, 1070-1076.

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

Michel AP, Sima S, Powella THQ, Taylora MS, Nosil P, Feder JF (2010) Widespread genomic divergence during sympatric speciation. *Proceedings of the National Academy of Sciences USA, 10.1073/pnas.1000939107.*

Mihola O, Trachtulec Z, Vlcek C, Schimenti JC, Forejt JA (2008) Mouse speciation gene encodes a meiotic histone H3 methyltransferase. *Science,* **323**, 373-375.

Miller CT, Beleza S, Pollen AA, Schluter D, Kittles RA, Shriver MD, Kingsley DM (2007) cis-Regulatory Changes in Kit Ligand Expression and Parallel Evolution of Pigmentation in Sticklebacks and Humans. *Cell,* **131**, 1179-1189.

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

Milos P (2008) Helicos BioSciences. *Pharmacogenomics* **9**, 477-480.

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.

Moehring AJ, Teeter KC, Noor MA (2007) Genome-wide patterns of expression in *Drosophila* pure species and hybrid males. II. Examination of multiple-species hybridizations, platforms, and life cycle stages. *Molecular Biology and Evolution*, **24**, 137-145.

Moen T, Hayes B, Baranski M, Berg PR, Kjøglum S, Koop BF, Davidson WS, Omholt SW, Lien S (2008) A linkage map of the Atlantic salmon (*Salmo salar*) based on EST-derived SNP markers, *BMC genomics*, **9**, 223.

Morin PA, Luikart G, Wayne RK, SNP workshop group (2004) SNPs in ecology, evolution and conservation. Trends *in Ecology and Evolution*, **19**, 208-216.

Muller HJ (1940) Bearings of the *Drosophila* work on systematics. In: *The New Systematics* (ed. Huxley JS), pp. 185–268. Clarendon Press, Oxford, UK.

Muller HJ (1942) Isolating mechanisms, evolution, and temperature. *Biological Symposia*, **6**, 71-125.

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.

Nielsen R (2005) Molecular Signatures of Natural Selection. *Annual Review of Genetics* **39**, 197-218.

Nolte AW, Freyhof J, Tautz D (2006) When invaders meet locally adapted types: rapid moulding of hybrid zones between sculpins (*Cottus*, Pisces) in the Rhine system. *Molecular Ecology*, **15**, 1983-1993.

Nolte AW, Renaut S, Bernatchez L (2009a) 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.

Nolte AW, Gompert Z, Buerkle CA (2009b) Variable patterns of introgression in two sculpin hybrid zones suggest that genomic isolation differs among populations. *Molecular Ecology*, **18**, 2615-2627.

Noor MAF, Feder JL (2006) Speciation genetics: evolving approaches. *Nature Reviews Genetics*. **7**, 851-861.

Nosil P (2008) Speciation with gene flow could be common. *Molecular Ecology*, **17**, 2103-2106.

Nosil P, Crespi BJ, Sandoval CP (2002) Host-plant adaptation drives the parallel evolution of reproductive isolation. *Nature*, **417**, 440-443.

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

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

Orr HA, Presgraves D (2000) Speciation by postzygotic isolation: forces, genes and molecules. *BioEssays*, **22**, 1085-1094.

Orr HA, Turelli M (2001) The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities. *Evolution*, **55**, 1085-1094.

Orr HA (2001) The genetics of species differences. *Trends in Ecology and Evolution*, **6**, 343-350.

Orr HA (2005) The genetic theory of adaptation: A brief history. *Nature Reviews Genetics,* **6**, 119-127.

Ortíz-Barrientos D, Counterman BA, Noor MAF (2007) Gene expression divergence and the origin of hybrid dysfunctions. *Genetica*, **129**, 71-81.

Ortiz-Barrientos D, Reiland J, Hey J, Noor MAF (2002) Recombination and the divergence of hybridizing species. *Genetica*, **116**, 167-178.

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.

Pourquié O (2003) The Segmentation Clock: Converting Embryonic Time into Spatial Pattern. *Science*, **301**, 328-330.

Presgraves DC (2003) A Fine-Scale Genetic Analysis of Hybrid Incompatibilities in *Drosophila*. *Genetics*, **163**, 955-972.

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

Price TD, Bouvier MM (2002) The evolution of F1 postzygotic incompatibilities in birds. *Evolution*, **56**, 2083-2089.

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

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

Ranz JM, Namgyal K, Gibson G, Hartl D (2004) Anomalies in the expression profile of interspecific hybrids of *Drosophila melanogaster* and *Drosophila simulans*. *Genome Research*, **14**, 373-379.

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* sp.). *Molecular Ecology*, **19** (Suppl. 1), 115-131.

Renaut S, Bernatchez L. Transcriptome-wide signature of hybrid breakdown associated with intrinsic reproductive isolation in lake whitefish species pairs (*Coregonus* spp. Salmonidae). *Heredity. accepted*

Renaut S, Nolte AW, Rogers SM, Derome N, Bernatchez L. Gradients of ecological speciation, SNP signature of selection on standing genetic variation, and association with adaptive phenotypes in lake whitefish species pairs (*Coregonus* spp.). *Mol. Ecol. accepted*

Reshetnikov JS (1988) Coregonid fishes in recent conditions. *Finnish Fisheries Research*, **9**, 11-16.

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

Richardson MK (1999) Vertebrate evolution: the developmental origins of adult variation. *BioEssays*, **21**, 604-613.

Rieseberg LH, Archer MA, Wayne RK (1999) Transgressive segregation, adaptation and speciation. *Heredity*, **83**, 363-372.

Rieseberg LH, Raymond O, Rosenthal DM, Lai Z, Livingstone K, Nakazato T, Durphy JL, Schwarzbach AE, Donovan LA, Lexer C (2003) Major Ecological Transitions in Wild Sunflowers Facilitated by Hybridization. *Science*, **301**, 1211-1216.

Rieseberg LH, Willis JH (2007) Plant Speciation, *Science,* **317**, 910-914.

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.

Roberge C, Normandeau E, Einum S, Guderley H, Bernatchez L (2008) Genetic consequences of interbreeding between farmed and wild Atlantic salmon: insights from the transcriptome. *Molecular Ecology*, **17**, 314-324.

Roche: http://www.454.com/ Accédé en août 2010

Rockman MV, Kruglyak L (2006) Genetics of global gene expression. *Nature Reviews Genetics* 7, 862-872.

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, 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, 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.

Rottscheidt R, Harr B (2007) Extensive additivity of gene expression differentiates subspecies of the house mouse. *Genetics,* **177**, 1553-1567.

Rundell RJ, Price TD (2009) Adaptive radiation, nonadaptive radiation, ecological speciation and nonecological speciation. *Trends in Ecology & Evolution*, **24**, 394-399.

Rundle HD, Nagel L, Boughman JW, Schluter D (2000) Natural selection and parallel speciation in sympatric sticklebacks. *Science*, **287**, 306-308.

Rundle HD, Nosil P (2005) Ecological Speciation. *Ecology Letters,* **8**, 336-352.

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Science USA*, **74**, 5463-5467.

Sauvage C, Derôme N, Normandeau E, St-Cyr J, Audet C, Bernatchez L (2010) Fast Transcriptional Responses to Domestication in the Brook Charr *Salvelinus fontinalis*. *Genetics*, **185**, 105-112.

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

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

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

Shan X, Liu Z, Dong Z, Wang Y, Chen Y, Lin X, Long L, Han F, Dong Y, Liu B (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.

Slack JMW, Holland PWH, Graham CF (1993) The zootype and the phylotypic stage. *Nature,* **361**, 490-492.

Sladek R, Hudson TJ (2006) Elucidating cis and trans-regulatory variation using genetical genomics. *Trends in Genetics,* **22**, 245-250

Stanek MT, Cooper TF, Lenski RE (2009) Identification and dynamics of a beneficial mutation in a long-term evolution experiment with *Escherichia coli. BMC evolutionary biology*, **9**, 302.

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.

Steiner CC, Römpler H, Boettger LM, Schöneberg T, Hoekstra HE (2009) The genetic basis of phenotypic convergence in beach mice: similar pigment patterns but different genes. *Molecular Biology and Evolution*, **26**, 35-45.

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

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.

Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology*, **14**, 671-688.

Storz JF, Wheat CW (2010) Integrating evolutionary and functional approaches to infer adaptation at specific loci. *Evolution* doi:10.1111/j.1558-5646.2010.01044.x

Svanbäck R, Persson L (2004) Individual diet specialization, niche width and population dynamics: implication for trophic polymorphisms. *Journal of Animal Ecology,* **73**, 973-982.

Swanson-Wagner RA, Jia Y, DeCook R, Borsuk LA, Nettleton D, Schnable PS (2006) All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. *Proceedings of the National Academy of Sciences USA*, **103**, 6805-6810.

Tamura K, Subramanian S, Kumar S (2004) Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Molecular Biology and Evolution*, **21**, 36-44.

Tautz D (2002) A genetic uncertainty problem. *Trends in Genetics*, **16**, 475-477.

Tautz D, Schmid KJ (1998) From genes to individuals: developmental genes and the generation of the phenotype. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, **353**, 231-240.

Tautz D, Ellegren H, Weigel W (2010) Next Generation Molecular Ecology, *Molecular Ecology*, **19** (Suppl. 1), 1-3.

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.

Townsend JP, Cavalieri D, Hartl DL (2003) Population genetic variation in genome-wide gene expression. *Molecular Biology and Evolution*, **20**, 955-963.

Townsend JP, Hartl DL (2002) Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments. *Genome Biology,* **3**, RESEARCH0071.

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.

True JR, Haag ES (2001) Developmental system drift and flexibility in evolutionary trajectories. *Evolution & Development*, **3**, 109-119.

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 der Sluijs I, van Dooren TJM, Seehausen O, van Alphen JJM (2008) A test of fitness consequences of hybridization in sibling species of Lake Victoria cichlid fish. *Journal of Evolutionary Biology*, **21**, 480-491.

Van Tassell CP, Smith TP, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, Haudenschild CD, Moore SS, Warren WC, Sonstegard TS (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods*, **5**, 247-252.

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

Venter JC, Adams MD, Myers EW, *et al.* (2001) The sequence of the human genome. *Science,* **291,** 1304-1351.

Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, **17**, 1636-1647.

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

Via S (2009) Natural selection in action during speciation. *Proceedings of the National Academy of Science USA*, **106**, 9939-9946.

von Baer KE (1828) *Über Entwicklungsgeschichte der Thiere: Beobachtung und Reflexion.* Königsberg: Bornträger, Dl.

von Schalburg KR, Cooper GA, Leong J, Robb A, Lieph R, Rise ML, Davidson WS, Koop BF (2008) Expansion of the genomics research on Atlantic salmon *Salmo salar* L. project (GRASP) microarray tools. *Journal of Fish Biology*, **72**, 2051-2070.

von Schalburg KR, Rise ML, Cooper GA, Brown GD, Gibbs AR, Nelson, CC, Davidson WS, Koop BF (2005) Fish and Chips: Various methodologies demonstrate utility of a 16,006-gene salmonid microarray. *BMC Genomics*, **15**, 126.

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

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution,* **38**, 1358-1370.

White BJ, Cheng C, Simard F, Costantini C, Besansky NJ (2010) Genetic association of physically unlinked islands of genomic divergence in incipient species of *Anopheles gambiae*. *Molecular Ecology*, **19**, 925-939.

Whitehead A, Crawford DL (2006) Variation within and among species in gene expression: raw material for evolution. *Molecular Ecology,* **15**, 1197-1211.

Whiteley AR, Derome N, Rogers SM, St-Cyr J, Nolte AW, Renaut S, Jeukens J, Laroche J, Labbe A, Bernatchez L (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.). *Can J of Zoo*, **87**:566-572.

Whitfield CW, Cziko AM, Robinson GE (2003) Gene Expression Profiles in the Brain Predict Behavior in Individual Honey Bees. *Science,* **302**, 296-299.

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

Wilding CS, Butlin RK, Grahame J (2001) Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers. *Journal of Evolutionary Biology*, **14**, 611-619.

Wray GA (2007) The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*, **8**, 206-216.

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA (2003) The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution*, **20**, 1377-1419.

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 Reviews 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.

Yatabe Y, Kane NC, Scotti-Saintagne C, Rieseberg LH (2007) Rampant gene exchange across a strong reproductive barrier between the annual sunflowers, *Helianthus annuus* and *H. petiolaris*. *Genetics*, **175**, 1883-1893.

Yoon HS, David A, Baum DA (2004) Transgenic study of parallelism in plant morphological evolution. *Proceedings of the National Academy of Science USA*, **101**, 6524-6529.

Yvert G, Brem RB, Whittle J, *et al.* (2003) Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nature Genetics,* **35**, 57-64.

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

Zeldith ML, Swiderski DL, Sheets HD, Fink WL (2003) *Geometric Morphometrics for Biologists*: *A Primer.* Academic Press, London, UK.

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 Science USA,* **106**, 1886-1891.

Chapitre 8 : Annexes

BMC Evolutionary Biology

BioMed Central

Research article

Open Access

Divergence in gene regulation at young life history stages of whitefish (*Coregonus* sp.) and the emergence of genomic isolation

Arne W Nolte*, Sébastien Renaut and Louis Bernatchez

Address: Département de Biologie, Pavillon Charles-Eugène-Marchand, Université Laval, Ste-Foy, Québec, G1V 0A6, Canada

Email: Arne W Nolte* - arne.nolte.1@ulaval.ca; Sébastien Renaut - sebastien.renaut@gmail.com;
Louis Bernatchez - louis.bernatchez@bio.ulaval.ca

* Corresponding author

Published: 16 March 2009

BMC Evolutionary Biology 2009, **9**:59 doi:10.1186/1471-2148-9-59

Received: 6 May 2008
Accepted: 16 March 2009

This article is available from: http://www.biomedcentral.com/1471-2148/9/59

Abstract

Background: The evolution of barriers to reproduction is of key interest to understand speciation. However, there may be a current bias towards studying intrinsic postzygotic isolation in old species pairs as compared to the emergence of barriers to gene flow through adaptive divergence. This study evaluates the relative importance of both processes in the evolution of genomic isolation in incipient species of whitefish (*Coregonus clupeaformis*) for which preliminary data suggest that postzygotic isolation emerges with intrinsic factors acting at embryo stages but also due to extrinsic factors during adult life.

Results: Gene expression data were screened using cDNA microarrays to identify regulatory changes at embryo and juvenile stages that provide evidence for genomic divergence at the underlying genetic factors. A comparison of different life history stages shows that 16-week old juvenile fish have 14 times more genes displaying significant regulatory divergence than embryos. Furthermore, regulatory changes in juvenile fish match patterns in adult fish suggesting that gene expression divergence is established early in juvenile fish and persists throughout the adult phase. Comparative analyses with results from previous studies on dwarf-normal species pairs show that at least 26 genetic factors identified in juvenile fish are candidate traits for adaptive divergence in adult fish. Eight of these show parallel directions of gene expression divergence independent of tissue type or age of the fish. The latter are associated with energy metabolism, a complex trait known to drive adaptive divergence in dwarf and normal whitefish.

Conclusion: Although experimental evidence suggests the existence of genetic factors that cause intrinsic postzygotic isolation acting in embryos, the analysis presented here provided few candidate genes in embryos, which also corroborate previous studies showing a lack of ecological divergence between sympatric dwarf and normal whitefish at the larval stage. In contrast, gene expression divergence in juveniles can be linked to adaptive traits and seems to be driven by positive selection. The results support the idea that adaptive differentiation may be more important in explaining the emergence of barriers to gene flow in an early phase of speciation by providing a broad genomic basis for extrinsic postzygotic isolation rather than intrinsic barriers.

Background

The evolution of reproductive isolation is of fundamental interest in evolutionary biology because it represents a key step in speciation processes and the generation of biological diversity [1]. Merging of divergent lineages can be prevented by prezygotic barriers that reduce heterospecific mating or by decreased offspring fitness (postzygotic isolation). Some of the most inclusive studies on postzygotic isolation have focussed on taxa that have been separated for millions of years. For instance, hybrids among species of *Drosophila* are often completely sterile or inviable, which can be explained by Dobzhansky-Müller incompatibilities [2,3]. Postzygotic isolation results from genetic changes in the parental lineages that, while functional on their normal genetic backgrounds, reduce the viability or fertility when recombined in hybrids. Intrinsic postzygotic isolation is likely to manifest as soon as the respective factors are expressed, *i.e.* during early development [1], whereas effects on reproductive traits are naturally associated with the reproductive phase. Such intrinsic barriers to reproduction are thought to evolve slowly through a stochastic accumulation of genetic incompatibilities [4]. However, when young species have split only recently, extrinsic postzygotic isolation can also be effective through a more subtle effect. Alleles that reduce the fitness in a given genetic background can be removed by externally (e.g. ecological) caused natural selection. Here, heterospecific allele combinations are not lethal but perform worse than pure parental genotypes in dependence of the ecological context. Differentially adapted genes can be instrumental to generate initial patterns of genetic divergence and are thought to govern the divergence and merging of young evolutionary lineages [5-7]. At least under conditions of gene flow, speciation will be driven by natural selection imposed by external ecological factors [5,8,9] and models generally agree that intrinsic hybrid inviability is not an initial event that drives speciation [1].

There may be a bias in our perception of the contribution of intrinsic and extrinsic postzygotic isolation to speciation processes. This is because it is usually more straightforward to analyse intrinsic barriers than to grasp extrinsic barriers experimentally, since the latter will most likely depend on unknown ecological interactions. Therefore, traits that could provide a basis for genomic isolation in young lineages remain insufficiently explored. A possible approach is given by transcriptome analysis. Here, gene expression data may help identifying key genes involved in speciation since regulatory evolution is hypothesized to be a key factor in microevolutionary processes [10-12]. Genes that are regulated differently are likely to loose compatibility with the genetic environments of alternative lineages. Microarray approaches offer the potential to study genome-wide patterns of divergence, and can be considered as an inventory of characters that could serve as a basis for genome divergence. Although this does not provide evidence that selection acts on each of the particular genes under study, it will reveal the processes and functions that may be affected.

In this study, we explore by means of transcriptomics the regulatory divergence between incipient species of lake whitefish (*Coregonus clupeaformis* (Mitchill, 1818)) in order to identify candidate traits that could contribute to barriers to gene flow. This system is of particular interest to study the emergence of postzygotic isolation as the diverging lineages are of recent, most likely postglacial, origin (15 000 ya) [13]. Dwarf and normal whitefish have evolved multiple times in response to ecological selection pressures [14] and genome scans and mapping projects demonstrated that natural selection drives this divergence in multiple genomic regions [15,16] while also suggesting that the lineages are at a phase of speciation where gene flow is still occurring. On the other hand, Rogers and Bernatchez [17] have found evidence for genetic factors causing postzygotic isolation in developing eggs. The actual genes and functions involved in these processes are largely unknown due to a use of anonymous genetic markers. However, the application of transcriptome data offers a promising approach to identify candidate genes. Microarrays made for salmon (*Salmo salar, Onchorynchus mykiss*) can be readily used in whitefishes [18,19]. Derome *et al.* [20,21] and St-Cyr *et al.* [22] have identified a suite of candidate adaptive traits that display parallel changes in gene expression between adult dwarf and normal whitefish in replicated lakes.

Here, we tie in with the above studies, which suggest than both intrinsic and extrinsic barriers to reproduction play a role in the divergence of dwarf and normal whitefish at the embryo stage and during the adult life respectively. Our main objective was to compare regulatory changes at different life history stages to obtain an insight into the processes that may contribute to genomic divergence. Our results indicate that there is little regulatory divergence in embryos in sharp contrast with evidence that numerous genes display regulatory divergence in juvenile fish. Given that the latter patterns can be partially linked to ecological divergence, we conclude that extrinsic postzygotic barriers may be more important to explain early evolutionary divergence of dwarf and normal whitefish than intrinsic barriers to reproduction.

Results

Number and types of genes analysed

The number of features (spotted EST clones) for which we obtained gene expression data of sufficient quality for subsequent analyses was 7004 for the embryos and 5787 for the juvenile dataset. This discrepancy is correlated with technical aspects of the experiments. The number of spots that were excluded because they had a bad quality flag (obvious artefacts) after visual editing was 3209 in the juvenile dataset vs. 1055 in the embryo dataset. Furthermore, the average background in the embryo experiments was lower than in the juvenile experiments (744 vs. 849 relative fluorescence units). A total of 4293 features were common to both datasets. Accordingly, the embryo data contained 2711 and the juvenile data contained 1494 unique features. Those features of the whole embryo dataset that were associated with a GO term could be linked to 2034 unique unigene clusters and the features of the whole juvenile dataset represented 1549 unique unigene clusters.

The overall representation of gene functional groups among the expressed features between the two datasets differed significantly according to the ease score provided by the EASE software. The juvenile fish dataset contained a significant relative excess (ease score < 0.05) of unigene clusters representing three GO-Biological processes: Catabolism (55 genes, ease score = 0.004), lipid metabolism (26 genes, ease score = 0.011), proteolysis and peptidolysis (37 genes, ease score = 0.025). In contrast, the embryo dataset contained an almost significant relative excess of unigene clusters representing two GO-Biological processes: Cell cycle (44 genes, ease score = 0.095) and nucleobase\, nucleoside\, nucleotide and nucleic acid metabolism (139 genes, ease score = 0.096). These trends in the representation of genes in the two life history stages reflect the importance of metabolism and growth processes in the juvenile stage while gene transcription regulation and development predominate in the embryos.

Genes displaying significant differences

After applying a FDR correction for multiple testing, only 33 EST clones showed significant differential expression between the embryos of dwarf and normal whitefish [see Additional file 1]. In contrast, a total of 502 EST clones displayed significant differences in gene expression between dwarf and normal whitefish in the juvenile fish dataset. This difference in the proportion of EST clones that display significant differentiation in gene expression in the two datasets was highly significant (Fisher's Exact test, $p < 0.001$). For the embryos, 350 out of the 7004 features tested would be expected to have false positive tests according to our significance criterion ($p < 0.05$). However, the number of raw significant results in the embryo analysis was 590 (8,4%). The corresponding number of significant genes in the juvenile dataset was 988 (17%), while 289 false positives would be expected. This indicates for both datasets that the number of significant tests is not explainable by the expected false positive rate. The true number of genes with differential patterns of expression was higher than the list obtained after the FDR procedure, but the trend that there was much more differentiation in the juvenile compared to the embryonic stage remains independent of the FDR procedure.

