Génomique de la spéciation chez le Grand Corégone (Coregonus clupeaformis): Caractérisation des bases génomiques associées à la différenciation phénotypique

Thèse

Clément Rougeux

Doctorat en biologie
Philosophiæ doctor (Ph. D.)

Québec, Canada

Génomique de la spéciation chez le Grand Corégone (*Coregonus clupeaformis*):
Caractérisation des bases génomiques associées à la différenciation phénotypique

Thèse

Clément Rougeux

Sous la direction de :

Louis Bernatchez, directeur de recherche

Pierre-Alexandre Gagnaire, codirecteur de recherche

Résumé

L'évolution répétée et indépendante de la différenciation phénotypique entre espèces divergentes, suggérant des pressions sélectives similaires, constitue un contexte propice à l'étude de l'architecture génomique de la spéciation parallèle. L'objectif principal de cette thèse est d'apporter des éléments de réponses concernant les bases génomiques impliquées dans la différenciation phénotypique, et leur influence sur l'évolution de la divergence entre deux complexes d'espèces apparentés, le Grand Corégone (*Coregonus clupeaformis*) et le Corégone Lavaret (*C. lavaretus*). Plus précisément, il était nécessaire d'élucider l'origine du polymorphisme de chacune des populations, le rôle de cette variation génétique, son maintient durant la divergence et la différenciation phénotypique entre paires d'espèces. Une analyse génomique a permis de réaliser des inférences démographiques historiques, mettant en évidence la contribution simultanée de processus démographiques et sélectifs qui ont façonné les paysages génomiques de différenciation entre paires d'espèces. Ensuite, une analyse transcriptomique a permis d'identifier des bases polygéniques partagées impliquées dans la différenciation phénotypique parallèle entre paires d'espèces. De plus, ces bases polygéniques indiquent une forte rétention du polymorphisme ancestral sous l'action de la sélection divergente. Finalement, un parallélisme de régions d'ADN différentiellement méthylées entre espèces a été identifié. Bien que cette méthylation repose sur des bases génomiques, ces régions différentiellement méthylées sont associées à une différenciation transcriptionnelle entre espèces des complexes d'espèces du Corégone Lavaret et du Grand Corégone. Ces travaux montrent que la sélection naturelle est contrainte par certains génotypes permettant d'acquérir un parallélisme phénotypique de façon indépendante, et agir sur du polymorphisme ancestral, notamment dans un contexte de spéciation parallèle. Enfin, cette thèse permet de lever le voile et de contribuer à la compréhension des mécanismes génomiques associés à la divergence adaptative pouvant mener à la spéciation écologique, notamment en utilisant une approche intégrative.

Abstract

Repeated evolution of phenotypic differentiation between diverging species pairs provides an ideal context for the study of the genomic architecture of parallel speciation. The main objective of this thesis is to provide evidence concerning the genomic bases involved in phenotypic differentiation, and their influence on the evolutionary potential of species complexes belonging to two related lineages, the Lake Whitefish (*Coregonus clupeaformis*) and European Whitefish (*C. lavaretus*). Specificaly, it is necessary to elucidate the origin of the genetic polymorphism of each population from both whitefish lineages, and to which extend this polymorphism was involved in the genetic divergence and phenotypic differentiation between species pairs. A genome-wide analysis allowed to infer the divergence history combining the effects of historical demograhy and selective pressure that collectively shape the genomic landscape of differentiation between species pairs. Then, transcriptomic analyses revealed parallel polygenic bases involved in the phenotypic differentiation of species pairs, and such genes were enriched in shared ancestral polymorphism. Finaly, a parallel differential methylation level has been identified between species. Although this methylation is genomicaly based, these differentially methylated regions are associated with a transcriptional differentiation between the limnetic and benthic species. This work shows that selection is constrained by some genotypes which could lead to an independent parallel phenotypic aquisition, but also act on the maintainance of ancestral genetic polymorphism, particularly in a context of parallel speciation. This thesis allows to highlight and to contribute to the understanding of the genomic mechanisms generating biodiversity, notably by using an integrative approach.

Table des matières

Liste des Tableaux

Liste des figures

Liste des abréviations

QTL, eQTL: Quantitative trait loci (‘Locus de trait quantitatif’), expression QTL

ADN, mtDNA: Acide désoxyribonucléique, ADN mitochondrial

ARN: Acide ribonucléique

Ne: Taille efficace de population

SNP: Single nucleotide polymorphism (‘polymorphisme simple nucléotidique’)

SGV: Standing genetic variation (‘polymorphisme ancestral maintenu’)

BSC: Biological species concept (‘concept d’espèce biologique’)

IR: Isolement reproducteur

LD: Linkage disequilibrium (‘déséquilibre de liaison’)

DH: Divergence hitchhiking (‘divergence par auto-stop’)

Fst: Indice de différenciation génétique

Dxy: Indice de mesure de divergence génétique absolue entre lignées par comparaison de polymorphisme

CDS: Séquence codante de gène

Dn, Ds: Substitution non-synonyme, substitution synonyme

UTR: Untranslated region (‘région non traduite’)

siRNA: small interfering RNA (‘petits ARN interférants’)

TEs: Transposable elements (‘éléments transposables’)

RNAseq: Séquencage d’ARN

RADseq: Restriction Associated DNA sequencing (‘séquencage d'ADN associé à un site de restriction')

WGBS: Whole genome bisulfite sequencing (‘séquencage de génome entier convertit au bisulfite’)

JAFS: Joint allele frequency spectrum (‘spectre joint des frequencies alléliques’)

CIs: Confident intervals (‘Intervalle de confiance’)

se: standard-error (‘erreur standard’)

AIC: Akaike information criterion (‘critère d’information d’Akaike’)

GO: Gene ontology (‘Ontologie génique’)

VCF: Variant call format (‘format de genotypes’)

DEG: Differentially expressed genes (‘gène différentiellement exprimé’)

RDA: Redundancy analysis (‘Analyse de redondance’)

PCA, PC: Principal component analysis (‘Analyse en composantes principales’), composante principale

ANOVA: Analyse of variance (‘analyse de variance’)

GIm: Generalized linear model (‘modèle linéaire généralisé’)

PPIs: Protein-protein interactions (‘Intéractions protéine-protéine’)

DML: Differentially methylated loci (‘locus différentiellement méthylé’)

DMR: Differentially methylated region (‘région différentiellement méthylée’)

« *L'alpiniste est un homme qui conduit son corps là où,*

un jour, ses yeux ont regardé… »

Gaston Rébuffat – 1973 – Le Massif du Mont Blanc

Remerciements

En premier lieu, je tiens à remercier chaleureusement mon directeur de thèse Louis Bernatchez. Louis en tant que doctorant, merci infiniement de m'avoir intégré à ton laboratoire. Merci de l'opportunité que tu m'as offert initialement lors d'un stage. Merci de la confiance que tu m'as accordée en m'offrant la possibilité de rester au labo et me confiant la gestion de plusieurs projets dans le laboratoire. Surtout, merci pour ce projet de doctorat fascinant sur le Grand Corégone! Durant ce projet, tu as su me laisser l'entière libertée d'explorer et de comprendre les données. Tu as également su mettre les ressources nécessaires pour répondre à nos problématiques, et palier aux differents imprévus. Tu m'as offert et apporté énormément durant mon séjour au labo Bernatchez, j'ai beaucoup appris et t'en serais toujours reconnaissant! Je tiens également à te remercier sur un plan plus personnel. Merci pour ces moments informels, liants discussions, échanges et fun!! Pour tous ces aspects et bien d'autres, merci.

Je tiens à remercier mon co-directeur de these Pierre-Alexandre Gagnaire. Je te remercie d'avoir accepté d'embarquer sur ce projet de doctorat. Tu as très rapidement été présent dans mon projet de doctorat, et tu as su prendre le temps d'échanger avec moi régulièrement ou lorsque c'était nécessaire. J'ai appris énomément lors de chacun de nos échanges et lors de mon séjour à la station marine de Sète. Être à ton contact est vraiment enrichissant, anthousiasment et stimulant. Merci pour tout ce que tu m'as apporté et ce que tu as apporté à ce projet de doctorat. Merci également pour tous les aspects non-scientifiques. Je garde d'excellents souvenirs de mon séjour à Sète, notamment des sessions de chasse sous-marine dans l'étang de Thau, de la dégustation de vins et autres petits plaisirs. Merci pour tout!

Je souhaite remercier mon comité d'encadrement: Julie Turgeon et Christian Landry. Merci pour le recul que vous avez pu m'apporter durant nos rencontres, ainsi que les autres échanges tant scientifiques que personnels. Merci a vous!

Je voudrais également remercier mes collaborateurs qui ont joué un role important dans ce projet de doctorat, bien que certains fruits de nos travaux ne sont pas inclus dans cette these. Merci à Martin Laporte pour son apport et son support. Merci à Éric Normandeau pour sa richesse de ressources. Merci à Ole Seehausen et Kim Praebel pour le partage

d'échantillons provenant d'Europe. Merci également à Sigbjørn Lien et Torfinn Nome pour leur aide avec l'assemblage du genome du Grand Corégone.

Un projet et un étudiant ne seraient pas les mêmes sans les membres du laboratoire Bernatchez. Merci a tous pour votre aide, le support et les aspects sociaux! Merci particulier à Jean-Sébastien Moore, Anne Dalziel, Anne-Marie Dion-Côté, Charles Perrier, Thierry Gosselin, Anne-Laure Ferchaud et Laura Bénestan vous m'avez accueilli et encouragé dans les premières étapes et durant ce doctorat. Merci à Martin Laporte, Jérémy Le Luyer pour la folie, la philosophie et l'intéraction entre les deux la plupart du temps. Merci à Anne-Laure Ferchaud pour ce que tu as pu m'apporter en tant que personne. Le labo serait different si il n'y avait pas une équipe de rêve pour le gérer; un gros merci à Guillaume côté *alias* 'barbu' pour ton aide et ton support à mes débuts dans le laboratoire et pour les game de hockey chez Girard! Merci à Cécilia Hernandez pour l'aide dans le labo durant l'optimisation du GBS (t'es arrivée au bon moment!) et pour les discussions permettant de reconnecter à la réalité. Merci à Alysse Perreault-Payette pour tout ce que tu fais dans le laboratoire pour chacun des membres, ton dynamisme et ta bonne humeur. Merci à Bérénice Bougas, ton oreille attentive est vraiment appréciée dans les périodes moins façile et tu as la capacité à nous remotiver ou nous faire prendre du recul. Merci également à Éric Normandeau, en plus d'aider volontier les étudiants et de les soulager en leur apportant des alternatives/solutions, tu es capable de déceler des baisses de forme et de discuter avec nous. Meilleur qu'un psychologue, un bio-info-psy! Merci à Claire Mérot, Hugo Cayuela, vous m'apportez un support et un autre regard sur la science. Merci à Yann Dorant, Kyle Wellband, Ben Sutherland, Scott pavey, Vincent Bourret, Geneviève Ouellet-Cauchon, Damien Boivin-Delisle, Simon Bernatchez, Charles Babin puis Madoka Krick, pour un super terrain corégone.

Une pensée particulière pour le local 1151 et les personnes qui l'ont partagé avec moi au quotidien: Charles Perrier, Anne-Marie Dion-Côté, Madoka Krick, Olivier Morissette, Martin Laporte et Hugo Cayuela. Merci pour ces bons moments!

Lors de mon séjour en France, j'ai l'immense plaisir de recontrer Nicolas Bierne. Merci pour les conversations autour de la speciation et du Grand Corégone. Merci à Christelle Fraïsse pour les échanges, ainsi qu'à Florentine Riquet, Ahmed Souissi, Alban le Moan et à François Bonhomme.

Un merci special à ceux qui ont offert de leur temps à la relecture de cette these: Anne-Laure ferchaud, Anne-Marie Dion-Côté, Claire Mérot, Élodie Serre et Pierre-Alexandre Gagnaire.

C'est avec bonheur que j'adresse une pensée particulière à mes amis de toujours. Merci pour votre patience, pour votre présence malgré la distance: Baptiste Suzzoni, Benjamin Labelle, Nicolas Luzé, Fabien Lanave, Mathieu Bezecourt, Anouk Roussely, Simon Favre, La Urle, Laurent Lataste,… Vous me manquez!

Merci aux organismes subventionnaires qui m'ont soutenu et m'ont permis les stages et autres congrès à l'étranger: le RAQ, le département de biologie de l'Université Laval et la foundation Richard-Bernard, l'AÉLIÉS et l'IBIS.

Pour finir, un immense merci à ceux sans qui je ne serais certainement pas là, ma famille. Vous m'avez toujours supporté, encouragé, motivé dans toutes les étapes malgré la distance d'un ocean ou de plusieurs continents selon les périodes. Vous avez toujours été là et ca continue aujourd'hui encore. Je remercie mon père, Christian. Je remercie mes grands parents, Yvonne et Gérard pour l'exemple de générosité et de joie de vivre qu'ils sont. Merci a tatie. Merci à celle qui a toujours été là, contre vents et marées, ma mère Marie-Jeanne. Merci à mon grand frère Elric et ma grande soeur Elfi, vous avez clairement influencé ma vie et contribute à tous ça et leur petites familles, Mathéo, Yanis, Lalie, Fantine et Léane. Je ne pourrais conclure ces remerciements sans addresser ma reconnaissance affectueuse à celle qui m'accompagne au quotidien, Élodie.

Merci!

Avant-propos

Cette thèse est organisée en 6 chapitres, incluant l'introduction générale (Chapitre 1) et la conclusion (Chapitre 6). Les chapitres 2, 3, 4 et 5 sont publiés ou en voie de l'être dans des revues scientifiques. D'autre part, j'ai contribué à 5 autres articles publiés ou soumis par des membres du laboratoire, ou des chercheurs en visite au laboratoire (listés dans le chapitre 8) dont deux portant sur le Grand Corégone. Ceux-ci sont inclus en Annexe (Chapitre 9).

Le Chapitre 2 est publié sous la référence: Rougeux C, Bernatchez L, Gagnaire P-A (2017) Modeling the Multiple Facets of Speciation-with-Gene-Flow toward Inferring the Divergence History of Lake Whitefish Species Pairs (Coregonus clupeaformis). Genome Biology and Evolution, 9, 2057–2074.

CR, LB et PAG ont conçu le projet. PAG a supervisé le projet. PAG et CR ont produit les données. CR a analysé les données. CR a écrit le manuscrit en collaboration avec PAG et LB.

Le Chapitre 3 a été soumis à la revue *Journal of Evolutionnary Biology*: Rougeux C, Gagnaire P-A, Bernatchez L (en révision) Differential introgression following postglacial secondary contact in European whitefish species pairs.

CR, LB et PAG ont conçu le projet. CR a produit les données. CR a analysé les données. CR a écrit le manuscrit en collaboration avec PAG et LB

Le Chapitre 4 a été soumis à la revue *Evolution*, et est actuellement disponible en version preprint: Rougeux C, Gagnaire P-A, Praebel K, Seehausen O, Bernatchez L (2018) Convergent transcriptomic landscapes under polygenic selection accompany inter-continental parallel evolution within a Nearctic Coregonus (Salmonidae) sister-species complex. bioRxiv, 1–27.

CR et LB ont conçu le projet. LB a supervisé le projet. KP et OS ont fournit des échantillons. CR a produit les données. CR a analysé les données. CR a écrit le manuscrit en collaboration avec PAG, KP, OS et LB.

Le Chapitre 5 a été soumis à la revue *Molecular Ecology*: Rougeux C, Laporte M, Gagnaire P-A, Bernatchez L (en révision) Genomic, epigenomic and transcriptomic differentiation in inter-continental parallel speciation of whitefish (Coregonus sp.) sympatric pairs.

CR et LB ont conçu le projet. LB a supervisé le projet. CR a produit les données. CR et ML ont analysé les données. CR a écrit le manuscrit en collaboration avec ML, PAG, et LB.

Introduction générale

1.1 Biodiversité et évolution

L'immersion dans la compréhension du monde qui nous entoure, pour la plupart des biologistes, résulte généralement de l'émerveillement et de l'intérêt envers notre environnement. Par environnement j'entends ici tout écosystème (naturel, aménagé, rural ou urbain) composé d'un biotope, le milieu de vie, et d'une biocénose. La biocénose correspond à la composition et l'organisation des organismes vivant dans le biotope, ainsi que leurs interactions. La variabilité existante entre organismes de différents biotopes (terrestres, marins et aquatiques) et les complexes écologiques auxquels ils appartiennent, définissent la biodiversité. De plus, la variabilité observée au sein d'une espèce, entre espèces et dans les écosystèmes illustrent les différentes dimensions dans lesquelles la biodiversité se décline.

Ainsi, chaque *biota* peut être caractérisé par sa diversité taxonomique, écologique, phénotypique et génétique, mais aussi par la variation de ces diversités dans le temps et dans l'espace. De tels changements spatiotemporels de biodiversité peuvent être, par exemple, le résultat de changements de conditions d'un milieu donné, ou de manière intrinsèque, une réduction de la variation génétique au sein d'une population. Pour chacune de ces situations, nous pouvons observer une perte de biodiversité. Au contraire, d'un point de vue évolutif, dont nous discuterons plus loin, il est concevable que de tels changements peuvent aussi générer de la biodiversité par l'intermédiaire de processus de diversification, parmi lesquels la spéciation qui correspond à la naissance de nouvelles espèces à partir d'une espèce ancestrale. La plupart des caractéristiques associés à des taux de spéciation élevés, incluant la spécialisation pour une niche écologique, des populations isolées et de petite taille, sont aussi liés à un taux élevé d'extinctions (Greenberg and Mooers 2017). Finalement, la biodiversité observée dans un environnement restreint, serait la résultante d'un équilibre entre les phénomènes d'extinction et de spéciation.

Bien que les concepts d'extinction et de spéciation soient indissociables, dans le domaine de la biologie de la conservation, la prévention de l'extinction notamment, est fondamentale. En revanche, en biologie de l'évolution beaucoup d'efforts sont alloués à la compréhension de la diversification d'espèces à partir d'une espèce ancestrale. En effet, l'évolution correspond aux changements de traits héritables au sein d'une population au fil des générations. Ces modifications phénotypiques peuvent montrer un avantage, pour une

sous-population dans un habitat contrasté, laquelle finira éventuellment par se différencier de la population d'origine sous l'action de la sélection divergente. Ce processus graduel peut aboutir au processus de spéciation écologique si il est associé à la mise en place d'un isolement reproducteur. On parle du continuum de la spéciation (Shaw and Mullen 2014). Ce continuum peut être assimilé aux modifications génétiques associées à une divergence entre deux groupes d'individus (lignées), jusqu'à la mise en place de barrières à la reproduction entre les lignées. Cela correspond à la notion d'espèces biologiques (Mayr 1963). Par conséquent, la compréhension d'un tel processus évolutif nécessite de s'attarder dans un premier temps sur les étapes initiales de ce continuum. Je présenterais alors les concepts fondamentaux associés à différentes étapes de processus de spéciation, illustrés par des exemples empiriques. Nous verrons un cas possible d'étape initiale du continuum de spéciation, en discutant (i) des bases génétiques associées à la divergence évolutive en réponse à des pressions évolutives menant par exemple à l'adaptation, (ii) de la mise en place de barrières reproductives, (iii) les différents mécanismes évolutifs responsables de l'établissement de telles barrières de par la compréhension de paysages génomiques, ainsi que (iv) les contextes propices à l'étude de la génomique de la spéciation.

1.2 L'adaptation

Des populations d'une même espèce mais occupant des habitats différents peuvent se différencier sous l'effet de la sélection locale. En l'absence de toute forme de contrainte évolutive autre que la sélection naturelle, ces populations soumises à des conditions environnementales (habitats) différentes peuvent évoluer des traits phénotypiques différents qui leur confèrent un avantage sélectif dans leurs habitats respectifs. Ainsi, chacun des génotypes – composition allélique de l'ensemble des gènes d'un individu – résidents aura une valeur sélective supérieure dans son habitat d'origine, par rapport aux génotypes issus d'autres habitats. La valeur sélective d'un individu correspond à la contribution relative de ces gènes à la génération suivante. Cela signifie qu'un individu dont la valeur sélective est élevée pour un habitat donné a des chances de survie, de reproduction et de survie de sa progéniture supérieures à celles d'un individu dont la valeur sélective est plus faible (Hereford 2009; Savolainen, Lascoux, and Merilä 2013a). Par

conséquent, on parle d'adaptations locales différentielles lorsque, sous l'effet de la sélection naturelle, des populations locales évoluent des phénotypes à valeur sélective différentes, leur permettant alors de se maintenir sous des conditions biotiques ou abiotiques différentes (Wlliams 1966).

Outre la sélection naturelle divergente, d'autres contraintes évolutives peuvent cependant intervenir et influencer le niveau de divergence génétique entre populations et le niveau de différentiation phénotypique. Parmi ces contraintes, le flux génique entre populations peut empêcher l'adaptation locale et limiter la différentiation. Un manque de variation génétique, agissant comme véritable substrat à la sélection naturelle, peut aussi limiter l'adaptation locale. De même que l'architecture – nombre, répartition des gènes impliqués et leur relations entre eux – du trait phénotypique, selon son niveau de complexité. Finalement, des patrons similaires de divergence et de différenciation peuvent être obtenus par des processus stochastiques – tel que la dérive génétique qui est le processus d'échantillonnage aléatoire d'un sous ensemble d'allèles dans une population. Les forces évolutives et contraintes génétique énoncées précédemment seront détaillées dans les sections suivantes.

La variation génétique, qui correspond aux différents génotypes présents dans une population, est nécessaire à l'évolution en réponse aux pressions de sélection. En effet, c'est cette variation génétique pour une population soumise à des changements de conditions environnementales qui va conditionner l'évolvabilité de cette population. Un large éventail d'altérations génétiques est impliqué dans des changements évolutifs incluant les substitutions, les délétions et les insertions dans des gènes, le remaniement de séquences régulatrices, mais aussi la duplication de gènes, par exemple (Xie et al. 2019). Cependant, de nombreux variants alléliques n'affecteront pas nécessairement le trait phénotypique sous sélection, mais potentiellement un trait phénotypique qui n'est pas soumis à une forte pression de sélection. De plus, certains allèles délétères dans la plupart des environnements ou dans des fonds génétiques particuliers, ont tendance à être rapidement purgés de la population. Il est également possible qu'une diversité génétique substantielle au sein d'une population ne réponde pas à une pression sélective par une modification phénotypique. Cela est notamment possible si le trait phénotypique est limité par des contraintes physiologiques, développementales ou de valeur sélective négative ou nulle (Le Rouzic and Carlborg 2008). Par conséquence, une proportion restreinte de la diversité

génétique d'une population a le potentiel de participer à l'adaptation locale. Outre l'origine de la variation génétique d'une population, l'ampleur de l'effet d'une mutation, le nombre de variants et la distribution de ces variants affectant un trait phénotypique dans le génome, sont des aspects fondamentaux lorsqu'on s'intéresse à l'architecture génomique de l'adaptation locale (Yeaman 2015).

1.2.1 Architecture génétique de l'adaptation locale

L'architecture génétique de l'adaptation (locale) est au centre des considérations en biologie de l'évolution. Le premier développement théorique s'intéressant à la distribution de la taille des effets (*i.e.*, effet sur la variation phénotypique) de mutations adaptatives, c'est-à-dire le remplacement dans une population d'un allèle par un autre de valeur sélective supérieure, remonte aux années 1930. Le modèle géométrique de Fisher suggère que l'adaptation se met en place essentiellement par la fixation d'un grand nombre d'allèles à effets faibles et additifs (Fisher 1930). Ce modèle infinitésimal alloue donc peu de place pour des modifications génétiques dont les effets sont majeurs sur le phénotype. Selon le raisonnement de Fisher, une mutation à effet important sur la variance phénotypique a peu de chance de participer à l'évolution adaptative du fait de son effet pléiotrope – impliquée dans plusieurs traits phénotypiques – délétère. Ce modèle fut modifié par la suite de façon à intégrer des processus évolutifs. Ainsi, Kimura proposa un modèle intégrant des effets stochastiques associés à la dérive génétique. Son modèle prédit que les mutations de taille d'effet intermédiaire auraient tendance à être majoritaires face à des mutations à effet mineur qui seraient potentiellement perdues par dérive génétique (Kimura 1983). Ces deux modèles impliquent que la variation causée par les mutations conduit à l'expression et la sélection du phénotype optimal dans les nouvelles conditions de l'habitat. Ainsi, Orr (1998) émit l'hypothèse que l'effet relatif des mutations, qui seront fixées durant le processus d'adaptation locale, est dépendant de la distance entre la population initiale et le phénotype optimum dans le nouvel habitat. Le modèle sous-jacent considère l'ensemble des étapes – taille de la variation associée à chaque mutation – pour acquérir un phénotype optimal dans un paysage adaptatif et prédit une distribution exponentielle négative de la taille des effets associés aux substitutions adaptatives (Orr 1998). Par conséquent, des mutations à effets importants sont attendues en grand nombre dans les premières étapes du processus

d'adaptation à un nouvel environnement, puis de nombreuses mutations à moindre effet lorsque le phénotype tend vers l'optimum. Ce modèle est supporté par de nombreuses études de cartographie de locus de caractère quantitatifs (QTL : « Quantitative trait loci ») – régions génomiques associées à un trait phénotypique quantitatif. On retrouve notamment des cas pour lesquels les études de cartographie ont identifié des locus à forts effets. C'est le cas de la coévolution de traits floraux dans le genre *Mimulus* durant une transition de pollinisateur (Bradshaw and Schemske 2003), du changement de coloration de la robe chez la Souris des sables (*Peromyscus polionotus*) (Hoekstra et al. 2006), de la présence ou l'absence de plaques latérales chez l'épinoche à trois épines (*Gasterosteus aculeatus*) (Colosimo et al. 2005; Jones et al. 2012). Néanmoins, certaines études n'ont pu mettre en évidence des QTLs à effet majeur potentiellement impliqués dans la divergence de système d'appariements pour certaines espèces de *Mimulus* (Fishman et al. 2002) ou encore dans le changement d'hôte pour un insecte spécialiste (Oppenheim et al. 2012), suggérant l'implication de nombreux QTLs à faibles effets individuels et agissant de manière additive. Les exemples précédents illustrent la panoplie d'architectures génomiques de l'adaptation, en termes de nombre et variance associée aux QTLs. Ils recouvrent le modèle monogénique – un seul QTL gouverne l'effet sur le caractère quantifiable (exemple : trait dit mendélien) –, le modèle oligogénique pour lequel un faible sous-ensemble de QTLs explique la majorité de la variance associée au trait, puis le modèle polygénique ou un ensemble de gènes sont impliqués dans la variance du caractère quantifiable. Plus récemment, un modèle « omnigénique » a été proposé et stipule que l'ensemble du génome est associé à la variation d'un trait phénotypique complexe (Boyle et al. 2017). Dans l'ensemble, ces observations indiquent que les modèles théoriques historiques sont inappropriés pour capturer l'intégralité des scénarios biologiques pertinents pour comprendre la distribution de la taille d'effet de QTLs, celle-ci étant modulée selon les processus évolutifs et contrainte par les traits phénotypiques (Matuszewski et al. 2015).