For comparisons between datasets, EST clones were assigned to EST clone groups based on; i) unigene cluster or accession numbers (latest annotation following cGRASP) unless there was no known function, and ii) unique patterns of gene expression divergence. In doing this, significant patterns were integrated over replicate clones and overrepresentation of different genes by multiple clones was corrected. Among the features displaying significant differentiation in the embryo dataset, 20 can be assigned to one of twelve unigene clusters. Eleven out of the twelve unigene clusters that display significant differentiation in the embryo dataset also appear in the list of significant unigene clusters of the juvenile dataset [see Additional file 1]. A total of 191 of the significant EST clones of the juvenile fish dataset were assigned to one of 127 unigene clusters. Accounting for the different numbers of unigene clusters that were represented by the raw data in the egg (2034) and juvenile (1549) datasets, this indicates that roughly fourteen times more genes as represented by distinct unigene clusters display overall significant differentiation in the juvenile fish than in the embryos.

Comparisons with results from previous studies on adult fish

Some of the EST clones displaying significant divergence of gene expression in the analyses presented above have already been demonstrated to display differential expression in dwarf and normal whitefish. Only one EST clone for which significant differentiation was detected in the embryo dataset (CA057378; Accession AY872256; *Oncorhynchus mykiss* IgH.A locus) was previously found to be differentially expressed in white muscle between laboratory dwarf and normal whitefish [21]. In contrast, 108 EST clones that can be assigned to 44 different EST clone groups (identical accessions and unique patterns of expression) identified in whole juvenile fish [see Additional file 2] also show significant differentiation in gene expression in white muscle of adult fish of the same dwarf and normal strains in a controlled common environment [21]. Although different tissues and life history stages were compared, there was a significant excess of 31 out of 44 EST clone groups (Fishers' test, $p = 0.0403$) where gene expression divergence between dwarf and normal whitefish was congruent in the pattern of up or down regulation in dwarfs relative to normals [see Additional file 2] as compared to a random distribution of changes in both directions. Likewise, a comparison of the set of EST clones displaying significant gene expression divergence in juvenile fish reveals matches with candidate features that have been identified in independent natural lakes [20,22]. It should be noted that the study by Derome et al. [20] used a less inclusive microarray containing five times less features, which reduces the relative power of that study to identify genes that were found here or by St-Cyr *et al.* [22]. A total of 96 EST clones that displayed significant divergence in whole juvenile dwarf and normal whitefish of the strains studied here could be matched with EST clones displaying parallel adaptive divergence in gene expression in liver or white muscle of adult dwarf and normal whitefish from Cliff Lake (Maine, USA) and Indian Pond (Maine, USA) [see Additional file 3]. Ten out of 26 different EST clone groups (grouped as described above.) show regulatory changes in the controlled environment that were congruent with the patterns of candidate adaptive traits as observed in adult tissues in natural lakes. Among these, there appears to be a bias in that eight of ten affected genes are related to energy metabolism, which shows that regulatory changes between dwarf and normal whitefish related to this function are more constant across different environments, life history stages and tissues than those related to other functions [see Additional file 3].

In contrast to the observations for juvenile and adult fish of the experimental strains in controlled common environment [see Additional file 2], the direction of upregulation vs. downregulation of the transcript level of dwarf whitefish relative to normal whitefish shows less congruence when samples from natural environments and a new tissue (liver) are included in the comparison. Thus, 16 out of 26 groups of features that display significant divergence of gene expression in candidate adaptive features differ in the direction of the change between juveniles and adults (16) or between adult muscle and liver tissue (3) [see Additional file 3].

Discussion

Our results revealed a pronounced pattern of gene expression divergence for 502 EST clones between 16-week old juvenile dwarf and normal lake whitefish (*Coregonus clupeaformis* complex) as compared to embryos of the same experimental groups, which displayed little divergence in gene expression (33 EST clones). Although the number of evolutionary changes causing the observed differences is currently unknown, a fourteen-fold excess of unigene clusters displaying significant differentiation in the juvenile dataset suggests that multiple regulatory changes take effect only after development has passed the embryo stage. If gene expression divergence were the result of random accumulation of evolutionary differences between the studied populations, roughly equal proportions of gene expression differences would be expected to occur in both life history stages. The much more likely scenario is that evolutionary change in gene expression plays a greater role at the juvenile stages than the embryonic stage.

The general pattern observed for gene expression divergence and regulatory changes resembles the ontogeny of morphological features across the animal kingdom. Briefly, early developmental stages are usually extremely well conserved, whereas adult phenotypes vary as a consequence of evolutionary divergence, a classical observation made by early developmental biologists [23,24]. This observation has often been made for distantly related taxa. This study suggests that the same evolutionary pattern may not only apply to morphological characters, but also to transcriptomic divergence at the level of recently evolved lineages of fish. Below, the observed changes in gene expression are discussed in relation to life history divergence of dwarf and normal whitefish. We propose that the excess of gene expression divergence in juvenile fish relative to embryos can be attributed to selective pressures that are related to ecological adaptation in the juvenile and the adult phase rather than an evolutionary constraint on divergence in embryonic stages.

A key concern in the analyses was that the absence of differentiation observed in the embryos could represent an artefact. Developmental processes and therefore gene expression in embryos can be expected to change rapidly throughout ontogeny. The problem such heterogeneity in gene expression imposes for the analysis resembles that of allometry in studies of body shape [25] and the relevance of heterogeneity for the analysis of gene expression data has recently been pointed out by Leek and Storey [26]. If different stages with accordingly changed patterns of gene expression were sampled, the variance in gene expression would be inflated. Intra group variance could then exceed the between group variance to a point where the latter is not detected as significant in ANOVA based statistical approaches. The extent to which such variation occurs in the transcriptome can be inferred from a study by Arbeitman *et al.* [27] who performed a very complete analysis of patterns of gene expression throughout development of *Drosophila melanogaster* and found that significant heterogeneity was observed for 52% of all studied genes during the embryogenesis but few genes displaying developmental heterogeneity in adult *Drosophila*. Moreover, some classes of genes that are expressed during the segmentation phase of fishes show highly dynamic and cyclical patterns of expression through short periods of time [28]. The sampling of a relatively well-defined segmentation stage as done in this study is merely a snap-shot of the whole embryogenesis and should therefore contain considerably more genes with relatively constant expression patterns that are consequently useful for ANOVA. Accordingly, the inter sample variance in gene expression for eggs and juveniles was in the same order of magnitude (mean inter-sample variance 0.0085 and 0.0066 respectively) as estimated from a more inclusive dataset of the same stages including technical replication (Renaut *et al.* unpublished). Even if only half of all genes (comp. above; [27]) in the embryo dataset displayed homogeneous gene expression, one would still expect to detect considerably more significant genes if the proportion of significant genes in the embryo data was identical to that in the juvenile data. Although an effect of developmental heterogeneity cannot be excluded, it is unlikely to explain the observed excess in gene expression divergence in the juveniles.

The ontogeny and evolution of gene expression differentiation

Dwarf whitefish from Lake Témiscouata and normal whitefish from Lake Aylmer studied here have a similar reproductive biology. Adult fish live in the lake throughout the year and enter tributaries only for a brief spawning period. Spawning migrations are short and occur on a daily basis at night from late October to early November. Eggs are dispersed in currents and settle into rock and gravel substrates where they are left unattended to develop. Ninety-one percent of the unigene clusters that displayed significant regulatory divergence in the embryos were also significantly divergent in juvenile fish and in one case, found to be differentially expressed in white muscle tissue of adults of the same laboratory strains used here. However, none of the significantly different genes of the embryo stage was found to be a candidate for adaptive divergence in previous studies. The fact that ecological divergence of the egg stages of whitefishes has not been discovered to date together with the relative lack of gene expression differentiation suggests strongly that little or no adaptive evolutionary divergence has occurred specifically at the embryo stage and that most of the gene expression differentiation between dwarf and normal whitefish must develop at a later phase of the ontogeny.

Upon hatching whitefish larvae are washed from their natal river into the lakes were they spent their entire life. To date it is unknown at which phase of the life history the ecological differentiation into the dwarf (limnetic) and normal (benthic) lifestyles occurs in nature. In a study of dwarf and normal whitefish in Cliff Lake, larvae of dwarf and normal populations did not differ in their hatching time, diet, distribution and vertical migration within the lake [29]. Unlike their parents, the larvae lived in total syntopy and there was no evidence for differential trophic ecology or circadian vertical migration, suggesting that ecological divergence of the two forms must begin after the larval stage [29]. The experimental populations used in this study had a total age of 16 weeks and had morphologically transformed into juvenile fish (development of finrays and scales) for approximately 8 weeks before sampling was done. At the level of the transcriptome, the transformation into the juvenile stage is accompanied by an emergence of gene expression divergence between the two forms that was absent at the embryo stage. Also, gene expression divergence at this juvenile stage could be matched with patterns observed in adult fish. 108 EST clones representing 44 differentially regulated genes or accessions [see Additional file 2] also displayed regulatory divergence in muscle tissue of adult fish belonging to the same experimental groups and kept in a controlled common environment [21]. There was a significant excess of congruent regulatory change, which suggests that gene expression divergence between dwarf and normal whitefish is of a similar nature in juvenile and adults. The notion that regulatory divergence does not change much after the development of the adult morphology has finished is in line with the observation that many patterns of gene expression change little throughout adult life in *Drosophila* [27]. Hence, the juvenile stage studied here is useful to study the transition of life histories from non-differentiated larval fish [29] to adult dwarf and normal whitefish with pronounced differential adaptation [13].

A total of 96 of the EST clones identified here matched with 26 accessions or genes that were also described by Derome *et al.* [20] and St-Cyr *et al.* [22] [see Additional file 3] who found a recurrent association of divergence in gene expression and parallel adaptive differentiation in multiple lakes (including Cliff lake) for these genes. Briefly, dwarf whitefish tend to have a shorter lifespan and begin to reproduce earlier than normal whitefish. They are specifically adapted to the open water where they specialise on a zooplankton diet as opposed to the more benthic lifestyle of normal whitefish [13]. According to Trudel *et al.* [30] this ecological differentiation is driven by differences in metabolic rate and energy allocation between dwarf and normal whitefish. Derome *et al.* [20] and St-Cyr *et al.* [22] have inferred candidate adaptive traits bases on patterns of parallel divergence in independent lake systems each containing dwarf and normal whitefish. In agreement with the experimental results of Trudel *et al.* [30], the biological function of a part of these candidates implies a role in energy metabolism. Our results on the gene expression divergence between juvenile dwarf and normal whitefish has revealed ten out of 26 genes which match adaptive regulatory changes in the adult stage irrespective of the fact that different tissues were used. Most conspicuously, eight of these ten genes can be associated with energy metabolism. The comparison of different studies reveals a recurrent pattern of up regulation of transcripts for Glyceraldehyde-3-phosphate dehydrogenase, Fructose-bisphosphate aldolase A, Beta-enolase and Trypsin-1 precursor as well as a down regulation of transcripts for a mitochondrial precursor of Cytochrome c oxidase polypeptide VIa, Nucleoside diphosphate kinase and Nucleoside diphosphate kinase A in dwarf relative to normal whitefish [see Additional file 3] in different tissues and life history stages. This adds evidence for the hypothesis that energy metabolism as a complex trait plays an ubiquitous role in driving the adaptive divergence between dwarf and normal whitefish while other candidate adaptive traits play more tissue and context specific roles.

The results obtained here for juvenile fish and the comparison of the direction of the regulatory changes in different tissues or the controlled common environment vs. natural lakes shows that regulatory changes at candidate adaptive traits are not always congruent in terms of the direction of the change. It must be emphasized though, that the inference of adaptive divergence from parallel regulatory changes relies on the use of comparable samples. If tissue type and environmental context vary, the regulatory response can be expected to vary as well. Accordingly, there was more agreement in the direction of the regulatory changes when comparison based on liver tissue or natural environments were excluded [see Additional file 2]. This implies that the inference of an adaptive value of parallel divergent traits by Derome *et al.* [20] and St-Cyr *et al.* [22] is not invalidated by contrasting patterns of regulatory divergence for a given candidate gene in juvenile fish and vice versa. Although the relationship of the direction of regulatory change according to tissue context, age and environmental factors needs to be addressed in future studies, the fact remains that many candidate genes for adaptive divergence in adult fish also show regulatory differentiation at the juvenile stage. In contrast to the embryo stage analysed here, these results suggest that the divergence in gene expression in juvenile fish is subject to directional selection related to adaptive divergence.

The identification of candidate adaptive gene expression divergence still draws an incomplete picture of the processes that ultimately lead to the life history differentiation into dwarf and normal whitefish. If adaptive differences at the transcriptome level are expressed as early as young (16 weeks) juvenile stage, then it is likely that these genetic factors may initiate the development of life history divergence. Although data on juvenile fish from natural lakes are missing, inferences can be made from our laboratory populations. Experimental fish were kept in a controlled common environment and the candidate adaptive traits remain prevalent at the transcriptome level suggesting that juvenile fish already display differential adaptation. On the other hand, the phenotypes of our experimental populations bred in the laboratory seem to contradict a persistent ecophenotypic differentiation between dwarf and normal whitefish. Although morphological features distinguishing dwarf and normal ecotypes are heritable [16], the differentiation of life histories under laboratory conditions is less pronounced than in nature. Dwarf whitefish remain only slightly smaller than normal whitefish and grow older than their natural counterparts (unpublished observation). This strongly suggests that there is an environmental component that interacts with a genetic one to shape the ecophenotypic differentiation of dwarf and normal whitefish.

A potential role for stabilizing selection

Aside directional selection driving divergence at the juvenile and adult stages, an alternative explanation for the relative lack of embryo gene expression may be derived from developmental biology. The embryos during the segmentation phase that were studied here correspond closely to the phylotypic stage, a developmental phase that corresponds to an archetype bauplan of all representatives of a given phylum. This stage is generally extremely conserved [31] suggesting that strong evolutionary constraints prevent divergence. While the morphological conservation of phylotypic stage embryos remains undisputed, it has been found that patterns of gene expression need not be constrained. Selection on adult genotypes may alter gene regulation in embryos [32] and developmental system drift [33], a process whereby gene regulatory networks evolve without changing the expression level of a gene, have been demonstrated for even closely related taxa [34]. This casts doubt on the validity of the phylotypic stage concept at the transcriptome level [34]. Given these alternatives, the similarity between embryos of dwarf and normal whitefish could imply the action of strong evolutionary constraints that preserve identical patterns of gene expression relative to what we observed at the juvenile stage. Under an evolutionary constraints hypothesis, genes that determine the expression level of a transcript may still evolve as long as the sum of their effects would not be changed. F1 – hybrids or segregating backcrosses would prove useful to reveal such cryptic regulatory divergence. They would combine the altered regulatory elements into a common genetic background, which can result in misexpression of genes [35-37]. Very much in line with this, crossing experiments by Lu and Bernatchez [38] and Rogers and Bernatchez [17] suggest that F1 hybrids or backcrosses of the same populations studied here suffer from raised embryonic mortality in the same stage studied here. This would indicate the presence of alleles or genes that malfunction and should manifest as misexpression at the transcriptome level (Renaut et al. *in prep*). In any case, the scarcity of regulatory divergence observed here for the embryo stage suggests that regulatory divergence is rare and may have a narrow genomic basis involving only a small number of genetic factors. This view would agree with the fact Rogers and Bernatchez [16] have only found a limited number of QTL associated with hybrid embryo mortality in their mapping analysis.

Implications for the emergence of genomic isolation and speciation in whitefish

At the level of the transcriptome, only traits causing regulatory divergence produce a phenotype that differs between two diverging lineages. These could be affected by natural selection and are therefore candidate traits that may reduce the fitness of individuals of mixed ancestry. Thus, a screen for regulatory divergence can identify traits that could have fitness effects under conditions of gene flow and serve as a genetic basis for the emergence of genomic isolation. Models suggest that recombination can oppose genomic divergence and speciation because loci that are not selected for may still be exchanged relatively freely among populations [5,39]. Even if single genetic factors have a strong isolating effect, they may not be sufficient to reduce gene flow throughout the genome if hybrids are partially viable, as it is the case in *Coregonus* [17,38]. It is therefore evident that the more loci are divergent between two lineages, the easier it becomes to explain how these could, as a whole, provide a basis for evolutionary divergence and reproductive isolation. Few transcripts are regulated differently at the embryonic stage analysed here, which points towards a comparably small number of genes that may be differentially regulated between dwarf and normal whitefish embryos, at least relative to later stages. Although Rogers *et al.* [17] provided evidence for the presence of genetic factors causing considerable hybrid embryo mortality, it may be concluded that the regulatory changes that could cause postzygotic isolation at the embryo stage are comparably rare. Admittedly, the question to which degree this would be sufficient to prevent gene flow in nature cannot be answered at present. In any case, it would be easier to explain a reduction in gene flow if there was a broad genomic basis to reproductive isolation, i.e. effects of multiple traits that could be selected for, rather than few. Our results suggest strongly that the largest part of regulatory divergence occurs throughout the juvenile and adult phases. A pattern of more pronounced divergence in juvenile and adult fish corresponds well with what has been observed in studies on interspecific regulatory divergence [35]. Moreover, a suite of traits contributing to a diversified adult phenotype is expected to play a more important role in evolutionary divergence as compared to genes producing the conserved phylotypic stage phenotype [40].

A gap in the sampling of potentially relevant life history stages of this study is that the reproductive phase could not be analysed here. Studies on *Drosophila* [36] and *Mus* [41] have demonstrated an above average rate of regulatory divergence in genes associated with reproduction and in addition, have shown that these genes may be among the first to cause reproductive isolation. Future studies will have to show whether reproductive characters may play the same crucial role in evolutionary divergence of incipient species of fish or if ecologically selected traits are among the first to cause genomic isolation.

Conclusion

If the relative rarity of gene expression divergence in embryo stages was a general pattern in recently diverged species, then a focus on embryo dysgenesis in studies of the early evolution of postzygotic isolation could be misleading in that they would distract from a large pool of adult characters. While tests for embryonic mortality are straightforward, it is much more difficult to test the role of particular genes on complex adult phenotypes as those may have small effects [42] or their effects may depend on unknown environmental components. Still, the transcriptomic patterns in whitefish suggest that a full understanding of how gene flow is reduced among incipient species may depend more heavily on genes affecting adult phenotypes than developmental phenotypes. In agreement with this, van der Sluijs *et al.* [43] suggest, for closely related cichlids from Lake Victoria, that postzygotic reproductive isolation is mediated by extrinsic selection rather than intrinsic hybrid dysgenesis. Furthermore, studies across a broader phylogenetic range of taxa show that intrinsic genomic incompatibility evolves slowly and after the point of speciation between diverging species [44,45]. Our results together with these studies support the view that subtle selective pressures and ecological interactions that are related to specific complex environments may be the key in explaining incipient genomic divergence [5,43,45].

Methods

Strains, crosses and fish maintenance

Eggs of *Coregonus clupeaformis* were obtained from lab strains kept at the LARSA (Laboratoire de Recherche en Sciences Aquatiques, Université Laval) or harvested from wild fish that were caught on their natural spawning grounds. Normal whitefish used here originate from Lake Aylmer (Basin of the St. Lawrence River, southern Quebec) and were sampled at the spawning site in the St. Francois River in Disraeli (45° 54'N, 71° 20'W) in 1996 (as detailed in Lu and Bernatchez, 1998). Since then, they were kept in the laboratory as an outbred lab strain. Dwarf whitefish originate from Lake Témiscouata (St. Johns' river system in southern Quebec) and were caught on their spawning grounds in the Touladi River (47° 41'N, 68° 47'W). Like the normal whitefish, dwarfs were maintained as an outbred laboratory strain. We also included new wild caught material from Lake Témiscouata dwarf whitefish collected in October 2006.

Eggs and semen were stripped from deafened fish, fertilized *in vitro* and incubated on grids that were submerged in slowly flowing water of a temperature of 4,5–5,5°C. All egg batches were incubated in the same flow through system and were thus subjected to compartments of the same environment. Weekly treatments with malachite green oxalate were performed to inhibit growth of fungi. Morbid eggs or embryos were removed on a daily basis. After hatching, free-swimming larvae were transferred into aquaria (50 × 25 × 30 cm) and fed *ad libitum* with *Artemia* nauplii and complemented with commercial fish food (Epac CW 4/6, Epac CW 6/8; INVE AQUACULTURE Inc., Salt Lake City, Utah, U.S.A.). All aquaria were aerated and connected by a flow through and filtering system that fed each aquarium from a common pool. This permitted constant water exchange and near identical temperature and chemical conditions. The temperature in the rearing tanks was kept at 8°C for the first 8 weeks, raised to 10°C for 3 weeks and finally adjusted to 12°C.

To capture the within population variance in gene expression and reduce family specific effects, we generally used crosses that were composed of several parents depending on the availability of mature fish at a given time. We have created two independent experimental groups for each biological group (dwarf, normal) studied here. The group DD-E was derived from the lab strain of the dwarf whitefish from Lake Témiscouata and was created using one female and five different males. DD-G was created by crossing wild caught dwarf whitefish (leg. Nolte, Renaut and Bernatchez, 25th Oct. 2006) from the same lake using multiple females and multiple males. Two groups of normal whitefish (NN-C and NN-I) were created from one and five as well as two and three females and males of the lab strain of normal whitefish from Lake Aylmer, respectively.