1.2.2 Processus évolutif et architecture génétique de l'adaptation

1.2.2.1 La sélection naturelle

Si les modèles énoncés précédemment considèrent généralement que plusieurs QTLs sont impliqués dans la variation phénotypique d'un trait donné, il est concevable qu'une seule substitution puisse agir sur plusieurs traits simultanément (pléiotropie). Selon Fisher, il est vraisemblable que de telles mutations – à effet important – aient un effet pléiotropique délétère, éloignant la population de l'optimum phénotypique et par conséquent ne contribuent pas à l'adaptation (locale) de la population. Cette théorie est étayée par le fait que les mutations à effet important ont tendance à avoir des effets pléiotropiques antagonistes (Wagner *et al.* 2008) et que le niveau de pléiotropie d'un QTL est positivement corrélé à la taille de son effet (Wagner *et al.* 2008; Wang *et al.* 2010). De plus, des études théoriques plus récentes étayent la prédiction selon laquelle la pléiotropie favoriserait des mutations à faible effet lorsque beaucoup de traits phénotypiques sont sous sélection (Tenaillon 2014). Toutefois, ce concept est nuancé par des modèles intégrant de la modularité pléiotropique. Ainsi, une mutation affecte un sous-ensemble des traits phénotypiques sous sélection, permettant alors à la population de répondre aux contraintes adaptatives et réduisant le niveau de complexité (Welch and Waxman 2003). En contre-partie, des études empiriques ont illustré l'adaptation locale facilitée par pléiotropie synergique, où les effets de la mutation sur les traits phénotypiques sont avantageux (Wang et al. 2010; Paaby and Rockman 2013). En effet, de la pléiotropie est associée à de l'adaptation locale 1) chez l'épinoche à trois épines avec un effet du gène codant pour l'ectodysplasine (*EDA*) sur les plaques latérales et la distribution d'organes sensoriels (MacPherson *et al.* 2015), 2) entre espèces de plantes dans la divergence pour des traits floraux (Smith 2015), ou encore 3) dans l'adaptation virale sous évolution expérimentale (McGee *et al.* 2016). Finalement, outre l'effet d'une mutation sur un trait phénotypique, la valeur sélective associée à cette mutation est déterminante selon la distance que le phénotype initial doit parcourir pour aboutir à l'optimum dans le nouvel habitat.

La théorie considère généralement qu'une population est initialement éloignée de l'optimum phénotypique à la suite d'une perturbation environnementale ou d'un

changement d'habitat (Orr 1998). Cela implique que dans les premières étapes de l'adaptation, les mutations à effets faibles et importants permettront à la population de se rapprocher de l'optimum mais dès que la population est proche et tend vers l'optimum, les mutations à effet majeur auront une valeur sélective le plus souvent négative. Par conséquent, il existe une corrélation positive entre la distance à l'optimum et la taille moyenne des effets des mutations succeptibles d'améliorer la valeur sélective de la population. Cette relation a été illustrée en évolution expérimentale sur des bactériophages (McGee *et al.* 2016), des bactéries (*Escherichia coli*) (Sousa *et al.* 2011), mais aussi en étudiant des populations naturelles (Rogers *et al.* 2012). L'utilisation de populations différentiellement adaptées ou d'espèces sœurs peut permettre d'étudier les bases génomiques de traits adaptatifs et d'identifier les allèles qui auront été fixés durant la période d'adaptation. Par ce principe, des cas de QTLs à effet important ont été mis en évidence chez la drosophile (Jones 2002), la fleur-singe (Bradshaw and Schemske 2003) et l'épinoche à trois épines (Colosimo *et al.* 2004). Cependant, il serait pertinent de suivre l'évolution de la variance du trait phénotypique et de la distribution de la taille d'effet avec l'accumulation de mutations au fil du temps (Rockman 2012). Chez l'humain en revanche, la taille individuelle est contrôlée par de nombreux allèles à faibles effets. Dans une population européenne contemporaine, les allèles à faibles effets prédominent. Il est intéressant de considérer que la taille d'effet de ces allèles est liée à la distance à l'optimum (Dittmar *et al.* 2016). En effet, des analyses comparatives entre une population pygmée et une population voisine d'agriculteurs ont mis en évidence une fréquence supérieure d'allèles à forts effets dans la population pygmée (Perry *et al.* 2014). Or, les Pygmées ont évolué à plusieurs reprises de manière indépendante à la suite de la colonisation de la forêt tropicale (Perry *et al.* 2014). Cette architecture associée à la taille individuelle pourrait donc représenter les stades précoces d'une adaptation à un nouvel environnement. Ces études illustrent l'hétérogénéité des architectures dont les deux extrêmes sont définies par un QTL majeur où un ensemble d'allèles impliqués dans le phénotype optimal, ainsi que la complexité du trait phénotypique étudié.

Il est possible que les conditions du nouvel habitat soient variables, induisant une architecture génétique de l'adaptation locale dynamique. Certains scénarios, tels que l'adaptation à des modifications graduelles et continues des conditions environnementales, peuvent induire l'adaptation d'une population à un optimum phénotypique dynamique. Pour

une population de taille finie initialement adaptée à son habitat, la théorie prédit une distribution intermédiaire des tailles d'effets, avec peu de mutations à effets faibles et importants (Kopp and Hermisson 2009; Matuszewski et al. 2015). La magnitude des changements de taille d'effets varie selon la rapidité à laquelle les conditions changent. En effet, si les conditions de l'habitat changent rapidement, la distribution se déporte vers une augmentation des mutations à forts effets (Matuszewski *et al.* 2014). De plus, dans un tel cas, la pléiotropie peut ajouter un temps de latence jusqu'à l'obtention d'une mutation favorable, augmentant les probabilités d'adaptation par l'intermédiaire de mutations à forts effets (Matuszewski *et al.* 2014). Par conséquent, il est théoriquement possible qu'une population s'adapte à un environnement dont les conditions changent rapidement grâce à des mutations à effets importants, notamment en présence de pléiotropie. Au contraire, dans une dynamique lente du déplacement de l'optimum phénotypique – faible vitesse des changements environnementaux – l'adaptation locale implique des mutations à faibles effets. Une évolution expérimentale a permis d'illustrer cette association entre la distribution des tailles d'effets d'allèles adaptatifs et le taux de dynamisme de l'optimum chez *Chlamydomonas* (Collins and de Meaux 2009). Finalement, la sélection naturelle imposée par les conditions de l'habitat aura des conséquences différentes sur des traits phénotypiques selon leurs architectures génétiques, et en fonction de l'intensité de sélection associée aux changements environnementaux et à leur dynamique.

1.2.2.2 Mutation et dérive génétique

La mutation – modification de l'information d'une séquence d'ADN – est un évènement stochastique, dont la fréquence varie en fonction des régions génomiques et de la taille efficace de la population (*Ne*). Le taux de mutation et la distribution de la taille d'effets des mutations, sur les traits phénotypiques associés, vont déterminer et contribuer à la variation génétique totale de la population, qui sera ensuite soumise à des pressions évolutives (*i.e.*, sélection naturelle et dérive génétique). Le taux d'évolution d'une population sera, quant à lui, déterminé par un équilibre entre la fréquence des mutations bénéfiques et la fréquence des mutations délétères (Whitlock 2003), bien que la majorité des mutations qui affecte une population soient délétères (Eyre-Walker 2006). Malgré un avantage sélectif, des mutations ont une faible probabilité de fixation, même dans une grande

population. Cette probabilité dépend directement du coefficient de sélection de la mutation (Haldane 1927), et diminue avec le *Ne* de la population comme conséquence directe de l'effet de la dérive génétique (Whitlock 2003). Ainsi, la dérive génétique peut directement influencer la distribution de taille d'effet des substitutions adaptatives. En effet, dans le modèle original d'Orr (1998), la diminution du *Ne* induit une perte des allèles à effets faibles et intermédiaires par dérive génétique (plus importante dans les petites populations que dans les grandes). Par conséquent, de petites populations vont plus vraisemblablement porter des mutations adaptatives à effets importants, bien que des contraintes évolutives puissent contrecarrer la dérive génétique et maintenir des allèles à faibles effets (Barrett *et al.* 2006).

1.2.2.3 La variabilité génétique ancestrale

Les substitutions peuvent avoir différentes origines. Elles peuvent apparaître par nouvelle mutation, tel que dans les cas présentés précédemment, ou être présente initialement dans la variabilité génétique préexistente. On parle d'adaptation à partir de polymorphisme ancestral préexistant dans la population (*standing genetic variation* : SGV, (Hermisson and Pennings 2005)). Outre le polymorphisme d'une population, la SGV est composée essentiellement d'allèles neutres et délétères maintenus par un équilibre mutation-sélection-dérive génétique (Hermisson and Pennings 2005; Matuszewski et al. 2015). Lorsque les conditions de l'habitat changent, les allèles de la SGV peuvent devenir favorables et confèrent le substrat nécessaire pour l'adaptation à un nouvel optimum phénotypique. De plus, la sélection sur la SGV peut être plus efficace que sur les mutations *de novo*, car les allèles adaptatifs sont d'ores et déjà présents dans la population et sont maintenus à des fréquences supérieures à la fréquence initiale d'une mutation. Il est moins probable que ces allèles soient perdus par dérive génétique. Par conséquent, l'adaptation peut être plus rapide dans ces conditions (Barrett and Schluter 2008). La vraisemblance d'adaptation à partir de la SGV est dépendante du coefficient de sélection des allèles avant et après les changements de conditions (Matuszewski *et al.* 2015). Comme mentionné précédemment, les allèles initialement neutres ont une probabilité supérieure de fixation que les allèles provenant de substitutions, à taille d'effet similaire sur le phénotype, du fait de leur maintien dans la population au dépend de la dérive génétique (Hermisson and

Pennings 2005). De plus, plus les effets des allèles sont délétères dans les conditions d'habitat initiales, plus ils tendent à être avantageux dans les nouvelles conditions et donc à se fixer rapidement (Hermisson and Pennings 2005). Bien qu'il y ait peu d'exemples de variations de coefficients de sélection d'allèles suite à un changement d'habitat de la population, une étude a montré que des mutations délétères chez la bactérie *Pseudomonas fluorescens* confèrent une résistance à un antibiotique en présence de l'antibiotique (Kassen and Bataillon 2006). De même, chez la levure *Saccharomyces cerevisiae*, des mutations délétères en milieu de culture de référence ont montré des avantages sélectifs en conditions de stress salin (Hietpas *et al.* 2013). Des cas d'adaptation à partir de la variabilité génétique ancestrale ont été documentés dans des populations naturelles. Citons par exemple le cas de l'épinoche à trois épines, où l'allèle associé à l'absence de plaques latérales (trait caractéristique de populations d'eaux douces) est présent dans les populations marines pourvues de plaques où il est délétère (Barrett and Schluter 2008; Nelson and Cresko 2018). En outre, la résistance aux antibiotiques dans le cas du virus de l'immunodéficience humain (VIH) serait vraisemblablement associée à la SGV (Pennings 2012). Il est important de noter que l'adaptation par SGV et par mutation ne sont pas mutuellement exclusives. Le maintien du polymorphisme ancestral (équilibre mutation-sélection-dérive génétique) est également dépendant de l'effectif efficace de la population (*Ne*) et du coefficient de sélection (*s*), correspondant au produit *Ne*s*.

L'adaptation à partir de la SGV nécessite que la variabilité génétique constituée d'allèles neutres et délétères soient maintenue dans la population jusqu'au changement de conditions de l'habitat, moment où ces allèles deviendront avantageux. Ce maintien du potentiel évolutif est facteur du *Ne* et du taux de mutation (μ). La modélisation de l'adaptation à partir de la SGV a permis de tester l'influence de *Ne* et μ dans une population (Hermisson and Pennings 2005). Ces travaux ont mis en évidence l'importance de la SGV dans l'adaptation locale, mais aussi une très faible probabilité de fixation, dans une population de grande taille (Hermisson and Pennings 2005). Dans des populations de taille intermédiaire (*Ne* moins important), μ pourrait augmenter (*e.g.*, hotspot mutationel) de sorte à maintenir la probabilité d'adaptation à partir de la SGV, et contribuer à compenser l'effet croissant de la dérive génétique sur la perte de diversité génétique et la diminution de recrutement de mutations bénéfiques. En outre, la dérive génétique est un processus qui agit comme une contrainte évolutive à l'adaptation. Par exemple, des populations de

plantes, dont l'histoire démographique se compose d'une succession de goulots d'étranglements ('*bottleneck*'), accumulent davantage de mutations non-synonymes délétères, et maintiennent ainsi des QTLs maladaptés (Bustamante *et al.* 2002; Ågren *et al.* 2013). Par conséquent, l'adaptation pour des espèces de faible taille efficace nécessite de nouvelles mutations à large effet, pour échapper à la dérive génétique, ou bien de la variation génétique maintenue grâce au flux génique entre populations.

1.2.2.4 Flux génique

Des populations soumises à des conditions d'habitats différentes ou encore, sous sélection divergente peuvent cependant échanger des gènes *via* la migration d'individus d'une population à une autre. Lorsque les individus migrants se reproduisent dans une population receveuse, le flux génique qui en résulte peut influencer la composition génétique (outre la mutation et la SGV) de la population receveuse et ainsi, impacter la distribution de la taille d'effets de mutations adaptatives. Contrairement aux scénarios théoriques historiques, l'adaptation locale en présence de flux géniques serait favorisée par des locus à effets importants sur les traits phénotypiques adaptatifs (Griswold 2006; Yeaman and Whitlock 2011). En effet, comme observé en génétique des populations, une augmentation du flux génique entre populations tend à homogénéiser les fréquences alléliques entre dèmes. Par conséquent, les mutations à faibles effets sont homogénéisées entre populations et seuls les locus à effets importants sont maintenus dans la population où ils étaient initialement avantageux (Griswold 2006). Ces locus caractérisés par de forts coefficients de sélection perdurent davantage dans le temps et dans leur population d'origine malgré l'existence de flux génique – sous un modèle flux génique-sélection-dérive génétique – et ont plus de chance de contribuer à l'adaptation locale (Yeaman and Whitlock 2011). En général, la condition pour qu'un allèle localement avantageux résiste au remplacement par un autre allèle immigrant est que son coeficient de sélection soit supérieur à la migration ($s>m$) (Lenormand 2002). En présence de flux génique, la taille d'effet des allèles localement adaptés est donc supérieure à celle en l'absence de migration car les allèles à faible coefficient de sélection sont progressivement remplacés par des allèles immigrants, ce qui génère un déplacement de la distribution de la taille d'effet vers les allèles à effets majeurs (Yeaman and Whitlock 2011). Cependant, lorsque le flux

génique est important, que la sélection divergente entre populations est forte et que le taux de mutation est réduit, alors la sélection peut favoriser des groupes d'allèles à faibles effets liés (Yeaman and Whitlock 2011). De tels allèles se comportent ainsi comme un allèle à effet fort et sont maintenus au fil du temps selon l'avantage conféré à la population. Les conséquences du flux de gènes entre populations ont été peu étudiées d'un point de vue empirique car il est complexe de dissocier les effets sélectifs des effets stochastiques dans l'évolution des populations.

Comme nous l'avons vu, l'adaptation locale repose sur l'effet variable d'allèles, d'origines différentes, et qui confèrent un avantage phénotypique à une population dans des conditions d'habitat établies. Bien que fondamentalement importante en biologie de l'évolution, la distribution de la taille d'effet d'allèles adaptatifs dans une population est toujours largement inconnue. Il y a d'une part des évidences d'adaptation locale à partir d'un locus à effet majeur (Colosimo *et al.* 2004; Hoekstra *et al.* 2006; Jones *et al.* 2012), mais d'autre part un nombre grandissant d'évidences d'architectures génomiques d'adaptation locale fondées sur une multitude d'allèles à faibles effets (Berg and Coop 2014a; Yeaman, Hodgins, et al. 2016; Gouy et al. 2017). Ces contrastes, surtout causés initialement par des limites d'identification de la sélection polygénique, semblent s'estomper face à l'acceptation croissante d'un modèle général incluant à la fois des mutations à effet larges et faibles en proportions variables en fonction de l'histoire évolutive. En effet, considérant les variations de biotopes, démographiques et génétiques auxquelles sont soumises les populations naturelles, il est raisonnable d'assumer une grande diversité de distributions des tailles d'effets des allèles et donc d'architectures génomiques de l'adaptation locale selon les organismes et les circonstances.

Il est désormais fondamental, dans le cadre de cette thèse, d'intégrer l'évolution de populations divergentes le long du continuum de la spéciation jusqu'à la mise en place de barrières aux flux géniques.

1.3 La spéciation

L'origine, l'évolution et le maintien de la biodiversité sont des aspects fondamentaux sur lesquels des sujets de recherches se portent, tant pour un intérêt fondamental qu'appliqué. Comprendre le processus qui permet à deux lignées de se diversifier puis de se différencier pour finalement former un complexe de deux espèces, relève de l'étude du concept de spéciation (Felsenstein 1981). La difficulté survient alors rapidement lorsqu'il est nécessaire de définir ce qu'est une espèce. En effet, il existe différentes définitions pour le même terme, mais chacune présentant des caractéristiques distinctes permettant alors d'utiliser le concept d'espèce adéquat selon le système étudié (de Queiroz 1998; Coyne and Orr 2004; de Queiroz 2007). Cependant, dans le cadre de cette thèse, je m'intéresse particulièrement aux bases génétiques de la différenciation phénotypiques adaptative, à leur maintien et leurs conséquences en cas de flux génique entre individus de deux populations divergentes. Par conséquent, nous utiliserons le concept biologique d'espèce (*biological species* concept, BSC), qui définit une espèce comme un ensemble de populations naturelles interfécondes, ou partiellement interfécondes et qui sont reproductivement isolées d'autres groupes similaires (Mayr 1942). Le concept biologique d'espèce tel que défini, implique que la spéciation est le processus évolutif par lequel des barrières – écologiques et/ou génétiques – à la reproduction se mettent en place progressivement au fil du temps. Ainsi, la première question qui suscite l'intérêt des biologistes de l'évolution en spéciation concerne la mise en place et le maintien d'un isolement reproducteur (IR) entre espèces (Coyne and Orr 2004; Marie Curie SPECIATION Network et al. 2012). Ces barrières tendent à accentuer la divergence au cours du processus de spéciation (Turelli *et al.* 2001).

1.3.1 Les barrières à la reproduction

L'isolement reproductif peut être présenté sous une dichotomie selon laquelle il y a possibilité ou non de se reproduire et de former une descendance viable et fertile entre espèces divergentes. En fonction du stade auquel l'isolement reproductif oppère, on parle d'IR pré-zygotique ou post-zygotique.

1.3.1.1 Les barrières pré-zygotiques

Les barrières pré-zygotiques tendent à limiter le flux de gènes entre populations, avant qu'il n'y ait reproduction, accentuant la divergence entre celles-ci (Coyne and Orr 2004). Un choix différentiel d'habitat peut induire une divergence entre populations, notamment lorsque la reproduction est liée à l'habitat, tel qu'observé entre les espèces de timema de Cristina (*Timema cristina*) et de Sierra Nevada (*T. podura*), qui évoluent sur deux hôtes différents (Soria-Carrasco et al. 2014). Une autre barrière, celle-ci temporelle, peut limiter les échanges entre populations pourtant originaires d'une même localité. C'est le cas du saumon bossu (*Onchorhynchus gorbuscha*), anadrome obligatoire (qui se reproduit en rivières et croît en eaux salées), dont les cohortes des années paires ne peuvent se reproduire avec les cohortes des années impaires du fait de leur cycle de vie (Aspinwall 1974). Outre les aspects spatiaux et temporels, il est possible d'observer des comportements de reproduction restreignant les flux géniques entre populations, comme l'homogamie ('*Assortative mating*') tel qu'observé chez l'épinoche à trois épines, par exemple (McKinnon *et al.* 2004). Finalement, lorsque la reproduction est possible, des conflits peuvent survenir avant et/ou pendant l'acte copulatoire entre deux partenaires hétérospécifiques dont les organes reproducteurs évoluant indépendamment peuvent conduire à des incompatibilités par exemple mécaniques (Gagnon and Turgeon 2011; Barnard et al. 2017).

1.3.1.2 Les barrières post-zygotiques

Par ailleurs, la reproduction entre populations divergentes peut être possible, bien que cette divergence génétique puisse affecter la valeur sélective des hybrides, ou affecter leur viabilité. Ces barrières post-zygotiques peuvent être exogènes – elles dépendent de l'écologie – ou endogènes – indépendantes de l'écologie de la population mais dépendantes du fond génétique. La distinction entre des deux types d'isolements reproducteurs post-zygotiques n'est pas toujours évidente et peut former un continuum de degrés de dépendance à l'environnement (Robison *et al.* 2001).

Un isolement reproductif post-zygotique exogène implique que les individus hybrides, issus de populations parentales adaptées à leurs habitats respectifs, voient leur valeur sélective amoindrie. Ce type d'isolement a été illustré dans le système *Heliconius*,

entre les espèces *H. melpomene* et *H. cydno*, où à la fois la survie et le succès reproducteur des hybrides sont réduits. Cette diminution de la valeur sélective a été associée à une contre-sélection naturelle de la part des prédateurs (Merrill et al. 2015) et une contre-sélection sexuelle de la part des lignées parentales, basée sur des patrons de couleur des ailes différents des parents (Jiggins *et al.* 2001; Naisbit *et al.* 2001).

Les conséquences d'un isolement reproductif post-zygotique endogène se traduisent par des phénomènes de stérilité et/ou de non-viabilité chez les hybrides (Coyne and Orr 2004; Gompert et al. 2012). Les mécanismes endogènes associés à la dépression d'hybridation sont causés par des incompatibilités génétiques, des réarrangements chromosomiques, ou encore des niveaux de ploïdie différents entre les lignées parentales (Coyne and Orr 2004). Bien qu'il soit difficile d'estimer la localisation d'une population hybride dans une vallée adaptative, entre deux optimums correspondant aux phénotypes parentaux, Dobzhansky et Muller ont proposé un modèle théorique (Dobzhansky 1937; Muller 1942) qui propose que deux populations qui évoluent sans flux de gène (en allopatrie) peuvent accumuler des mutations indépendamment à des locus épistatiques au fil du temps (Orr and Turelli 2001). Les allèles dérivés apparaissant dans chaque population n'étant dans ce cas pas testés par la selection pour leur niveau de compatibilité, des mutations incompatibles peuvent se fixer neutralement entre les deux populations. En cas de contact secondaire ou d'hybridation entre lignées parentales, les interactions épistatiques négatives qui en résultent induisent une stérilité ou une mortalité des individus hybrides. Au fil des générations hybrides (première génération: F1, secondes générations: F2 et rétrocroisés, etc), l'effet de la recombinaison méiotique va dissocier les combinaisons parentales d'allèles coadaptés, exacerbant l'effet des incompatibilités génomiques. Un cas particulier d'érosion de la valeur sélective chez les hybrides est causée par la stérilité des hybrides de sexe hétérogamétique (XY ou ZW), selon la règle d'Haldane (Haldane 1922). Ce modèle est désormais fondamental dans l'étude de la spéciation. De nombreuses études ont cherché à identifier des gènes impliqués dans les incompatibilités de Dobzhansky-Muller, tel que chez un copépode marin (Edmands 1999; Ellison and Burton 2008), chez la drosophile (Barbash et al. 2003) ou encore chez la souris (Mihola *et al.* 2009).

Des variations de caryotype entre deux populations divergentes, causées par des réarrangements chromosomiques (inversions, fusions, fissions), peuvent induire une

barrière post-zygotique endogène. Par exemple, les inversions chromosomiques maintiennent un déséquilibre de liaison entre gènes compris au sein de l'inversion, en limitant la recombinaison de l'intégralité de l'inversion (Noor et al. 2001; Rieseberg 2001). Ainsi, les individus hétérozygotes (hybrides) auront une réduction locale du taux de recombinaison augmentant la divergence entre populations (Faria and Navarro 2010). De plus, la présence de gènes barrières dans l'inversion tendra à réduire localement le flux génique entre populations. Par conséquent, ces réarrangements chromosomiques impliqués dans la spéciation, amplifient la divergence et maintiennent l'isolement reproducteur entre deux espèces (lorsqu'une inversion est présente uniquement chez une des deux espèces, par exemple). Ce modèle a notamment été illustré chez la drosophile (Noor et al. 2001). Outre les inversions chromosomiques, il est possible d'observer un nombre de copies de chromosomes variable. En effet, il est fréquemment documenté chez les plantes, des cas de lignées stériles triploïdes, issues du croisement entre une lignée parentale diploïde et l'autre tétraploïde (Ramsey and Schemske 1998; Wood et al. 2009), indiquant le maintien de l'isolement reproducteur entre les deux lignées parentales. Le niveau intermédiaire de copies de chromosomes induit des problèmes de ségrégation chromosomique chez les hybrides (Coyne and Orr 2004). Finalement, les barrières post-zygotiques – essentiellement endogènes – sont vraisemblablement irréversibles par rapport aux barrières pré-zygotiques lorsque le flux génique est possible (Muller 1942). Ainsi, l'isolement reproductif peut résulter de l'interaction entre différentes barrières à la reproduction, dont l'établissemment peut notamment dépendre du contexte spatial, écologique et temporel de la divergence.

1.3.2 Contextes spatiaux-temporels de la spéciation

Le processus de spéciation se complète dans le temps, durant lequel les populations divergentes évoluent, mais aussi dans l'espace où des processus démographiques et sélectifs affectent les niveaux de divergence inter-populationnelle. Ainsi, la spéciation est un processus multidimensionnel (Abbott et al. 2016; Mérot et al. 2017). Historiquement, l'étude de la spéciation était liée à l'étude de la phylogéographie, soit l'étude des processus historiques ayant façonné la distribution actuelle de la biodiversité (Avise 2000; Hewitt 2001). Ainsi, les barrières exogènes, associées à la

géographie, qui tendaient à limiter les échanges géniques étaient considérée comme la source majeure d'évènements de spéciation (Mayr 1954). Par conséquent, l'emphase a été mise sur les niveaux d'isolement géographique, allant d'un isolement entre populations complet (allopatrie), partiel (parapatrie) ou inexistant (sympatrie) (Coyne and Orr 2004). En considérant deux populations issues d'une lignée ancestrale commune, l'allopatrie va induire une évolution indépendante de chacune des populations qui divergeront au fil du temps, soit sous l'effet de pressions de sélection locale, soit par stochasticité. Ainsi, il n'est pas rare d'observer des espèces sœurs dont la répartition géographique décrit des barrières naturelles. C'est le cas de la tique bleue asiatique, parasite du bétail, qui comprend deux espèces sœurs ayant divergées en allopatrie (l'une en Australie et l'autre sur les continents Américains et Africain) et dont les hybrides (en conditions expérimentales) sont stériles (Labruna *et al.* 2009). D'autres cas, dans des contextes d'îles océaniques jeunes, indiquent que la colonisation de nouveaux habitats peut conduire à l'établissement et la diversification de la population insulaire par rapport à la population source continentale. C'est notamment le cas d'espèces d'orchidées épiphytes de l'ile de La Réunion (Mallet *et al.* 2014). De tels cas mettent en avant le changement d'aires de répartitions contraintes par les variations de conditions environnementales. En effet, durant les multiples périodes glaciaires du Pléistocène, des populations d'espèces se sont retrouvées isolées dans des refuges glaciaires, résultant en des périodes d'allopatrie. En Amérique du Nord, la région des Grands Lacs constitue une zone de suture entre lignées glaciaires ayant recolonisé les habitats à nouveau disponibles à la suite du dernier retrait glaciaire (*i.e.*, contact secondaire entre lignées glaciaires initialement de la même espèce), où l'on observe désormais des contacts entre paires d'espèces sœurs (April *et al.* 2013).

La spéciation sympatrique correspond à la mise en place d'un isolement reproductif au sein d'une même population. La spéciation parapatrique consiste en la mise en place d'un isolement reproductif entre deux populations dont les répartitions géographiques se chevauchent. Dans ces cas, les échanges génétiques entre populations divergentes sont donc possibles et diminuent cette divergence. Par conséquent, un mécanisme de mise en place d'isolement reproducteur nécessite l'évolution d'adaptations à des habitats différents – différentes niches écologiques – sous la contrainte de sélection divergente, induisant la mise en place de déséquilibres de liaisons entre les allèles/gènes de l'adaptation locale à l'habitat (impliqués dans l'isolement pré-zygotique puis post-zygotique; (Nosil 2012),

(Yeaman and Whitlock 2011). De telles architectures seront brisées dès les évènements de recombinaison entre appariement d'individus divergents (Felsenstein 1981). Cependant, il est possible de maintenir ces organisations géniques afin de conserver les génotypes tendant vers un optimum phénotypique dans un habitat donné. En effet, différents modèles suggèrent que si les allèles impliqués dans l'homogamie ou le choix d'habitat sont liées à ceux impliqués dans l'adaptation locale, alors le polymorphisme peut être maintenu (Felsenstein 1981; Kirkpatrick and Ravigné 2002). L'évolution entre de tels traits phénotypiques (choix du partenaire et performance dans l'habitat) peut également se produire si les bases génétiques sont les mêmes. Ces traits pléiotropiques identifiés comme « traits magiques » ('*magic traits*') (Servedio *et al.* 2011; Haller *et al.* 2012), limitant les flux géniques entre populations, peuvent mener à la mise en place d'un isolement reproducteur. C'est le cas chez des populations naturelles en conditions expérimentales d'un parasite de pigeon (Villa *et al.* 2018) et chez l'épinoche à trois épines (Nagel and Schluter 1998; Conte and Schluter 2013).

Les exemples empiriques de spéciation sympatrique sont relativement rares, en comparaison aux démonstrations théoriques, mais persistent dans la littérature car souvent les modèles de spéciation alternatifs ne sont pas explorés et donc ne peuvent être rejetés. C'est souvent le cas dans ces systèmes où les effets démographiques historiques sont négligés, par exemple chez l'épinoche à trois épines (Taylor and Donald McPhail 2000), chez les cichlidés (Meier, Sousa, et al. 2017), ou un autre exemple classique de divergence sympatrique chez la mouche de la pomme (*Rhagoletis pomonella*) qui semble impliquer des évènements répétés d'introgression à partir d'une tierce population, favorisant la divergence locale (Feder *et al.* 2005). Cependant certains auteurs minimisent les rôles de contacts secondaires, et donc de phases allopatriques, identifiées *a posteriori* dans des systèmes décrits comme étant le résultat de spéciation sympatrique (Richards *et al.* 2017). Finalement, les variations cycliques climatiques et de distributions d'habitat dans le temps, induisent des modes de divergence entre populations qui varient entre la sympatrie exclusive et l'allopatrie complète, influençant notamment les flux géniques (Jiggins 2006). Cela sous-entend que l'alternance entre phase d'allopatrie et de sympatrie soit possible et puisse accélérer le processus de spéciation par des mécanismes de renforcement par exemple (Smadja and Butlin 2011). Les contextes spatiaux-temporels de la spéciation sont régis par l'interaction entre les forces sélectives et démographiques. Ainsi, au fil du temps

les processus démographiques, sélectifs et la contingence historique vont modeler le patrimoine génétique des populations divergentes et donc, l'architecture génomique des espèces naissantes.

1.3.3 Architecture génomique de la spéciation

1.3.3.1 Gène(s) de spéciation

D'un point de vue génétique, l'isolement reproductif entre espèces divergentes peut être causé par des gènes dits « de spéciation » (Wu 2001). Ce type de gène est défini comme un gène contribuant à augmenter le niveau d'isolement reproducteur entre populations divergentes (Nosil and Schluter 2011). L'identification de tels gènes est moins difficile dans les stades précoces de la spéciation car leur nombre est plus limité. Malheureusement, les systèmes modèles de spéciation semble comparer des paires d'espèces dont la divergence est importante (*e.g.*, *Rhagoletis pomonella*). De plus, des gènes de spéciation peuvent être associés à n'importe quelles barrières à la reproduction – pré- ou post-zygotique – bien qu'ils soient essentiellement documentés dans le cadre de dysfonctions hybrides (Presgraves 2010; Maheshwari and Barbash 2011). C'est le cas notamment d'incompatibilités cyto-nucléaires (Dion-Côté and Barbash 2017), telles qu'observées chez les plantes (Fishman and Willis 2006), chez les poissons (Gagnaire *et al.* 2012) ou chez les insectes (Gibson *et al.* 2010). Les conséquences de ces incompatibilités génétiques sur la dépression des génotypes hybrides peuvent être mesurées. Les causes de divergence entre gènes de spéciation sont relativement peu connues, mais souvent associées à l'adaptation locale avec des effets majeurs sur le phénotype. Cependant, certains systèmes ont permis de souligner l'effet neutre de certaines caractéristiques génomiques impliquées dans des incompatibilités hybrides (Presgraves 2010). Par conséquent, outre l'effet majeur d'un gène de spéciation initialement impliqué dans l'adaptation locale (Colosimo *et al.* 2005; Hoekstra *et al.* 2006), il est possible que le nombre de gènes de spéciation impliqués soit important, avec des effets faibles sur le phénotype (Marie Curie SPECIATION Network et al. 2012). Certains travaux se sont intéressé à identifier des bases polygéniques de la spéciation en ciblant des gènes d'intérêt, souvent en restreignant leurs approches à l'analyse de QTLs, tout en mesurant la taille des effets sur l'isolement reproducteur (Smadja and Butlin 2009; Ellison et al. 2011),

alors que d'autres études tendent à identifier le caractère polygénique de la divergence pouvant mener à la spéciation, comme chez l'épinoche à trois épines (Bay *et al.* 2017).

L'isolement reproducteur peut se construire dans des régions génomiques associées aux gènes de la spéciation – initialement impliqués dans l'adaptation locale – (Turner *et al.* 2005; Servedio *et al.* 2011). Au fil du temps, de tels gènes positivement sélectionnés vont se retrouver en déséquilibre de liaison (LD) avec d'autres gènes du même type. Ces blocs d'allèles divergents augmentent en taille et en nombre dans l'ensemble du génome par réduction locale du taux de recombinaison entre populations divergentes. Par conséquent, le LD est maintenu et amplifié par la sélection naturelle qui tend à renforcer les barrières pré-zygotiques (Bolnick and Fitzpatrick 2007), ou bien en associant des gènes indépendants (non-liés initialement) impliqués dans les barrières pré- et post-zygotiques (Barton and de Cara 2009). Avec l'augmentation du LD, l'efficacité de la sélection pour maintenir les barrières à la reproduction augmente (Via 2009; Via 2012). En effet, le coefficient de sélection de chaque locus est alors associé aux coefficients de l'ensemble des locus sous sélection, ainsi que celui des locus périphériques à l'allèle initialement impliqué dans l'adaptation locale (Barton 1983; Kruuk *et al.* 1999). Ces coefficients de sélection localement supérieurs, augmentent alors la barrière aux flux géniques localement au sein du génome (Barton and Bengtsson 1986).

1.3.3.2 Des barrières perméables

D'un point de vue génomique, le concept d'espèce biologique de Ernst Mayr suppose que le nombre de locus impliqués dans la différenciation entre espèces est important, et que l'intégralité du génome évolue avec cohésion en tant qu'unité dans la divergence en absence flux de gènes. Ainsi, dans un contexte de divergence en présence de flux génique, la recombinaison détruirait cette unité fonctionnelle (Felsenstein 1981; Wu 2001). Par exemple, si deux populations divergentes entrent en contact et échangent leurs gènes à un taux élevé, alors les allèles de l'adaptation locale recombinent entre eux, et les fréquences alléliques tendent à s'homogénéiser. Ainsi, les deux populations n'en sont en réalité qu'un seule (Wu 2001). Cependant, des groupes de gènes différentiellement adaptés entre populations et impliqués dans l'IR de ces espèces naissante, persisteront uniquement dans l'habitat dans lequel leur valeur sélective est favorable (Wu 2001). Cela signifie que les allèles des gènes de spéciation seront contre-sélectionnés et purgés de la

population de laquelle ils ne sont pas originaires. En revanche, les allèles neutres seront échangés par flux de gènes et dissociés des gènes de IR par recombinaison génétique – destruction du LD mis en place par sélection naturelle – selon leur distance à l'allèle initialement impliqué dans l'adaptation locale (Felsenstein 1981; Barton and Bengtsson 1986; Wu 2001). Cette interaction entre processus évolutifs et divergence reflète la semi-perméabilité des barrières aux flux de gènes (Bengtsson 1994), et se traduit par un degré variable d'IR selon les régions génomiques considérées.

1.3.3.3 L'émergence d'îlots de divergence

Des prédictions théoriques à l'avènement des technologies de séquençages et d'assemblages de génomes, les études portant sur la génomique de la spéciation ont permis de décrire des paysages de différenciation génomiques hétérogènes (Ellegren et al. 2012; Gagnaire, Pavey, et al. 2013; Martin et al. 2013; Burri et al. 2015; Burri 2017a), en accord avec la notion de perméabilité des barrières (Harrison and Larson 2014; Harrison and Larson 2016a). Ainsi, certaines régions montrent un niveau de différenciation variable par rapport au niveau génomique moyen. Certaines régions hautement différenciées par rapport à la moyenne, sont décrites comme étant des îlots de différenciation (Harr 2006). Outre le niveau de différenciation entre populations divergentes, certaines études ont mis en évidence un nombre et des tailles variables d'îlots de différenciation. Certaines paires de comparaisons ont ainsi révélé qu'un nombre limité de régions de taille importante étaient impliquées dans la différenciation génomique (Turner et al. 2005; Jones et al. 2012; Martin et al. 2013), alors que d'autres ont identifié de nombreuses régions de tailles réduites le long du génome (Renaut et al. 2013; Soria-Carrasco et al. 2014; Burri et al. 2015; Burri 2017a).

De tels patrons de divergence, dont le niveau de différenciation est variable, la taille et le nombre de régions impliquées ont mené à l'élaboration du concept de '*divergence hitchhiking*' (DH) (Via 2012). En théorie, l'effet d'une barrière génétique permet la divergence locale en présence de flux génique, par action combinée de la sélection divergente et de l'homogamie réduisant alors le taux de recombinaison, entre populations divergentes, générant ainsi des îlots de différenciation. Ce processus de réduction locale de la migration efficace (ou de la recombinaison efficace, concept equivalent) est progressif dans le temps car il dépend de l'accumulation du nombre de locus impliqués dans l'IR et

l'évolution du déséquilibre de liaison. En effet, le stade initial correspondant à la différenciation génétique des locus initialement impliqués dans l'adaptation de chacune des populations divergentes. Puis, la réduction du taux de recombinaison inter-populations autour des allèles localement adaptés induit une augmentation de la fréquence des variants neutres aux sites liés. Finalement, cette perte de diversité locale entraine une hausse de la différenciation génétique jusqu'à obtenir un niveau de différenciation globale élevé le long du génome suite à la mise en place de nouveaux locus d'IR (Feder, Egan, and Nosil 2012a), s'illustrant comme l'immersion des îlots génomiques dans la différenciation globale. Par ailleurs, l'émergence d'îlots de différenciation peut également se former par l'action de la sélection en liaison dans les régions à faible taux de recombinaison, indépendemment de tout effet d'IR (Burri 2017b).

L'identification des îlots génomiques comme étant des îlots de « spéciation » en se basant uniquement sur la différenciation génétique (*Fst*) est controversée. L'indice génétique de *Fst* est un indice de divergence relative, qui est dépendent du niveau de diversité génétique local (Charlesworth et al. 1997) et donc du niveau de recombinaison (Noor and Bennett 2010). Tel qu'attendu, des îlots de différenciation ont été observés dans des régions à faible recombinaison dans plusieurs systèmes empiriques (Roesti et al. 2012; Renaut et al. 2013; Burri et al. 2015; Burri 2017a; Duranton et al. 2018).

Des estimations de divergence absolue ont été proposées, notamment en estimant la distance évolutive à partir de comparaisons de séquences entre populations divergentes (*Dxy*) (Nei 1972). Le *Dxy* indique le nombre moyen de différences observées par site entre deux haplotypes aléatoires provenant de deux lignées différentes (Nei 1972). Des régions divergentes à *Dxy* élevé, seront définies comme des îlots de divergence (Nosil et al. 2009; Burri 2017b). Contrairement à la différenciation génétique, la divergence absolue est essentiellement dépendante du niveau de polymorphisme de la population ancestrale et du niveau de polymorphisme ancestral contenu initialement dans chacune des populations divergentes (Guerrero and Hahn 2017; Burri 2017b). Cependant, le paysage génomique de divergence absolue est également modelé par l'hétérogénéité des taux de migration et de recombinaison le long du génome, ainsi que par les processus sélectifs, tels que la sélection aux sites liés ('*linked selection*') (Burri 2017b). Ainsi, les indices de divergence absolue (*Dxy*) et relative (*Fst*) peuvent être corrélés. C'est le cas dans un contexte de semiperméabilité du génome au flux génique, qui tend à réduire la divergence absolue et

relative entre espèces en cours de spéciation, générant une hétérogénéité de divergence entre locus dont le niveau de flux de gènes efficace est hétérogène (Nadeau et al. 2012; Cruickshank and Hahn 2014). Au contraire, une corrélation négative entre ces deux types de divergence peut se produire dans les régions génomiques sous sélection aux sites liés ou dans les régions à faible recombinaison. De telles régions montreront une faible divergence absolue, ainsi qu'une diminution de la diversité génétique intra-populationnelle induisant alors une hausse de la divergence relative (Cruickshank and Hahn 2014; Burri 2017a; Burri 2017b). Cependant, il est aussi possible de voir des patrons plus complexes de divergence absolue. C'est le cas lorsque du polymorphisme ancestral partagé est maintenu en tant que polymorphisme balancé entre les lignées divergentes (Guerrero and Hahn 2017). Par conséquent, les études sur la nature des îlots de différenciation tendent à combiner ces approches complémentaires. Bien que les modèles discutés précédemment impliquent la présence ou l'absence de flux géniques, ils sont essentiellement basés sur un socle de divergence primaire.

Initialement, l'observation d'îlots de différenciation était associée à l'identification de régions génomiques contenant des gènes impliqués dans l'isolement reproducteur (Turner et al. 2005). Comme nous l'avons vu, les îlots génomiques peuvent être façonnés par différents processus sélectifs et/ou démographiques, et leurs tailles sont variables selon le stade de divergence et le niveau de spéciation entre espèces. Cependant, dans un contexte de spéciation avec flux de gènes tel qu'un contact secondaire entre deux populations initialement en allopatrie, le paysage génomique de différenciation et de divergence peut être modifié au cours du temps dans une dynamique opposée à celle habituellement considérée dans les modèles de divergence primaire. En effet, le caractère semi-perméable des barrières génomiques à la reproduction induit un flux de gènes hétérogène entre les locus de spéciation et les locus neutres en fonction de leur degrès de liaison. Puisque le flux génique tend à homogénéiser les fréquences alléliques des allèles neutres, il diminuera également le niveau de différenciation dans ces régions dont le coefficient de sélection est moindre par rapport à l'intensité de la migration. Par conséquent, ces flux géniques éroderont progressivement les îlots de divergence (Tine et al. 2014; Rougeux et al. 2017), complexifiant la distinction entre un mode de divergence primaire, d'un mode de divergence secondaire, puisque les deux méchanismes peuvent aboutir à des partons de divergence simillaires.

Finalement, l'architecture génétique de la spéciation indique que de nombreux processus évolutifs interdépendants peuvent modeler le paysage génomique. L'étude de la spéciation en biologie évolutive est une tâche complexe qui nécessite de déterminer les contextes démographiques historiques afin de dissocier les effets confondants de la démographie, des effets stochastiques et sélectifs. En complément aux approches génomiques, des approches comparatives dans des systèmes d'évolution parallèles permettent de dissocier les effets démographiques des processus sélectifs et de localiser les bases génétiques associées à l'évolution parallèle dans chacune des populations.

1.4 Associations génotype-phénotype

Pour étudier l'évolution de la biodiversité, il est nécessaire de comprendre l'adaptation ainsi que la diversité génétique et phénotypique intra-spécifique associée, puis la spéciation. L'association génotype-phénotype est donc cruciale pour établir le lien entre les processus globaux (*i.e.*, adaptation, divergence et isolement reproducteur) associés à des variations de traits phénotypiques, et leurs bases génétiques (Feder and Mitchell-Olds 2003). L'identification des bases génétiques de traits divergent repose sur des méthodes de génétiques quantitatives et de détection de la variation génétique sous-jacente à un trait phénotypique. Il s'agit ici de l'identification de QTLs, dont la valeur sélective influence l'atteinte de l'optimum phénotypique.

1.4.1 Le potentiel évolutif d'une population

Parmi la variation génétique d'une population, les QTLs constituent la partie détectable de la variation. Cependant, certains effets alléliques ne sont pas mesurables soit car il s'agit de mutations neutres, de substitutions dans des régions non-codantes, soit parce qu'ils ne sont pas dans des conditions génétiques ou environnementales nécessaires pour que leur effet soit visible. On parle alors de variation génétique cryptique (Hermisson and Wagner 2004). Généralement, la plupart des variants génétiques n'affectent pas de traits phénotypiques, ou affectent un trait qui n'est pas sous une forte pression sélective (Le Rouzic and Carlborg 2008). De plus, il est reconnu que la majorité des mutations sont de nature délétère et seront donc purgées de la population si leur effet est suffisamment

délétère dans un contexte démographique donné ($Ne*s>1$). Par conséquent, seulement une partie de la diversité génétique d'une population a le potentiel de contribuer à l'adaptation locale. Il est couramment admis que le potentiel évolutif d'une population est corrélé à la diversité génétique de cette population, ou du moins, que la diversité génétique est requise pour qu'une population évolue (Le Rouzic and Carlborg 2008). Ainsi, le polymorphisme génétique, maintenu dans une population, fournit un socle évolutif à cette population. Ce socle évolutif est composé d'une importante hétérogénéité de la taille d'effet de chacun des variants dans les conditions environnementales de l'habitat, et est sensible aux caractéristiques génétiques d'arrière plan. Finalement, le polymorphisme génétique, variable le long du génome, apporte une source de variance dans les traits phénotypiques associés (Stranger et al. 2007).

1.4.2 Les cibles de la sélection

L'identification de QTLs et la compréhension de l'architecture génétique associée à un trait phénotypique complexe nécessitent de déterminer la localisation précise des variants associés à des variations phénotypiques. Il a été démontré que le polymorphisme de séquence dans les régions régulatrices et codantes peut modeler le niveau d'expression génique, cette variation phénotypique permettant alors d'alimenter la sélection naturelle (Wray 2007; Pai et al. 2015). Cette variation transcriptionnelle intra-population peut être initiée par du polymorphisme dans la séquence codante (CDS) du gène et/ou dans la région non-traduite (UTR) du gène, qui contient la machinerie de régulation de l'expression du gène auquel elle est rattachée (Barrett et al. 2012). Les CDS font l'objet d'études afin d'inférer de la sélection positive – en déterminant des ratios de mutations non-synonymes (Dn) par rapport à des mutations synonymes (Ds) – (McDonald and Kreitman 1991). Cependant, la théorie prédit que les régions géniques codantes sont essentiellement sous sélection stabilisatrice afin de préserver la fonction protéique (Signor and Nuzhdin 2018). Ainsi, l'accumulation de variants à forts effets dans la CDS des gènes est limitée (Glassberg et al. 2018). De manière alternative et complémentaire, dans les régions promotrices UTRs d'un gène, des variants liés (*–cis*) peuvent induire une hausse de la variance transcriptionnelle intra-population pour le gène considéré (Fay and Wittkopp 2007; Glassberg et al. 2018; Signor and Nuzhdin 2018). Cependant, la théorie autour de

l'accumulation de variants à effets de tailles variables peut être transposée aux régions UTRs. En effet, bien que les variants à forts effets permettent une importante variance d'expression génique et donc une probabilité accrue d'être soumise à la sélection, les variants à faibles effets en *–cis* sont favorisés car ils n'éloignent pas la population de l'optimum phénotypique – niveau d'expression génique – et permettent tout de même des innovations phénotypiques (Glassberg et al. 2018).

Certains systèmes ont révélé une relation directe entre la taille de l'effet d'un variant en *–cis* et la fréquence de ce variant dans la population considérée (Simons et al. 2018). De plus, d'autres systèmes ont mis en évidence une corrélation positive entre la fréquence d'un variant et la taille de l'effet sur l'expression génique associée au QTL (eQTL) (Battle et al. 2014; Tung et al. 2014). Par ailleurs, les variants en *–cis* tendent à diminuer l'effet pléiotrope du gène associé, réduisant ainsi les contraintes évolutives si ce gène agit sur des traits phénotypiques anti-corrélés (Signor and Nuzhdin 2018). Par conséquent les mutations affectant un gène en *–cis* sont un bon substrat génétique pour conférer un potentiel évolutif à une population.