Stages and samples

It was our goal to analyse gene expression in a developmental stage corresponding to the phase for which Lu and Bernatchez [38] and Rogers and Bernatchez [17] observed increased embryo mortalities. We found that this corresponds to the beginning of the segmentation period. In this phase of development, an anterior-posterior axis has developed and undergoes divisions into body segments (for a detailed account on phases of fish development see [46]). Progress of development in this phase can be evaluated by counting body segments, which are added successively. This task can be performed on live eggs with the help of a binocular. Due to slight batch-to-batch variation in the precise timing of the segmentation process, we assessed developmental stages of embryos entirely by morphological features, rather than age. In our experiments, the process of segmentation began after roughly 16 days of development and ended after 29–31 days. Embryos were examined once or twice daily, from the moment that the tail bud of the embryo detached from the yolk sac (at day 20–22). It was easier to count only those segments in the part of the tail that was detached rather than all segments of the live embryo within the intact eggs. For our experiments we chose embryos that had formed approximately 20 segments in the detached portion of their tail. This stage appeared to be relatively easy to identify as at the same time the tail started moving and bent to an angle of approximately 30° (tail curvature). Furthermore, in this stage the optic primordium begins to hollow thus initiating the formation of the lens in the eye. This developmental stage corresponds best with the 20–25 somite stage observed in *Danio rerio* after only 19 hours of development at 28.5°C (compare http://zfin.org/zf_info/zfbook/stages/stages.html) [46]. Eggs that were chosen for experiments were individually examined using binoculars. Only apparently viable eggs with well-formed embryos were used. In these, the number of segments was counted through the chorion. This introduces some uncertainty in the somite count but preserves intact eggs and embryos for RNA extraction. Whole eggs were preserved in RNA later (Ambion) and frozen at -20°C for storage.

Juvenile fish were chosen as the next sampling stage as these represent an immature adult phenotype. All hatched larvae were transferred to basins and started external feeding in mid January 2007. The larvae had developed fin rays by the end of January 2007. We sampled juvenile fish at an age of approximately 16 weeks (May 10th 2007), when these attained a weight of approximately 800 mg (540–1190 mg). At this stage, the development of morphological features is finished and the young whitefish resemble their parents. Individuals chosen for gene expression analysis were well developed and in good general shape (vs. slow growing and meagre, as observed in some specimens). Sampling was done in the morning following an 18 hour fast. Fish were then sacrificed with a blow, kept on ice and homogenized in TRIzol reagent for RNA extraction as quickly as possible (waiting time no longer than 20 min). The homogenate was stored at -80°C prior to RNA extraction.

Experimental design and choice of samples

The gene expression analysis for this study focuses on the divergence at different life history stages of dwarf and normal whitefish. Eight pairwise (dwarf vs. normal) comparisons for both the embryonic and the juvenile fish stage were performed resulting in two sets of eight microarrays per stage (see Table 1). Initial testing had shown that even the more sensitive Gene Array 350 Kit (see below) ideally requires 5 μg of total RNA per sample and experiment. Given that only 2,5 – 3 μg of total RNA could be extracted from a single embryo, pools containing the total RNA of five embryos were used for the embryo experiments. This pooling approach integrates patterns of gene expression over a larger number of individuals but would nevertheless reveal differences between group means as tested for in an analysis of variance. Juvenile fish extractions yielded large quantities of total RNA and were used individually.

The same representation of experimental groups was used in both the embryo and juvenile fish experiments. Thus four replicates used samples from the groups DD-E and NN-C (see Table 1). Tests for differentiation among groups require that there is a homogeneous distribution of traits within groups. However, transitory developmen-

Table 1: Experimental design and types of biological samples used in microarray experiments.

Pairwise comparison	Dwarf whitefish experimental groups and properties of samples		Normal whitefish experimental groups and properties of samples	
1 Embryo	18 segments, tail curved up to 30°	DD-E	20 segments, tail curved up to 30°	NN-C
2 Embryo	23–25 segments, tail curved up to 30°.		20 segments, tail curved up to 30°	
3 Embryo	18 segments, tail curved up to 30°		20 segments, tail curved up to 30°	
4 Embryo	18 segments, tail curved up to 30°		20 segments, tail curved up to 30°	
5 Embryo	15–20 segments in detached tail.	DD-G	tail partially segmented, curved up to 30°.	NN-I
6 Embryo	15–20 segments in detached tail		tail partially segmented, curved up to 30°.	
7 Embryo	15–20 segments in detached tail.		tail partially segmented, curved up to 30°.	
8 Embryo	10–20 segments in detached tail.		tail partially segmented, curved up to 30°	
9 Juvenile	1.04 g	DD-E	1.06 g	NN-C
10 Juvenile	0.96 g		0.63 g	
11 Juvenile	0.96 g		0.89 g	
12 Juvenile	0.84 g		1.19 g	
13 Juvenile	0.99 g	DD-G	1.03 g	NN-I
14 Juvenile	0.54 g		0.85 g	
15 Juvenile	0.91 g		0.78 g	
16 Juvenile	0.80 g		0.87 g	

Analysis of gene expression was performed pair wise for a set of embryo experiments (1–8) using pools of 5 eggs while single juveniles were used in experiments 9–16. Staging of embryos and juveniles is assessed by key developmental parameters as described in the text. For embryos the number of somites in the part of the tail that is detached from the yolk as well the observed tail curvature is given. The number of somites is an approximation as it was determined *in vivo*. For juvenile fishes we provide the total body weight in gram.

tal stages, by their very nature, change constantly which may introduce biases when the sampling is not balanced. In order to reduce artefacts a set of samples that are similar with respect to developmental features (segment count) was chosen (Table 1). The sampling of juvenile fish is less difficult as they have finished their morphogenesis. Their ontogeny is also slower and probably reduced to relatively constant growth processes. Juvenile fishes were chosen to represent a similar body mass range (Table 1).

Analysis of gene expression

Total RNA was extracted using the TRIzol Reagent (Invitrogen) according to the protocol of the vendor. For the embryo experiments, five whole embryos preserved in RNAlater were homogenized using a bead mill (Quiagen) while for the juvenile fish experiments a single whole juvenile fish was homogenized using a polytron homogenizer. Crude total RNA was further cleaned by ultra filtration using microcon (Millipore) spin columns (embryo experiments) or a combination of a lithium chloride precipitation (addition of 1 Volume 5 M LiCl, incubation at -20°C for 2 hours, centrifugation at 16.000 g at 4°C for 30 min, final wash of the resulting pellet in 70% ethanol) and subsequent ultra filtration (juvenile fish experiments). Total RNA was quantified and quality checked using the Experion™ RNA StdSens Analysis Kit (BIO RAD). Total RNA was stored in pure water supplemented with Superase-In™ RNase Inhibitor (Ambion) at -80°C.

Gene expression analysis was performed using the 16 K v2.0 Salmon cDNA microarray as provided by the cGRASP consortium (von Schalburg *et al.* 2005). This microarray comprises 16 006 different cDNA features derived from salmonids as well as control spots that have previously been demonstrated to be useful to analyse gene expression in whitefish [20-22]. In each experiment, two different RNA samples are transcribed into strands of cDNA that are end-labelled with a specific oligonucleotide sequence. Samples are co hybridised to the microarray and the relative quantity of the hybridised product is assessed via fluorescent detection reagents that are specific to the end labels of a given sample. These experiments were made using the Genisphere 3DNA Array Detection Array 350™ Kit (Cy3/Alexa647) and Genisphere 3DNA Array Detection Array 50™ Kit (Cy3/Cy5) for the embryos and juveniles respectively and followed the protocols of the vendor. Dyes were swapped between different pairwise comparisons. Per sample and slide, we have used approximately 4–5 μg of total RNA in the embryo experiments and 18–20 μg of total RNA in the juvenile fish experiments. Reverse transcription reactions were performed using the Superscript II Kit (Invitrogen).

Microarrays were scanned using a ScanArrayTM Express scanner (Packard Bioscience) and quantified using the Quantarray software. The positioning of all grids was checked manually for both dye channels. Suspicious spots of inconsistent shape or obvious artefacts were marked with a bad quality flag. Raw data were quantified using the histogram method and exported into text files. Microarray data are deposited at ArrayExpress, a public repository, under the following experiment accession number (ArrayExpress accession: E-MEXP-1973). Input for the statistical analyses was generated from separate text files using a Perl script. For each spot, local background was subtracted from the PMT value. Data were subsequently used in the statistical analyses only if there was no more than 12,5% missing or unusable data per gene (e.g. a single value in a series of eight pairwise comparisons). Here, unusable may mean: i) a bad quality flag or ii) a gene expression value that is lower than the average background + 2 times its standard deviation for both samples measured on a given spot. The average background was determined from 800 empty spots or blank wells on the 16 K salmon chip, provided that these spots were not excluded due to artefacts after visual inspection. Statistical analysis of the data was performed in R version 2.6.1 (The R Foundation for Statistical Computing Copyright (C) 2007 ISBN 3-900051-07-0) using the R package R/maanova Version 1.4.1 [47,48]. Raw data was imported and missing data were imputed (KNN method, 10 nearest neighbours). Data was log2 transformed and normalized using the lowess algorithm. An ANOVA model including the following terms as fixed sources of variance: Type (population, the term of interest), Dye (Fluorescent dye) and Sample (biological sample) while Array (individual microarray) was included as random term. Statistical testing for divergence in gene expression between groups is based on an F test (Fs test option in R/maanova). P-values were determined by comparing observed values to a distribution obtained by randomly shuffling values of samples (1000 permutations). The FDR procedure as implemented in R/maanova was used to correct for multiple testing using an FDR cut off value of 5%.

The populations studied here have previously served to study gene expression divergence at the adult stage between specific tissues of dwarf and normal whitefish using a similar ANOVA. EST clones or genes for which significant differentiation was found in this study were compared with lists of candidate genes from studies on white muscle and liver tissue [20-22]. This comparison was based on EST clone ID numbers for comparisons with the studies that used the same microarray [21,22]. Derome *et al.* [20] used an earlier and less inclusive salmon microarray. In order to match results with the current dataset, EST clone sequences of candidate genes described by Derome *et al.* [20] were compared to all features on the 16 K v2.0 Salmon cDNA microarray using the BLAST algorithm [49] as implemented in BioEdit [50]. All sequences could be unequivocally matched to features on the new arrays based on the longest 100% identical sequence. The annotation of all features reported here follows the latest version of the gene annotation file for the 16 K Salmon microarray (as of Feb 13th 2008) provided on the cGRASP homepage http://web.uvic.ca/grasp/microarray/array.html. Analyses of the representation of functional categories of genes among datasets was based on the DAVID/EASE programs http://david.niaid.nih.gov/david/ease.htm[51].

Authors' contributions

AWN conceived the study, collected the data, performed the analyses and wrote the manuscript. SR contributed to all parts of the experimental design, data collection and analysis. LB conceived the study, coordinated the work and edited the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

EST clones for which significant divergence was detected between embryos or juveniles of dwarf and normal whitefish. *List of all EST clones for which significant divergence of gene expression was detected for dwarf and normal whitefish (33 and 502 EST clones for the Embryo and Juvenile datasets). EST clone and annotation, the Log2-fold change of dwarfs relative to normals and the FDR adjusted Permutation p-value (significance criterion in this study) are given.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2148-9-59-S1.xls]

Additional file 2

Regulatory changes between juvenile dwarf and normal whitefish in a common environment. *Similarity of regulatory changes between dwarf and normal whitefish at different life history stages in controlled common environments. All comparisons are based on the same strains (Lake Témiscouata "dwarf" and Lake Aylmer "normal" respectively). 108 EST clones that display significant differentiation in gene expression in whole juvenile fish (this study) were previously found to be differentially expressed in muscle tissue from adult fish [21]. When genes are represented by several EST clones these are grouped according to their annotation (accession number and gene name) and patterns of gene expression. "up" or "down" regulation describes the direction of the change of gene expression in the dwarf whitefish relative to normal whitefish. Biological functions are as given in Derome et al. [21]. EM = energetic metabolism; IR = immune response; OB = oxygen binding; OF = other function; PS = protein synthesis; RPD = reproduction; MCR = muscle contraction regulation; PM = protein metabolism. Despite the fact that different tissues were used, 31 out of 45 genes show congruent directions of significant regulatory changes in juvenile and adult fish (last column), which points towards ubiquitous patterns of expression divergence at juvenile and adult stages in a common laboratory environment.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2148-9-59-S2.doc]

Additional file 3

Regulatory changes between juvenile dwarf and normal whitefish at candidate adaptive traits. *Comparison of significant regulatory changes between juvenile dwarf and normal whitefish with patterns observed at candidate adaptive traits. Derome et al. [20] and St-Cyr et al. [22] analysed gene expression in white muscle and liver tissue of adult fish from species pairs in natural lakes (dwarfs and normals from Cliff Lake and Indian Pond) and identified candidate adaptive traits based on patterns of parallel divergence. 96 EST clones that match candidate adaptive traits displayed significant patterns of divergence in whole juvenile dwarf and normal whitefish in this study (Lake Témiscouata "dwarf" and Lake Aylmer "normal" respectively). Column contents correspond with Additional file 2 [see Additional file 2], with an additional column describing regulatory changes found by St-Cyr et al. [22] for adult fish liver tissue and the following additional biological functions as described in St-Cyr et al. [22]: DT = detoxification; LM = lipid metabolism; BT = blood and transport; GLF = germ-line formation; PD = Protein degradation. A minus (-) indicates that gene expression divergence was not tested or detected in the corresponding adult tissue. Ten out of 26 genes clones show regulatory changes in the controlled environment that are congruent with the patterns observed in candidate adaptive traits in independent dwarf normal pairs in natural lakes. Eight of these ten genes are related to energy metabolism, which shows that regulatory changes between dwarf and normal whitefish related to this function are more constant across different environments, tissues and life history stages than those related to other functions.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2148-9-59-S3.doc]

Acknowledgements

We would like to thank the Ministère des Ressources Naturelles et de la Faune du Québec (MRNF) for issuing scientific collecting permits for our fieldwork. Jean-Christophe Therrien and Serge Higgins (LARSA) helped to spawn and maintain fish in the laboratory. We thank the members of the Bernatchez' lab for help with the experiments and analysis. Christian Roberge provided helpful suggestions to improve MS. This project was funded by a postdoctoral research stipend to AN (NO 762/1-1) from the German Research Foundation, a postgraduate scholarship from the Natural Science and Engineering Research (NSERC) to SR and an NSERC strategic program to LB. The supporting agencies had no role in planning or any other aspect of this study.

References

1. Coyne JA, Orr HA: **Speciation.** Sinauer Associates, Sunderland, MA; 2004.
2. Orr HA, Presgraves D: **Speciation by postzygotic isolation: forces, genes and molecules.** *BioEssays* 2000, **22:**1085-1094.
3. Presgraves DC: **A Fine-Scale Genetic Analysis of Hybrid Incompatibilities in Drosophila.** *Genetics* 2003, **163:**955-972.
4. Orr HA, Turelli M: **The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities.** *Evolution* 2001, **55:**1085-1094.
5. Wu C-I: **The genic view of the process of speciation.** *J Evol Biol* 2001, **14:**851-865.
6. Wilding CS, Butlin RK, Grahame J: **Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers.** *J Evol Biol* 2001, **14:**611-619.
7. Nolte AW, Freyhof J, Tautz D: **When invaders meet locally adapted types: rapid moulding of hybrid zones between sculpins (Cottus, Pisces) in the Rhine system.** *Molecular Ecology* 2006, **15:**1983-1993.
8. Dieckmann U, Doebeli M: **On the origin of species by sympatric speciation.** *Nature* 1999, **400:**354-357.
9. Dieckmann U, Doebeli M, Metz JAJ, Tautz D: *Adaptive speciation* Cambridge University Press. Cambridge, UK; 2004.
10. King MC, Wilson AC: **Evolution at two levels in humans and chimpanzees.** *Science* 1975, **188:**107-16.
11. Carroll S: **Homeotic genes and the evolution of arthropods and chordates.** *Nature* 1995, **376:**479-485.
12. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA: **The evolution of transcriptional regulation in eukaryotes.** *Mol Biol Evol* 2003, **20:**1377-419.
13. Bernatchez L: **Ecological theory of adaptive radiation: an empirical assessment from coregonine fishes (Salmoniformes).** In *Evolution illuminated: salmon and their relatives* Edited by: Hendry AP, Stearns SC. Oxford Univ. Press, Oxford; 2004:175-207.
14. Landry L, Vincent WF, Bernatchez L: **Parallelism between limnological features and phenotypic evolution of lake whitefish dwarf ecotypes.** *J Evol Biol* 2007, **20:**971-984.
15. Campell D, Bernatchez L: **Genomic scan using AFLP markers as a means to assess the role of directional selection in the divergence of sympatric whitefish ecotypes.** *Mol Biol Evol* 2004, **21:**945-956.
16. Rogers SM, Bernatchez L: **The Genetic Architecture of Ecological Speciation and the Association with Signatures of Selection in Natural Lake Whitefish (*Coregonus* sp. Salmonidae) Species Pairs.** *Mol Biol Evol* 2007, **24:**1423.
17. Rogers SM, Bernatchez L: **The genetic basis of intrinsic and extrinsic postzygotic isolation jointly promoting speciation in the lake whitefish species complex (*Coregonus clupeaformis*).** *J Evol Biol* 2006, **19:**1979-1994.
18. Rise ML, von Schalburg KR, Brown GD, Mawer MA, Devlin RH, Kuipers N, Busby M, Beetz-Sargent M, Alberto R, Gibbs AR, Hunt P, Shukin R, Zeznik JA, Nelson C, Jones SMR, Smailus DE, Jones SJM, Schein JE, Marra MA, Butterfield YSN, Stott JM, Ng SHS, Davidson WS, Koop BF: **Development and Application of a Salmonid EST Database and cDNA Microarray: Data Mining and Interspecific Hybridization Characteristics.** *Genome Research* 2004, **14:**478-490.
19. von Schalburg KR, Rise ML, Cooper GA, Brown GD, Gibbs AR, Nelson CC, Davidson WS, Koop BF: **Fish and chips: Various methodologies demonstrate utility of a 16,006-gene salmonid microarray.** *BMC Genomics* 2005, **6:**126.
20. Derome N, Duchesne P, Bernatchez L: **Parallelism in gene transcription among sympatric lake whitefish ecotypes (*Coregonus clupeaformis* Mitchill).** *Molecular Ecology* 2006, **15:**1239-1250.
21. Derome N, Bougas B, Rogers SM, Whiteley A, Labbe A, Laroche J, Bernatchez L: **Pervasive sex-linked effects on transcription regulation as revealed by eQTLmapping in lake whitefish species pairs (*Coregonus* sp, Salmonidae).** *Genetics* 2008, **179(4):**1903-1917.
22. St-Cyr J, Derome N, Bernatchez L: **The transcriptomics of life-history trade-offs in whitefish species pairs (Coregonus sp.).** *Molecular Ecology* in press.

23. Haeckel E: *Generelle Morphologie der Organismen. Bd. 2: Allgemeine Entwickelungsgeschichte der Organismen* Georg Reimer: Berlin; 1866.
24. von Baer KE: *Über Entwicklungsgeschichte der Thiere: Beobachtung und Reflexion* Königsberg: Bornträger; 1828.
25. Zeldith ML, Swiderski DL, Sheets HD, Fink WL: *Geometric Morphometrics for Biologists: A Primer* Academic Press, London; 2003.
26. Leek JT, Storey JD: **Capturing heterogeneity in gene expression studies by "surrogate variable analysis.".** *PLoS Genetics* 2007, **3:**e161.
27. Arbeitman MN, Furlong EEM, Imam F, Johnson E, Null BH, Baker BS, Krasnow MA, Scott MP, Davis RW, White KP: **Gene Expression During the Life Cycle of *Drosophila melanogaster*.** *Science* 2002, **297:**2270-2275.
28. Pourquié O: **The Segmentation Clock: Converting Embryonic Time into Spatial Pattern.** *Science* 2003, **301:**328-330.
29. Chouinard A, Bernatchez L: **A study of trophic niche partitioning between larval populations of reproductively isolated whitefish (Coregonus sp.) ecotypes.** *Journal of Fish Biology* 1998, **53:**1231-1242.
30. Trudel M, Tremblay A, Schetagne R, Rasmussen JB: **Why are dwarf fish so small? An energetic analysis of polymorphism in lake whitefish (Coregonus clupeaformis).** *Can J Fish Aquat Sci* 2001, **58:**394-405.
31. Slack JMW, Holland PWH, Graham CF: **The zootype and the phylotypic stage.** *Nature* 1993, **362:**490-492.
32. Richardson MK: **Vertebrate evolution: the developmental origins of adult variation.** *BioEssays* 1999, **21:**604-613.
33. True JR, Haag ES: **Developmental system drift and flexibility in evolutionary trajectories.** *Evolution & Development* 2001, **3:**109-119.
34. Hall BK: **Phylotypic stage or phantom: is there a highly conserved embryonic stage in vertebrates?** *Trends Ecol Evol* 1997, **12:**461-463.
35. Moehring AJ, Teeter KC, Noor MAF: **Genome-Wide Patterns of Expression in Drosophila Pure Species and Hybrid Males. II. Examination of Multiple-Species Hybridizations, Platforms, and Life Cycle Stages.** *Mol Biol Evol* 2007, **24:**137-145.
36. Ortíz-Barrientos D, Counterman BA, Noor MAF: **Gene expression divergence and the origin of hybrid dysfunctions.** *Genetica* 2007, **129:**71-81.
37. Landry CR, Hartl DL, Ranz JM: **Genome clashes in hybrids: insights from gene expression.** *Heredity* 2007, **99:**483-493.
38. Lu G, Bernatchez L: **Experimental evidence for reduced hybrid viability between dwarf and normal ecotypes of lake whitefish (*Coregonus clupeaformis* Mitchill).** *Proc R Soc Lond B Biol Sci* 1998, **265:**1025-1030.
39. Ortiz-Barrientos D, Reiland J, Hey J, Noor MAF: **Recombination and the divergence of hybridizing species.** *Genetica* 2002, **116:**167-178.
40. Tautz D, Schmid KJ: **From genes to individuals: developmental genes and the generation of the phenotype.** *Philos Trans R Soc Lond B Biol Sci* 1998, **353:**231-240.
41. Rottscheidt R, Harr B: **Extensive additivity of gene expression differentiates subspecies of the house mouse.** *Genetics* 2007, **177:**1553-1567.
42. Tautz D: **A genetic uncertainty problem.** *Trends in Genetics* 2000, **16:**475-477.
43. van der Sluijs I, van Dooren TJM, Seehausen O, van Alphen JJM: **A test of fitness consequences of hybridization in sibling species of Lake Victoria cichlid fish.** *J Evol Biol* 2008, **21(2):**480-491.
44. Bolnick DI, Near TJ: **Tempo of hybrid inviability in centrarchid fishes (Teleostei: Centrachidae).** *Evolution* 2005, **59:**1754-1767.
45. Price TD, Bouvier MM: **The evolution of F1 postzygotic incompatibilities in birds.** *Evolution* 2002, **56:**2083-2089.
46. Kimmel CB, Ballard WW, Kimmel SE, Ullmann B, Schilling TF: **Stages of Embryonic Development of the Zebrafish.** *Developmental Dynamics* 1995, **203:**253-310.
47. Kerr MK, Martin M, Churchill GA: **Analysis of variance for gene expression microarray data.** *Journal of Computational Biology* 2000, **7:**819-837.
48. Kerr MK, Afshari CA, Bennett L, Bushel P, Martinez J, Walker NJ, Churchill GA: **Statistical analysis of a gene expression microarray experiment with replication.** *Statistica Sinica* 2002, **12:**203-217.
49. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
50. Hall TA: **BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT.** *Nucl Acids Symp Ser* 1999, **41:**95-98.
51. Hosack DA, Dennis G Jr, Sherman BT, Lane HC, Lempicki RA: **Identifying Biological Themes within Lists of Genes with EASE.** *Genome Biology* 2003, **4(10):**R70.