Outre les variations de séquences nucléotidiques, d'autre mécanismes moléculaires tels que l'épigénétique sont à même d'influencer le phénotype d'individus. En effet, l'épigénétique réfère à des changements de fonctions de gènes qui ne sont pas attribués à des changements dans la séquences des gènes (Youngson and Whitelaw 2008). Les modifications épigénétiques correspondent essentiellement à la méthylation de l'ADN, à des modifications d'histones et à la régulation des petits ARN (siRNA) (Duncan et al. 2014). La méthylation de l'ADN implique une modification biochimique de l'ADN par ajout d'un groupement méthyl à une cytosine au sein de dinucléotides CpG chez les animaux (Jones 2012). Une telle modification peut engendrer un changement de conformation de la molécule d'ADN résultant alors en un ADN condensé lorsque la molécule est méthylée. Au contraire, l'ADN est décondensé en absence de méthylation. Cela contribue à former l'euchromatine (ADN décondensé) et l'hétérochromatine (ADN condensé). De telles conformations de la séquence d'ADN sont notamment asosciées à la modulation de l'expression de gènes. Par exemple, la méthylation de la région promotrice d'un gène est souvent associée à l'expression du gène auquel elle appartient. Ainsi, une hyperméthylation, par rapport au niveu de méthylation de référence, est associé à une répression de l'expression génique, alors qu'une hypométhylation est associée à une sur-

expression du gène. Au contraire, la méthylation de la séquence du corps du gène peut induire des épissages alternatifs du gène, ainsi que l'activation de l'expression du gène (Jones 2012). Cela est d'importance primordiale, pour assurer la régulation de gènes du développement ou encore, l'intégrité du génome en réprimants les éléments transposables (TEs), riches en CpGs, dans les séquences géniques (Rey et al. 2016). Ces variations de méthylation sont stables dans le temps et au fil des générations, on parle de méthylation constitutive. L'origine de la variation épigénétique est multiple. Elle peut être obligatoire, dépendante de bases génétiques; facilitée, directement ou indirectement associée à un génotype; pure, typiquement générée par des processus stochastiques et dépourvue de bases génétiques (Richards 2006). Les deux extrêmes (*i.e.*, variation épigénétique obligatoire et pure) peuvent avoir d'importantes conséquences sur la plasticité phénotypiques – capacité d'un génotype à produire différents phénotypes selon les conditions environnementales – et par conséquent sur l'évolution de populations. Les variations épigénétiques obligatoires sont essentiellement liées à la régulation de TEs qui, suite à des changements de conditions environnementales, peuvent générer un polymorphisme d'insertion de TEs pouvant alors affecter l'expression d'un gène (Rey et al. 2016). En outre, la théorie prédit que les variants épigénétiques (épimarques) pure peuvent permettre une exploration plus rapide du paysage adaptatif dans des conditions environnementales changeantes, du fait du taux de mutations supérieurs pour les épimarques, et par conséquent contribuer à générer des phénotypes adaptatifs si la sélection agit sur les gènes affectés (Klironomos et al. 2013). Bien qu'une étude portant sur *Arabidospsis thaliana* a mis en évidence une contribution de variants épigénétiques limitée car les épimarques étaient majoritairement obligatoires, aucune étude empirique n'a pu mettre en évidence la dynamique de variants épigénétiques pure au sein de populations dans des environnements changeants. Par ailleurs, il existe plusieurs cas de variations épigénétiques obligatoires, tels que l'effet de la variation de méthylation sur la variation d'expression génique en corrigeant pour les effets -*cis* chez *A. thaliana* (Meng et al. 2016), ou chez l'humain (Zaghlool et al. 2016).

Finalement, il est admis que les variations génétiques, et indirectement épigénétiques, en -*cis* peuvent être associées à des variations phénotypiques menant à l'adaptation locale ou encore, la divergence entre populations pouvant aboutir à la spéciation (Johnson and Porter 2000; Wray 2003; Fay and Wittkopp 2007; Manceau et al.

2011; Hosaka and Kakutani 2018). Les bases génétiques associées à des variations phénotypiques peuvent notamment être identifiées avec robustesse lorsqu'elles sont décrites dans des systèmes indépendants montrant des variations phénotypiques similaires.

1.5 Évolution parallèle et convergente

1.5.1 Évolution parallèle

L'évolution parallèle se définit comme l'évolution indépendante d'un même trait phénotypique dans des populations ou des espèces différentes (Elmer and Meyer 2011). De plus, lorsque l'évolution parallèle est causée par des modifications similaires de l'environnement, elle est considérée comme une résultante de la sélection naturelle (Schluter et al. 2004; Orr 2005a; Losos 2011). L'évolution indépendante de phénotypes similaires a été illustrée par plusieurs systèmes (Pigeon, Chouinard, and Bernatchez 1997a; Rundle et al. 2000; Anon 2014; Comeault et al. 2016). Certains de ces modèles tels que l'épinoche à trois épines montrent des réplicats indépendants de paires de populations divergentes (Rundle et al. 2000). De tels systèmes sont alors tout à fait pertinents pour élucider les modes de divergence entre paires d'espèces, inférer les processus évolutifs en jeu et identifier les bases génomiques de la différenciation phénotypique parallèle. De façon générale, il existe plusieurs modes génotypiques associés au parallélisme de la différenciation phénotypique entre paires de populations. En effet, la source génétique de la divergence entre deux populations peut survenir d'une mutation *de novo* indépendante au même locus (Pearce et al. 2009; Manceau et al. 2011), peut avoir une origine unique et être diffusée par flux géniques entre les populations divergentes de chacune des paires de populations (Schluter and Conte 2009; Abbott et al. 2013; Bierne et al. 2013), ou finalement être recruté de manière répétée entre les populations à partir du polymorphisme ancestral préexistant (comme vu précédemment) (Colosimo et al. 2005; Yeaman, Hodgins, et al. 2016; Rougeux et al. 2018).

L'évolution parallèle peut, par définition, être étendue à différents niveaux biologiques. Considérant le niveau d'expression génique comme un phénotype, les études comparatives de transcriptomique entre réplicats de paires de populations divergentes

permettent d'identifier les changements parallèles d'expression associés au parallélisme de traits. Ceci est bien documenté dans des systèmes d'évolution expérimentale soumettant des réplicats de populations à des pressions sélectives identiques (Cooper et al. 2003), ainsi que dans des systèmes de populations naturelles (Zhao et al. 2015; Zhao and Begun 2017). Ces travaux ont notamment mis en évidence que le niveau de parallélisme transcriptomique est commun dans les contextes de parallélisme phénotypique, et que le taux de parallélisme est corrélé négativement au temps de divergence entre espèces (Losos 2011). Cependant, peu d'études ont étudié l'étendue des changements d'expression génique parallèle, contribuant à la divergence au sein de chacune des paires de populations, entre réplicats indépendants de paires de populations (McGirr and Martin 2018). Ceci résulte du fait que peu de systèmes offrent des réplicats indépendants de différenciation phénotypique. Bien que les études se soient cumulées au fil du temps pour déterminer les facteurs influençant la probabilité d'une évolution parallèle (Conte, Arnegard, Peichel, and Schluter 2012a) 2012) peu d'aspects théoriques associent le niveau de variation génétique parallèle avec le degré de changements transcriptomiques parallèles dans un contexte de spéciation (Schluter et al. 2004; Pavey et al. 2010). Finalement, outre l'identification des bases génomiques impliquées dans l'adaptation et/ou la divergence parallèle de populations, un champ de recherche croissant consiste à identifier les bases génétiques (*i.e.*, gènes) qui sont impliquées dans l'adaptation d'espèces de différentes distances phylogénétiques en réponse à des conditions similaires.

1.5.2 Répétabilité de l'adaptation

La carte du génotype au phénotype est riche en itinéraires alternatifs pour lier un phénotype à des bases génétiques (*e.g.*, pléiotropie, polygénie). En effet, les routes liant le génotype au phénotype optimal peuvent être uniques ou multiples. Cela nécessite de comprendre non seulement l'architecture génétique associée aux phénotypes sous pression sélective, mais aussi le potentiel d'évolution d'organismes complexes, au sein de populations soumises à des conditions d'habitats variables (Wagner and Zhang 2011). Or, ces conditions peuvent être observées à différentes localités et affecter différentes populations. Il est alors nécessaire de s'interroger sur les conséquences de cette pression de sélection identique, sur des populations indépendantes. Notamment, la variabilité

phénotypique intra-populationnelle sera-t-elle suffisante pour pouvoir répondre à une pression de sélection, et quelles sont les bases génétiques sous-jacentes? L'adaptation se fera-t-elle par action de la sélection sur le même gène dans chacune des populations, ou sur des routes génétiques différentes menant au même phénotype? Selon les traits phénotypiques étudiés et leur complexité, la sélection naturelle peut être contrainte par les génotypes associés. C'est le cas lorsque la sélection naturelle répond en agissant sur des gènes hautement contraints et donc peu variables (*e.g.*, gènes du développement), ce qui tend à augmenter la probabilité de répétabilité à l'adaptation. Alternativement, la sélection naturelle est contrainte par la valeur sélective du trait phénotypique qui elle-même limite le nombre de génotypes capables de mener au phénotype optimal (Chevin 2013; Yeaman et al. 2018). Dans un tel contexte d'association des bases génétiques à la valeur sélective du trait phénotypique considéré pour une pression de sélection identique, la notion de parallélisme adaptatif (discuté précédemment) ou de convergence adaptative (*i.e.*, cas de parallélisme entre lignées distantes, ou cas dont les bases moléculaires de l'adaptation locale sont différentes pour un même phénotype) (Elmer and Meyer 2011), est directement sous-entendue. Ainsi, les routes viables de la répétabilité à l'adaptation peuvent être élucidées par des approches de génomique comparatives de différentes populations pour lesquelles un parallélisme phénotypique (en réponse à une pression de sélection similaire) est observé. Plusieurs études ont permis d'identifier les bases génétiques (*i.e.*, gènes, polymorphisme) soumis à la sélection et donc au moins partiellement impliquées dans la convergence adaptative à des conditions environnementales telles que la température chez des conifères (Yeaman, Hodgins, et al. 2016), à l'altitude chez des populations humaines (Gouy et al. 2017) ou à des niches écologiques particulières chez des poissons (Rougeux et al. 2018). Outre l'identification des bases génétiques, des modèles théoriques permettent de déterminer la probabilité qu'un gène soit impliqué dans la convergence adaptative, afin d'étayer les observations en nature et de s'affranchir de l'identification de faux positifs (Conte, Arnegard, Peichel, and Schluter 2012a; Bailey et al. 2015; Yeaman et al. 2018). Le développement récent de tels modèles devrait améliorer les études génomiques de l'architecture de l'adaptation, notamment dans des contextes d'approches comparatives en amplifiant la robustesse des inférences réalisées. Dans de telles conditions, il est tout à fait pertinent de comparer les bases génétiques de systèmes exposant du parallélisme de différenciation phénotypique entre réplicats indépendants de populations divergentes mais

aussi d'y coupler la probabilité qu'un gène donné soit associé à la répétabilité de la divergence.

1.6 Deux espèces soeurs dans le genre *Coregonus*

1.6.1 Le Grand Corégone

Le Grand Corégone (*Coregonus clupeaformis*) est un poisson lacustre appartenant à la famille des Salmonidés, dont l'aire de répartition s'étend en Amérique du Nord. Le système à l'étude se localise dans le bassin versant de la Rivière Saint-Jean, à cheval entre l'État du Maine et le sud de la province de Québec. Cette région est particulièrement intéressante car on y retrouve deux formes de Grand Corégone. En effet, cette région en particulier correspond à une zone de suture entre deux lignées glaciaires. Il y a environ 60,000 ans, durant la période glaciaire, l'avancée des glaciers a isolé des lignées dans des refuges glaciaires (Bernatchez and Dodson 1990; Bernatchez and Dodson 1991; Jacobsen et al. 2012). Durant cette période d'allopatrie, chacune des lignées glaciaires a alors pu accumuler des incompatibilités génomiques. Suite au retrait de la calotte glacière, il y a environ 12,000 ans, deux lignées glaciaires Atlantique/Mississipienne et Acadienne ont colonisé des lacs postglaciaires nouvellement formés, se retrouvant alors en sympatrie suite à un contact secondaire général dans la zone d'étude (Rougeux et al. 2017). Le système à l'étude est composé de cinq lacs dans lesquels on retrouve des réplicats indépendants de contacts secondaires entre lignées glaciaires. Les incompatibilités génomiques ainsi que les conditions écologiques de sélection et de compétition pour les ressources au sein de niches écologiques identiques entre lignées sympatriques, a entrainé un déplacement de caractères (Bernatchez and Dodson 1990; Bernatchez and Dodson 1991; Pigeon, Chouinard, and Bernatchez 1997b; Lu and Bernatchez 1999; Landry and Bernatchez 2010; Rougeux et al. 2017). Ainsi, les populations sympatriques de corégone présentent une forme naine limnétique ayant divergé de la forme normale benthique, constituant alors des réplicats indépendants de divergence phénotypique entre lacs. Outre l'occupation de niches écologiques différentes et de traits phénotypiques impliqués dans l'adaptation à ces niches écologiques, tels qu'un nombre différent de branchicténies (Bernatchez et al. 1999), ces deux formes diffèrent par un large éventail de traits

phénotypiques, tels que la taille, le poids et l'âge à la maturité sexuelle, et physiologiques, comme la phosphorylation oxidative induisant un production d'ATP supérieure chez l'espèce limnétique ou encore une activité glycolitique accrue chez cette même espèce (Bernatchez et al. 2010; Evans and Bernatchez 2012; Laporte, Rogers, Dion-Côté, Normandeau, Gagnaire, Dalziel, Chebib, and Bernatchez 2015a; Dalziel, Laporte, Guderley, et al. 2017). De plus, des analyses transcriptomiques ont permis de mettre en évidence des différentiels d'expression importants entre les formes naines et normales de façon répétée entr lacs (Derome et al. 2006; St-Cyr et al. 2008; Jeukens et al. 2010; Jeukens and Bernatchez 2011), indiquant des pressions de sélection similaires entre lacs.

Des analyses de l'ADN génomique, associées à une étude écologique, ont mis en évidence une corrélation négative entre la disponibilité en habitats et en ressources, au sein d'un lac, et la divergence phénotypique mesurée (Landry et al. 2007; Landry and Bernatchez 2010). Cette divergence phénotypique est associée à une augmentation des indices de différenciation génétique selon les lacs (Renaut et al. 2010; Gagnaire, Pavey, et al. 2013). De plus, le paysage génomique de la différenciation entre paires sympatriques montre des niveaux de différenciation hétérogènes le long du génome, avec notamment des îlots de différenciation de tailles croissantes dans les lacs ayant les populations les plus divergentes, en lien avec des taux de flux de gènes hétérogènes (Gagnaire, Pavey, et al. 2013). Les lacs Cliff, Indian, Webster, Est et Témiscouata représentent des réplicats indépendants le long d'un continuum de différenciation génétique (niveau de différenciation décroissant). Ce contexte est particulièrement intéressant pour étudier les bases génomiques de la spéciation.

Des barrières pré-zygotiques entre les deux formes de Grand Corégone, pourrait être associées à de l'homogamie tel qu'observé chez l'épinoche à trois épines (McKinnon et al. 2004). Par ailleurs, des croisements entre nains et normaux en conditions expérimentales ont identifié des barrières post-zygotiques endogènes fortes. Par exemple le taux de survie des individus hybrides issus de rétrocroisements est inférieur de 80% à la survie des embryons issus de lignées parentales pures (Rogers and Bernatchez 2006). Parmi ces larves issues de rétrocroisements, une proportion importante présentait des malformations phénotypiques (Renaut and Bernatchez 2010; Dion-Côté et al. 2015). Des approches transcriptomiques ont révélé des phénomènes de dérégulation transcriptionelle affectant des gènes essentiels du développement chez des hybrides issus de

rétrocroisements (Renaut and Bernatchez 2011), ainsi que la réactivation d'éléments transposables et de l'aneuploidie chez les hybrides soulignant le choc entre génomes des lignées parentales (Dion-Côté et al. 2014; Dion-Côté et al. 2015). Par conséquent, les deux formes de Grand Corégone semblent remplir les critères de l'espèce biologique. Nous réfèrerons donc aux Grand Corégone nain et Grand Corégone normal comme étant deux espèces distinctes, dans le cadre de cette thèse de doctorat.

Finalement, le complexe d'espèces du Grand Corégone est un excellent système pour étudier les bases moléculaires de l'isolement reproducteur, ainsi que l'architecture génomique de la divergence adaptative pouvant mener à la spéciation écologique, du fait de réplicats naturels de divergence entre espèces sympatriques.

1.6.2 Le Lavaret

À une autre échelle géographique et taxonomique, le Grand Corégone possède une espèce sœur sur le continent européen, le Lavaret (*Coregonus lavaretus*), elle-même particulièrement intéressante. Ces deux lignées continentales ont divergé il y a quelques 500 000 ans, et sont complètement isolées géographiquement. Le Lavaret possède une répartition géographique circumpolaire. Il occupe les lacs Alpins et scandinaves (Hudson et al. 2010), et présente au travers de cette importante distribution de grandes variations morphologiques, écologiques et biologiques similaires à celles observées en Amérique du Nord (Østbye et al. 2006). Certaines études ont décrit jusqu'à six formes de Lavaret, associées à l'utilisation de niches écologiques différentes, notamment la profondeur des lacs (Douglas et al. 1999; Østbye et al. 2005). Ces formes sont associées à une origine hybride provenant du contact secondaire entre lignées glaciaires européennes, et montrent des évènements indépendants de divergence sympatrique entre espèces limnétiques et benthiques (Hudson et al. 2010). Le complexe d'espèces du Lavaret se compose lui aussi d'un parallélisme de différenciation phénotypique entres espèces limnétiques et benthiques. Par conséquent, les systèmes composés des deux espèces soeurs montre un parallélisme intercontinental de différenciation phénotypiques entre espèces limnétiques et benthiques.

Bien que peu d'études se soient attardées sur les composantes génomiques de la spéciation dans ces systèmes, des études basées sur de l'ADN mitochondrial ont indiqué

une diversité génétique supérieure chez le Lavaret par rapport au Grand Corégone. Ceci pourrait s'expliquer par un moindre impact de la dernière glaciation sur la réduction de la taille efficace des lignées de Lavaret (Bernatchez et al. 1989). De plus, des études comparatives de transcriptomique ont révélé un parallélisme intercontinental de différentiel d'expression génique entre les espèces limnétiques et benthiques (Jeukens et al. 2010). Ces études ont également illustré le rôle du polymorphisme génétique dans l'évolution de la régulation génique, impliquée dans la divergence entre espèces limnétiques et benthiques, notamment en Europe (Jeukens and Bernatchez 2011).

Un tel système, présentant un parallélisme d'adaptation et de spéciation, à l'échelle de plusieurs lacs et à l'échelle de deux continents offre d'intéressantes perspectives de génomique comparative. Il est concevable d'étudier les bases génomiques associées à un tel niveau de parallélisme de différenciation phénotypique, d'élucider l'origine de l'architecture génomique associée à la divergence entre espèces sympatriques, ou encore d'identifier les différences de polymorphismes et l'impact de ces derniers, afin de comprendre ce qui différencie les lignées de Grand Corégone et de Lavaret (*e.g.*, différences dans le nombre de morphes).

1.7 Objectifs de la thèse

L'objectif général de ma thèse est de mieux comprendre l'architecture génomique de la spéciation dans un système d'espèces naissantes. Mon approche consiste à définir l'origine des patrons génomiques hétérogènes résultant de la divergence et associés à la différenciation entre espèces, à inférer les bases génomiques communes impliquées dans la différenciation génomique parallèle, puis à identifier les différences de polymorphismes associées au potentiel d'évolution défini comme proxy de potentiel de divergence entre espèces. Plus spécifiquement, j'investigue l'origine démographique et sélective du polymorphisme génétique existant entre espèces, comment celui-ci est maintenu dans des contextes sélectifs similaires et, outre mesure, comment les différences propres à chaque système peuvent induire des niveaux de variance phénotypiques.

Dans le second chapitre, je cherche à définir l'histoire démographique pour chacun des lacs étudiés dans le système du Grand Corégone (*Coregonus clupeaformis*), en

Amérique du Nord. En effet, des travaux sur la description de l'architecture génomique de la spéciation ont mis en évidence des patrons de différenciation génétique autour de potentiels gènes de spéciation (Gagnaire, Pavey, et al. 2013). Par une approche de séquençage d'ADN génomique (RADseq), nous montrons ici que les paires d'espèces sympatriques de Grand Corégone ont comme origine un contact secondaire. Cette étude montre également l'importance d'intégrer des effets sélectifs indirects dans les modèles d'inférence démographique. Elle révèle que les patrons de différenciation entre espèces limnétiques et benthiques sont issus d'une érosion différentielle de la différenciation autour des gènes de spéciation, mais aussi à des accélérations du tri du polymorphisme ancestral sous l'effet de la sélection en liaison dans les régions à recombinaison réduite.

Le troisième chapitre, constitue une étude complémentaire au second chapitre afin d'accroitre notre compréhension sur le système étudié. Dans ce chapitre, j'infère l'histoire démographique pour chacun des lacs étudiés dans le système du Lavaret (*Coregonus lavaretus*), en Europe. Bien que des études de phylogéographies ont identifié différents haplotypes mitochodriaux (Bernatchez and Dodson 1991), que des études de génétiques des populations ont analysé la structure intra-lac des populations (espèces) pour identifier de l'admixture (Hudson et al. 2010), elles ne se sont jamais intéressées à l'histoire évolutive de la diversification phénotypique dans ce système. Par une approche de séquençage d'ARNm codant (RNAseq), nous montrons ici que les paires d'espèces sympatriques de Lavaret ont comme origine un contact secondaire. Cette étude montre également l'importance d'intégrer des effets sélectifs indirects dans les modèles d'inférence démographique, en appliquant les modèles développés dans le second chapitre de cette thèse. Comme chez le Grand Corégone, cette étude révèle que les patrons de différenciation entre espèces limnétiques et benthiques sont issus d'une érosion différentielle de la différenciation autour des gènes de spéciation, mais aussi à des accélérations du tri du polymorphisme ancestral sous l'effet de la sélection en liaison dans les régions à recombinaison réduite.

Le quatrième chapitre teste un potentiel parallélisme génétique associé au parallélisme de différenciation phénotypique entre espèces limnétiques et benthiques observées à la fois dans l'espèce américaine et l'espèce européenne, le Grand Corégone

et le Lavaret. Par une approche de séquençage d'ARN, nous avons identifié des gènes différentiellement exprimés entre espèces limnétiques et benthiques de façon parallèle dans l'ensemble du système, suggérant des bases génomiques similaires soumises à des pressions de sélection équivalentes. De plus, nous avons quantifié un taux de polymorphisme ancestral partagé entre toutes les populations, plus haut que l'attendu dans ces mêmes gènes, indiquant un processus évolutif commun pour ces gènes impliqués dans la différenciation entre espèces limnétiques et benthiques. Ceci suggère que les processus sélectifs de divergence entre espèces maintiennent le polymorphisme ancestral dans le système à l'échelle transcontinentale. Ainsi, dans le *système C. clupeaformis - C. lavaretus* du polymorphisme ancestral partagé entre ces lignées serait à l'origine d'une proportion du parallélisme phénotypique observé.

Le cinquième chapitre s'intéresse aux rôles potentiels de l'épigénétique, et notamment de la méthylation, dans la divergence adaptative entre espèces limnétiques et benthiques. Je cherche à caractériser des patrons de différences de méthylation entre paires d'espèces, et documenter un potentiel parallélisme de différentiel de méthylation maintenu chez le Grand Corégone et le Lavaret. En associant le niveaux de méthylation à l'information transcriptionelle et génomique, j'ai constaté que la méthylation repose sur des bases génomiques (méthylation obligatoire), et que le degrés de méthylation affecte les niveaux d'expression de gènes. Cependant, la différenciation d'expression génique est plus importante pour les gènes affectés par des variants génétiques (*cis*-eQTL) que ceux affectés par des variants de méthylation. Finalement, je discute d'un possible rôle joué par l'assimilation génétique dans la réponse commune à des pressions de sélections divergentes entre espèces limnétiques et benthiques pour les différents lacs. Cette étude a donc permis d'identifier des différences de méthylations maintenues entre espèces dans l'ensemble du système, de définir l'origine des variations de méthylation, et de quantifier la contribution respective de la méthylation et de la génétique à la variation d'expression entre les espèces.

Chapitre 1: Modeling the multiple facets of speciation-with-gene-flow towards inferring the divergence history of Lake Whitefish species pairs (*Coregonus clupeaformis*).

2.1 Résumé

Le parallélisme de divergence entre réplicats de paires d'espèces dans des environnements contrastés peut survenir via des scenarios évolutifs différents. Il est nécessaire de déterminer si un tel niveau de parallélisme reflète une divergence parallèle causée par la répétition de l'action de la sélection divergente et/ou un évènement unique suivit de flux géniques. La reconstruction de la démographie historique est ainsi fondamentale pour comprendre comment la démographie et la sélection ont interagit pour modeler la divergence génomique durant le processus de spéciation. Dans cette étude, nous avons utilisé une approche par modélisation pour explorer les multiples facettes de la spéciation en présence de flux génique, avec des modèles de divergence démo-génétiques qui capturent les variations à la fois temporelles et génomiques de la taille effective de la population et du taux de migration. Nous avons investigué l'histoire de la divergence entre paires d'espèces de Grand Corégone (normal, benthique et nain, limnétique), qui se caractérisent par différents degrés de divergence écologique et d'isolement reproducteur. Des marqueurs génomiques de type *SNP* ont été utilisés pour documenter l'étendue de la différenciation génétique dans chaque paire d'espèces, et 26 modèles de divergence ont été ajustés et comparés au spectre joint déplié des fréquences alléliques de chaque paire. Ces données ont mis en évidence qu'un contact secondaire asymétrique récent (environ 3000-4000 générations) entre populations en expansion démographique, a accompagné la diversification du Grand Corégone. Nos résultats suggèrent que la différenciation génomique a émergé sous les effets combinés de la sélection aux sites liés, générant des taux variables de tris de lignées au sein du génome durant l'isolement géographique, et de l'introgression hétérogène ayant érodé la divergence le long du génome suite au contact secondaire. Cette étude fournit ainsi un nouvel aperçu rétrospectif des processus historiques et sélectifs démographiques qui ont formé un continuum de divergences associé à la spéciation écologique.

2.2 Abstract

Parallel divergence across replicated species pairs occurring in similar environmental contrasts may arise through distinct evolutionary scenarios. Deciphering whether such parallelism actually reflects repeated parallel divergence driven by divergent selection or a single divergence event with subsequent gene flow needs to be ascertained. Reconstructing historical gene flow is therefore of fundamental interest to understand how demography and selection jointly shaped genomic divergence during speciation. Here, we use an extended modeling framework to explore the multiple facets of speciation-with-gene-flow with demo-genetic divergence models that capture both temporal and genomic variation in effective population size and migration rate. We investigate the divergence history of replicate sympatric species pairs of Lake Whitefish (normal benthic and dwarf limnetic) characterized by variable degrees of ecological divergence and reproductive isolation. Genome-wide SNPs were used to document the extent of genetic differentiation in each species pair, and 26 divergence models were fitted and compared to the unfolded joint allele frequency spectrum of each pair. We found evidence that a recent (circa 3,000-4,000 generations) asymmetrical secondary contact between expanding post-glacial populations has accompanied Whitefish diversification. Our results suggest that heterogeneous genomic differentiation has emerged through the combined effects of linked selection generating variable rates of lineage sorting across the genome during geographical isolation, and heterogeneous introgression eroding divergence at different rates across the genome upon secondary contact. This study thus provides a new retrospective insight into the historical demographic and selective processes that shaped a continuum of divergence associated with ecological speciation.

2.3 Introduction

Historical changes in the geographical distribution of species have been an important driver of diversification across many taxa (Coyne and Orr 2004). In particular, the pronounced climatic variations that occurred during the late Pleistocene caused major shifts in the distribution ranges of many species. These shifts are responsible for the divergence of ancestral lineages that survived in different glacial refugia, and then possibly came into secondary contact during interglacial periods (Bernatchez and Wilson 1998; Avise 2000; Hewitt 2001). The signature of post-glacial recolonization is still apparent in well-known terrestrial and aquatic suture zones, where multiple contacts between expanding post-glacial lineages tend to overlap and form hybrid zones hotspots (Hewitt 1996; Hewitt 2000; Hewitt 2004; Swenson and Howard 2005; Bierne et al. 2011; April et al. 2013).

In some cases, secondary contacts have resulted in the sympatric enclosure of previously allopatric, partially reproductively isolated lineages, for instance within post-glacial lakes (reviewed by Taylor 1999). This sympatric coexistence should have facilitated gene flow compared to parapatric populations, eventually leading to complete genetic homogenization of the original glacial lineages. This is not the case, however, for several north temperate freshwater fishes in which sympatric glacial lineages have further diverged into phenotypically differentiated and reproductively isolated species pairs following secondary contact (Bernatchez and Dodson 1990; McPhail 1992; Taylor and Bentzen 1993; Schluter 1996; Wood and Foote 1996; Taylor 1999). These cases of ecological speciation have been hypothesized to reflect adaptive responses to minimize competitive interactions and outbreeding depression through ecological niche segregation and hybridization avoidance among previously allopatric lineages (Bernatchez et al. 2010).

The evolutionary processes responsible for the phenotypic diversification of these incipient sympatric species remain contentious, especially with regards to the relative contributions of genetic differences that evolved in allopatry compared to more recent genetic changes occurring in sympatry (Welch and Jiggins 2014). To gain a more thorough understanding of how divergence unfolds at the molecular level, it is crucial to simultaneously take into account the historical demographic events that accompanied divergence and the subsequent genetic exchanges that occurred in sympatry. Genome-wide polymorphism data now provide the opportunity to infer complex demographic

histories (Gutenkunst et al. 2009; Excoffier et al. 2013; Butlin et al. 2014) and investigate the evolutionary processes leading to the formation of nascent sympatric species.

Many aspects of populations' evolutionary history are influenced by demography, such as the rate of lineage sorting and gene exchange (Sousa and Hey 2013). Several approaches have been developed to infer the history of population divergence from genetic data obtained from contemporary populations. These methods usually rely on demographic models capturing the effects of population size, splitting time and migration between two populations exchanging genes (Hey and Nielsen 2004; Becquet and Przeworski 2007; Hey and Nielsen 2007). An important facet of the speciation process which is usually not taken into account by demographic models is that a significant proportion of the genome may be affected by selection (Barton and Bengtsson 1986; Nosil 2008; Feder, Egan, and Nosil 2012b; Harrison and Larson 2016b; Wolf and Ellegren 2017). Two different selective processes that generate heterogeneous genome divergence can be distinguished. One occurs during contact episodes and corresponds to selection acting on genomic regions involved in reproductive isolation and local adaptation (Harrison 1990; Wu 2001; Payseur 2010). The second occurs through the action of background selection and selective sweeps (Hill and Robertson 1966; Smith and Haigh 1974; Charlesworth et al. 1997), which remove linked neutral diversity within populations during periods of reduced gene flow (Cruickshank and Hahn 2014). While the barrier effect of speciation genes is equivalent to a local reduction in effective migration rate (m_e) (Barton and Bengtsson 1986; Feder and Nosil 2010), linked selection rather corresponds to a reduction in effective population size (N_e) that locally accelerates lineage sorting in the genome (Charlesworth 2006). Based on this, it is possible to capture the barrier effect of speciation genes by allowing for varying rates of introgression among loci (Roux et al. 2013; Sousa and Hey 2013; Tine et al. 2014; Roux, Fraïsse, Romiguier, Anciaux, Galtier, and Bierne 2016a), and to capture the effect of linked selection by allowing loci to experience varying rates of genetic drift (Sousa and Hey 2013; Roux, Fraïsse, Romiguier, Anciaux, Galtier, and Bierne 2016a). This provides a framework in which the effects of heterogeneous selection and gene flow can be considered separately or simultaneously to identify the simplest divergence scenario that best explain the data, thus avoiding overparametrization issues.

North American Lake Whitefish (*Coregonus clupeaformis*) represents a valuable model to study the role of past allopatric isolation on recent, sympatric ecological

divergence. The Saint-John River drainage (southeastern Québec, northeastern Maine), where benthic (normal) and limnetic (dwarf) Whitefish sympatric species pairs occur in different lakes, corresponds to a suture zone where two glacial lineages (Atlantic/Mississippian and Acadian) have been hypothesized to have come into secondary contact during the last glacial retreat on the basis of mitochondrial DNA phylogeography (Bernatchez and Dodson 1990). Given the historical hydrology of the region whereby there was a limited temporal window during which different lakes could be colonized by fish before becoming isolated (Curry 2007), and the absence of limnetic (dwarf) Whitefish in allopatry, the most likely scenario is that of an independent phenotypic divergence that occurred in each lake (Bernatchez 2004). In some lakes, phenotypic divergence between sympatric dwarf (limnetic) and normal (benthic) populations is still partly associated with the mitochondrial DNA lineages characterizing the different glacial origins of the sympatric populations (Pigeon, Chouinard, and Bernatchez 1997a). Dwarf whitefish are most often associated with the Acadian mitochondrial lineage and are only found in sympatry with the normal species. Moreover, fish from the Acadian lineage have a normal phenotype outside the contact zone, which supports the hypothesis that the dwarf phenotype has been derived postglacially from an Acadian genetic background within the contact zone and independently in each lake (Bernatchez and Dodson 1990; Bernatchez and Dodson 1991; Bernatchez et al. 2010). The evolution of further phenotypic divergence in sympatry suggests that character displacement may have been facilitated by the contact between genetically differentiated lineages (Bernatchez 2004). Moreover, the different dwarf-normal species pairs found in the contact zone are arrayed along a continuum of phenotypic differentiation, where smaller lakes exhibit higher morphological differentiation, which closely mirrors the potential for niche segregation and exclusive interactions within lakes (Lu and Bernatchez 1999; Rogers et al. 2002; Landry et al. 2007; Rogers and Bernatchez 2007b; Landry and Bernatchez 2010). This continuum is also evident at the genomic level, with increased baseline genetic differentiation and larger genomic islands of differentiation being found from the least to the most phenotypically and ecologically differentiated species pair (Renaut et al. 2012; Gagnaire, Pavey, et al. 2013). Finally, quantitative trait loci (QTL) underlying adaptive phenotypic divergence on complex and polygenic quantitative traits (*e.g.* behavioural, morphological, physiological and life history traits) map preferentially to genomic islands of differentiation (Renaut et al. 2012; Gagnaire, Pavey, et al. 2013), suggesting that selection acting on these traits contribute to the barrier to gene flow.

Despite such detailed knowledge on this system, previous studies did not allow clarifying how the genomic landscape of dwarf-normal divergence in each lake has been influenced by the relative effects of directional selection on these QTLs and post-glacial differential introgression. Consequently, it is fundamental to elucidate the demographic history of the dwarf-normal whitefish species pairs to disentangle the evolutionary mechanisms involved in their diversification.

The main goal of this study was to use a genome-wide single nucleotide polymorphism (SNP) dataset to infer the demographic history associated with the recent phenotypic diversification of replicate sympatric dwarf and normal Lake Whitefish species pairs. Using RAD-seq SNP data to document the Joint Allele Frequency Spectrum (JAFS) in each species pair, we specifically test for the role of temporal and genomic variations in the rate of gene flow for each species pair separately, controlling for both effective population size and migration. We then performed historical gene flow analyses among lakes to determine how the different scenarios independently inferred within each lake collectively depict a parsimonious evolutionary scenario of diversification. Finally, we document how the complex interplay between historical contingency, demography and selection jointly shaped the continuum of divergence among sympatric whitefish species pairs.

2.4 Materials and Methods

Sampling and genotyping

We used RAD-sequencing data from Gagnaire et al. (Gagnaire, Pavey, et al. 2013) to generate a new genome-wide polymorphism dataset. Previous studies based on these data (Gagnaire, Pavey, et al. 2013; Laporte, Rogers, Dion-Côté, Normandeau, Gagnaire, Dalziel, Chebib, and Bernatchez 2015b) only focused on a subset of 3438 SNPs that were included in the Lake Whitefish linkage map (Gagnaire, Normandeau, et al. 2013). Here, we used the total amount of sequence data (1.7×10^9 reads of 101 bp) to document genome-wide variation in all five sympatric species pairs that are still occurring in five isolated lakes from the Saint-John River basin (Figure 2.1). For each pair, 20 normal and 20 dwarf individuals were used for RAD-sequencing, but five individuals that received poor sequencing coverage were removed from the dataset. Consequently, the following analyses were performed with 195 individuals, each having an average number of $8.7\text{x}10^6$ of reads.

We also sequenced six European Whitefish (*Coregonus lavaretus*), a sister species closely related to the North American Lake Whitefish, to provide an outgroup for identifying ancestral and derived alleles at each polymorphic site within the Lake Whitefish. European Whitefish were sampled in Skrukkebukta Lake (Norway, 69°34'11.6"N-30°02'31.9"E), which also harbors postglacial sympatric whitefish species pairs (Amundsen, Bøhn, and Våga 2004a; Østbye et al. 2006). RAD libraries were prepared for three individuals from each ecotype population, using the same procedure as for American Lake Whitefish (Amundsen, Bøhn, and Våga 2004a; Østbye et al. 2006).

The *C. lavaretus* raw sequence dataset was filtered using the same criteria as for *C. clupeaformis* (Gagnaire, Pavey, et al. 2013). After sequence de-multiplexing, the reads were trimmed to a length of 80 bp to avoid sequencing errors due to decreasing data quality near the end of reads. We then used the *Stacks* pipeline (v1.24) for *de novo* RAD-tags assembly and individual genotyping (Catchen et al. 2013). We empirically determined an optimal set of assembly parameters to *Ustacks*, in order to (i) adjust haplotype divergence between alleles at the same locus to the within-population diversity, (ii) take into account the possibility of introgression between differentiated populations, while (iii) controlling for false allelic variation due to hidden paralogy. A minimal coverage depth of 5x per allele

(m=5) and a maximal number of six mismatches between two haplotypes at a same locus within individuals (M=6) were set. We then allowed a maximal number of six mismatches between individuals in *Cstacks* (n=6) to merge homologous RAD-tags from different samples. Finally, we used the program *Populations* to export a VCF file containing the genotypes of all individuals.

Several filtering steps were then performed with *VCFtools* v0.1.13 (Danecek et al. 2011) to remove miscalled and low-quality SNPs, as well as false variation induced by the merging of paralogous loci. We first removed SNPs with more than 10% missing genotypes in each *C. clupeaformis* population. A lower exclusion threshold of 50% was used for *C. lavaretus* to retain a maximum of orthologous loci in the outgroup. We then filtered for Hardy-Weinberg disequilibrium within each population using a p-value exclusion threshold of 0.01. Finally, we merged the filtered datasets of dwarf and normal populations within each lake together with the European whitefish outgroup and kept only loci that passed the previous filters in all three samples. This resulted in five lake-outgroup datasets containing 14812, 22788, 5482, 26149, and 14452 SNPs for Témiscouata, East, Webster, Indian and Cliff lakes, respectively. Finally, we determined the most parsimonious ancestral allelic state for loci that were monomorphic in the outgroup but polymorphic in *C. clupeaformis*, defining the allelic state in the outgroup as the ancestral allele (Tine et al. 2014). The resulting oriented SNP datasets contained 11985, 11315, 5080, 13905 and 9686 SNPs for Témiscouata, East, Webster, Indian and Cliff lakes respectively, that were used to build the unfolded JAFS of each lake, using custom *perl* and *R* scripts. Because the amount of SNPs was limited for Webster Lake possibly due to lower quality samples, this species pair was not considered for the subsequent demographic inferences to avoid potential biases due to a lack of resolution of its JAFS. In the four retained lakes, the JAFS was projected down to 13 individuals (*i.e.*, 26 chromosomes) per population to avoid remaining missing genotypes and optimize the resolution of the JAFS.

Inferring the history of divergence with gene flow and selection

Because the different lakes have been isolated from each other for approximately 12,000 years (Curry 2007), we analyzed their JAFS separately. The demographic and selective histories of the four species pairs were inferred using a custom version of the software *∂a∂i* v1.7 (Gutenkunst et al. 2009). We considered 26 models (Figure 2.2) that

were built to extend four basic models representing alternative modes of divergence: Strict Isolation (SI), Isolation-with-Migration (IM), Ancient Migration (AM), and Secondary Contact (SC). Each model consists of an ancestral population of size N_{ref} that splits into two populations of effective size N_1 and N_2 during T_S (SI, IM), $T_{AM}+T_S$ (AM), or T_S+T_{SC} (SC) generations, possibly exchanging migrants during T_S (IM), T_{AM} (AM), or T_{SC} (SC) generations at rate m_{e12} from population 2 (normal) into population 1 (dwarf), and m_{e21} in the opposite direction. These models were extended to integrate temporal variation in effective population size (-G) by enabling exponential growth using current-to-ancient population size ratio parameters b_1 (for dwarf) and b_2 (for normal) to account for expansions or bottlenecks. Variation in effective population size across the genome due to Hill-Robertson effects (Hill and Robertson 1966) - *i.e.* local reduction in N_e at linked neutral sites due to the effect of background (Charlesworth et al. 1993) and positive selection (Smith and Haigh 1974) - was modeled by considering two categories of loci (-2N) occurring in proportions Q and 1-Q in the genome. In order to quantify a mean effect of selection at linked sites, we defined a Hill-Robertson scaling factor (*hrf*), relating the effective population size of loci influenced by selection ($N'_1=hrf\times N_1$ and $N'_2=hrf\times N_2$) to that of neutral loci (N_1 and N_2). Then, models of divergence with gene flow were extended to account for heterogeneous migration across the genome by considering two categories of loci (-2m). In addition to a first category of loci evolving under effective migration rates m_{e12} and m_{e21} and occurring in proportion P in the genome, we considered a second category of loci that occur in proportion 1-P, experiencing different effective migration rates $m_{e'12}$ and $m_{e'21}$ (Tine et al. 2014). The proportion P was identical in the two diverging populations (same for Q). Because migration and drift influence gene flow during the whole divergence period in the IM model, the effects of heterogeneous migration and population effective size are difficult to dissociate. Therefore, these effects were estimated jointly only within the AM and SC models, in which a strict isolation period helps uncoupling the effects of migration and drift (i.e. -2N2m extensions were considered in addition to -2N and -2m models, Figure 5.2). In these cases, the proportions 1-P and 1-Q corresponded to different sets of loci in the genome. All models with heterogeneous gene flow were also implemented to allow for population growth (*i.e.* -2NG, -2mG and -2N2mG). Finally, in order to take into account potential errors in the identification of ancestral allelic states, predicted JAFS were constructed using a mixture of correctly oriented and mis-orientated SNPs occurring in proportions O and 1-O,

respectively. The JAFS of mis-oriented variants was obtained by reversing the model spectrum along its two axes (Tine et al. 2014).

The 26 models were fitted independently for each lake using successively a hot and a cold simulated annealing procedure followed by 'BFGS' optimization (Tine et al. 2014). We ran 25 independent optimizations for each model in order to check for convergence and retained the best one (Supplementary Table S2.S3) to perform comparisons among models based on Akaike information criterion (AIC). Our comparative framework thus addresses overparametrization issues by penalizing models which contain more parameters. By allowing comparisons among nested models of increasing complexity, it also provides a means to independently evaluate the effect of accounting for temporal or genomic variation in migration rate or effective population size. A conservative threshold was applied to retain models with $\Delta AIC_i = AIC_i - AIC_{min} \leq 10$, since the level of empirical support for a given model with a $\Delta AIC_i > 10$ is essentially none (Burnham and Anderson 2002). For each lake, the difference in AIC between the worst and the best model $\Delta_{max} = AIC_{max} - AIC_{min}$ was used to obtain a scaled score for each model using:

$$model\ score = \frac{(\Delta_{max} - \Delta AIC_i)}{\Delta_{max}} \quad (1)$$

such that for each lake the worst model takes a score of 0 and the best model takes a score of 1. Therefore, the model score could be used to more easily compare the retained models among lakes. In order to evaluate the relative probabilities of the different models within each lake, we also computed Akaike weights (w_{AIC}) following equation (2), where *R* corresponds to the total number of models considered (*R* = 26).

$$w_{AIC} = \frac{e^{\frac{(-\Delta AIC_i)}{2}}}{\sum_{i=1}^{R} e^{\frac{(-\Delta AIC_i)}{2}}} \quad (2)$$

To estimate parameter uncertainty, we used the *Godambe* information matrix method from *∂a∂i* v1.7. Non-parametric bootstrapping was used to generate 1000 bootstrapped datasets

to estimate confidence intervals (CIs) using the standard-error of maximum likelihood estimates (*se*).

Finally, we converted estimated demographic parameters into biologically meaningful units in order to compare informative parameter values among lakes (*e.g.* timing and strength of gene flow). These estimates were only used for indication and comparative purpose, since we were missing crucial information about the per generation mutation rate in Lake Whitefish. We used the optimal multiplicative scaling factor *theta* between model and data to estimate the ancestral effective population size (N_{ref}) before split for each lake:

$$N_{ref} = \frac{\theta}{4L\mu} \quad (3)$$

with μ being the mutation rate (fixed at 10^{-8} mutations/site/generation) and L the effective length of the genome explored by our RAD-Seq experiment and estimated as:

$$L = \frac{zy80}{x} \quad (4)$$

where x is the number of SNPs that were originally detected from y RAD-tags of 80 bp present in the initial dataset, and z is the number of SNP retained for *∂a∂i* analyses in the lake considered. Estimated times in units of $2N_{ref}$ generations were converted into years assuming a generation time of 3.5 years (*i.e.*, the average between 3 and 4 years for sexual maturity in dwarf and normal whitefish, respectively) (Chebib et al. 2015). Estimated migration rates were divided by $2N_{ref}$ to obtain the proportion of migrants received by each population every generation.

Patterns of shared ancestry and admixture among lakes

We also searched for signatures of shared ancestry and gene flow among replicate species pairs to provide a broader understanding of the divergence history in whitefish. The four lake-specific datasets used for demographic inferences were merged together with the Webster Lake dataset, added here to get a more thorough understanding of the system as a whole. Only polymorphic loci that were retained after filtering in all five lakes were considered (42,558 SNPs in total, without the outgroup). For each lake, we determined the

fraction of private polymorphisms, the fraction of SNPs shared with at least one other lake or shared across all five lakes. We then measured the proportion of SNPs that were shared between species and private to each species within each lake, as well as among lakes.

To visualize the overall genetic structure and relationships among lakes and species, we performed a discriminant analysis of principal components (dAPC) in *Adegenet* v2.0.0 (Jombart et al. 2010). We first imputed missing genotypes within each population using a Random Forest regression approach (Poland et al. 2012). Imputation was performed using ten iterations with 150 trees using the *randomForestSRC* v1.6.1 package in *Stackr* v.0.1.5 (Gosselin & Bernatchez 2016) and imputed sub-datasets were subsequently merged to perform the dAPC.

Finally, we used *TreeMix* v1.12 (Pickrell and Pritchard 2012) to infer historical relationships among populations. This method uses the covariance structure of allele frequencies between populations and a Gaussian approximation for genetic drift to build a maximum likelihood graph relating sampled populations to their common ancestor, taking migration events into account to improve the fit to the inferred tree. Migration events between populations are modeled in *TreeMix* by discrete mixture events. Such events may either reflect gene exchange between populations within lakes and/or genetic correlations between geographically isolated populations of the same species, due to the retention of shared ancestral polymorphism among populations from different lakes following their geographic isolation. In order to avoid interpreting spurious migration signals, we focused on the main events of gene flow that received the highest weights (Pickerel et al. 2012), which likely correspond to the largest admixture proportions. We thus allowed a maximum of six migration events to be inferred among branches of the whitefish population tree. For this analysis, we used a 20% missing genotype rate per population without imputing missing genotypes to avoid potential biases in the covariance matrix.

2.5 Results

Comparisons among divergence models

For each of the four sympatric whitefish species pairs retained for historical divergence analyses (Figure 2.1), 26 alternative divergence models of increasing complexity were fitted to polymorphism data and compared to each other. This comparative framework enabled us to account for four different aspects of gene flow, that were considered separately or in combination. This includes temporal variation in effective population size (N_e) and migration rate (m), but also genomic heterogeneity in these parameters to capture selective effects (Figure 2.2). The four JAFS obtained highlighted the continuum of divergence existing among lakes (Figure 2.3A, Supplementary Figure S2.1). Namely, the density of shared polymorphisms located along the diagonal decreased from the least divergent (Témiscouata and East) to the most divergent (Indian and Cliff) species pairs, while the variance of SNP density around the diagonal increased accordingly. In addition, non-shared polymorphisms (*i.e.*, private SNPs) located on the outer frame of the spectrum were mostly found in Indian and Cliff lakes, which were also the only lakes where differentially fixed SNPs between dwarf and normal whitefish were observed.

The comparison of model scores within and among lakes showed the importance of considering temporal changes in effective population size. Models including population growth (-G models) generally provided better fits to the data for Témiscouata, Indian and Cliff lakes (Mann-Whitney U test, p = 0.002; p = 0.011 and p = 0.0001, Figure 2.4A). Similarly, accounting for heterogeneous migration rates across the genome (-2m models) improved the average model scores for each lake, although not significantly in East Lake (U test, p = 0.016; p = 0.063 and p = 0.029 for Témiscouata, Indian and Cliff lakes, respectively) (Figure 2.4B). Moreover, models integrating heterogeneous effective population size at the genomic level (-2N models) provided significant improvements for the two most divergent species pairs from Indian and Cliff lakes (U test, p = 0.013; p = 0.010 and p = 0.005, respectively) (Figure 2.4C).