8.2 Matériel supplémentaire mentionné dans les chapitres deux à cinq.

Chapitre 2 : Nolte *et al.* 2009

Additional file 1: *Excel spreadsheet.* EST clones for which significant divergence was detected between embryos or juveniles of dwarf and normal whitefish.

http://www.biomedcentral.com/content/supplementary/1471- 2148-9-59-S1.xls

Additional file 2: *Word document.* Regulatory changes between juvenile dwarf and normal whitefish in a common environment.

http://www.biomedcentral.com/content/supplementary/1471- 2148-9-59-S2.doc

Additional file 3: *Word document.* Regulatory changes between juvenile dwarf and normal whitefish at candidate adaptive traits.

http://www.biomedcentral.com/content/supplementary/1471- 2148-9-59-S3.doc

Chapitre 2 : Renaut *et al.* 2009

Supplementary Table 1: *Excel spreadsheet.* List of significant transcripts for the different parental and hybrid group comparisons, FDR corrected permutated P values, Fold changes and relative gene expression

http://mbe.oxfordjournals.org/cgi/content/full/msp017/DC1?maxtoshow=&hits=10&RESULTFORMAT=&fulltext=renaut&searchid=1&FIRSTINDEX=0&resourcetype=HWCIT

Chapitre 3 : Renaut & Bernatchez (accepté)

Supplementary Table 1: *Excel spreadsheet.* List of all transcripts identified in the ANOVA as differentially expressed (q-values < 0.05), including fold changes ($\log_2$ values), annotations, clone sequences printed on the microarray and q-values for all six t-tests performed *a posteriori*.

Gene expression data deposited at Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo, Series accession GSE23095).

Chapitre 4 : Renaut *et al.* 2010

454 sequences archived at the National Center for Biotechnology Information http://www.ncbi.nlm.nih.gov/sites/sra (then search for SRA009800)

Chapitre 5 : Renaut *et al.* (*accepté*)

Supplementary table 5.1: *Excel spreadsheet.* Summary of genotyping frequency, amplicon sequences and BLAST annotations for all 96 SNPs retained for analyses of genetic differentiation in natural populations of sympatric dwarf and normal whitefish.

Supplementary table 5.2: *Excel spreadsheet.* Summary of genotyping frequency, amplicon sequences and BLAST annotations for all 87 SNP retained for analyses of association with phenotypic traits.

Supplementary table 5.3: *Excel spreadsheet.* Summary of observed and expected heterozygosities and F_{ST} for all 96 SNPs retained for analyses of genetic differentiation in natural populations.

Supplementary table 5.4: Summary of linkage disequilibrium for all loci in the association family and loci identified as outliers in the genome scan.

Suplementary table 5.1

SNP_marker	Cliff_Normal	Cliff_Normal	Cliff_Normal	Cliff_Normal	Cliff_Dwarf	Cliff_Dwarf	Cliff_Dwarf	Cliff_Dwarf	Webster_Normal	Webster_Normal	Webster_Normal	Webster_Normal	Webster_Dwarf
BU965641_re5_27	AA: 4	AG:11	GG:11	NA: 1	AA:16	AG: 9	GG: 5	NA: 1	AA:8	AG:4	GG:1	-	AA:16
CA037452_re5_29	AA: 1	AG:14	GG:11	NA: 1	GG:30	NA: 1	-	-	AA:7	AG:6	-	-	AA:16
CA037647_ca4_26	CC: 2	CT:18	NA: 2	TT: 5	CC: 5	CT:20	NA: 1	TT: 5	CC:9	CT:4	-	-	CC: 8
CA037876_re3_19	CC: 4	CT:14	NA: 1	TT: 8	CT: 6	NA: 1	TT:24	-	CC:4	CT:8	TT:1	-	CC: 6
CA038170_NDBC	AA:17	AG: 9	NA: 1	-	GG:29	NA: 2	-	-	AA: 1	AG: 1	GG:11	-	AA: 5
CA038790_re5_57	CC: 3	CT:14	NA: 1	TT: 9	NA: 1	TT:30	-	-	CC:3	CT:7	TT:3	-	CT: 4
CA039055_ND	GG:26	NA: 1	-	-	GG:18	GT:10	NA: 2	TT: 1	GG:13	-	-	-	GG:28
CA042392_re5_28	AA:10	AG:13	GG: 3	NA: 1	AA:26	AG: 4	NA: 1	-	AA:2	AG:7	GG:4	-	AA: 6
CA042792_622W	AA:26	NA: 1	-	-	AA:29	NA: 2	-	-	AA:10	AT: 3	-	-	AA:27
CA042951_ND	AA:25	CA: 1	NA: 1	-	AA: 3	CA:16	CC: 8	NA: 4	AA:5	CA:6	CC:2	-	AA: 7
CA044550_ca20_4	GG:26	NA: 1	-	-	GG:30	NA: 1	-	-	GG:8	TG:2	TT:3	-	GG:18
CA045465_re5_30	AA: 6	AG:16	GG: 1	NA: 4	AA:24	AG: 6	NA: 1	-	AG:5	GG:7	NA:1	-	AG:11
CA049476_re5_101	AA: 3	AG:23	NA: 1	-	AA: 8	AG:21	GG: 1	NA: 1	AA:7	AG:6	-	-	AA:20
CA051860_BC	CC:25	NA: 1	TT: 1	-	CC: 2	CT: 6	NA: 4	TT:19	CC:11	CT: 2	-	-	CC:14
CA052650_148M	AA: 1	CA:16	CC: 8	NA: 2	CC:29	NA: 2	-	-	CA:8	CC:5	-	-	CA: 2
CA053246_re5_89	GG: 6	GT:14	NA: 6	TT: 1	GT:25	NA: 1	TT: 5	-	GT:9	NA:1	TT:3	-	GT:11
CA053896_D	GG: 3	GT: 8	NA: 2	TT:14	GT:11	NA: 2	TT:18	-	TG:4	TT:9	-	-	NA: 1
CA054079_re5_14	GG:17	GT: 8	NA: 2	-	GG:11	GT:19	NA: 1	-	GG:4	GT:9	-	-	GG: 7
CA054630_BC	AA:15	AC: 7	CC: 4	NA: 1	CC:28	NA: 3	-	-	CC:13	-	-	-	CC:28
CA054959_380R	CC: 4	CT:10	NA: 4	TT: 9	NA:31	-	-	-	CC: 1	CT: 2	TT:10	-	CT: 5
CA056473_re3_17	CC: 4	CT:16	NA: 7	-	CC: 4	CT:24	NA: 2	TT: 1	CT:12	NA: 1	-	-	CC: 3
CA057176_NDBC	CG:13	GG:13	NA: 1	-	CG: 9	GG:19	NA: 3	-	CC:5	CG:8	-	-	CC: 4
CA057603_ND	AA:23	AT: 3	NA: 1	-	NA: 2	TT:29	-	-	AA:4	AT:8	TT:1	-	AA: 4
CA057987_156K	AA:10	AC:15	CC: 1	NA: 1	CC:30	NA: 1	-	-	AA:2	AC:7	CC:4	-	AA: 4
CA058340_re5_16	AA:24	AG: 1	NA: 2	-	AA:16	AG:13	GG: 1	NA: 1	AA:2	AG:5	GG:6	-	AA: 5
CA058958_re5_26	NA: 1	TT:26	-	-	NA: 1	TT:30	-	-	TT:13	-	-	-	TT:28
CA060324_154M	CA: 3	CC:23	NA: 1	-	CA:16	CC:14	NA: 1	-	CA:12	CC: 1	-	-	CA:16
CA061393_ND	GT:12	NA: 2	TT:13	-	GG:18	GT: 9	NA: 3	TT: 1	GG:4	GT:7	NA:1	TT:1	GG:17
CA062071_NDBC	GG:25	GT: 1	NA: 1	-	GG:26	GT: 4	NA: 1	-	GG:12	GT: 1	-	-	GG:18
CA063046_219W	AA:26	NA: 1	-	-	AA:30	NA: 1	-	-	AA:13	-	-	-	AA:28
CA063623_352M	GG:11	GT: 10	NA: 1	TT: 5	GT: 1	NA: 1	TT:29	-	GG:5	TG:2	TT:6	-	GG:12
CB492682_re5_136	GG:20	GT: 5	NA: 1	TT: 1	GG: 8	GT:10	NA: 1	TT:12	GG:10	TG: 3	-	-	GG:23
CB492725_ca3_21	CC:25	CT: 1	NA: 1	-	CC:23	CT: 7	NA: 1	-	CC:11	CT: 2	-	-	CC:23
CB492813_188K	GG:26	NA: 1	-	-	GG:21	GT: 8	NA: 1	TT: 1	GG:4	TG:8	TT:1	-	GG: 9
CB492855_ca4_11	AA:26	NA: 1	-	-	AA:30	NA: 1	-	-	AA:2	AC:5	CC:6	-	AA: 1
CB494318_BC	AA:25	NA: 2	-	-	AA:11	AG:14	GG: 1	NA: 5	AA:13	-	-	-	AA:28

CB496486_re5_37	AA: 6	CA:11	CC: 9	NA: 1	AA:30	NA: 1	-	-	AA:4	CA:6	CC:3	-	AA: 2
CB497584_re3_2	AG:25	GG: 1	NA: 1	-	AA: 7	AG:22	GG: 1	NA: 1	AA:4	AG:8	GG:1	-	AA: 1
CB497894_re5_154	CC: 6	CT:11	NA: 2	TT: 8	NA: 1	TT:30	-	-	CC:1	CT:6	TT:6	-	CT: 5
CB498771_re5_67	CC: 3	CT: 9	NA: 1	TT:14	CC:12	CT:12	NA: 1	TT: 6	CC:1	CT:5	TT:7	-	CC: 2
CB500248_502K	AA: 1	AC: 8	CC:17	NA: 1	CA: 1	CC:29	NA: 1	-	AA:4	AC:8	CC:1	-	AA:21
CB509509_397K	CC:23	NA: 4	-	-	NA:31	-	-	-	CC:13	-	-	-	CC:28
CB509723_re5_75	GG:14	GT: 8	NA: 1	TT: 4	GG:21	GT: 7	NA: 1	TT: 2	GG:12	TG: 1	-	-	GG:20
CB510585_ND	AA: 2	AG: 8	GG:16	NA: 1	AG: 1	GG:29	NA: 1	-	AA:1	AG:8	GG:4	-	AA:12
CB511030_339K	GG: 5	GT:11	NA: 1	TT:10	NA: 1	TT:30	-	-	GT:5	TT:8	-	-	GT: 2
CB512085_538R	CC: 3	CT:14	NA: 1	TT: 9	NA: 2	TT:29	-	-	CC:2	CT:5	TT:6	-	CC: 2
CB512493_ND	CC: 6	CT:12	NA: 3	TT: 6	CC: 8	CT:16	NA: 2	TT: 5	CC:4	CT:8	TT:1	-	CC: 9
CB514545_re5_125	NA: 1	TT:26	-	-	GT: 7	NA: 1	TT:23	-	GG:6	TG:7	-	-	GG: 2
CB516392_126M	AA: 4	CA:19	NA: 4	-	NA:31	-	-	-	AA:4	CA:9	-	-	AA: 4
CB516686_131R	AA:26	NA: 1	-	-	AA:30	NA: 1	-	-	AA:13	-	-	-	AA:28
cc000085_01AC	CA: 1	CC:26	-	-	CC:30	NA: 1	-	-	CC:13	-	-	-	CC:28
cc000102_01AG	CC:27	-	-	-	CC: 5	CT:17	NA: 1	TT: 8	CC:8	CT:5	-	-	CC:14
cc000129_01AT	AA:23	AT: 4	-	-	AA: 1	AT:27	NA: 1	TT: 2	AA: 1	AT:10	TT: 2	-	AA: 3
cc000167_01AC	GG:22	NA: 1	TG: 3	TT: 1	GG:24	NA: 1	TG: 6	-	GG:6	TG:4	TT:3	-	GG:19
cc000225_01AC	CA: 3	CC:23	NA: 1	-	CC:30	NA: 1	-	-	CA:5	CC:8	-	-	CA: 7
cc000236_01CT	CC:21	CT: 5	TT: 1	-	CC:30	NA: 1	-	-	CC:1	CT:6	TT:6	-	CC: 2
cc000240_01CT	AA: 5	AG:13	GG: 9	-	AA: 1	AG:12	GG:17	NA: 1	AG:5	GG:8	-	-	AA: 1
cc000258_01AC	GG:27	-	-	-	GG:30	NA: 1	-	-	GG:1	TG:8	TT:4	-	TG:11
cc000270_01CT	CC:27	-	-	-	CC:30	NA: 1	-	-	CC:8	CT:5	-	-	CC:22
cc000303_01CT	NA: 1	TT:26	-	-	CC:26	CT: 4	NA: 1	-	CT:5	TT:8	-	-	CC:13
cc000541_01AT	AA:25	NA: 2	-	-	AA: 7	AT:17	NA: 7	-	AA:3	AT:9	NA:1	-	AA: 8
cc000873_01AG	GG:25	NA: 2	-	-	GG:29	NA: 2	-	-	AG:5	GG:8	-	-	AG:10
cc000952_01GT	NA: 1	TT:26	-	-	NA: 1	TT:30	-	-	TT:13	-	-	-	TT:28
cc001175_01AG	AA: 3	AG:14	GG: 9	NA: 1	GG:30	NA: 1	-	-	AA:5	AG:6	GG:2	-	AA: 2
cc001365_01AG	AA:13	AG:14	-	-	AA:17	AG:13	NA: 1	-	AA:9	AG:4	-	-	AA: 4
cc001404_04GT	CC:27	-	-	-	AC:17	CC:11	NA: 3	-	AC:8	CC:5	-	-	AC:22
cc001422_01CT	CC:19	CT: 6	TT: 2	-	CC: 2	CT:15	NA: 2	TT:12	CC:12	CT: 1	-	-	CC:27
cc001461_02AT	AT:27	-	-	-	AT:29	NA: 2	-	-	AT:13	-	-	-	AT:28
cc001462_01CT	TT:27	-	-	-	CC:30	NA: 1	-	-	CC:4	TT:9	-	-	CC:18
cc001468_01CT	CT: 1	TT:26	-	-	NA: 2	TT:29	-	-	CT: 2	TT:11	-	-	TT:28
cc001516_01CT	AG:25	GG: 2	-	-	AG:17	GG:12	NA: 2	-	AA: 3	AG:10	-	-	AG:22
cc001576_01CT	CC:25	CT: 2	-	-	CC: 3	CT:14	NA: 1	TT:13	CC:13	-	-	-	CC:27
cc001605_01AG	CC:10	CT:17	-	-	CC: 7	CT:18	NA: 1	TT: 5	CC:7	CT:5	TT:1	-	CC: 4
cc001682_02CT	CC:27	-	-	-	CC:29	NA: 2	-	-	CC:13	-	-	-	CC:28
cc001705_02AG	CT:21	TT: 6	-	-	CC: 2	CT:21	NA: 1	TT: 7	CT:12	TT: 1	-	-	CC: 5
cc001720_01CT	AA: 7	AG:19	NA: 1	-	AA: 1	AG:26	GG: 3	NA: 1	AA: 1	AG:12	-	-	AG:17
cc001763_01GT	GG: 5	GT:16	TT: 6	-	GG:30	NA: 1	-	-	GG:5	GT:5	TT:3	-	GG:16
cc001810_03AG	CT:26	TT: 1	-	-	CT:29	NA: 1	TT: 1	-	CT:9	TT:4	-	-	CT:23
cc001837_07AT	AA: 1	AT:23	NA: 3	-	AA:11	AT:17	NA: 3	-	AT:13	-	-	-	AT:27

cc001862_03AG	CC:27	-	-	-	NA: 1	TT:30	-	-	CC:9	TT:4	-	-	CC:10
cc001934_01AT	AA: 1	AT:26	-	-	AA:11	AT:17	NA: 3	-	AT:13	-	-	-	AT:28
cc001937_01AG	CC:21	CT: 6	-	-	CC: 5	CT:11	NA: 2	TT:13	CC:8	CT:2	TT:3	-	CC: 7
cc002110_01AG	AG:27	-	-	-	AG:22	GG: 8	NA: 1	-	AG:12	GG: 1	-	-	AG:26
cc002119_06AC	CC:27	-	-	-	CC:30	NA: 1	-	-	CA: 2	CC:11	-	-	CA: 1
cc002133_01CT	CT:14	NA: 1	TT:12	-	CC: 2	CT:27	NA: 1	TT: 1	CT:8	TT:5	-	-	CC: 2
cc002148_01AG	AA:27	-	-	-	AA:30	NA: 1	-	-	AA:13	-	-	-	AA:28
cc002181_03CT	CC: 1	CT:12	TT:14	-	CC: 2	CT:12	NA: 1	TT:16	CT: 2	TT:11	-	-	CT: 4
cc002195_01AG	CC:11	CT:14	TT: 2	-	CC: 2	CT:10	NA: 1	TT:18	CC:2	CT:6	TT:5	-	CC: 5
CK990730_re5_31	NA: 1	TT:26	-	-	NA: 1	TT:30	-	-	TT:13	-	-	-	TT:28
CK990756_425M	GG: 3	NA: 1	TG: 7	TT:16	NA: 1	TT:30	-	-	GG:2	TG:7	TT:4	-	GG: 5
CX030416_re5_72	CC: 7	CT:16	NA: 1	TT: 3	NA: 1	TT:30	-	-	CT:4	TT:9	-	-	CT: 1
DW553899_re5_130	CC:15	CT: 9	NA: 1	TT: 2	CC: 1	CT:12	NA: 1	TT:17	CC:7	CT:4	TT:2	-	CC: 1
DY738545_re5_40	CC:15	CT: 9	NA: 1	TT: 2	NA: 1	TT:30	-	-	CC:1	CT:6	TT:6	-	CT: 5
EG755964_BC	CA: 2	CC:24	NA: 1	-	AA: 7	CA:17	CC: 5	NA: 2	CA: 2	CC:11	-	-	AA: 7
EG782906_re5_56	NA: 1	TT:26	-	-	NA: 1	TT:30	-	-	TT:13	-	-	-	TT:28
EG811444_re3_6	AA:26	NA: 1	-	-	AA:29	CA: 1	NA: 1	-	AA:13	-	-	-	AA:28