For each lake, we performed model selection based on the AIC to penalize model likelihood by the number of parameters to avoid over-fitting. Applying a criterion of $\Delta AIC_i \leq 10$ for model selection, we retained two best models for Témiscouata (SC2mG and

SIG), Indian (SC2N2mG and SC2NG) and Cliff (SC2N2mG and SC2mG) lakes and four best models for East Lake (SC2N2mG, IM2mG, AM2N2m, AMG) (Supplementary Table S2.1, Table S2.2). Akaike weights (w_{AIC}) were higher than 0.5 for the highest ranked model of each lake (Figure 2.4D, Supplementary Table S2.1, Table S2.2). For Témiscouata, the best model was a secondary contact with heterogeneous migration contemporary to population size change (SC2mG; $w_{AIC} = 0.68$). The other three species pairs received the highest support for a secondary contact model with heterogeneous migration rate and effective population size contemporary to population size change (SC2N2mG; $w_{AIC} = 0.77$ in East, $w_{AIC} = 0.94$ in Indian and $w_{AIC} = 0.53$ in Cliff). Comparisons between the JASF predicted under the best models and the data showed variable patterns in the distribution of residuals depending on lakes (Supplementary Figure S2.1), which were most likely due to model departure from the real (and probably more complex) evolutionary history of species pairs.

Inference of model parameters

The inferred proportion of correctly oriented markers in the unfolded JAFS (parameter *O*) ranged from 95.4% to 99%, suggesting that the vast majority of ancestral allelic states were correctly inferred using the European Whitefish as an outgroup (Supplementary Table S2.1, 2).

Considering only the highest ranked model for each lake, some general patterns emerged from the comparisons of inferred model parameters among lakes. First, differences in effective population sizes between dwarf and normal whitefish were inferred after splitting from the ancestral population, especially in Indian and Cliff lakes where inferred N_e showed no overlap in CIs between dwarf and normal (Supplementary Table S2.1, Table S2.2). When such differences were observed, N_e was larger for dwarf compared to normal whitefish. Taking into account population growth in the four lakes revealed quite similar patterns of temporal population size changes with recent demographic expansions being found in all populations except for normal whitefish in Lake Témiscouata. A more pronounced demographic expansion was generally inferred for dwarf compared to normal whitefish, and the contemporary effective population size was also larger in dwarf than in normal populations in the four lakes (non-overlapping CIs, Supplementary Table S2.4). The contemporary number of migrants exchanged per

generation from one population to the other, estimated using the weighted mean effective migration rate in each direction (*i.e.*, average effective gene flow = $N \times b \times (P \times me + 1 - P \times me')$), revealed more pronounced gene flow from dwarf to normal populations in all lakes except Cliff (Figure 2.5).

The highest ranked model for East, Indian and Cliff lakes included heterogeneous effective population size at the genomic level (Figure 2.3C, Figure 2.4, Supplementary Table S2.1). The fraction of the genome with a reduced N_e (Q) was estimated to about 16% in East Lake, 40% in Indian and Cliff lakes, and the proportion of reduction in N_e in those fractions of the genome (*i.e.*, the Hill-Robertson factor, *hrf*) was about 11%, 20% and 17% for East, Indian and Cliff lakes, respectively.

Time parameters, namely the duration of allopatric isolation (T_S) and secondary gene flow (T_{SC}), were converted into years. Estimated durations of allopatric isolation periods revealed a recent and similar divergence history in all five lakes (*Ts* was approximately 30,000; 36,000; 29,000; 32,000 yrs for Témiscouata, East, Indian and Cliff lakes respectively, Supplementary Table S2.2). The inferred time of secondary contact coincided roughly with the last glacial retreat following the Wisconsinian glaciation, which happened between 18,000 and 11,000 yrs before present (7,200; 19,600; 8,500 and 9,200 ybp for Témiscouata, East, Indian and Cliff lakes respectively, Supplementary Table S2.2).

Comparisons of genetic variation among lakes

Only a small proportion (2.5%) of the 42,582 SNPs that were genotyped in all lakes (including a fifth species pair from Webster Lake) corresponded to polymorphic loci that are shared across all five lakes (*i.e.*, 'shared across all five lakes' category; Figure 2.1). Reciprocally, about 25% of the SNPs were private to Témiscouata or East lakes, whereas Webster, Indian and Cliff lakes each had ~10% of private SNPs. The majority of the loci were segregating in at least two (but less than five) lakes (Figure 2.1). When combined over the five lakes, a higher proportion of the loci were private to normal compared to dwarf populations. Within lakes, the highest proportions of SNPs private to normal whitefish were found in Témiscouata (51%) and Webster (69%). The three other lakes displayed the opposite pattern with higher proportions of SNPs private to the dwarf populations. Shared variation within lakes represented only 6% to 16% of the SNPs.

Partitioning genetic variation within and among lakes using a dAPC revealed distinct signals along the four first axes (Figure 2.6A). On the first axis (LD1, explaining 39.5% of the variance), populations clustered by lakes according to their geographical distribution, roughly separating the three southernmost lakes (Webster, Indian and Cliff, negative coordinates) from the two northernmost lakes (Témiscouata and East, positive coordinates). The second axis (LD2, explaining 22.5% of the variance, not shown here but see Figure S2.2) mostly separated dwarf and normal whitefish from Cliff. The third axis (LD3, explaining 16% of the variance) tended to separate species pairs by shifting normal whitefish of Mississippian/Atlantic origin towards positive coordinates and dwarf whitefish of Acadian origin towards negative coordinates, except for East Lake. The two most extreme populations values on that axis corresponded to normal species from Cliff Lake, the least introgressed relict of the Mississippian/Atlantic lineage (Gagnaire, Pavey, et al. 2013), and the dwarf species from Cliff Lake associated with the Acadian lineage. Finally, the fourth axis (LD4, explaining 11% of the variance) separated the dwarf and normal species from each lake, thus illustrating the shared ancestry between species among lakes (Figure 2.6B).

The genetic relationships among populations analyzed with *TreeMix* revealed two levels of signal (Figure 2.7, Figure S2.3). The first level was directly linked with genetic distance between the different populations of dwarf and normal whitefish. The population tree rooted with the normal population from Cliff (the most divergent population that best reflects the ancestral state of the Mississippian/Atlantic lineage (Lu et al. 2001), clearly separated normal whitefish from Cliff (CN) and Indian (IN) and dwarf whitefish from Cliff (CD) and Indian (ID), which were grouped together separately from all other populations. The clustering of CN with IN as that of CD and ID most likely reflect their shared ancestral polymorphism associated with their glacial lineage origin (Mississippian/Atlantic lineage for CN and IN; Acadian for CD and ID) (Lu et al. 2001). The second level of signal (geographic signal) grouped population pairs by lake in the remaining part of the tree, most likely reflecting the effect of gene flow following secondary contact between glacial lineages in each lake (Gagnaire et al. 2013a).

Inferred migration links were represented by arrows, the color of which indicates their relative weights (Figure 2.7). Migrations links between sympatric species pairs for Cliff and Indian lakes suggested contemporary gene flow (*i.e.*, consecutive to the colonization of

the post-glacial lakes) between dwarf and normal populations within each of these two lakes. Other migration links between allopatric populations of the same species (*i.e.*, populations from different lakes which have been isolated since the lakes formation) illustrated the genetic proximities (*i.e.* shared ancestry) of species from distinct lakes. For instance, dwarf whitefish from Webster (WD) was related to dwarf whitefish from Indian lake (ID), and the same link was found between the normal populations of these lakes (WN and IN). Finally, the ancestral population of East Lake was related with dwarf whitefish from Indian Pond, whereas normal whitefish from East Lake (EN) was linked to normal whitefish from Indian Lake, thus supporting a common genetic background between the normal populations of East and Indian lakes.

2.6 Discussion

The observed patterns of genomic differentiation among replicated dwarf/normal Lake Whitefish species pairs provide new insights into the divergence history of a well-studied model of ecological speciation (Bernatchez et al. 2010). The first approach implemented here relied on inferring the divergence history of each species pair separately, using the JAFS as a summary statistics of genome-wide differentiation. In order to maximize the amount of available information, each JAFS was oriented using the closely related European whitefish as an outgroup species, thus providing increased power to detect demographic processes that generate asymmetric distributions of derived variants around the diagonal of the JAFS.

The secondary contact (SC) scenario provided the best fit to the observed data for all of the four species pairs used for historical divergence analyses. Therefore, our ability to detect the secondary contact was apparently not affected by the small degree of genetic divergence between the least differentiated species pairs. Using simulations, Roux et al. (2016) recently showed that the power to detect the SC scenario can be high when the period of isolation is long relative to the duration of secondary contact, which was the case here. Moreover, secondary contacts are expected to leave detectable signatures on the JAFS (Alcala et al. 2016), since the erosion of past allopatric divergence by secondary gene flow typically generates an excess of shared intermediate frequency alleles.

A scenario of secondary contact is concordant with past phylogeograhic studies performed in the early days of mtDNA studies (Bernatchez and Dodson 1990; Bernatchez and Dodson 1991; Pigeon, Chouinard, and Bernatchez 1997b), but provided much deeper insights into the evolutionary history of whitefish radiation. The geographic area where sympatric whitefish species pairs occur corresponds to a well-known suture zone where glacial lineages have come into contact in several freshwater species, as they were recolonizing from different refugia after the Wisconsinian glaciation (Curry 2007; April et al. 2013). In Lake Whitefish, this zone corresponds to a phylogeograhic transition between Acadian and Atlantic/Mississippian mitochondrial lineages (Bernatchez and Dodson 1990; Pigeon, Chouinard, and Bernatchez 1997b). Interestingly, the Allegash River basin (including the studied lakes), which represents the core of this contact zone, is the only area where sympatric populations of Lake Whitefish are observed. Moreover, no dwarf

population, either in allopatry or sympatry, has been reported outside this region (Bernatchez and Dodson 1990). Therefore, phenotypic and ecological divergence, and in particular the occurrence of dwarf whitefish, is tightly linked to this secondary contact zone.

The frequency of Acadian and Atlantic/Mississippian mitochondrial lineages within lakes was partly associated with the level of phenotypic divergence between dwarf and normal whitefish, with variable amounts of mitochondrial introgression being found among lakes (Bernatchez and Dodson 1990; Pigeon, Chouinard, and Bernatchez 1997b). At one extreme, the least phenotypically divergent pair from East Lake is fixed for the Acadian mitochondrial haplogroup in both dwarf and normal whitefish, which has previously been interpreted as a support for sympatric divergence in this particular lake (Pigeon et al. 1997). Although our inferences based on the JAFS could not definitely rule out the IM model (IM2mG, $w_{AIC} = 0.19$), we obtained much stronger evidence in favor of the secondary-contact scenario in this lake (SC2N2mG, $w_{AIC} = 0.77$). A possible explanation for the loss of the Atlantic/Mississippian haplogroup in East Lake involves a stronger recent demographic expansion in the dwarf population following secondary contact, which may have contributed to the fixation of the Acadian lineage. Indeed, the preferential direction of introgression between hybridizing populations with asymmetrical N_e is expected to occur from the larger to the smaller population (Beysard et al. 2011). This is consistent with the inferred asymmetrical direction of the effective gene flow from dwarf to normal populations (Figure 2.5). This is also supported by a similar scenario in the neighboring Témiscouata Lake, which harbors the second least divergent species pair. Témiscouata Lake is also dominated by the Acadian haplogroup, but a small proportion of normal whitefish in this lake is still associated with the Atlantic/Mississippian lineage. Since we also inferred an expansion of the dwarf population (but not in normal whitefish) following secondary contact in this lake, it is likely again that this demographic imbalance explains the predominance of Acadian mitochondrial haplotypes in the northern part of the contact zone. At the other extreme, Cliff Lake where species divergence is the most pronounced shows differential fixation of Acadian and Atlantic/Mississippian haplotypes in dwarf and normal populations, respectively (Bernatchez & Dodson 1990). Thus, there is a perfect association in this lake between glacial lineage origin and phenotypic divergence, which was also attributed to a secondary contact in our demographic inferences. Similarly to East Lake, Indian Lake harbored both dwarf and normal populations fixed for the Acadian haplogroup (Lu et al.

2001). However, our analysis of the JAFS also confirmed that two distinct glacial lineages have come into contact in this lake. Such a partial concordance is typically expected when secondary contact occurs between incompletely reproductively isolated species (Taylor and Donald McPhail 2000).

A shared history of divergence before independent evolution within lakes

The global analysis including all five pairs simultaneously helped clarifying the extent to which replicate whitefish species pairs share a common history of divergence. The secondary contact scenario implies that the different species pairs are derived from the same two glacial lineages, as partly supported by mitochondrial data (Bernatchez and Dodson 1990; Pigeon, Chouinard, and Bernatchez 1997b). However, whether whitefish species pairs share a common history before secondary contact has never been assessed using nuclear markers.

Grouping populations based on their overall genetic similarities with *Treemix* produced two different types of grouping in the population tree. Populations from the three least divergent species pairs were grouped by lake (*i.e.*, TN grouped with TD, EN with ED, and WN with WD), while populations from the two most divergent species pairs were grouped by ecotypes (*i.e.*, IN with CN, and ID with CD). This complex picture likely reflects the relative importance of gene flow between species within lakes and genetic drift among lakes, and is in itself insufficient to distinguish contemporary admixture from shared ancestry during lakes colonization. Inferring migration events among populations enabled us to detect contemporary gene flow between sympatric dwarf and normal whitefish within Indian and Cliff lakes. However, the other inferred links connecting populations of the same species but from different lakes (*i.e.*, isolated populations) rather indicated shared genetic variation due to common ancestry. Namely, inferred links between Webster and Indian indicated the sharing of ancestral variation between WN and IN (and therefore with CN), as well as between WD and ID (and therefore with CD). This supports the view that the different populations of each species in these three lakes, which are not physically connected today thus hampering any gene flow, were genetically similar before being isolated in their respective lakes. An additional link inferred between EN and IN (IN being connected with WN and CN) confirmed that normal whitefish from East Lake share ancestral variation with other normal populations from the southern part of the contact zone.

This provides further evidence that the secondary contact inferred in East Lake has occurred between the same two glacial lineages as for the other lakes, despite the lack of Atlantic/Mississippian mitochondrial lineage in this lake. Finally, the ancestral population from East Lake was linked to the dwarf population from Indian (and therefore to WD and CD), indicating that both populations from East Lake share much of the ancestral variation originating from the Acadian lineage. This is also consistent with the genetic swamping hypothesis proposed for explaining the lack of mitochondrial polymorphism in this lake.

The analysis of overall diversity patterns performed with the dAPC (Figure 2.6) was a complementary way to disentangle remaining signals of genetic differentiation between glacial lineages (axis 3) from genetic differentiation among lakes (axis 1). On the third axis, the projection of dwarf and normal populations from Cliff Lake indicated the positions of the two least introgressed populations (and closest to their ancestral genetic background) of our dataset. Therefore, they could be used to define an Acadian (negative coordinates) and an Atlantic/Mississippian (positive coordinates) reference for comparisons with other lakes. Interestingly, both populations from East and Indian lakes occupied intermediate positions, which is concordant with a higher proportion of Acadian ancestry in these lakes, as suggested by mitochondrial data (Pigeon, Chouinard, and Bernatchez 1997b) .

In summary, the most parsimonious overall scenario supported by our analyses corresponds to a secondary contact for all lakes, with variable contributions of Acadian and Atlantic/Mississippian lineages due to demographic contingencies. The secondary contact was concomitant to population expansions in both glacial lineages, which were detected for most lakes. This is broadly consistent with the idea that the two glacial lineages were undergoing spatial expansions after the last glacial retreat, which provoked a secondary contact at the origin of parallel genetic divergence patterns across whitefish species pairs. Our results also support that population expansions were generally more pronounced for dwarf relative to normal populations, still reflected today by the higher contemporary abundance of dwarf whitefish in all lakes (L. Bernatchez, unpubl. data). This demographic imbalance also impacted the main direction of gene flow, which was more pronounced from dwarf to normal populations than the reverse. As a consequence, an important amount of shared ancestral polymorphism between dwarf and normal populations (Figure 2.1) most likely correspond to genetic variation coming from gene exchanges due to introgression between lineages, in addition to incomplete lineage sorting.

An extended framework for inferring speciation-with-gene-flow

The concept of speciation-with-gene-flow embraces a large diversity of divergence scenarios with regards to the timing of gene flow, which in turn pertains to different modes of speciation that have long been recognized in the speciation literature (Coyne & Orr 2004). Diverging populations can experience temporal variations in effective size and migration rate, which both influence the temporal dynamics of gene flow. Consequently, demographic inference methods that account for these temporal variations have the potential to provide an improved understanding of the historical demographic events that shaped the unfolding of speciation.

For the Lake Whitefish as for other species with a pan-Arctic distribution, the history of divergence has been strongly impacted by quaternary climatic oscillations (Bernatchez and Wilson 1998). Glaciations have drastically restricted the area of species distribution provoking geographic isolation among bottlenecked populations (Bernatchez et al. 1989; Ambrose 1998; Aoki et al. 2008), while inter-glacial periods have allowed secondary contacts between populations expanding from their glacial refugia (Hewitt 2001). Here, accounting for temporal variation in migration rate and N_e allowed us to determine that the secondary contact between whitefish glacial lineages has occurred contemporarily with population expansions. This later point is of prime importance for understanding the evolution of reproductive isolation, since bottlenecked populations undergoing demographic expansions are more likely to carry and even fix deleterious alleles (Luikart et al. 1998; Peischl et al. 2013; Lohmueller 2014), which could later translate into substrate for genetic incompatibilities upon secondary contact. Indeed pronounced postzygotic incompatibilities between dwarf and normal whitefish representing different glacial lineages have been documented despite relatively small overall genomic divergence between them (Lu and bernatchez 1998; Rogers & bernatchez 2006; Dion-Côté et al. 2014). Moreover, such genetic incompatibilities may associate by coupling to form stronger barriers to gene flow (Barton & de Cara 2009; Bierne et al. 2011), as proposed in Gagnaire et al. (2013a).

Another important aspect of divergence-with-gene-flow relates to the extent to which the previously described demographic effects interact with selection. The speciation genomics literature has been increasingly integrating the influence of selective processes in

historical divergence models (*e.g.*, (Roux et al. 2013; Sousa and Hey 2013; Tine et al. 2014), and more generally, in the analytical approaches to relate genomic divergence patterns to the underling evolutionary processes (Cruickshank and Hahn 2014). These selective effects can be separated in two broad categories. First, genetic barriers caused by local adaptation and reproductive isolation loci can resist introgression, hence reducing the effective migration rate at linked loci (Barton and Bengtsson 1986; Feder and Nosil 2010). The second category embraces the effect of positive (e.g. selective sweeps) (Smith and Haigh 1974) and background selection (Charlesworth et al. 1993), which cause local reductions in genetic diversity at both selected sites and linked neutral sites. These later selective effects rather correspond to a reduction in the N_e of the genomic regions influenced by selection, irrespective to the role that they play in the speciation process. Since gene flow depends both on N_e and migration rate, both types of selective effects are likely to impact genomic divergence patterns during speciation. Here, we captured these effects separately using divergence-with-gene-flow models that take into account in a simple way the effects of genetic barriers and linked selection.

Accounting for variation in effective migration rate across the genome generally improved the fits to empirical data whatever the model considered (Figure 2.4B), and the best models for all lakes also included heterogeneous migration rates. This suggests that the rate of introgression between whitefish glacial lineages has been highly variable across their genome since the beginning of secondary contact, as reported previously in other species (Tine et al. 2014; Le Moan et al. 2016; Rougemont et al. 2017). Moreover, integrating heterogeneous N_e in the models also improved model scores for the two most divergent species pairs (Cliff and Indian, Figure 2.4C). Therefore, our results also support the view that linked selection has influenced the patterns of genomic divergence in whitefish sympatric species pairs. As proposed in earlier studies, this mechanism may be particularly efficient in low-recombining chromosomal regions (Cruickshank and Hahn 2014). Some of our models (*e.g.* SC2m2N and AM2m2N) combined both genome-wide variation in N_e and m_e, as already developed within an ABC framework (Roux et al. 2016). The rationale behind this is that only models that both contain a period of isolation and gene flow enable to dissociate the influence of both sources of chromosomal variation, since only linked selection is at play during periods of geographic isolation. However, it is currently unclear

how much the signal contained in empirical polymorphism data can retain distinct signatures for these two selective effects. This will need to be addressed using simulations.

In summary, as for most models, our models remain simplifications of a probably more complex reality. Yet our approach illustrates the need to take into account both temporal and genomic variations in effective population size and migration rates when inferring the history of speciation. The 26 divergence models considered here enabled us to evaluate a large diversity of scenarios, taking each effect separately an in combination with other to improve the inference of the divergence history while controlling for model complexity.

Understanding the divergence continuum in whitefish

Lake Whitefish nascent species pairs offer a rare opportunity to understand the influence of selection and historical demography on a continuum of phenotypic and genomic divergence associated with speciation. Previous works have provided mounting evidence for the role of selection in shaping genetic and phenotypic divergence across this continuum (Rogers et al. 2002; Landry et al. 2007; Bernatchez et al. 2010; Renaut et al. 2012; Gagnaire, Pavey, et al. 2013; Laporte, Rogers, Dion-Côté, Normandeau, Gagnaire, Dalziel, Chebib, and Bernatchez 2015b). However, the role played by historical demography has never been fully resolved since previous studies largely depended on mitochondrial DNA alone.

Our study brings new evidence supporting previous findings based on mitochondrial DNA that the onset of this young radiation matched the last glacial period (Bernatchez and Dodson 1990; Bernatchez and Dodson 1991). Using a mutation rate of 10^{-8} mutations/site/generation and a generation time of 3.5 years, the average divergence date between glacial lineages was 41,600 ybp (s.d. 8,100). This is consistent with a previous study based on mitogenome sequencing (Jacobsen et al. 2012) that estimated the divergence time between 20,000 and 60,000 ybp. This corresponds to the late Wisconsin glacial episode, starting 85,000 years ago, during which the Laurentide ice sheet covered the studied region in eastern North America, with a maximum ice extent occurring around 25,000 ybp (Curry 2007). The average time of secondary contact obtained here, was dated to 11,200 years (s.d. 5,700), which also corresponds to the glacial retreat period, at which the lakes were colonized by the two glacial lineages from eastern (Acadian) and western

(Atlantic/Mississipian) glacial refuges (Curry 2007). Therefore, the inferred timing of divergence and secondary contact between glacial lineages matches relatively well the chronology of the climatic events in eastern North America.

Our results suggest that demographic differences among lakes have contributed to shaping the divergence continuum observed among the five lakes. Introgression rates tended to be higher in the least divergent species pairs, resulting into a weaker genetic differentiation. Yeaman et al. (Yeaman, Aeschbacher, et al. 2016) recently showed that the formation of genomic islands by erosion of divergence following secondary contact depends on the amount of linkage disequilibrium (LD) among selected loci and the intensity of effective migration. Here, we showed that effective migration rate was generally higher in the least differentiated lakes (Témiscouata and East), while at the same time, increased LD among genomic islands has been documented in the most divergent lakes (Gagnaire, Pavey, et al. 2013). Therefore, the divergence continuum likely implies both the antagonistic effects of divergent selection maintaining LD and introgression eroding past divergence.

Our study also provides new insights on the role of linked selection in shaping patterns of genomic divergence observed among the whitefish species pairs. Namely, we inferred that some genomic regions have experienced a reduction in N_e, as predicted under the effect of selection at linked sites (Cruickshank and Hahn 2014). The increasing proportion of genomic regions affected by Hill-Robertson effects, from the least to the most divergent lakes, indicated that the divergence continuum among lakes was also influenced by linked selection.

In the light of those observations, along with previous studies on this system, we propose that the continuum of genetic divergence in whitefish species pairs is the evolutionary result of a complex interplay between (*i*) genetic divergence between glacial lineages through lineage sorting and mutation accumulation, (*ii*) reduced introgression in genomic regions involved in reproductive isolation due to the accumulation of incompatibilities, (*iii*) divergent selection on phenotypes maintaining LD, and (*iv*) the independent contingency of demographic events among lakes. The heterogeneous landscape of species divergence in the whitefish system was thus likely built by a combination of selective and demographic factors. Our inferences allowed us to disentangle part of this complex interplay, although many aspects remain to be clarified. In particular,

whether selection at linked sites also plays a role in facilitating the accumulation of incompatibilities during allopatry isolation will need to be scrutinized into more details, as well as the role of such incompatibilities in facilitating the divergence of quantitative polygenic traits following secondary contact. This could be achieved by testing the effect of divergence on quantitative traits with and without the joint action of selection against hybrids. Indeed, a model mixing components of allopatric speciation, with the accumulation of genetic incompatibilities (*e.g.*, under-dominant mutations) and sympatric speciation (*i.e.*, local adaptation involving divergent selection on quantitative traits), would differ from the coupling hypothesis model which mostly considers local adaptation loci of relatively strong effect (Anon 2011). We argue that this kind of models could be relevant for some systems in which sympatric speciation after admixture, or parallel hybrid speciation, has been inferred without explicitly testing a single divergence event with recent secondary gene flow (Kautt et al. 2016; Meier et al. 2016). We also believe that demographic inferences approaches should systematically include basic scenarios of divergence, extended by models with increasing levels of complexity to address demographic and selective effects separately and then combined, and not only focus on the *a priori* history of the system. To conclude, this study illustrates the potential benefits of applying a modeling framework to disentangle the relative role of demography and selection, towards elucidating the complexity of species divergence in any other taxonomic group.

2.7 Acknowledgements

We would like to thank Nicolas Bierne, Anne-Laure Ferchaud, Martin Laporte, Anne-Marie Dion-Côté and Charles Perrier for insightful discussions, Thierry Gosselin for *stackr* inputs, as well as Anne C. Dalziel for commenting an earlier version of this manuscript. We are grateful to Guillaume Côté, Melissa L. Evans, William Adam and the staff of Maine Department of Inland Fisheries & Wildlife (David J. Basley and Jeremiah Wood) for all of their help sampling whitefish and sharing information about watersheds histories, and to Kim Præbel for *Coregonus lavaretus* RAD-seq data. We are also grateful to Editor Bill Martin and Associate Editor Judith E. Mank as well as three anonymous reviewers for their constructive inputs, which improved a previous version of this paper. This research was supported by a discovery research grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to L.B. L.B also holds the Canadian Research Chair

in genomics and conservation of aquatic resources, which funded the research infrastructure for this project.

2.8 Figures

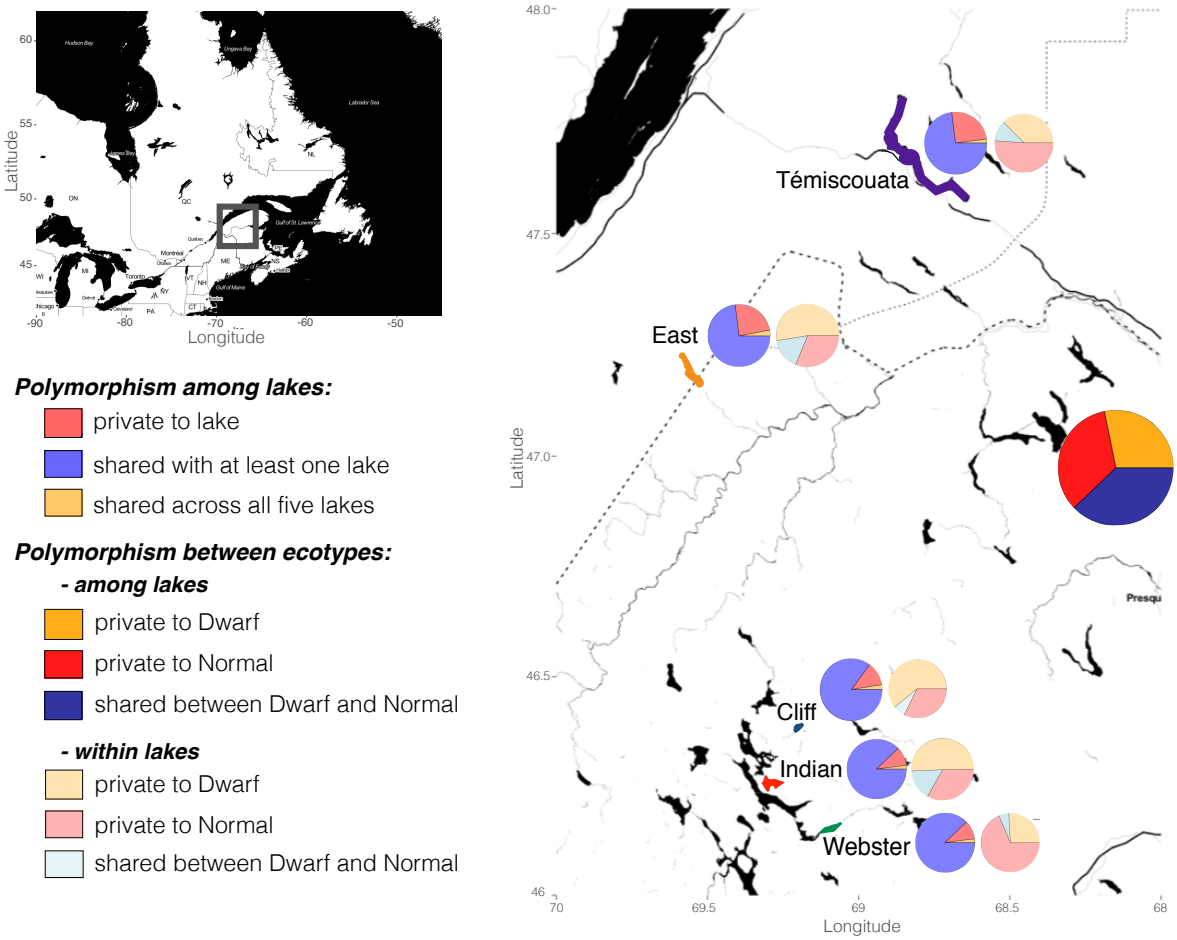


Figure 2.1: Geographic locations of the lakes where sympatric whitefish species pairs were sampled, and overview of the extent of shared vs. private polymorphism. Pie charts illustrate the amount of shared and private SNPs among lakes as well as between species among or within lakes.

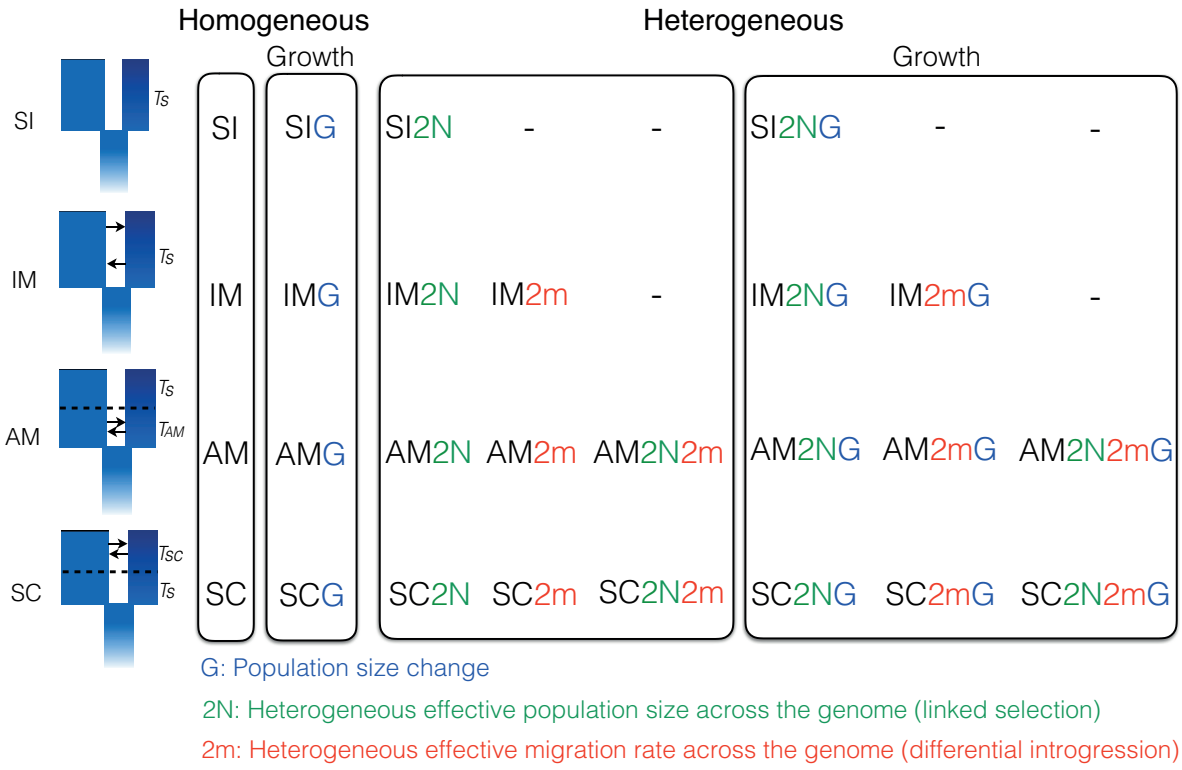


Figure 2.2: The 26 models implemented in this study. The models implemented here represent extensions of the four classical models of divergence: 'Strict Isolation' (-SI), 'Isolation with Migration' (-IM), 'Ancient Migration' (-AM) and 'Secondary Contact' (-SC). Briefly, T_S corresponds to the duration of complete isolation between diverging populations and T_{AM} and T_{SC} correspond to the duration of gene flow in AM and SC models, respectively. The first extension of these four models accounts for temporal variation in effective populations size (-G models), allowing independent expansion/contraction of the diverging populations. The last categories correspond to 'Heterogeneous gene flow' models, which integrate parameters allowing genomic variations in effective migration rate (-2m), effective population size (-2N) or both (-2m2N) to account for genetic barriers and selection at linked sites.

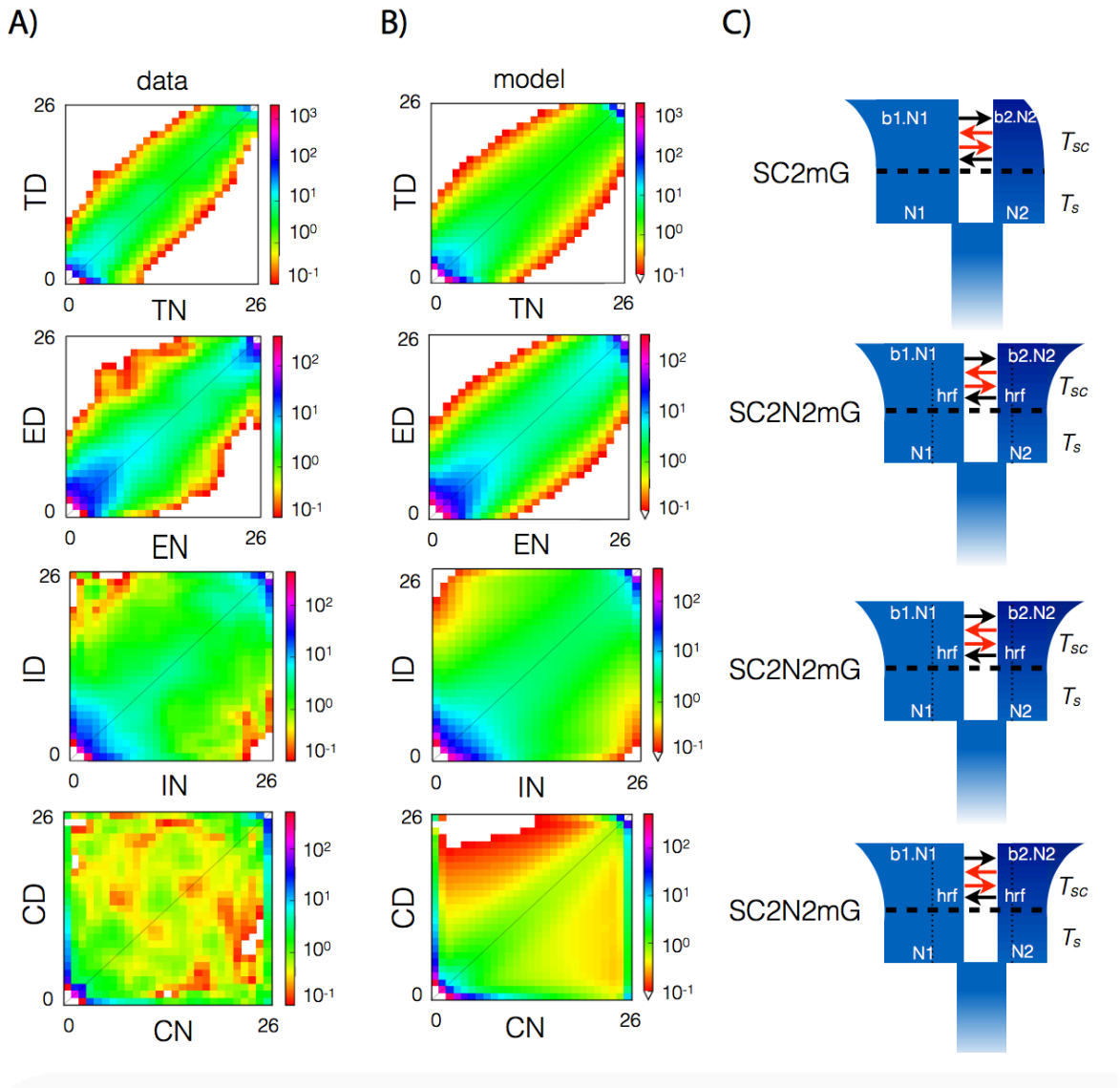


Figure 2.3: Historical demography of the Lake Whitefish species pairs. (**A**) Observed joint allele frequency spectrum (JAFS) for normal (-N; x-axis) and dwarf (-D; y-axis) populations for each of four lakes (T=Témiscouata, E=East, I=Indian and C=Cliff), obtained by projection of empirical data to 13 diploid individuals per population. For each JAFS, the color scale indicates the number of SNPs falling in each bin defined by a unique combination of the number of derived allele observed in normal and dwarf populations. (**B**) Predicted JAFS of the fittest model for each lake. (**C**) Representation of the fittest model for each lake.

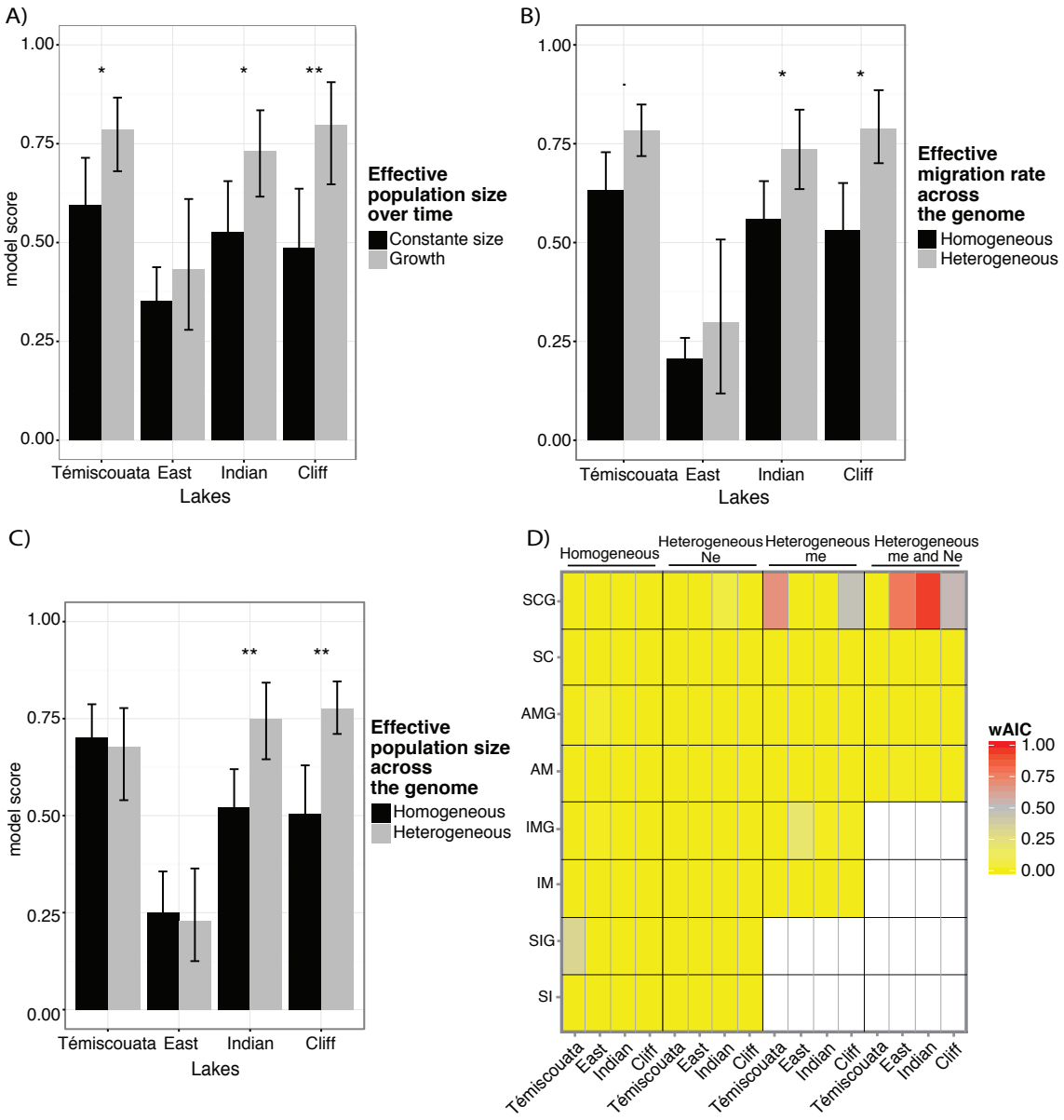


Figure 2.4: Model comparisons. Barplots showing the effect of taking into account a particular demographic or selective aspect in the models, assessed using model scores, with (**A**) the effect of including temporal variation in population effective size (-G), (**B**) heterogeneous migration rates among loci (-2m) and (**C**) heterogeneous effective population sizes among loci (-2N). The vertical bars indicate the variance of model scores within a given category of models and asterisks represent significant differences in average model scores between the compared categories of models. (**D**) Heat-map of the weighted AIC (w_{AIC}) showing the relative weights of the 26 models for each lake. The color scale corresponds to w_{AIC} values ranging from 0 to 1. Warmer colors indicate the best models.

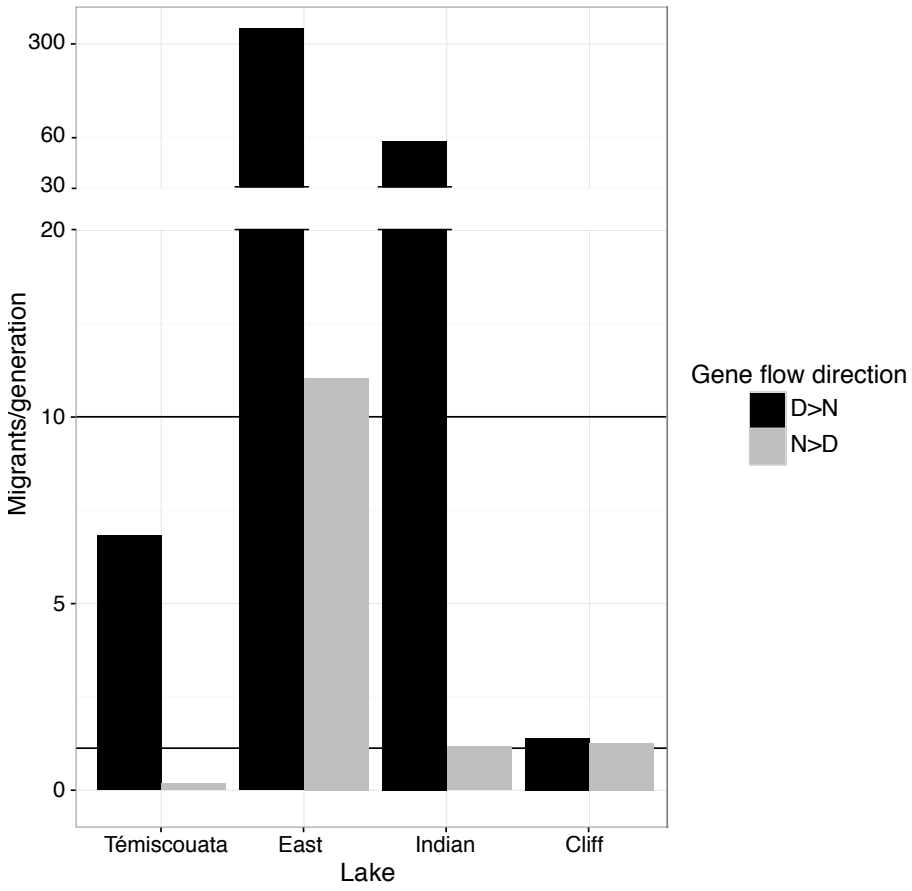


Figure 2.5: Asymmetrical effective gene flow between normal and dwarf whitefish within lakes. Bar plot of the effective number of migrants per generation in both directions from dwarf to normal (black) and reciprocally (gray), obtained from estimated parameters of the fittest model for each lake, using the average gene flow formula (average-gene flow = $N^*b^*(P^*m_e+(1-P)^*m_{e'})$ in each direction.

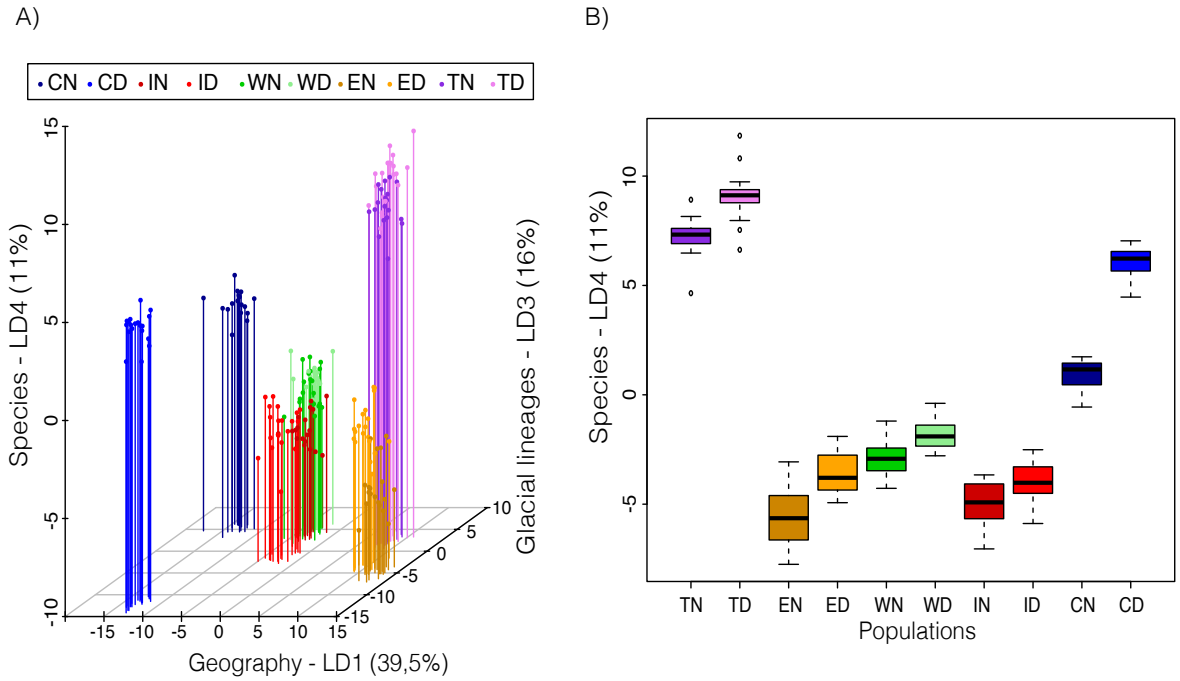


Figure 2.6: Genetic structure and relationship among lakes and species. (**A**) Discriminant analysis of principal components (dAPC) of the different lakes (Cliff, Indian, Webster, East, Témiscouata) for either Dwarf or Normal whitefish (D or N), representing three dimensional relationships among populations. The first axis (LD1, 39.5%) captures the geographical signal of differentiation among lakes. The third dAPC axis (LD3, 16%) separates species pairs according to their residual genetic promixity to ancient glacial lineages represented by the least introgressed populations from Cliff Lake. Positive coordinates represent populations with high proportions of Atlantic ancestry whereas negative coordinates reflect increased proportions of Acadian ancestry. The fourth axis (LD4, 11%) tends to separate species pairs within each lake. The second dAPC axis LD2 is not shown here to avoid partial redundancy with LD1, but is provided as a supplementary material. (**B**) Boxplot of individual coordinates along the fourth dAPC axis (LD4), highlighting the divergence parallelism between species among lakes.