Webster_Dwarf	Webster_Dwarf	Webster_Dwarf	Indian_Normal	Indian_Normal	Indian_Normal	Indian_Normal	Indian_Dwarf	Indian_Dwarf	Indian_Dwarf	Indian_Dwarf	East_Normal	East_Normal	East_Normal	East_Normal	East_Dwarf
AG: 9	GG: 2	NA: 1	AA:14	AG:11	GG: 1	-	AA:15	AG: 6	NA: 1	-	AA:17	AG: 7	-	-	AA:17
AG: 6	GG: 5	NA: 1	AA:10	AG:11	GG: 5	-	AA: 7	AG:12	GG: 3	-	AA:11	AG:11	GG: 2	-	AA:13
CT:17	NA: 1	TT: 2	CC:20	CT: 5	TT: 1	-	CC:20	CT: 2	-	-	CC:21	CT: 3	-	-	CC:23
CT:18	TT: 4	-	CC:10	CT:10	TT: 6	-	CC: 5	CT:12	TT: 5	-	CC: 4	CT:12	TT: 8	-	CC: 2
AG:10	GG:13	-	AA: 6	AG:17	GG: 3	-	AA: 3	AG:11	GG: 8	-	AA: 9	AG:13	GG: 2	-	AA:12
TT:24	-	-	CC: 1	CT:10	TT:15	-	CC: 2	CT:12	TT: 8	-	CC: 2	CT:11	TT:11	-	CT: 6
-	-	-	GG:26	-	-	-	GG:22	-	-	-	GG:24	-	-	-	GG:24
AG:14	GG: 8	-	AA: 6	AG:11	GG: 9	-	AA: 9	AG:10	GG: 3	-	AA: 2	AG:13	GG: 9	-	AA: 9
AT: 1	-	-	AA:22	AT: 3	TT: 1	-	AA:21	AT: 1	-	-	AA:24	-	-	-	AA:22
CA:15	CC: 6	-	AA:20	CA: 5	CC: 1	-	AA:18	CA: 4	-	-	AA:15	CA: 8	CC: 1	-	AA:17
TG:10	-	-	GG:26	-	-	-	GG:21	TG: 1	-	-	GG:7	TG:9	TT:8	-	GG:11
GG:15	NA: 2	-	AA: 2	AG:13	GG: 9	NA: 2	AG: 6	GG:15	NA: 1	-	AA: 2	AG:12	GG: 6	NA: 4	AG: 7
AG: 7	NA: 1	-	AA: 5	AG:21	-	-	AA:14	AG: 8	-	-	AA:13	AG:10	NA: 1	-	AA:14
CT:12	TT: 2	-	CC:23	CT: 2	TT: 1	-	CC:8	CT:7	TT:7	-	CC:14	CT: 9	TT: 1	-	CC:21
CC:26	-	-	CA: 9	CC:17	-	-	CC:22	-	-	-	CA: 8	CC:16	-	-	CA: 8
NA: 1	TT:16	-	GT:16	NA: 6	TT: 4	-	GT: 5	NA: 4	TT:13	-	GT: 9	TT:15	-	-	GG: 1
TG: 1	TT:26	-	GG: 1	NA: 3	TG: 8	TT:14	GG: 1	NA:14	TG: 5	TT: 2	GG: 1	NA: 2	TG: 6	TT:15	NA: 3
GT:16	NA: 4	TT: 1	GG: 8	GT:18	-	-	GG: 8	GT:14	-	-	GG:13	GT:11	-	-	GG:13
-	-	-	AC: 1	CC:25	-	-	CC:21	NA: 1	-	-	AA: 1	AC: 4	CC:19	-	CC:23
TT:23	-	-	CT: 2	TT:24	-	-	TT:22	-	-	-	CT: 5	TT:19	-	-	CT: 2
CT:17	NA: 6	TT: 2	CC: 2	CT:12	NA:10	TT: 2	CC: 1	CT: 4	NA:17	-	CC: 3	CT:17	NA: 2	TT: 2	CC: 3
CG:21	GG: 3	-	CC: 2	CG:17	GG: 6	NA: 1	CC:14	CG: 8	-	-	CC: 1	CG:17	GG: 6	-	CG:18
AT: 7	TT:17	-	AA:23	AT: 3	-	-	AA: 7	AT:12	TT: 3	-	AA:16	AT: 7	TT: 1	-	AA:14
AC:16	CC: 8	-	AA: 2	AC: 6	CC:17	NA: 1	AA:4	AC:9	CC:9	-	AA:16	AC: 6	CC: 2	-	AA:18
AG:18	GG: 5	-	AA:13	AG: 9	GG: 4	-	AA: 3	AG: 7	GG:12	-	AA: 2	AG:14	GG: 8	-	AA: 2
-	-	-	TT:26	-	-	-	TT:22	-	-	-	TT:24	-	-	-	TT:24
CC:12	-	-	CA:21	CC: 5	-	-	CA:11	CC:10	NA: 1	-	CA:21	CC: 3	-	-	CA:10
GT: 9	TT: 2	-	GG:10	GT:12	NA: 1	TT: 3	GG:15	GT: 7	-	-	GG:13	GT:10	NA: 1	-	GG:13
GT: 9	TT: 1	-	GG:20	GT: 6	-	-	GG:15	GT: 7	-	-	GG:21	GT: 3	-	-	GG:20
-	-	-	AA:26	-	-	-	AA:22	-	-	-	AA:24	-	-	-	AA:24
TG:14	TT: 2	-	GG:12	TG:10	TT: 4	-	GG:5	TG:9	TT:8	-	GG: 2	TG:14	TT: 8	-	GG: 3
TG: 5	-	-	GG:24	TG: 2	-	-	GG: 9	TG:11	TT: 2	-	GG:22	TG: 2	-	-	GG:23
CT: 4	TT: 1	-	CC:21	CT: 4	TT: 1	-	CC:18	CT: 3	TT: 1	-	CC:16	CT: 7	NA: 1	-	CC:12
TG:11	TT: 8	-	GG:23	TG: 3	-	-	GG:20	TG: 2	-	-	GG:10	TG:11	TT: 3	-	GG: 4
AC: 7	CC:20	-	AA:11	AC:13	CC: 2	-	AA:7	AC:8	CC:6	NA:1	AA: 1	AC:11	CC:12	-	AA: 3
-	-	-	AA:24	NA: 2	-	-	AA:18	NA: 4	-	-	AA:24	-	-	-	AA:23

CA:10	CC:16	-	AA:17	CA: 9	-	-	CA:16	CC: 6	-	-	AA:10	CA: 8	CC: 6	-	AA:10
AG:24	GG: 3	-	AA: 9	AG:17	-	-	AA: 1	AG:17	GG: 3	NA: 1	AA: 4	AG:20	-	-	AA: 4
TT:23	-	-	CC: 5	CT:14	TT: 7	-	CT: 1	NA: 1	TT:20	-	TT:24	-	-	-	CT: 2
CT:11	TT:15	-	CC: 6	CT:18	TT: 2	-	CC:22	-	-	-	CC:10	CT: 9	TT: 5	-	CC: 4
AC: 7	-	-	AA: 1	AC: 5	CC:20	-	AA:14	AC: 8	-	-	AA:17	AC: 7	-	-	AA:20
-	-	-	CC:26	-	-	-	CC:22	-	-	-	CC:24	-	-	-	CC:24
TG: 6	TT: 2	-	GG:14	TG:12	-	-	GG:12	NA: 7	TG: 2	TT: 1	GG:23	NA: 1	-	-	GG:24
AG:12	GG: 4	-	AA: 1	AG:13	GG:12	-	AA: 7	AG:14	GG: 1	-	AA:18	AG: 6	-	-	AA:15
TT:26	-	-	GG: 1	GT: 2	TT:23	-	TT:22	-	-	-	GT: 8	TT:16	-	-	GT: 1
CT: 9	TT:17	-	CC: 3	CT:14	TT: 9	-	CT: 3	TT:19	-	-	CC: 2	CT:16	TT: 6	-	CT: 7
CT:13	NA: 2	TT: 4	CC:10	CT:14	NA: 1	TT: 1	CC:15	CT: 6	NA: 1	-	CC:10	CT: 7	TT: 7	-	CC: 1
TG:12	TT:14	-	GG: 5	TG:12	TT: 9	-	TG: 7	TT:15	-	-	GG:13	TG:10	TT: 1	-	GG:19
CA:23	NA: 1	-	AA: 5	CA:18	CC: 1	NA: 2	CA:16	CC: 1	NA: 5	-	AA: 2	CA:21	CC: 1	-	AA: 9
-	-	-	AA:26	-	-	-	AA:21	NA: 1	-	-	AA:24	-	-	-	AA:24
-	-	-	CC:26	-	-	-	CC:21	NA: 1	-	-	CC:24	-	-	-	CC:24
CT:12	TT: 2	-	CC:13	CT:13	-	-	CC:11	CT:11	-	-	CC:23	CT: 1	-	-	CC:24
AT:25	-	-	AA:10	AT:16	-	-	AA: 5	AT:16	NA: 1	-	AA:10	AT:14	-	-	AA:13
TG: 9	-	-	GG:18	TG: 7	TT: 1	-	GG: 5	NA: 1	TG:10	TT: 6	GG:16	TG: 8	-	-	GG:22
CC:21	-	-	CA: 1	CC:25	-	-	CA: 8	CC:14	-	-	AA: 6	CA: 6	CC:12	-	AA: 5
CT:12	TT:14	-	CC: 9	CT:13	TT: 4	-	CC:20	CT: 2	-	-	CC: 6	CT:12	TT: 6	-	CC: 3
AG:13	GG:14	-	AG: 7	GG:19	-	-	AA: 6	AG:11	GG: 5	-	AA: 2	AG:13	GG: 9	-	AA: 3
TT:17	-	-	GG:16	TG: 9	TT: 1	-	GG: 8	TG:11	TT: 3	-	GG: 5	TG:10	TT: 9	-	GG: 2
CT: 6	-	-	CC:16	CT: 7	TT: 3	-	CC:10	CT:11	TT: 1	-	CC: 8	CT:11	TT: 5	-	CC: 6
CT:14	TT: 1	-	CC: 9	CT:10	TT: 7	-	CC: 8	CT:11	TT: 3	-	CC: 7	CT:11	TT: 6	-	CC: 2
AT:20	-	-	AT:26	-	-	-	AA: 4	AT:10	NA: 8	-	AA: 7	AT:15	NA: 1	TT: 1	AA: 3
GG:18	-	-	AG: 4	GG:22	-	-	GG:22	-	-	-	AG:11	GG:12	NA: 1	-	AA: 1
-	-	-	TT:26	-	-	-	NA: 1	TT:21	-	-	TT:24	-	-	-	TT:24
AG: 8	GG:18	-	AA: 6	AG:13	GG: 7	-	AG: 1	GG:21	-	-	AA:15	AG: 5	GG: 3	NA: 1	AA: 4
AG:24	-	-	AA:14	AG:11	NA: 1	-	AA: 6	AG:15	NA: 1	-	AA:11	AG:13	-	-	AA:14
CC: 6	-	-	AC: 2	CC:24	-	-	AC: 4	CC:18	-	-	AC: 8	CC:15	NA: 1	-	AA: 1
CT: 1	-	-	CC:25	CT: 1	-	-	CC: 9	CT:11	TT: 2	-	CC:21	CT: 2	TT: 1	-	CC:24
-	-	-	AT:25	TT: 1	-	-	AT:20	NA: 1	TT: 1	-	AT:24	-	-	-	AT:24
TT:10	-	-	CC: 2	TT:24	-	-	CC: 2	TT:20	-	-	CC: 1	TT:23	-	-	TT:24
-	-	-	CT: 1	TT:25	-	-	TT:22	-	-	-	CT: 1	TT:23	-	-	CT: 1
GG: 6	-	-	AA: 1	AG:25	-	-	AG:19	GG: 3	-	-	AA: 6	AG:18	-	-	AA: 3
CT: 1	-	-	CC:17	CT: 9	-	-	CC:22	-	-	-	CC:22	CT: 2	-	-	CC:22
CT:21	TT: 3	-	CC:11	CT:14	NA: 1	-	CC: 5	CT:14	TT: 3	-	CC: 8	CT:14	TT: 2	-	CC:12
-	-	-	CC:26	-	-	-	CC:22	-	-	-	CC:24	-	-	-	CC:24
CT:23	-	-	CT:20	NA: 1	TT: 5	-	CC: 1	CT:18	TT: 3	-	CC: 1	CT: 3	TT:20	-	CT: 8
GG:11	-	-	AA:10	AG:16	-	-	AG:16	GG: 6	-	-	AA: 1	AG:20	GG: 2	NA: 1	AA: 2
GT:12	-	-	GG: 4	GT:11	TT:11	-	GG:17	GT: 3	TT: 2	-	GG: 6	GT:10	NA: 2	TT: 6	GG: 1
TT: 5	-	-	CT:26	-	-	-	CT:22	-	-	-	CT:20	NA: 1	TT: 3	-	CT:23
NA: 1	-	-	AT:26	-	-	-	AT:21	NA: 1	-	-	AT:24	-	-	-	AT:23

TT:18	-	-	CC:25	TT: 1	-	-	CC:20	TT: 2	-	-	CC:22	NA: 1	TT: 1	-	CC:24
-	-	-	AT:26	-	-	-	AT:22	-	-	-	AT:24	-	-	-	AT:24
CT:13	TT: 8	-	CC:22	CT: 4	-	-	CT: 8	NA: 1	TT:13	-	CT: 2	TT:22	-	-	CT: 1
GG: 2	-	-	AG:20	GG: 5	NA: 1	-	AG:22	-	-	-	AG:24	-	-	-	AG:24
CC:27	-	-	CA: 3	CC:23	-	-	CC:22	-	-	-	AA: 1	CA: 3	CC:20	-	CA: 4
CT:25	NA: 1	-	CT: 4	NA: 3	TT:19	-	CT: 3	NA:18	TT: 1	-	CC: 2	CT:21	NA: 1	-	CC: 2
-	-	-	AA:25	NA: 1	-	-	AA:21	NA: 1	-	-	AA:24	-	-	-	AA:24
TT:24	-	-	CT: 1	TT:25	-	-	CT: 7	TT:15	-	-	CT:10	TT:14	-	-	CC: 1
CT:17	TT: 6	-	CC: 5	CT:12	TT: 9	-	CC: 7	CT:12	TT: 3	-	CC:11	CT:11	TT: 2	-	CC:10
-	-	-	TT:26	-	-	-	TT:22	-	-	-	TT:24	-	-	-	TT:24
TG:16	TT: 7	-	TG: 5	TT:21	-	-	TG: 6	TT:16	-	-	GG: 3	TG:14	TT: 7	-	GG: 7
TT:27	-	-	CT: 7	TT:19	-	-	CC: 2	CT: 3	TT:17	-	CC: 8	CT:12	TT: 4	-	CC: 5
CT:11	TT:16	-	CC: 8	CT:13	TT: 5	-	TT:22	-	-	-	CC:19	CT: 4	TT: 1	-	CC:20
TT:23	-	-	CC:10	CT: 6	TT:10	-	CT: 1	NA:12	TT: 9	-	CC: 3	CT:12	NA: 2	TT: 7	CT: 9
CA:12	CC: 9	-	AA: 1	CA: 9	CC:16	-	AA:5	CA:8	CC:9	-	CA:11	CC:13	-	-	CA: 5
-	-	-	TT:26	-	-	-	TT:22	-	-	-	TT:24	-	-	-	TT:24
-	-	-	AA:26	-	-	-	AA:22	-	-	-	AA:24	-	-	-	AA:24

East_Dwarf	East_Dwarf	East_Dwarf	Témiscouata	Témiscouata	Témiscouata	Témiscouata	Témiscouata	Témiscouata	Témiscouata	Témiscouata	Pohénégamook	Pohénégamook	Pohénégamook	Pohénégamook	amplicon_sequence
AG: 7	-	-	AA:18	AG: 5	GG: 1	-	AA:14	AG: 9	GG: 1	-	AA:10	AG: 7	NA: 7	-	GCTTA
AG: 9	GG: 1	NA: 1	AA: 5	AG:12	GG: 7	-	AA:13	AG: 8	GG: 3	-	AA: 3	AG:10	GG: 2	NA: 9	GGACG
CT: 1	-	-	CC:18	CT: 5	NA: 1	-	CC:23	CT: 1	-	-	CC:11	CT: 3	NA:10	-	TCTCC
CT:10	TT:12	-	CC: 1	CT:14	TT: 9	-	CC: 2	CT:17	TT: 5	-	CC: 2	CT:12	NA: 7	TT: 3	AAACC
AG:10	GG: 1	NA: 1	AA: 2	AG: 7	GG:15	-	AA: 1	AG:11	GG:12	-	AG: 4	GG:11	NA: 9	-	GAAgG
TT:18	-	-	CT: 6	TT:18	-	-	CT: 4	TT:20	-	-	CC:1	CT:6	NA:8	TT:9	TGAAA
-	-	-	GG:24	-	-	-	GG:24	-	-	-	GG:17	NA: 7	-	-	GCGGK
AG:10	GG: 5	-	AA: 3	AG:14	GG: 7	-	AA: 7	AG:11	GG: 6	-	AA:9	AG:7	GG:1	NA:7	GCAGC
AT: 2	-	-	AA:23	AT: 1	-	-	AA:21	AT: 3	-	-	AA:17	NA: 7	-	-	CACGC
CA: 7	-	-	AA:12	CA: 8	CC: 4	-	AA: 8	CA:14	CC: 1	NA: 1	AA:7	CA:9	CC:1	NA:7	TGGCC
TG:11	TT: 2	-	GG:15	TG: 7	TT: 2	-	GG:14	TG: 9	TT: 1	-	GG:6	NA:7	TG:8	TT:3	ACTTC
GG:17	-	-	AG: 8	GG:15	NA: 1	-	AG: 9	GG:13	NA: 2	-	AA:2	AG:9	GG:4	NA:9	AGGAC
AG:10	-	-	AA: 2	AG:22	-	-	AA: 8	AG:16	-	-	AA: 5	AG:12	NA: 7	-	CCCAT
CT: 2	TT: 1	-	CC:15	CT: 9	-	-	CC:11	CT: 9	TT: 4	-	CC:9	CT:7	NA:8	-	TTGAG
CC:16	-	-	AA: 2	CA:12	CC:10	-	CA:13	CC:10	NA: 1	-	CA: 4	CC:10	NA:10	-	CGCAG
GT:15	TT: 8	-	GG: 1	GT:16	NA: 1	TT: 6	GT:22	NA: 1	TT: 1	-	GT: 6	NA: 7	TT:11	-	AGTAC
TG: 6	TT:15	-	GG: 4	TG:10	TT:10	-	GG: 1	TG:14	TT: 9	-	GG:5	NA:8	TG:7	TT:4	ACTAT
GT:11	-	-	GG: 7	GT:17	-	-	GG:18	GT: 6	-	-	GG: 9	GT: 5	NA:10	-	GGACT
NA: 1	-	-	AC: 4	CC:20	-	-	AC: 5	CC:19	-	-	AC: 1	CC:16	NA: 7	-	CGACG
NA: 1	TT:21	-	CT: 3	TT:21	-	-	CT: 1	TT:23	-	-	CC: 1	CT: 3	NA: 7	TT:13	GCACC
CT:18	NA: 3	-	CC: 1	CT:21	TT: 2	-	CT:24	-	-	-	CT: 7	NA:14	TT: 3	-	GCTCT
GG: 6	-	-	CC: 4	CG:13	GG: 7	-	CC: 4	CG:15	GG: 5	-	CG:12	GG: 5	NA: 7	-	TCCGC
AT:10	-	-	AA: 9	AT:11	TT: 4	-	AA:12	AT:10	TT: 2	-	AA:3	AT:6	NA:8	TT:7	TAGAT
AC: 6	-	-	AA: 7	AC:10	CC: 7	-	AA: 6	AC:12	CC: 6	-	AA: 2	AC:10	CC: 4	NA: 8	GCATT
AG:16	GG: 6	-	AA: 3	AG:12	GG: 9	-	AA: 3	AG: 8	GG:13	-	AA:4	AG:6	GG:5	NA:9	AGGCG
-	-	-	TT:24	-	-	-	TT:24	-	-	-	NA: 7	TT:17	-	-	TCGTA
CC:14	-	-	CA:15	CC: 9	-	-	AA: 2	CA:17	CC: 4	NA: 1	AA:1	CA:7	CC:7	NA:9	GCTGG
GT:10	TT: 1	-	GG:11	GT: 9	TT: 4	-	GG:13	GT:11	-	-	GG:1	GT:8	NA:8	TT:7	GCCCT
GT: 4	-	-	GG:19	GT: 4	TT: 1	-	GG:23	GT: 1	-	-	GG:17	NA: 7	-	-	GGCGT
-	-	-	AA:24	-	-	-	AA:24	-	-	-	AA:17	NA: 7	-	-	AAGGC
NA: 1	TG: 9	TT:11	TG: 8	TT:16	-	-	GG: 4	TG:11	TT: 9	-	GG: 1	NA:10	TG: 6	TT: 7	CCTGT
NA: 1	-	-	GG:19	TG: 5	-	-	GG:21	TG: 3	-	-	GG:11	NA:10	TG: 3	-	CTCGC
CT: 9	NA: 3	-	CC:10	CT:11	NA: 1	TT: 2	CC:17	CT: 6	TT: 1	-	CC:9	CT:6	NA:7	TT:2	GGGAC
TG:13	TT: 7	-	GG:15	TG: 8	TT: 1	-	GG:11	TG:11	TT: 2	-	GG:11	NA: 8	TG: 2	TT: 3	GTCAG
AC:12	CC: 9	-	AA:10	AC:10	CC: 4	-	AA:10	AC: 9	CC: 5	-	AA:3	AC:9	CC:5	NA:7	TTCGA
NA: 1	-	-	AA:24	-	-	-	AA:24	-	-	-	AA:17	NA: 7	-	-	AGGCC

CA:12	CC: 2	-	AA:12	CA:11	CC: 1	-	AA:11	CA:13	-	-	CA:9	CC:6	NA:9	-	GTACA
AG:19	GG: 1	-	AA: 9	AG:15	-	-	AA: 9	AG:15	-	-	AA: 4	AG:13	NA: 7	-	AGGAC
TT:22	-	-	CC: 2	CT: 8	TT:14	-	CC: 1	CT:11	TT:12	-	CT:7	NA:8	TT:9	-	ATAAC
CT:15	TT: 5	-	CC: 1	CT:17	TT: 6	-	CC:11	CT: 8	TT: 5	-	CT: 6	NA: 7	TT:11	-	CGCGC
AC: 2	CC: 1	NA: 1	AA: 2	AC:13	CC: 9	-	AA: 5	AC:13	CC: 6	-	AA:8	AC:9	NA:7	-	CTTCA
-	-	-	AC: 3	CC:21	-	-	CC:24	-	-	-	AC: 1	CC:16	NA: 7	-	CGGCA
-	-	-	GG:21	TG: 3	-	-	GG:17	TG: 7	-	-	GG:13	NA: 7	TG: 3	TT: 1	CATTT
AG: 8	GG: 1	-	AA: 8	AG:11	GG: 5	-	AA:14	AG: 7	GG: 3	-	AA:9	AG:7	GG:1	NA:7	GCCAC
TT:23	-	-	GG: 1	GT: 7	TT:16	-	GT: 6	TT:18	-	-	GT: 5	NA: 8	TT:11	-	AGTCT
TT:17	-	-	CC: 2	CT: 7	TT:15	-	CT:10	TT:14	-	-	CT: 4	NA: 7	TT:13	-	TCTAC
CT: 6	TT:17	-	CC:16	CT: 8	-	-	CC:19	CT: 4	TT: 1	-	CC:13	CT: 3	NA: 8	-	GACTG
TG: 3	TT: 2	-	GG:16	TG: 6	TT: 2	-	GG:15	TG: 9	-	-	GG:13	NA:10	TG: 1	-	CCACA
CA:14	NA: 1	-	AA: 6	CA:18	-	-	AA: 2	CA:22	-	-	AA: 2	CA:12	NA:10	-	CACGC
-	-	-	AA:24	-	-	-	AA:24	-	-	-	AA:17	NA: 7	-	-	TAGGC
-	-	-	CC:24	-	-	-	CC:22	NA: 2	-	-	CC:15	NA: 9	-	-	CCCAC
-	-	-	CC:18	CT: 6	-	-	CC:20	CT: 4	-	-	CC: 7	CT: 5	NA:10	TT: 2	GCGGG
AT:11	-	-	AA: 7	AT:17	-	-	AA: 4	AT:19	TT: 1	-	AT:13	NA: 9	TT: 2	-	GGGGT
TG: 2	-	-	GG:22	TG: 1	TT: 1	-	GG:20	TG: 3	TT: 1	-	GG:15	NA: 9	-	-	AAGCC
CA:13	CC: 6	-	AA: 1	CA: 7	CC:16	-	AA: 1	CA: 8	CC:15	-	CA: 4	CC:11	NA: 9	-	CCAGA
CT:13	TT: 8	-	CC: 2	CT: 7	TT:15	-	CC: 1	CT: 3	TT:20	-	CC: 1	CT: 4	NA: 9	TT:10	GAGCT
AG: 8	GG:13	-	AG: 7	GG:17	-	-	AG: 4	GG:20	-	-	AA:4	AG:7	GG:4	NA:9	AAAAC
TG:12	TT:10	-	GG:15	TG: 7	TT: 2	-	GG:15	TG: 6	TT: 3	-	GG: 9	NA:10	TG: 4	TT: 1	GTTTC
CT:12	TT: 6	-	CC:12	CT:11	TT: 1	-	CC:15	CT: 7	TT: 2	-	CC: 4	CT: 7	NA:10	TT: 3	TACCC
CT:13	TT: 9	-	CC: 2	CT: 7	TT:15	-	CT: 7	TT:17	-	-	CC: 1	CT: 2	NA:10	TT:11	AATATA
AT:13	NA: 7	TT: 1	AA: 2	AT:20	NA: 1	TT: 1	AA: 1	AT:21	NA: 2	-	AA: 5	AT: 7	NA:12	-	TTCAC
AG:15	GG: 7	NA: 1	AA: 4	AG:13	GG: 5	NA: 2	AG:16	GG: 6	NA: 2	-	AA: 1	AG: 3	GG: 1	NA:19	CGGGG
-	-	-	NA: 1	TT:23	-	-	GT: 1	NA: 1	TT:22	-	NA: 9	TT:15	-	-	TTTTCA
AG:13	GG: 7	-	AA: 9	AG:13	GG: 2	-	AA:14	AG: 7	GG: 2	NA: 1	AA:12	AG: 2	NA:10	-	AAACT
AG:10	-	-	AA: 4	AG:20	-	-	AA:15	AG: 9	-	-	AA: 6	AG: 8	NA:10	-	GGGAC
AC: 4	CC:19	-	AA: 1	AC:15	CC: 8	-	AA: 1	AC:20	CC: 2	NA: 1	AC:8	CC:7	NA:9	-	TAGAA
-	-	-	CC:23	CT: 1	-	-	CC:23	CT: 1	-	-	CC:13	CT: 2	NA: 9	-	TTTTG
-	-	-	AT:24	-	-	-	AT:24	-	-	-	AT:15	NA: 9	-	-	GGAAT
-	-	-	CC: 1	TT:23	-	-	TT:24	-	-	-	NA:11	TT:13	-	-	CGTTA
TT:23	-	-	CT: 6	TT:18	-	-	CC: 1	CT: 5	TT:18	-	NA: 9	TT:15	-	-	GTTAG
AG:20	GG: 1	-	AA: 6	AG:18	-	-	AA: 7	AG:17	-	-	AA: 3	AG:11	NA:10	-	GGAGC
CT: 2	-	-	CC:19	CT: 5	-	-	CC:14	CT: 9	TT: 1	-	CC:12	CT: 3	NA: 9	-	TGGTG
CT:10	NA: 2	-	CC:19	CT: 4	NA: 1	-	CC:20	CT: 3	NA: 1	-	CC:14	NA:10	-	-	TAAAG
-	-	-	CC:24	-	-	-	CC:23	CT: 1	-	-	CC:15	NA: 9	-	-	TGGGC
TT:16	-	-	CC: 1	CT:14	TT: 9	-	CC: 1	CT:14	TT: 9	-	CT:11	NA: 9	TT: 4	-	TCAGG
AG:22	-	-	AA: 2	AG:19	GG: 3	-	AG:20	GG: 3	NA: 1	-	AA: 2	AG:13	NA: 9	-	TGCAG
GT: 6	TT:17	-	GG:18	GT: 4	NA: 2	-	GG:13	GT: 8	NA: 3	-	GG:12	NA:12	-	-	GCGGa
TT: 1	-	-	CT:19	TT: 5	-	-	CT:23	NA: 1	-	-	CT:15	NA: 9	-	-	GCACT
NA: 1	-	-	AT:22	NA: 2	-	-	AT:23	NA: 1	-	-	AT:12	NA:12	-	-	GTTTT

-	-	-	CC:23	TT: 1	-	-	CC:23	NA: 1	-	-	CC:15	NA: 9	-	-	CAGAA
-	-	-	AT:24	-	-	-	AT:23	NA: 1	-	-	AT:14	NA:10	-	-	TTTTTT
TT:23	-	-	CT: 1	TT:23	-	-	CT: 6	NA: 1	TT:17	-	CT: 3	NA:13	TT: 8	-	CGTTT
-	-	-	AG:23	GG: 1	-	-	AG:24	-	-	-	AG:15	NA: 9	-	-	AAAGC
CC:20	-	-	CA:11	CC:12	NA: 1	-	CA: 6	CC:18	-	-	AA: 3	CA: 4	CC: 7	NA:10	AGGGC
CT:20	NA: 2	-	CC: 1	CT:18	NA: 2	TT: 3	CT:19	NA: 4	TT: 1	-	CT: 2	NA:21	TT: 1	-	GCCAC
-	-	-	AA:24	-	-	-	AA:24	-	-	-	AA:15	NA: 9	-	-	AAAGG
CT: 6	TT:17	-	CT: 5	TT:19	-	-	CT: 1	TT:23	-	-	CT: 2	NA: 9	TT:13	-	CAAAA
CT:12	TT: 2	-	CC:13	CT:11	-	-	CC:12	CT:12	-	-	CC:8	CT:7	NA:9	-	ACAAA
-	-	-	TT:24	-	-	-	TT:24	-	-	-	NA: 7	TT:17	-	-	AGCTG
TG:13	TT: 4	-	GG: 2	TG: 6	TT:16	-	GG: 5	TG:15	TT: 4	-	GG: 4	NA: 7	TG:10	TT: 3	TATCC
CT:10	NA: 1	TT: 8	CC: 5	CT: 5	TT:14	-	CC: 2	CT: 4	TT:18	-	CC:6	CT:7	NA:8	TT:3	CCGTA
CT: 3	TT: 1	-	CC: 3	CT:14	TT: 7	-	CC: 4	CT:13	TT: 7	-	CC:17	NA: 7	-	-	ATTTG
NA: 1	TT:14	-	CC: 2	CT:10	TT:12	-	CC: 3	CT: 7	TT:14	-	CC: 2	CT: 6	NA:10	TT: 6	CATTA
CC:19	-	-	AA: 2	CA: 5	CC:17	-	AA: 2	CA:14	CC: 8	-	AA: 2	CA: 4	CC:11	NA: 7	CGCCC
-	-	-	TT:24	-	-	-	TT:24	-	-	-	NA: 7	TT:17	-	-	GTGCT
-	-	-	AA:24	-	-	-	AA:24	-	-	-	AA:17	NA: 7	-	-	TCGGG

BLAST annotation	BLAST evalue
Gasterosteus aculeatus clone CFW82-G02 mRNA sequence	8E-46
Danio rerio hypothetical protein LOC792049 (LOC792049), mRNA	2E-28
Salmo salar serum albumin 2 (LOC100136922), mRNA	8E-135
Salmo salar antithrombin protein (antithrombin), mRNA	1E-131
Oncorhynchus mykiss clone omyk-evn-513-055 26S protease regulatory subunit 4 put	3E-71
Salmo salar antithrombin protein (antithrombin), mRNA	1E-132
Oncorhynchus mykiss complement factor B (cfb), mRNA	2E-80
PREDICTED: Danio rerio glycerol-3-phosphate acyltransferase 1, mitochondrial-like	2E-46
Salmo salar clone ssal-rgh-511-049 Cytochrome c oxidase polypeptide VIa, mitochon	2E-62
Salmo salar clone ssal-rgb2-576-352 ATP synthase subunit gamma, mitochondrial pre	2E-154
Salmo salar clone ssal-rgf-534-261 BNIP2 motif-containing molecule at the C-termina	1E-76
Salmo salar clone ssal-rgg-505-144 Dynein light chain 1, cytoplasmic putative mRNA	2E-116
Salmo salar T-complex protein 1 subunit beta (tcpb), mRNA	7E-129
Salmo salar clone ssal-rgf-501-118 Sodium/potassium-transporting ATPase subunit be	9E-166
Salmo salar ATP-binding cassette, sub-family E (OABP), member 1 (abce1), mRNA	0
TSA: Hippoglossus hippoglossus all	1E-43
Salmo salar clone ssal-rgf-503-225 Coagulation factor V precursor putative mRNA, p	7E-161
Salmo salar clone ssal-evf-569-078 Retinol dehydrogenase 3 putative mRNA, comple	7E-110
Salmo salar clone ssal-rgf-522-241 Sodium/potassium-transporting ATPase subunit al	4E-119
serum amyloid P component [guinea pigs, Genomic, 1800 nt]	3E-07
Salmo salar clone ssal-rgf-528-109 NADH-ubiquinone oxidoreductase 75 kDa subuni	1E-55
Salmo salar clone ssal-rgf-537-032 Eukaryotic translation initiation factor 3 subunit M	6E-174
Salmo salar clone ssal-rgf-518-283 Glucose-6-phosphatase putative mRNA, pseudoge	5E-131
Salmo salar clone ssal-rgb2-575-183 Triosephosphate isomerase putative mRNA, com	0
Salmo salar clone ssal-eve-541-107 Transcriptional activator protein Pur-beta putative	2E-129
Salmo salar clone ssal-rgg-508-187 40S ribosomal protein S24 putative mRNA, comp	5E-156
Salmo salar formiminotransferase cyclodeaminase (ftcd), mRNA	1E-38
Salmo salar interleukin 1 receptor accessory protein (il-1racp), mRNA	1E-150
Salmo salar clone ssal-rgf-524-048 ATP synthase subunit beta, mitochondrial precurso	3E-128
Salmo salar Rhesus blood group, B glycoprotein (rhbg), mRNA	0.00004
Salmo salar clone ssal-rgf-535-350 Methylmalonate-semialdehyde dehydrogenase put	5E-144
Sparus aurata contig 33 unknown mRNA	1E-30
Salmo salar clone HM5	5E-131
Salmo salar clone ssal-rgh-517-363 Glyceraldehyde-3-phosphate dehydrogenase putat	2E-135
Salmo salar 40S ribosomal protein S8 (rs8), mRNA	0
Oncorhynchus mykiss mRNA for myosin heavy chain	2E-67

Salmo salar clone ssal-rgh-512-226 Proteasome subunit beta type-8 precursor putative	2E-129
Oncorhynchus mykiss mRNA for ATP synthase beta-subunit, partial cds	8E-103
Epinephelus coioides fibrinogen beta chain precursor	5E-55
Salmo salar T-complex protein 1 subunit epsilon (tcpe), mRNA	1E-95
Salmo salar clone ssal-evd-510-220 ATP synthase subunit e, mitochondrial putative mR	1E-17
Anguilla anguilla RBP mRNA, partial cds	6E-92
Thunnus maccoyii phospholipid hydroperoxide glutathione peroxidase mRNA, compl	3E-96
Salmo salar clone ssal-plnb-024-105 Apolipoprotein A-I-1 precursor putative mRNA,	2E-60
Zebrafish DNA sequence from clone CH211-120J21 in linkage group 8, complete seq	4E-12
Salmo salar ancient ubiquitous protein 1 (aup1), mRNA	4E-100
Salmo salar clone ssal-rgh-512-279 Ribulose-phosphate 3-epimerase putative mRNA,	9E-147
Osmerus mordax clone omor-rgc-514-091 Multifunctional protein ADE2 putative mRN	3E-53
Salmo salar Tubulin alpha-1A chain (tba1a), mRNA	0
Medicago truncatula clone mth2-30h19, complete sequence	0.95
PREDICTED: Danio rerio titin b (ttnb), mRNA	3E-52
Salmo salar kinesin light chain 1 (klc1), mRNA	0
Rattus norvegicus chromosome 1 clone RP32-323A19 map q33, complete sequence	0.002
Salmo salar clone ssal-rgf-519-307 unknown large open reading frame mRNA, novel	2E-67
Salmo salar clone ssal-rgf-511-285 Probable glutamate receptor precursor putative mR	0
Oncorhynchus mykiss clone omyk-evo-502-012 Calmodulin putative mRNA, complet	0
Salmo salar clone HM4	0
Salmo salar lactate dehydrogenase A4 (ldha), mRNA	0
Salmo salar clone ssal-rgf-536-029 Heterogeneous nuclear ribonucleoprotein D-like p	0
Salmo salar clone ssal-rgf-517-208 Probable ubiquitin carboxyl-terminal hydrolase C	1E-158
Mus musculus chromosome UNKNOWN clone RP24-122B7, complete sequence	0.95
Salmo salar Amiloride-sensitive cation channel 2, neuronal (accn2), mRNA	0.00004
Salmo salar mRNA for putative ISG12(3) protein (isg12(3) gene)	1E-145
Oncorhynchus mykiss heat shock 27kDa protein 1 transcript variant 1 (hspb1), mRNA	2E-42
Danio rerio cleavage and polyadenylation specific factor 6, mRNA (cDNA clone MG	7E-28
Salmo salar clone ssal-rgf-506-036 Succinyl-CoA ligase subunit alpha, mitochondrial	0
Danio rerio high density lipoprotein-binding protein (vigilin), mRNA (cDNA clone M	2E-27
Salmo salar clone ssal-rgb2-610-343 Claudin-6 putative mRNA, complete cds	9E-128
Coregonus lavaretus mitochondrial DNA, complete genome	0
Salmo salar phosphoglucomutase 1 (pgm1), mRNA	0
Oncorhynchus mykiss clone omyk-evo-506-216 Proteasome subunit alpha type 4 puta	1E-157
Salmo salar T-complex protein 1 subunit delta (tcpd), mRNA	1E-132
PREDICTED: Danio rerio zinc finger protein 638-like (LOC563749), mRNA	4E-06
Salmo salar clone ssal-rgb2-656-252 ATP synthase subunit delta, mitochondrial precu	0
Oncorhynchus mykiss S6 ribosomal protein (LOC100135859), mRNA	0
Salmo salar clone ssal-rgb2-617-226 C21orf51 putative mRNA, complete cds	3E-83
Salmo salar clone HM4 (cyclin I)	9E-71
Salmo salar clone ssal-rgf-503-219 Vacuolar ATP synthase catalytic subunit A putative	0
Salmo salar clone ssal-rgf-510-220 unknown large open reading frame mRNA, novel	0

Coregonus lavaretus mitochondrial DNA, complete genome	0
Salmo salar clone ssal-rgf-510-220 unknown large open reading frame mRNA, novel	0
Oncorhynchus mykiss heat shock 90kDa protein 1 beta isoform a (hsp90ba), mRNA	0
AF308735 Oncorhynchus mykiss 18S ribosomal RNA gene and internal transcribed sp	0
Salmo salar phosphorylase, glycogen (muscle) A (pygma), mRNA	6E-67
Salmo salar clone ssal-eve-543-105 Signal peptidase complex subunit 3 putative mRN	2E-99
Salmo salar partial mRNA for myosin regulatory light chain 2 (mlc-2 gene), splice va	0
TSA: Hippoglossus hippoglossus all	6E-73
Salmo salar clone ssal-rgh-519-111 Heterogeneous nuclear ribonucleoprotein A1 puta	0
Caligus rogercresseyi clone crog-evp-516-001 Activated RNA polymerase II transcrip	6E-143
Salmo salar clone ssal-rgb2-561-188 Fatty acid-binding protein, brain putative mRNA	4E-131
Oncorhynchus mykiss complement component C9 (LOC100136130), mRNA	4E-139
Salmo salar Angiotensinogen (angt), mRNA	4E-132
Salmo salar RED protein (red), mRNA	1E-151
Salmo salar IMP (inosine monophosphate) dehydrogenase 2 (impdh2), mRNA	4E-145
Salmo salar clone HM6	5E-137
Salmo salar clone ssal-evd-509-343 Mitochondrial 39S ribosomal protein L27 putativ	1E-107

Supplementary table 5.2

SNP name	homozygous1_frequency	heterozygous_frequency	homozygous2_frequency	failed assays_frequency	homozygous1_genotype	heterozygous_genotype	homozygous2_genotype	segregation distortion_qvalue	amplicon_sequence	BLAST annotation	BLAST evalue
BU965641_re5_27	48	89	46	25	G/G	A/G	A/A	0.75	GCTTA	Gasterosteus aculeatus clone CFW82-G02 n	3E-46
CA037452_re5_29	97	85	-	26	A/A	A/G	-	0.47	GGACC	Danio rerio hypothetical protein LOC79204	7E-29
CA037876_re3_19	54	93	36	25	C/C	C/T	T/T	0.28	AAACC	Salmo salar antithrombin protein (antithrom	5E-132
CA038170_NDBC	87	96	-	25	A/A	A/G	-	0.53	GAAgG	Oncorhynchus mykiss clone omyk-evn-513-	9E-72
CA038790_re5_57	86	97	-	25	T/T	C/T	-	0.48	TGAAA	Salmo salar antithrombin protein (antithrom	4E-133
CA042392_re5_28	44	90	43	31	A/A	A/G	G/G	0.77	GCAGC	PREDICTED: Danio rerio glycerol-3-phosp	9E-47
CA042951_ND	13	13	-	182	A/A	C/A	-	0.77	TGGCC	Salmo salar clone ssal-rgb2-576-352 ATP sy	8E-155
CA044550_ca20_4	45	95	43	25	T/T	T/G	G/G	0.74	ACTTC	Salmo salar clone ssal-rgf-534-261 BNIP2 n	4E-77
CA045465_re5_30	87	84	-	37	G/G	A/G	-	0.73	AGGAC	Salmo salar clone ssal-rgg-505-144 Dynein	8E-117
CA049476_re5_101	97	82	-	29	A/A	A/G	-	0.36	CCCAT	Salmo salar T-complex protein 1 subunit bet	3E-129
CA051860_BC	46	92	43	27	T/T	C/T	C/C	0.75	TTGAG	Salmo salar clone ssal-rgf-501-118 Sodium/	3E-166
CA052650_148M	84	87	-	37	C/C	C/A	-	0.73	CGCAC	Salmo salar ATP-binding cassette, sub-famil	0E+00
CA053246_re5_89	16	161	2	29	T/T	G/T	G/G	0.00	AGTAC	TSA: Hippoglossus hippoglossus all halibut	5E-44
CA053896_D	83	95	-	30	T/T	T/G	-	0.47	ACTAT	Salmo salar clone ssal-rgf-503-225 Coagula	3E-161
CA054630_BC	44	102	37	25	A/A	A/C	C/C	0.36	CGACC	Salmo salar clone ssal-rgf-522-241 Sodium/	2E-119
CA054959_380R	92	90	-	26	T/T	C/T	-	0.74	TAAAT	serum amyloid P component [guinea pigs, G	1E-07
CA056473_re3_17	27	152	-	29	T/T	C/T	-	0.00	GCTCT	Salmo salar clone ssal-rgf-528-109 NADH-	5E-56
CA057176_NDBC	39	144	-	25	C/C	C/G	-	0.00	TCCGC	Salmo salar clone ssal-rgf-537-032 Eukaryo	2E-174
CA057603_ND	90	93	-	25	T/T	A/T	-	0.73	TAGAT	Salmo salar clone ssal-rgf-518-283 Glucose-	2E-131
CA057987_156K	94	84	-	30	C/C	A/C	-	0.50	GCATT	Salmo salar clone ssal-rgb2-575-183 Triosep	0E+00
CA058340_re5_16	107	76	-	25	G/G	A/G	-	0.06	AGGCC	Salmo salar clone ssal-eve-541-107 Transcri	8E-130
CA058958_re5_26	96	87	-	25	T/T	C/T	-	0.53	TCGTA	Salmo salar clone ssal-rgg-508-187 40S ribo	2E-156
CA060324_154M	15	168	-	25	C/C	C/A	-	0.00	GCTGG	Salmo salar formiminotransferase cyclodear	4E-39
CA061393_ND	1	180	-	27	G/G	G/T	-	0.00	GCCCT	Salmo salar interleukin 1 receptor accessory	5E-151
CA062071_NDBC	33	149	1	25	G/G	G/T	T/T	0.00	GGCGT	Salmo salar clone ssal-rgf-524-048 ATP syn	9E-129
CA063623_352M	91	92	-	25	T/T	T/G	-	0.75	ATCTAT	Salmo salar clone ssal-rgf-535-350 Methylm	2E-87
CB492682_re5_136	81	101	-	26	G/G	T/G	-	0.24	CTCGC	Sparus aurata contig 33 unknown mRNA	5E-31
CB492725_ca3_21	89	89	1	29	T/T	C/T	C/C	0.00	GGGAC	Salmo salar clone HM5	2E-131
CB496739_ca4_2	86	89	-	33	G/G	T/G	-	0.73	CATTC	Salmo salar aldolase a, fructose-bisphosphat	1E-145
CB497584_re3_2	30	152	-	26	A/A	A/G	-	0.00	AGGAC	Oncorhynchus mykiss mRNA for ATP synth	3E-103
CB498771_re5_67	49	89	43	27	T/T	C/T	C/C	0.73	CGCGC	Salmo salar T-complex protein 1 subunit eps	4E-96
CB500248_502K	100	83	-	25	C/C	A/C	-	0.33	ATTAG	Salmo salar clone ssal-evd-510-220 ATP syn	4E-38
CB509509_397K	94	88	-	26	C/C	A/C	-	0.65	AATCT	Salmo salar clone ssal-eve-565-112 Plasma r	2E-92
CB510585_ND	50	82	50	26	A/A	A/G	G/G	0.48	GCCAC	Salmo salar clone ssal-plnb-024-105 Apolip	7E-61
CB511030_339K	92	90	-	26	T/T	G/T	-	0.74	TGTAT	Zebrafish DNA sequence from clone CH211	2E-12
CB512085_538R	83	98	-	27	T/T	C/T	-	0.36	CCACA	Salmo salar ancient ubiquitous protein 1 (au	2E-100
CB514545_re5_125	51	90	41	26	G/G	T/G	T/T	0.59	CCACA	TSA: Hippoglossus hippoglossus all halibut	1E-83
CB516686_131R	78	104	1	25	A/A	A/G	G/G	0.00	TAGGC	Medicago truncatula clone mth2-30h19, con	4E-01
CK990730_re5_31	97	85	-	26	T/T	A/T	-	0.47	AGCTG	Caligus rogercresseyi clone crog-evp-516-0	2E-143
CX030416_re5_72	2	176	-	30	T/T	C/T	-	0.00	CCGTA	Oncorhynchus mykiss complement compon	1E-139
DW553899_re5_130	49	84	46	29	T/T	C/T	C/C	0.65	ATTTG	Salmo salar Angiotensinogen (angt), mRNA	1E-132
DY738545_re5_40	92	86	-	30	T/T	C/T	-	0.65	CATTA	Salmo salar RED protein (red), mRNA	4E-152
EG755964_BC	103	77	-	28	C/C	C/A	-	0.12	CGCCC	Salmo salar IMP (inosine monophosphate) d	1E-145
EG782906_re5_56	74	109	-	25	T/T	C/T	-	0.03	GTGCT	Salmo salar clone HM6_0740 nebulin-like r	2E-137
CA042951_ND	13	13	-	182	A/A	C/A	-	0.77	TGGCC	Salmo salar clone ssal-rgb2-576-352 ATP sy	8E-155
CA052607_ca3_33	19	9	-	180	A/A	C/A	-	0.13	CCCAC	Salmo salar clone ssal-rgf-520-280 Elongati	7E-143
CA057774_ca4_36	15	7	-	186	A/A	A/G	-	0.19	GCTCC	Oncorhynchus tshawytscha beta actin mRN	2E-163
CA058008_re5_4	4	23	-	181	C/C	C/A	-	0.00	CGGAC	Salmo salar clone ssal-rgb2-639-018 40S rib	1E-107
CA058986_re5_146	10	16	-	182	C/C	C/T	-	0.36	AGAGA	Salmo salar formiminotransferase cyclodear	1E-145
CA063528_139R	12	13	-	183	C/C	C/T	-	0.74	TATATT	Oncorhynchus keta IT-II gene for isotocin, c	1E-134
CK991313_ca4_17	11	15	-	182	C/C	C/T	-	0.49	CGATC	Salmo salar clone HM4_4065 actin, alpha 1	6E-150
cc000085_01AC	77	101	1	29	C/C	C/A	A/A	0.00	CCCAC	PREDICTED: Danio rerio titin b (ttnb), mR	9E-53
cc000102_01AG	94	82	-	32	C/C	C/T	-	0.47	GCGGC	Salmo salar kinesin light chain 1 (klc1), mR	0E+00
cc000129_01AT	32	135	14	27	T/T	A/T	A/A	0.00	GGGGT	Rattus norvegicus chromosome 1 clone RP3	7E-04
cc000151_03CG	7	58	1	142	C/C	C/G	G/G	0.00	TGTAA	Osmerus mordax clone omor-eva-514-176 I	3E-08
cc000167_01AC	90	87	-	31	G/G	T/G	-	0.73	AAGCC	Salmo salar clone ssal-rgf-519-307 unknow	6E-68
cc000225_01AC	83	98	-	27	C/C	C/A	-	0.36	CCAGA	Salmo salar clone ssal-rgf-511-285 Probable	0E+00
cc000236_01CT	44	98	42	24	C/C	C/T	T/T	0.65	GAGCT	Oncorhynchus mykiss clone omyk-evo-502-	0E+00
cc000240_01CT	104	74	-	30	G/G	A/G	-	0.06	AAAAC	Salmo salar clone HM4_1638 solute carrier	0E+00
cc000258_01AC	85	96	-	27	G/G	T/G	-	0.48	CCCCA	Salmo salar lactate dehydrogenase A4 (ldha	0E+00

cc000270_01CT	45	92	42	29	T/T	C/T	C/C	0.74	TACCC	Salmo salar clone ssal-rgf-536-029 Heterog	0E+00
cc000303_01CT	85	95	-	28	T/T	C/T	-	0.50	AATATA	Salmo salar clone ssal-rgf-517-208 Probable	4E-159
cc000541_01AT	2	164	-	42	A/A	A/T	-	0.00	TTCAC	Mus musculus chromosome UNKNOWN cl	4E-01
cc000857_01AG	83	99	-	26	G/G	A/G	-	0.36	AAAAA	Gasterosteus aculeatus clone VMRC26-150	2E-22
cc000873_01AG	25	131	21	31	A/A	A/G	G/G	0.00	CGGGC	Salmo salar lactate dehydrogenase A4 (ldha)	2E-05
cc000952_01GT	39	96	36	37	T/T	G/T	G/G	0.36	TTTTC	Oncorhynchus mykiss mRNA for putative IS	4E-146
cc001175_01AG	79	101	-	28	A/A	A/G	-	0.21	AAACT	Oncorhynchus mykiss heat shock 27kDa pro	6E-43
cc001276_01CT	81	101	-	26	C/C	C/T	-	0.24	GCAAC	Tetraodon nigroviridis partial TY3/GYPSY-	7E-99
cc001365_01AG	59	105	17	27	A/A	A/G	G/G	0.00	GGGAC	Danio rerio cleavage and polyadenylation sp	3E-28
cc001404_04GT	33	146	-	29	A/A	A/C	-	0.00	TAGAA	Salmo salar clone ssal-rgf-506-036 Succinyl	0E+00
cc001461_02AT	2	178	-	28	T/T	A/T	-	0.00	GGAAT	Salmo salar clone ssal-rgb2-610-343 Claudi	3E-128
cc001477_15AG	101	81	-	26	G/G	A/G	-	0.24	TTATTA	Salmo salar clone BAC CHORI214-184H23	0E+00
cc001516_01CT	27	142	11	28	A/A	A/G	G/G	0.00	GGAGC	Oncorhynchus mykiss clone omyk-evo-506-	5E-158
cc001558_03AC	78	99	-	31	G/G	T/G	-	0.23	TACAA	Esox lucius clone eluc-evq-528-250 Tetrasp	1E-32
cc001576_01CT	105	77	-	26	C/C	C/T	-	0.09	TGGTG	Salmo salar T-complex protein 1 subunit del	4E-133
cc001605_01AG	86	96	-	26	C/C	C/T	-	0.50	TAAAG	PREDICTED: Danio rerio zinc finger protei	1E-06
cc001682_02CT	96	84	-	28	C/C	C/T	-	0.47	TGGGC	Salmo salar clone ssal-rgb2-656-252 ATP sy	0E+00
cc001705_02AG	18	156	7	27	C/C	C/T	T/T	0.00	TCAGG	Oncorhynchus mykiss S6 ribosomal protein	0E+00
cc001720_01CT	55	118	6	29	G/G	A/G	A/A	0.00	TGCAC	Salmo salar clone ssal-rgb2-617-226 C21orf	1E-83
cc001746_01AC	46	28	-	134	G/G	T/G	-	0.09	TCAAC	AF250196 Oncorhynchus mykiss clone Hot	9E-167
cc001882_04AC	101	81	-	26	A/A	C/A	-	0.24	CCCCG	Salmo salar EAP30 subunit of ELL complex	0E+00
cc001956_01CT	80	101	-	27	G/G	A/G	-	0.23	TACCA	Salmo salar clone ssal-rgg-508-255 Ferritin,	0E+00
cc002099_04CT	101	81	-	26	T/T	C/T	-	0.24	GGACC	Zebrafish DNA sequence from clone DKEY	1E-77
cc002119_06AC	82	99	-	27	C/C	C/A	-	0.33	AGGGC	Salmo salar phosphorylase, glycogen (musc	2E-67
cc002133_01CT	64	113	2	29	C/C	C/T	T/T	0.00	GCCAC	Salmo salar clone ssal-eve-543-105 Signal p	6E-100
cc002148_01AG	84	95	-	29	A/A	A/G	-	0.48	AAAGC	Salmo salar partial mRNA for myosin regula	0E+00
cc002195_01AG	94	88	-	26	C/C	C/T	-	0.65	ACAAA	Salmo salar clone ssal-rgh-519-111 Heterog	0E+00

Suplementary table 5.3

SNP_name	Cliff-Normal	Cliff-Normal	Cliff-Dwarf	Cliff-Dwarf	Cliff-N/D	Webster-Normal	Webster-Normal	Webster-Dwarf	Webster-Dwarf	Webster-N/D	Indian-Normal	Indian-Normal	Indian-Dwarf	Indian-Dwarf	Indian-N/D
	Ho	He	Ho	He	Fst	Ho	He	Ho	He	Fst	Ho	He	Ho	He	Fst
BU965641_re5_27	0.42	0.47	0.30	0.44	0.17	0.31	0.37	0.33	0.37	0.02	0.42	0.38	0.29	0.25	-0.03
CA037452_re5_29	0.54	0.43	0.00	0.00	0.31	0.46	0.37	0.22	0.42	-0.02	0.42	0.49	0.55	0.49	-0.03
CA037647_ca4_26	0.72	0.50	0.67	0.51	0.00	0.31	0.27	0.63	0.48	0.02	0.19	0.24	0.09	0.09	0.10
CA037876_re3_19	0.54	0.50	0.20	0.18	0.23	0.62	0.49	0.64	0.51	-0.01	0.38	0.50	0.55	0.51	-0.01
CA038170_NDBC	0.35	0.29	0.00	0.00	0.83	0.08	0.21	0.36	0.47	0.04	0.65	0.50	0.50	0.49	0.10
CA038790_re5_57	0.54	0.48	0.00	0.00	0.39	0.54	0.52	0.14	0.14	0.02	0.38	0.36	0.55	0.47	0.41
CA039055_ND	0.00	0.00	0.34	0.33	0.18	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA
CA042392_re5_28	0.50	0.47	0.13	0.13	0.23	0.54	0.51	0.50	0.51	0.05	0.42	0.50	0.45	0.47	-0.03
CA042792_622W	0.00	0.00	0.00	0.00	NA	0.23	0.21	0.04	0.04	0.02	0.12	0.18	0.05	0.05	0.07
CA042951_ND	0.04	0.04	0.59	0.49	0.55	0.46	0.49	0.54	0.51	-0.01	0.19	0.24	0.18	0.17	-0.01
CA044550_ca20_4	0.00	0.00	0.00	0.00	NA	0.15	0.44	0.36	0.30	0.00	0.00	0.00	0.05	0.05	0.02
CA045465_re5_30	0.70	0.49	0.20	0.18	0.21	0.42	0.34	0.42	0.34	0.09	0.54	0.47	0.29	0.25	-0.02
CA049476_re5_101	0.88	0.50	0.70	0.48	0.00	0.46	0.37	0.26	0.23	0.10	0.81	0.49	0.36	0.30	0.01
CA051860_BC	0.00	0.08	0.22	0.31	0.75	0.15	0.15	0.43	0.42	0.32	0.08	0.14	0.32	0.51	0.09
CA052650_148M	0.64	0.47	0.00	0.00	0.37	0.62	0.44	0.07	0.07	0.15	0.35	0.29	0.00	0.00	0.28
CA053246_re5_89	0.67	0.48	0.83	0.49	0.07	0.75	0.49	0.41	0.33	0.15	0.80	0.49	0.28	0.25	0.06
CA053896_D	0.32	0.41	0.38	0.31	0.00	0.31	0.27	0.04	0.04	0.08	0.35	0.35	0.63	0.53	0.12
CA054079_re5_14	0.32	0.27	0.63	0.44	0.05	0.69	0.47	0.67	0.48	-0.01	0.69	0.46	0.64	0.44	-0.02
CA054630_BC	0.27	0.42	0.00	0.00	0.71	0.00	0.00	0.00	0.00	0.00	0.04	0.04	0.00	0.00	NA
CA054959_380R	0.43	0.49	NA	NA	NA	0.15	0.27	0.18	0.17	0.02	0.08	0.08	0.00	0.00	-0.01
CA056473_re3_17	0.80	0.49	0.83	0.50	0.00	1.00	0.52	0.77	0.51	-0.01	0.75	0.52	0.80	0.53	-0.01
CA057176_NDBC	0.50	0.38	0.32	0.27	0.01	0.62	0.44	0.75	0.51	0.27	0.68	0.50	0.36	0.30	0.05
CA057603_ND	0.12	0.11	0.00	0.00	0.94	0.62	0.49	0.25	0.40	0.29	0.12	0.11	0.55	0.49	0.20
CA057987_156K	0.58	0.45	0.00	0.00	0.69	0.54	0.51	0.57	0.50	0.06	0.24	0.33	0.41	0.49	-0.03
CA058340_re5_16	0.04	0.04	0.43	0.38	0.18	0.38	0.47	0.64	0.51	0.23	0.35	0.45	0.32	0.43	0.02
CA058958_re5_26	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA
CA060324_154M	0.12	0.11	0.53	0.40	0.13	0.92	0.52	0.57	0.42	0.03	0.81	0.49	0.52	0.40	0.05
CA061393_ND	0.48	0.37	0.32	0.32	0.48	0.58	0.49	0.32	0.36	0.08	0.48	0.47	0.32	0.27	0.02
CA062071_NDBC	0.04	0.04	0.13	0.13	0.01	0.08	0.08	0.32	0.32	-0.01	0.23	0.21	0.32	0.27	0.07
CA063046_219W	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA
CA063623_352M	0.38	0.48	0.03	0.03	0.59	0.15	0.52	0.50	0.44	0.07	0.38	0.46	0.41	0.50	0.06
CB492682_re5_136	0.19	0.24	0.33	0.50	0.32	0.23	0.21	0.18	0.17	0.26	0.08	0.08	0.50	0.46	-0.02
CB492725_ca3_21	0.04	0.04	0.23	0.21	0.05	0.15	0.15	0.14	0.19	-0.03	0.15	0.21	0.14	0.21	-0.03
CB492813_188K	0.00	0.00	0.27	0.28	0.14	0.62	0.49	0.39	0.51	-0.02	0.12	0.11	0.09	0.09	-0.01
CB492855_ca4_11	0.00	0.00	0.00	0.00	NA	0.38	0.47	0.25	0.27	0.02	0.50	0.45	0.38	0.51	0.07
CB494318_BC	0.00	0.00	0.54	0.43	0.29	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA
CB496486_re5_37	0.42	0.50	0.00	0.00	0.57	0.46	0.52	0.36	0.38	0.36	0.35	0.29	0.73	0.47	0.14
CB497584_re3_2	0.96	0.51	0.73	0.49	0.02	0.62	0.49	0.86	0.51	0.09	0.65	0.45	0.81	0.51	0.03
CB497894_re5_154	0.44	0.51	0.00	0.00	0.47	0.46	0.44	0.18	0.17	0.38	0.54	0.51	0.05	0.05	0.14
CB498771_re5_67	0.35	0.42	0.40	0.49	0.16	0.38	0.41	0.39	0.40	0.40	0.69	0.50	0.00	0.00	-0.03
CB500248_502K	0.31	0.32	0.03	0.03	0.14	0.62	0.49	0.25	0.22	0.63	0.19	0.24	0.36	0.30	0.17
CB509509_397K	0.00	0.00	NA	NA	NA	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA
CB509723_re5_75	0.31	0.43	0.23	0.30	0.02	0.08	0.08	0.21	0.30	0.01	0.46	0.36	0.13	0.24	0.05
CB510585_ND	0.31	0.36	0.03	0.03	0.18	0.62	0.49	0.43	0.47	0.21	0.50	0.42	0.64	0.47	0.10
CB511030_339K	0.42	0.49	0.00	0.00	0.41	0.38	0.32	0.07	0.07	0.04	0.08	0.14	0.00	0.00	0.12
CB512085_538R	0.54	0.48	0.00	0.00	0.39	0.38	0.47	0.32	0.36	0.23	0.54	0.48	0.14	0.13	0.00
CB512493_ND	0.50	0.51	0.55	0.50	-0.01	0.62	0.49	0.50	0.49	0.07	0.56	0.44	0.29	0.25	-0.03
CB514545_re5_125	0.00	0.00	0.23	0.21	0.10	0.54	0.41	0.43	0.42	0.13	0.46	0.50	0.32	0.27	0.31
CB516392_126M	0.83	0.50	NA	NA	NA	0.69	0.47	0.85	0.50	0.02	0.75	0.50	0.94	0.51	0.00
CB516686_131R	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA
cc000085_01AC	0.04	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA
cc000102_01AG	0.00	0.00	0.57	0.50	0.53	0.38	0.32	0.43	0.42	-0.01	0.50	0.38	0.50	0.38	0.00

cc000129_01AT	0.15	0.14	0.90	0.51	0.37	0.77	0.52	0.89	0.50	0.00	0.62	0.43	0.76	0.48	0.01
cc000167_01AC	0.12	0.18	0.20	0.18	-0.02	0.31	0.49	0.32	0.27	0.23	0.27	0.29	0.48	0.51	0.10
cc000225_01AC	0.12	0.11	0.00	0.00	0.05	0.38	0.32	0.25	0.22	0.13	0.04	0.04	0.36	0.30	-0.01
cc000236_01CT	0.19	0.23	0.00	0.00	0.12	0.46	0.44	0.43	0.42	0.28	0.50	0.49	0.09	0.09	-0.03
cc000240_01CT	0.48	0.50	0.40	0.36	0.07	0.38	0.32	0.46	0.40	0.28	0.27	0.24	0.50	0.51	-0.01
cc000258_01AC	0.00	0.00	0.00	0.00	NA	0.62	0.49	0.39	0.32	0.05	0.35	0.34	0.50	0.49	0.07
cc000270_01CT	0.00	0.00	0.00	0.00	NA	0.38	0.32	0.21	0.19	-0.02	0.27	0.38	0.50	0.43	0.01
cc000303_01CT	0.00	0.00	0.13	0.13	0.93	0.38	0.32	0.50	0.42	-0.01	0.38	0.51	0.50	0.49	0.41
cc000541_01AT	0.00	0.00	0.71	0.47	0.35	0.75	0.49	0.71	0.47	0.04	1.00	0.51	0.71	0.48	-0.01
cc000873_01AG	0.00	0.00	0.00	0.00	NA	0.38	0.32	0.36	0.30	0.05	0.15	0.14	0.00	0.00	-0.02
cc000952_01GT	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA
cc001175_01AG	0.54	0.48	0.00	0.00	0.39	0.46	0.49	0.29	0.34	0.41	0.50	0.51	0.05	0.05	0.28
cc001365_01AG	0.52	0.39	0.43	0.35	-0.01	0.31	0.27	0.86	0.50	0.03	0.44	0.35	0.71	0.47	0.14
cc001404_04GT	0.00	0.00	0.61	0.43	0.29	0.62	0.44	0.79	0.49	0.00	0.08	0.08	0.18	0.17	0.00
cc001422_01CT	0.22	0.31	0.52	0.45	0.38	0.08	0.08	0.04	0.04	0.30	0.04	0.04	0.50	0.46	-0.02
cc001461_02AT	1.00	0.51	1.00	0.51	0.00	1.00	0.52	1.00	0.51	0.00	0.96	0.51	0.95	0.51	0.00
cc001462_01CT	0.00	0.00	0.00	0.00	1.00	0.00	0.44	0.00	0.47	-0.04	0.00	0.14	0.00	0.17	0.15
cc001468_01CT	0.04	0.04	0.00	0.00	0.00	0.15	0.15	0.00	0.00	0.00	0.04	0.04	0.00	0.00	0.09
cc001516_01CT	0.93	0.51	0.59	0.42	0.05	0.77	0.49	0.79	0.49	0.01	0.96	0.51	0.86	0.50	0.08
cc001576_01CT	0.07	0.07	0.47	0.45	0.59	0.00	0.00	0.04	0.04	0.15	0.35	0.29	0.00	0.00	-0.02
cc001605_01AG	0.63	0.44	0.60	0.51	0.04	0.38	0.41	0.75	0.51	0.05	0.56	0.41	0.64	0.51	0.07
cc001682_02CT	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA
cc001705_02AG	0.78	0.48	0.70	0.49	-0.01	0.92	0.52	0.82	0.49	0.00	0.80	0.49	0.82	0.51	0.03
cc001720_01CT	0.73	0.47	0.87	0.51	0.05	0.92	0.52	0.61	0.43	0.19	0.62	0.43	0.73	0.47	0.10
cc001763_01GT	0.59	0.51	0.00	0.00	0.53	0.38	0.51	0.43	0.34	0.36	0.42	0.47	0.14	0.27	0.08
cc001810_03AG	0.96	0.51	0.97	0.51	0.00	0.69	0.47	0.82	0.49	0.00	1.00	0.51	1.00	0.51	0.00
cc001837_07AT	0.96	0.51	0.61	0.43	0.06	1.00	0.52	1.00	0.51	0.00	1.00	0.51	1.00	0.51	0.00
cc001862_03AG	0.00	0.00	0.00	0.00	1.00	0.00	0.44	0.00	0.47	-0.02	0.00	0.08	0.00	0.17	0.15
cc001934_01AT	0.96	0.51	0.61	0.43	0.06	1.00	0.52	1.00	0.51	0.00	1.00	0.51	1.00	0.51	0.00
cc001937_01AG	0.22	0.20	0.38	0.47	0.44	0.15	0.44	0.46	0.51	0.71	0.15	0.14	0.38	0.32	0.05
cc002110_01AG	1.00	0.51	0.73	0.47	0.03	0.92	0.52	0.93	0.51	0.02	0.80	0.49	1.00	0.51	0.00
cc002119_06AC	0.00	0.00	0.00	0.00	NA	0.15	0.15	0.04	0.04	0.03	0.12	0.11	0.00	0.00	0.02
cc002133_01CT	0.54	0.40	0.90	0.51	0.11	0.62	0.44	0.93	0.51	0.25	0.17	0.16	0.75	0.54	0.09
cc002148_01AG	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA
cc002181_03CT	0.44	0.39	0.40	0.40	-0.02	0.15	0.15	0.14	0.14	0.10	0.04	0.04	0.32	0.27	-0.03
cc002195_01AG	0.52	0.45	0.33	0.36	0.31	0.46	0.49	0.61	0.51	0.03	0.46	0.50	0.55	0.49	-0.01
CK990730_re5_31	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA
CK990756_425M	0.27	0.38	0.00	0.00	0.25	0.54	0.51	0.57	0.51	-0.01	0.19	0.18	0.27	0.24	-0.02
CX030416_re5_72	0.62	0.50	0.00	0.00	0.59	0.31	0.27	0.04	0.04	-0.02	0.27	0.24	0.14	0.27	0.13
DW553899_re5_130	0.35	0.38	0.40	0.36	0.41	0.31	0.44	0.39	0.36	0.53	0.50	0.50	0.00	0.00	0.34
DY738545_re5_40	0.35	0.38	0.00	0.00	0.76	0.46	0.44	0.18	0.17	0.31	0.23	0.51	0.10	0.10	0.14
EG755964_BC	0.08	0.08	0.59	0.51	0.44	0.15	0.15	0.43	0.51	0.07	0.35	0.34	0.36	0.49	0.26
EG782906_re5_56	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA
EG811444_re3_6	0.00	0.00	0.03	0.03	0.00	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA

East-Normal	East-Normal	East-Dwarf	East-Dwarf	East-N/D	Témiscouata-Normal	Témiscouata-Normal	Témiscouata-Dwarf	Témiscouata-Dwarf	Témiscouata-N/D	Pohénégamook-Normal	Pohénégamook-Normal	BLAST annotation	BLAST evalue
Ho	He	Ho	He	Fst	Ho	He	Ho	He	Fst	Ho	He		
0.29	0.25	0.29	0.25	-0.02	0.38	0.36	0.21	0.25	0.00	0.41	0.34	Gasterosteus aculeatus clone CFW82-G	8E-46
0.46	0.44	0.39	0.37	-0.01	0.33	0.42	0.50	0.51	0.10	0.67	0.51	Danio rerio hypothetical protein LOC79	2E-28
0.13	0.12	0.04	0.04	0.00	0.04	0.04	0.22	0.20	0.04	0.21	0.20	Salmo salar serum albumin 2 (LOC100	8E-135
0.50	0.50	0.42	0.42	0.01	0.71	0.50	0.58	0.45	0.01	0.71	0.51	Salmo salar antithrombin protein (antith	1E-131
0.54	0.47	0.43	0.39	0.00	0.46	0.40	0.29	0.36	-0.02	0.27	0.24	Oncorhynchus mykiss clone omyk-evn-	3E-71
0.46	0.44	0.25	0.22	0.08	0.17	0.16	0.25	0.22	-0.01	0.38	0.39	Salmo salar antithrombin protein (antith	1E-132
0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA	0.00	0.00	Oncorhynchus mykiss complement fact	2E-80
0.54	0.47	0.42	0.50	0.08	0.46	0.51	0.58	0.50	0.00	0.41	0.40	PREDICTED: Danio rerio glycerol-3-p	2E-46
0.00	0.00	0.08	0.08	0.02	0.13	0.12	0.04	0.04	0.00	0.00	0.00	Salmo salar clone ssal-rgh-511-049 Cyt	2E-62
0.33	0.34	0.29	0.25	-0.01	0.61	0.46	0.33	0.45	-0.02	0.53	0.45	Salmo salar clone ssal-rgb2-576-352 AT	2E-154
0.38	0.51	0.46	0.44	0.06	0.38	0.36	0.29	0.36	-0.02	0.47	0.50	Salmo salar clone ssal-rgf-534-261 BN	1E-76
0.60	0.49	0.29	0.25	0.14	0.41	0.33	0.35	0.29	-0.01	0.60	0.51	Salmo salar clone ssal-rgg-505-144 Dy	2E-116
0.43	0.35	0.42	0.34	-0.02	0.67	0.45	0.92	0.51	0.03	0.71	0.47	Salmo salar T-complex protein 1 subuni	7E-129
0.38	0.36	0.08	0.16	0.06	0.38	0.47	0.38	0.31	0.05	0.44	0.35	Salmo salar clone ssal-rgf-501-118 Sodi	9E-166
0.33	0.28	0.33	0.28	-0.02	0.57	0.41	0.50	0.45	-0.01	0.29	0.25	Salmo salar ATP-binding cassette, sub-f	0
0.38	0.31	0.63	0.47	0.05	0.96	0.51	0.70	0.49	0.01	0.35	0.30	TSA: Hippoglossus hippoglossus all	1E-43
0.27	0.30	0.29	0.25	-0.02	0.58	0.45	0.42	0.48	-0.02	0.44	0.51	Salmo salar clone ssal-rgf-503-225 Coa	7E-161
0.46	0.36	0.46	0.36	-0.02	0.25	0.22	0.71	0.47	0.12	0.36	0.30	Salmo salar clone ssal-evf-569-078 Reti	7E-110
0.17	0.22	0.00	0.00	0.10	0.21	0.19	0.17	0.16	-0.02	0.06	0.06	Salmo salar clone ssal-rgf-522-241 Sodi	4E-119
0.21	0.19	0.09	0.09	0.01	0.04	0.04	0.13	0.12	0.00	0.18	0.26	serum amyloid P component [guinea pig	3E-07
0.77	0.51	0.86	0.50	0.00	1.00	0.51	0.88	0.51	0.00	0.70	0.48	Salmo salar clone ssal-rgf-528-109 NA	1E-55
0.71	0.49	0.75	0.48	-0.01	0.63	0.51	0.54	0.50	-0.01	0.71	0.47	Salmo salar clone ssal-rgf-537-032 Euk	6E-174
0.29	0.31	0.42	0.34	-0.02	0.42	0.42	0.46	0.49	0.00	0.38	0.48	Salmo salar clone ssal-rgf-518-283 Glu	5E-131
0.25	0.34	0.25	0.22	0.00	0.50	0.51	0.42	0.51	-0.02	0.63	0.51	Salmo salar clone ssal-rgb2-575-183 Tri	0
0.58	0.48	0.67	0.50	-0.01	0.33	0.42	0.50	0.48	-0.01	0.40	0.51	Salmo salar clone ssal-eve-541-107 Tra	2E-129
0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA	0.00	0.00	Salmo salar clone ssal-rgg-508-187 40S	5E-156
0.88	0.50	0.42	0.34	0.11	0.74	0.51	0.63	0.44	0.03	0.47	0.43	Salmo salar formiminotransferase cyclo	1E-38
0.43	0.35	0.42	0.38	-0.02	0.46	0.36	0.38	0.47	0.02	0.50	0.44	Salmo salar interleukin 1 receptor acces	1E-150
0.13	0.12	0.17	0.16	-0.02	0.04	0.04	0.17	0.22	0.05	0.00	0.00	Salmo salar clone ssal-rgf-524-048 ATP	3E-128
0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA	0.00	0.00	Salmo salar Rhesus blood group, B glyc	0.00004
0.58	0.48	0.39	0.45	-0.02	0.46	0.49	0.33	0.28	0.10	0.43	0.42	Salmo salar clone ssal-rgf-535-350 Met	5E-144
0.08	0.08	0.00	0.00	0.02	0.13	0.12	0.21	0.19	-0.01	0.21	0.20	Sparus aurata contig 33 unknown mRN	1E-30
0.30	0.26	0.43	0.34	-0.01	0.25	0.28	0.48	0.45	0.05	0.35	0.43	Salmo salar clone HM5	5E-131
0.46	0.47	0.54	0.50	0.06	0.46	0.44	0.33	0.34	0.01	0.13	0.39	Salmo salar clone ssal-rgh-517-363 Gly	2E-135
0.46	0.40	0.50	0.48	0.01	0.38	0.49	0.42	0.48	-0.02	0.53	0.51	Salmo salar 40S ribosomal protein S8 (	0
0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA	0.00	0.00	Oncorhynchus mykiss mRNA for myosi	2E-67
0.33	0.50	0.50	0.45	-0.01	0.54	0.40	0.46	0.40	-0.02	0.60	0.43	Salmo salar clone ssal-rgh-512-226 Pro	2E-129
0.83	0.50	0.79	0.50	-0.01	0.63	0.44	0.63	0.44	-0.01	0.76	0.49	Oncorhynchus mykiss mRNA for ATP s	8E-103
0.00	0.00	0.08	0.08	0.02	0.46	0.40	0.33	0.38	-0.02	0.44	0.35	Epinephelus coioides fibrinogen beta ch	5E-55
0.38	0.49	0.63	0.51	0.01	0.33	0.48	0.71	0.49	0.08	0.35	0.30	Salmo salar T-complex protein 1 subuni	1E-95
0.29	0.25	0.09	0.16	-0.01	0.54	0.51	0.54	0.47	0.01	0.53	0.40	Salmo salar clone ssal-evd-510-220 ATP	1E-17
0.00	0.00	0.00	0.00	NA	0.00	0.00	0.13	0.12	0.04	0.06	0.06	Anguilla anguilla RBP mRNA, partial c	6E-92
0.00	0.00	0.00	0.00	NA	0.29	0.25	0.13	0.12	0.02	0.18	0.26	Thunnus maccoyii phospholipid hydrop	3E-96
0.25	0.22	0.33	0.34	0.00	0.29	0.40	0.46	0.50	0.04	0.41	0.40	Salmo salar clone ssal-plnb-024-105 Ap	2E-60
0.33	0.28	0.04	0.04	0.10	0.25	0.22	0.29	0.31	-0.01	0.31	0.27	Zebrafish DNA sequence from clone C	4E-12
0.67	0.50	0.29	0.25	0.15	0.42	0.34	0.29	0.36	-0.02	0.24	0.21	Salmo salar ancient ubiquitous protein 1	4E-100
0.29	0.50	0.25	0.28	0.27	0.17	0.22	0.33	0.28	-0.01	0.19	0.18	Salmo salar clone ssal-rgh-512-279 Rib	9E-147
0.42	0.38	0.13	0.25	0.01	0.38	0.31	0.25	0.34	-0.02	0.07	0.07	Osmerus mordax clone omor-rgc-514-09	3E-53
0.88	0.51	0.61	0.43	0.05	0.92	0.51	0.75	0.48	0.01	0.86	0.51	Salmo salar Tubulin alpha-1A chain (tb	0
0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA	0.00	0.00	Medicago truncatula clone mth2-30h19,	0.95
0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA	0.00	0.00	PREDICTED: Danio rerio titin b (ttnb),	3E-52
0.04	0.04	0.00	0.00	0.00	0.17	0.16	0.25	0.22	-0.01	0.36	0.45	Salmo salar kinesin light chain 1 (klc1),	0

0.58	0.42	0.46	0.36	0.00	0.79	0.50	0.71	0.47	0.01	0.87	0.51	Rattus norvegicus chromosome 1 clone	0.002
0.33	0.28	0.08	0.08	0.06	0.13	0.19	0.04	0.12	-0.02	0.00	0.00	Salmo salar clone ssal-rgf-519-307 unk	2E-67
0.25	0.48	0.54	0.51	0.00	0.33	0.34	0.29	0.31	-0.02	0.27	0.24	Salmo salar clone ssal-rgf-511-285 Prob	0
0.50	0.51	0.54	0.49	0.00	0.13	0.19	0.29	0.36	0.03	0.27	0.33	Oncorhynchus mykiss clone omyk-evo-	0
0.54	0.47	0.33	0.42	-0.01	0.17	0.16	0.29	0.25	0.00	0.47	0.52	Salmo salar clone HM4	0
0.42	0.50	0.50	0.45	-0.01	0.25	0.38	0.29	0.36	-0.03	0.29	0.35	Salmo salar lactate dehydrogenase A4 (	0
0.46	0.50	0.50	0.51	-0.01	0.29	0.36	0.46	0.40	-0.02	0.50	0.52	Salmo salar clone ssal-rgf-536-029 Hete	0
0.46	0.51	0.54	0.47	0.04	0.29	0.25	0.29	0.36	0.00	0.14	0.25	Salmo salar clone ssal-rgf-517-208 Prob	1E-158
0.65	0.48	0.76	0.51	0.00	0.95	0.51	0.87	0.51	0.00	0.58	0.43	Mus musculus chromosome UNKNOW	0.95
0.48	0.37	0.65	0.48	0.03	0.73	0.47	0.59	0.51	0.01	0.60	0.56	Salmo salar Amiloride-sensitive cation	0.00004
0.00	0.00	0.00	0.00	NA	0.04	0.04	0.00	0.00	0.00	0.00	0.00	Salmo salar mRNA for putative ISG12(3	1E-145
0.22	0.37	0.54	0.50	0.18	0.30	0.37	0.54	0.47	0.01	0.14	0.14	Oncorhynchus mykiss heat shock 27kD	2E-42
0.54	0.40	0.42	0.34	0.00	0.38	0.31	0.83	0.50	0.11	0.57	0.42	Danio rerio cleavage and polyadenylati	7E-28
0.35	0.29	0.17	0.22	-0.01	0.87	0.51	0.63	0.47	0.02	0.53	0.40	Salmo salar clone ssal-rgf-506-036 Suc	0
0.08	0.16	0.00	0.00	0.05	0.04	0.04	0.04	0.04	-0.02	0.13	0.13	Danio rerio high density lipoprotein-bin	2E-27
1.00	0.51	1.00	0.51	0.00	1.00	0.51	1.00	0.51	0.00	1.00	0.52	Salmo salar clone ssal-rgb2-610-343 Cl	9E-128
0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	Coregonus lavaretus mitochondrial DN	0
0.04	0.04	0.04	0.04	-0.02	0.21	0.25	0.25	0.22	-0.02	0.00	0.00	Salmo salar phosphoglucomutase 1 (pg	0
0.75	0.48	0.83	0.51	0.01	0.71	0.47	0.75	0.48	-0.01	0.79	0.49	Oncorhynchus mykiss clone omyk-evo-	1E-157
0.08	0.08	0.08	0.08	-0.02	0.38	0.36	0.21	0.19	0.04	0.20	0.19	Salmo salar T-complex protein 1 subuni	1E-132
0.58	0.48	0.45	0.36	0.03	0.13	0.12	0.17	0.16	-0.02	0.00	0.00	PREDICTED: Danio rerio zinc finger p	4E-06
0.00	0.00	0.00	0.00	NA	0.04	0.04	0.00	0.00	0.00	0.00	0.00	Salmo salar clone ssal-rgb2-656-252 AT	0
0.13	0.19	0.33	0.28	-0.01	0.58	0.45	0.58	0.45	-0.01	0.73	0.48	Oncorhynchus mykiss S6 ribosomal pro	0
0.87	0.51	0.92	0.51	0.00	0.87	0.50	0.79	0.51	0.00	0.87	0.51	Salmo salar clone ssal-rgb2-617-226 C2	3E-83
0.45	0.51	0.25	0.28	0.21	0.38	0.32	0.18	0.17	0.02	0.00	0.00	Salmo salar clone HM4 (cyclin I)	9E-71
0.87	0.50	0.96	0.51	0.00	1.00	0.51	0.79	0.49	0.02	1.00	0.52	Salmo salar clone ssal-rgf-503-219 Vac	0
1.00	0.51	1.00	0.51	0.00	1.00	0.51	1.00	0.51	0.00	1.00	0.52	Salmo salar clone ssal-rgf-510-220 unk	0
0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	Coregonus lavaretus mitochondrial DN	0
1.00	0.51	1.00	0.51	0.00	1.00	0.51	1.00	0.51	0.00	1.00	0.52	Salmo salar clone ssal-rgf-510-220 unk	0
0.08	0.08	0.04	0.04	-0.01	0.26	0.23	0.04	0.04	0.07	0.27	0.25	Oncorhynchus mykiss heat shock 90kD	0
1.00	0.51	1.00	0.51	0.00	1.00	0.51	0.96	0.51	0.00	1.00	0.52	AF308735 Oncorhynchus mykiss 18S ri	0
0.13	0.19	0.17	0.16	-0.02	0.25	0.22	0.48	0.37	0.03	0.29	0.48	Salmo salar phosphorylase, glycogen (m	6E-67
0.91	0.51	0.91	0.51	0.00	0.95	0.51	0.82	0.51	0.00	0.67	0.53	Salmo salar clone ssal-eve-543-105 Sig	2E-99
0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA	0.00	0.00	Salmo salar partial mRNA for myosin r	0
0.42	0.34	0.25	0.28	-0.01	0.04	0.04	0.21	0.19	0.04	0.13	0.13	TSA: Hippoglossus hippoglossus all	6E-73
0.46	0.44	0.50	0.45	-0.02	0.50	0.38	0.46	0.36	-0.01	0.47	0.37	Salmo salar clone ssal-rgh-519-111 Het	0
0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA	0.00	0.00	Caligus rogercresseyi clone crog-evp-5	6E-143
0.58	0.50	0.54	0.50	0.02	0.63	0.51	0.25	0.34	0.17	0.59	0.51	Salmo salar clone ssal-rgb2-561-188 Fa	4E-131
0.50	0.50	0.43	0.50	0.02	0.17	0.28	0.21	0.44	0.03	0.44	0.50	Oncorhynchus mykiss complement com	4E-139
0.17	0.22	0.13	0.19	-0.03	0.54	0.50	0.58	0.50	-0.02	0.00	0.00	Salmo salar Angiotensinogen (angt), m	4E-132
0.55	0.49	0.39	0.32	0.09	0.29	0.40	0.42	0.42	-0.02	0.43	0.48	Salmo salar RED protein (red), mRNA	1E-151
0.46	0.36	0.21	0.19	0.04	0.58	0.48	0.21	0.31	0.06	0.24	0.37	Salmo salar IMP (inosine monophospha	4E-145
0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA	0.00	0.00	Salmo salar clone HM6	5E-137
0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	NA	0.00	0.00	Salmo salar clone ssal-evd-509-343 Mit	1E-107

Supplementary table 5.4

Analysis of linkage disequilibrium
Markers in LD with *q*-value < 0.05, calculated from 10000 permutations in Arlequin (v3.5.1.2)

a) analysis was done for the association family for all markers
(excluding markers showing <50% missing data, 9 markers removed)

b) analysis was done for the association family for all markers identified
as Fst outlier in natural population (15 markers)

a) all markers

Sample.Name	NUMBER OF MARKER IN LD	PERCENTAGE OF MARKER IN LD
BU965641_re5_	13	0.167
CA037452_re5_	11	0.141
CA037876_re3_	5	0.064
CA038170_NDB	16	0.205
CA038790_re5_	19	0.244
CA042392_re5_	13	0.167
CA042951_ND_	17	0.218
CA044550_ca20	3	0.038
CA045465_re5_	14	0.179
CA049476_re5_	15	0.192
CA051860_BC_	9	0.115
CA052650_148M	29	0.372
CA053246_re5_	52	0.667
CA053896_D__(	21	0.269
CA054630_BC_	6	0.077
CA054959_380R	18	0.231
CA056473_re3_	50	0.641
CA057176_NDB	55	0.705
CA057603_ND_	18	0.231
CA057987_156K	18	0.231
CA058340_re5_	11	0.141
CA058958_re5_	15	0.192
CA060324_154M	66	0.846
CA061393_ND_	65	0.833
CA062071_NDB	50	0.641
CA063623_352M	19	0.244
CB492682_re5_	12	0.154

CB492725_ca3_	8	0.103
CB496739_ca4_	16	0.205
CB497584_re3_	46	0.590
CB498771_re5_	7	0.090
CB500248_502K	16	0.205
CB509509_397K	13	0.167
CB510585_ND_	6	0.077
CB511030_339K	15	0.192
CB512085_538R	20	0.256
CB514545_re5_	8	0.103
CB516686_131R	22	0.282
CK990730_re5_	10	0.128
CX030416_re5_	65	0.833
DW553899_re5_	4	0.051
DY738545_re5_	16	0.205
EG755964_BC_	16	0.205
EG782906_re5_	22	0.282
cc000085_01AC	24	0.308
cc000102_01AG	28	0.359
cc000129_01AT	37	0.474
cc000167_01AC	21	0.269
cc000225_01AC	19	0.244
cc000236_01CT	4	0.051
cc000240_01CT	14	0.179
cc000258_01AC	15	0.192
cc000270_01CT	16	0.205
cc000303_01CT	29	0.372
cc000541_01AT	66	0.846
cc000857_01AG	36	0.462
cc000873_01AG	25	0.321
cc000952_01GT	28	0.359
cc001175_01AG	17	0.218
cc001276_01CT	34	0.436
cc001365_01AG	13	0.167
cc001404_04GT	41	0.526
cc001461_02AT	66	0.846
cc001477_15AG	23	0.295
cc001516_01CT	30	0.385
cc001558_03AC	34	0.436
cc001576_01CT	10	0.128
cc001605_01AG	32	0.410
cc001682_02CT	14	0.179
cc001705_02AG	37	0.474

cc001720_01CT	25	0.321
cc001882_04AC	24	0.308
cc001956_01CT	25	0.321
cc002099_04CT	24	0.308
cc002119_06AC	12	0.154
cc002133_01CT	31	0.397
cc002148_01AG	16	0.205
cc002195_01AG	24	0.308
	mean	**0.303**

b) outlier markers

Sample.Name	NUMBER OF MARKER IN LD	PERCENTAGE OF MARKER IN LD
CA038170_NDB	2	0.125
CA038790_re5_	2	0.125
CA051860_BC_	2	0.125
CA052650_148N	3	0.188
CA054630_BC_	2	0.125
CA054959_380R	5	0.313
CA057603_ND_	2	0.125
CA057987_156K	5	0.313
CB498771_re5_(	2	0.125
CB500248_502K	2	0.125
CB514545_re5_1	2	0.125
CX030416_re5_	11	0.688
DW553899_re5_	0	0.000
DY738545_re5_	2	0.125
cc000303_01CT_	5	0.313
cc001175_01AG	2	0.125
	mean	**0.191**