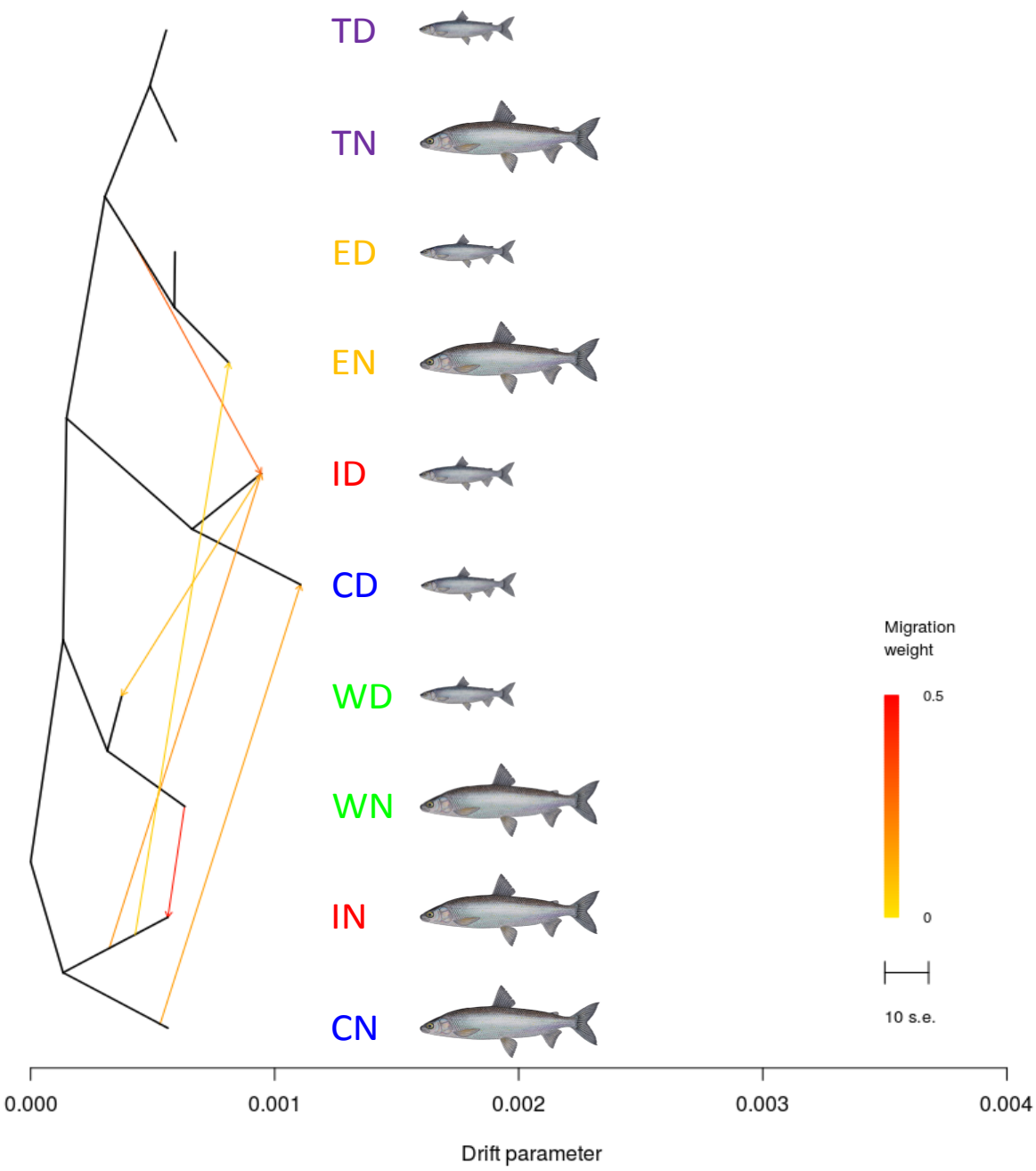


Figure 2.7: Shared ancestral genetic variation between allopatric populations and admixture events between sympatric species pairs. The least introgressed normal whitefish population from Cliff Lake was used to root the tree. Horizontal branch lengths are proportional to the amount of genetic drift in each branch, and the scale bar indicates 10 times the average standard error (s.e.) of the entries in the covariance matrix between pairs of populations. Color-scale indicates the weight of inferred migration events.

2.9 Supplementary tables

Table S2.1: Results of model fitting. For all lakes, details of statistics and demographic parameter values for the fittest models determined under the threshold of $\Delta AIC_i < 10$. The table contains in this order the maximum likelihood (MLE) for each model, the value of Akaike information criterion (AIC), the ΔAIC value and the weighted AIC (w_{AIC}). Then, the inferred raw demographic parameter values including: Theta (θ); the ancestral effective population size before population split (Nref); the effective population size after split for dwarf (N1) and normal (N2) populations; the growth coefficient for dwarf (b1) and normal (b2) populations. The b parameter is defined as the ratio of contemporary to ancestral effective population size (ancestral meaning after splitting time). Population exponential growth is associated with bi>1 and reduction in population effective size with bi<1. The Hill-Robertson factor (hrf) corresponds to the degree to which the effective population size of the diverging populations (not considering the ancestral population) is locally reduced due to the effect of linked selection. Migration parameters include migration rates from normal population to dwarf population (me12) and reciprocally (me21), and a second category of effective migration rates (me'12) and (me'21) applying to a second category of loci. Time parameters include the duration (in years) of the allopatric divergence period (Tsplit), and the duration of the migration period (i.e., TAM for the AM models and TSC for the secondary contact models). Finally, the table also contains proportion parameters such as the proportion (Q) of the genome with effective population sizes N1 and N2 (a second category of loci occupying a fraction 1-Q of the genome has effective population sizes b1N1 and b2N2). A proportion (P) of the genome is occupied by loci with effective migration rates me12 and me21 (and a second category of loci occupying a fraction 1-P of the genome has effective population sizes m'e12 and m'e21). The parameter (O) is the proportion of correct SNP orientation. Numbers into brackets denote 95% confidence intervals obtained using the MLE parameter values ±2 s.e.

LAKE	MODEL	MLE	AIC	ΔAIC	*w*AIC	*θ*	Nref	N_1	N_2	b_1	b_1	*hrf*	m_{e12}	m_{e21}	$m_{e'12}$	$m_{e'21}$	*Tsplit*	*Tpost-split*	*P*	*Q*	*O*
Témiscouata	SC2mG	-1656.47	3336.93	-	0.68	532.15	7611.39	110.8	42.1	20.3	0.3	-	0.011	0.000	0.000	0.042	0.56	0.14	0.93	-	0.99
								[32.7; 189.0]	[0.0; 84.6]	[0.0; 68.6]	[0.0; 32.4]	-	[0.000; 21.107]	[0.000; 21.680]	[0.000; 3.610]	[0.000; 3.858]	[0.00; 3.39]	[0.00; 0.91]	[0.22; 0.95]	-	[0.99; 0.99]
Témiscouata	SIG	-1663.23	3338.47	1.54	0.32	570.56	4350.13	33.4	49.2	95.4	0.2	-	-	-	-	-	0.97	-	-	-	0.99
								[21.7; 45.0]	[47.1; 51.4]	[81.0; 109.8]	[0.0; 1.5]	-	-	-	-	-	[0.91; 1.02]	-	-	-	[0.84; 0.99]
East	SC2N2mG	-909.87	1847.74	-	0.77	212.86	1622.96	0.4	0.4	37.0	2.2	0.1	18.900	37.344	1.979	1.249	3.16	1.73	0.60	0.16	0.98
								[0.0; 9.5]	[0.0; 6.8]	[0.0; 104.7]	[0.0; 41.1]	[0.0; 0.6]	[7.861; 29.940]	[15.639; 59.050]	[0.000; 7.340]	[0.000; 5.661]	[0.00; 7.50]	[0.93; 2.52]	[0.26; 0.94]	[0.01; 0.37]	[0.96; 0.98]
East	IM2mG	-914.25	1850.50	2.77	0.19	982.85	7493.63	1.0	0.1	1.8	84.9	-	14.780	0.007	0.046	5.821	0.28	-	0.77	-	0.97
								[0.0; 2.2]	[0.0; 1.2]	[0.0; 3.7]	[82.1; 87.7]	-	[14.570; 14.991]	[0.000; 4.195]	[0.000; 0.546]	[5.493; 6.148]	[0.00; 0.65]	-	[0.59; 0.95]	-	[0.94; 0.99]
East	AM2N2m	-916.30	1856.60	7.71	0.02	406.95	3102.73	8.6	4.3	-	-	0.0	26.409	2.112	0.001	4.454	0.19	0.00	0.06	0.50	0.96
								[0.0; 12.2]	[0.0; 5.1]	-	-	[0.0; 0.3]	[14.941; 37.877]	[0.000; 12.290]	[0.000; 3.853]	[0.000; 12.517]	[0.06; 0.33]	[0.00; 0.13]	[0.87; 0.86]	[0.26; 0.74]	[0.96; 0.96]
East	AMG	-918.72	1855.45	8.86	0.01	775.93	5915.95	0.6	0.2	25.8	96.3	-	11.344	0.037	-	-	0.21	0.04	-	-	0.96
								[0.0; 40.0]	[0.0; 11.2]	[7.0; 44.6]	[53.6; 139.0]	-	[0.260; 22.428]	[0.000; 1.202]	-	-	[0.08; 0.34]	[0.00; 0.15]	-	-	[0.95; 0.97]
Indian	SC2N2mG	-1089.03	2206.07	-	0.94	158.61	933.35	0.6	0.3	64.0	6.3	0.2	0.809	1.906	0.155	1.014	4.50	1.31	0.60	0.41	0.97
								[0.6; 0.6]	[0.0; 2.7]	[62.1; 65.9]	[4.1; 8.6]	[0.0; 0.4]	[0.000; 1.708]	[1.053; 2.760]	[0.000; 1.098]	[0.994; 1.033]	[3.65; 5.35]	[0.96; 1.65]	[0.19; 0.95]	[0.39; 0.42]	[0.94; 1.00]
Indian	SC2NG	-1094.92	2211.85	5.78	0.05	1201.86	7072.67	0.2	0.2	43.2	4.4	0.2	1.676	5.703	-	-	0.13	0.34	-	0.50	0.95
								[0.0; 28.7]	[0.0; 70.8]	[42.4; 44.0]	[0.0; 77.7]	[0.0; 0.4]	[0.000; 42.686]	[0.000; 48.432]	-	-	[0.11; 0.15]	[0.00; 8.00]	-	[0.48; 0.50]	[0.29; 0.99]
Cliff	SC2N2mG	-1006.59	2039.18	-	0.53	123.06	1482.41	4.1	0.9	24.7	19.6	0.2	0.088	0.024	0.074	0.008	3.12	0.89	0.48	0.40	0.99
								[1.9; 6.4]	[0.4; 1.4]	[20.9; 28.5]	[16.9; 22.3]	[0.0; 1.0]	[0.000; 1.159]	[0.000; 1.894]	[0.000; 0.494]	[0.000; 7.252]	[2.79; 3.45]	[0.84; 0.94]	[0.36; 0.61]	[0.00; 2.17]	[0.98; 0.99]
Cliff	SC2mG	-1007.72	2039.43	0.25	0.46	232.67	2802.70	7.3	1.0	3.6	10.7	-	14.629	34.813	0.085	0.006	0.65	0.49	0.05	-	0.96
								[0.0; 53.0]	[0.0; 12.1]	[0.0; 132.8]	[0.0; 86.9]	-	[0.000; 35.145]	[14.934; 54.692]	[0.000; 3.306]	[0.000; 2.323]	[0.11; 1.20]	[0.00; 0.99]	[0.00; 0.26]	-	[0.95; 0.97]

Table S2.2: Converted model parameter values for the best fit models in each lake. For all lakes, details of statistics and demographic parameter values for the fittest models determined under the threshold of $\Delta AIC_i < 10$. The table contains in this order the maximum likelihood (MLE) for each model, the value of Akaike information criterion (AIC), the ΔAIC value and the weighted AIC (w_{AIC}). Then, the inferred demographic parameter values converted with the estimate of Theta (θ): the ancestral effective population size before population split (Nref); the effective population size after split for dwarf (N_1) and normal (N_2) populations; the growth coefficient for dwarf (b_1) and normal (b_2) populations. The *b* parameter defined as a ratio of contemporary to ancestral effective population size (ancestral meaning after splitting time). Population exponential growth is associated with b_i>1 and reduction in population effective size with b_i<1. The Hill-Robertson factor (*hrf*) corresponds to the degree to which the effective population size of the diverging populations (not considering the ancestral population) is locally reduced due to the effect of linked selection. Migration parameters include migration rates from normal population to dwarf population (m_{e12}) and reciprocally (m_{e21}), and a second category of effective migration rates ($m_{e'12}$) and ($m_{e'21}$) applying to a second category of loci. Time parameters include the duration (in years) of the allopatric divergence period (Tsplit), and the duration of the migration period (*i.e.*, T_{AM} for the AM models and T_{SC} for the secondary contact models). Finally, the table also contains proportion parameters such as the proportion (*Q*) of the genome with effective population sizes N_1 and N_2 (a second category of loci occupying a fraction 1-*Q* of the genome has effective population sizes b_1N_1 and b_2N_2). A proportion (*P*) of the genome is occupied by loci with effective migration rates m_{e12} and m_{e21} (and a second category of loci occupying a fraction 1-*P* of the genome has effective population sizes m'_{e12} and m'_{e21}). The parameter (*O*) is the proportion of correct SNP orientation. The estimated value of each parameter was converted so that migration rates represent the fraction of a population replaced by migrants every generation, and temporal parameters appear in years. Numbers into brackets denote 95% confidence intervals obtained using the MLE parameter values ±2 s.e.

LAKE	MODEL	MLE	AIC	ΔAIC	wAIC	θ	Nref	N_1	N_2	b_1	b_1	*hrf*	m_{e12}	m_{e21}	$m_{e'12}$	$m_{e'21}$	*Tsplit*	*Tpost-split*	*P*	*Q*	*O*
Témiscouata	SC2mG	-1656.47	3336.93	-	0.68	532.15	7611.39	2952603.4	1120233.2	20.3	0.3	-	0.000	0.000	0.000	0.000	29996.49	7246.04	0.93	-	0.99
								[870315.0; 5034891.7]	[266.4; 2253298.1]	[0.0; 68.6]	[0.0; 32.4]	-	[0.000; 0.005]	[0.000; 0.005]	[0.000; 0.001]	[0.000; 0.001]	[0.00; 180578.93]	[0.00; 48478.87]	[0.22; 0.95]	-	[0.99; 0.99]
Témiscouata	SIG	-1663.23	3338.47	1.54	0.32	570.56	4350.13	507815.2	749808.8	95.4	0.2	-	-	-	-	-	29506.96	-	-	-	0.99
								[330990.9; 684639.4]	[717244.8; 782372.8]	[81.0; 109.8]	[0.0; 1.5]	-	-	-	-	-	[48741.01; 31157.00]	-	-	-	[0.84; 0.99]
East	SC2N2mG	-909.87	1847.74	-	0.77	212.86	1622.96	2325.3	2453.9	37.0	2.2	0.11	0.020	0.040	0.002	0.001	35907.02	19601.58	0.60	0.16	0.98
								[56.8; 54009.3]	[56.8; 38360.6]	[0.0; 104.7]	[0.0; 41.1]	[0.01; 0.56]	[0.008; 0.032]	[0.017; 0.064]	[0.000; 0.008]	[0.000; 0.006]	[0.00; 85224.95]	[10542.92; 28660.24]	[0.26; 0.94]	[0.01; 0.37]	[0.96; 0.98]
East	IM2mG	-914.25	1850.50	2.77	0.19	982.85	7493.63	27310.7	2249.4	1.8	84.9	-	0.003	0.000	0.000	0.001	14597.50	-	0.77	-	0.97
								[262.3; 57150.4]	[262.3; 30703.2]	[0.0; 3.7]	[82.1; 87.7]	-	[0.000; 0.004]	[0.000; 0.001]	[0.000; 0.000]	[0.001; 0.001]	[0.00; 33933.42]	-	[0.59; 0.95]	-	[0.94; 0.99]
East	AM2N2m	-916.30	1856.60	7.71	0.02	406.95	3102.73	93159.0	47124.4	-	-	0.04	0.015	0.001	0.000	0.003	4204.02	0.00	0.06	0.50	0.96
								[108.6; 131971.6]	[108.6; 55127.0]	-	-	[0.01; 0.34]	[0.008; 0.021]	[0.000; 0.007]	[0.000; 0.002]	[0.000; 0.007]	[3263.59; 7077.67]	[0.00; 2730.24]	[0.87; 0.86]	[0.26; 0.74]	[0.96; 0.96]
East	AMG	-918.72	1855.45	8.86	0.01	775.93	5915.95	12292.9	4330.7	25.8	96.3	-	0.003	0.000	-	-	8725.63	1536.75	-	-	0.96
								[207.1; 827979.7]	[207.1; 232806.3]	[7.0; 44.6]	[53.6; 139.0]	-	[0.000; 0.007]	[0.000; 0.000]	-	-	[4276.72; 14127.18]	[0.00; 6368.43]	-	-	[0.95; 0.97]
Indian	SC2N2mG	-1089.03	2206.07	-	0.94	158.61	933.35	2048.1	957.3	64.0	6.3	0.20	0.002	0.004	0.000	0.002	29401.99	8541.27	0.60	0.41	0.97
								[1984.1; 2112.1]	[32.7; 8971.6]	[62.1; 65.9]	[4.1; 8.6]	[0.01; 0.43]	[0.000; 0.003]	[0.002; 0.005]	[0.000; 0.002]	[0.002; 0.002]	[23833.27; 34970.72]	[6279.02; 10803.53]	[0.19; 0.95]	[0.39; 0.42]	[0.94; 1.00]
Indian	SC2NG	-1094.92	2211.85	5.78	0.05	1201.86	7072.67	5306.7	4851.4	43.2	4.4	0.19	0.000	0.001	-	-	6230.14	16705.74	-	0.50	0.95
								[247.5; 709290.9]	[247.5; 1753775.2]	[42.4; 44.0]	[0.0; 77.7]	[0.01; 0.38]	[0.000; 0.011]	[0.000; 0.012]	-	-	[5259.77; 7200.51]	[0.00; 396069.76]	-	[0.48; 0.50]	[0.29; 0.99]
Cliff	SC2N2mG	-1006.59	2039.18	-	0.53	123.06	1482.41	21376.3	4560.6	24.7	19.6	0.17	0.000	0.000	0.000	0.000	32375.74	9245.76	0.48	0.40	0.99
								[9640.8; 33111.8]	[1982.4; 7138.8]	[20.9; 28.5]	[16.9; 22.3]	[0.00; 1.00]	[0.000; 0.001]	[0.000; 0.002]	[0.000; 0.001]	[0.000; 0.009]	[28981.58; 35769.90]	[8722.07; 9769.46]	[0.36; 0.61]	[0.00; 2.17]	[0.98; 0.99]
Cliff	SC2mG	-1007.72	2039.43	0.25	0.46	232.67	2802.70	71589.0	9480.8	3.6	10.7	-	0.009	0.022	0.000	0.000	12818.87	9679.91	0.05	-	0.96
								[98.1; 519641.2]	[98.1; 118734.2]	[0.0; 132.8]	[0.0; 86.9]	-	[0.000; 0.022]	[0.009; 0.034]	[0.000; 0.002]	[0.000; 0.001]	[2190.82; 23446.92]	[0.00; 19505.76]	[0.00; 0.26]	-	[0.95; 0.97]

Table S2.3: Summary statistics and parameter estimations for the 26 models per lake. For all lakes, summary statistics and model parameter values are provided for the fittest run obtained from 25 independent runs of optimisation for each model for each lake. Models are ranked according to their Akaike information criterion (AIC) value. Also provided are the ΔAIC value for the corresponding model and the weighted AIC (w_{AIC}). Inferred demographic parameter values are scaled by Theta (θ): the ancestral effective population size before population split (Nref); the effective population size after split for dwarf (N_1) and normal (N_2) populations; the growth coefficient for dwarf (b_1) and normal (b_2) populations; the Hill-Robertson factor (*hrf*); effective migrations rates from normal population to dwarf population (m_{e12}) and reciprocally (m_{e21}) for the first category of loci, and for the second category ($m_{e'12}$ and $m_{e'21}$); Allopatric isolation duration parameter (*Ts*) and migration duration parameters (T_{AM} for AM models and T_{SC} for SC models). Finally, the table also contains proportion parameters for the effect of linked selection (*Q*), semi-permeability (*P*) and correct orientation (*O*).

LAKE	MODEL	MLE	AIC	THETA	nu1	nu2	b1	b2	hrf	m12	m21	me12	me21	Tsplit	Tpost-split	P	Q	O
TEMISCOUATA	SC2mG	-1656,47	3336,93	532,15	110,834	42,051	20,263	0,255	-	0,011	0,000	0,000	0,042	0,563	0,136	0,933	-	0,990
TEMISCOUATA	SIG	-1663,23	3338,47	570,56	33,353	49,247	95,361	0,225	-	-	-	-	-	0,969	-	-	-	0,990
TEMISCOUATA	SI2NG	-1715,82	3447,64	570,30	50,758	25,073	42,398	0,679	0,012	-	-	-	-	1,000	-	-	0,014	0,989
TEMISCOUATA	AMG	-1718,01	3454,02	587,19	27,899	32,769	35,792	0,522	-	0,092	0,867	-	-	0,003	0,934	-	-	0,989
TEMISCOUATA	SC2N2m	-1741,62	3507,24	440,01	0,339	97,904	-	-	0,209	5,078	56,919	1,410	0,734	7,082	0,507	0,715	0,445	0,990
TEMISCOUATA	AM2N2mG	-1743,68	3515,36	271,69	57,220	70,037	3,750	0,097	0,024	8,188	0,012	1,331	0,000	0,627	0,402	0,527	0,027	0,988
TEMISCOUATA	AM2m	-1753,04	3526,08	564,13	205,612	18,805	-	-	-	4,990	2,339	1,522	0,010	0,000	0,992	0,784	-	0,989
TEMISCOUATA	IM2mG	-1752,97	3527,94	611,96	42,608	23,046	56,470	0,489	-	0,000	13,899	0,060	0,006	0,879	-	0,082	-	0,990
TEMISCOUATA	IMG	-1761,00	3538,00	529,77	75,249	12,450	9,004	2,468	-	0,093	0,022	-	-	1,166	-	-	-	0,989
TEMISCOUATA	SC2NG	-1767,61	3557,21	661,81	87,355	23,160	28,158	0,141	0,953	0,464	0,025	-	-	0,586	0,128	0,020	0,990	
TEMISCOUATA	AM2mG	-1776,19	3576,38	571,39	20,183	28,587	22,612	0,691	-	3,718	0,000	4,101	0,146	0,002	0,975	0,946	-	0,990
TEMISCOUATA	SI2N	-1789,46	3590,92	601,11	98,815	22,679	-	0,000	0,222	-	-	-	-	0,920	-	-	-	0,996
TEMISCOUATA	IM2NG	-1804,32	3628,63	609,90	93,455	22,615	3,087	0,542	0,159	0,242	0,273	-	-	0,912	-	-	0,019	0,989
TEMISCOUATA	IM2m	-1813,77	3645,53	598,24	188,779	18,542	-	-	-	0,411	0,044	0,123	0,316	0,953	-	0,707	-	0,989
TEMISCOUATA	SI	-1829,71	3667,43	574,92	94,643	18,965	-	-	-	-	-	-	-	0,936	-	-	-	0,990
TEMISCOUATA	SC2N2mG	-1836,52	3701,04	327,43	60,519	13,571	15,374	0,844	0,543	0,302	0,304	0,000	0,062	0,000	0,796	0,586	0,055	0,990
TEMISCOUATA	AM2N	-1843,36	3704,72	595,25	77,826	18,625	-	-	0,541	4,287	0,000	0,360	-	0,542	-	-	0,026	0,989
TEMISCOUATA	SC2m	-1846,09	3712,18	595,10	93,013	16,693	-	-	-	8,002	1,324	0,031	0,011	0,873	0,000	0,271	-	0,990
TEMISCOUATA	AM	-1864,87	3743,74	595,95	74,613	16,193	-	-	-	0,288	0,017	-	-	0,823	0,063	-	-	0,990
TEMISCOUATA	AM2N2m	-1860,21	3744,42	282,46	89,717	19,353	-	-	0,370	0,584	0,000	12,295	0,968	0,928	0,058	0,210	0,079	0,990
TEMISCOUATA	SC2N	-1873,63	3765,26	641,56	116,008	14,745	-	-	0,236	1,821	2,407	-	-	0,763	0,000	-	0,015	0,988
TEMISCOUATA	SC	-1896,33	3806,66	613,99	66,019	16,112	-	-	-	0,033	0,000	-	-	0,008	0,805	-	-	0,989
TEMISCOUATA	SCG	-2061,74	4141,49	297,77	0,944	33,118	1,263	0,327	0,766	1,740	-	-	-	4,064	0,792	-	-	0,986
TEMISCOUATA	IM	-2188,12	4388,24	677,45	21,798	11,870	-	-	-	0,002	0,099	-	-	0,608	-	-	-	0,988
TEMISCOUATA	IM2N	-2219,56	4455,12	402,78	1,776	16,341	-	-	0,679	0,225	1,297	-	-	3,911	-	-	0,069	0,990
TEMISCOUATA	AM2NG	-1865,58	3753,16	514,41	19,774	22,151	21,374	0,604	0,737	13,488	0,936			0,193	0,960		0,183	0,989
EST	SC2N2mG	-909,87	1847,74	212,86	0,409	0,432	36,967	2,193	0,110	18,900	37,344	1,979	1,249	3,161	1,725	0,599	0,156	0,984
EST	IM2mG	-914,25	1850,50	982,85	1,041	0,086	1,795	84,878	-	14,780	0,007	0,046	5,821	0,278	-	0,768	-	0,966
EST	AM2N2m	-916,30	1856,60	406,95	8,579	4,339	-	-	0,035	26,409	2,112	0,001	4,454	0,194	0,000	0,056	0,498	0,962
EST	AMG	-918,72	1855,45	775,93	0,594	0,209	25,827	96,319	-	11,344	0,037	-	-	0,211	0,037	-	-	0,957
EST	IM2NG	-920,00	1860,00	1068,61	8,096	0,081	0,071	73,507	0,254	13,887	0,922	-	-	0,275	-	-	0,010	0,972
EST	SC2mG	-920,80	1865,61	700,13	6,957	8,791	8,089	0,011		35,227	0,000	0,001	0,001	0,211	0,000	0,122		0,966
EST	IMG	-920,91	1857,82	1307,48	10,156	0,062	0,044	85,351	-	18,401	0,000	-	-	0,253	-	-	-	0,975
EST	IM2m	-922,37	1862,73	716,03	6,059	3,188	-	-	-	55,951	63,428	0,029	0,089	0,138	-	0,083	-	0,958
EST	SC2NG	-924,11	1870,23	718,50	18,325	6,174	0,245	0,467	0,157	0,424	0,046	-	-	0,007	0,146	-	0,034	0,956
EST	IM2N	-924,14	1864,27	721,19	10,103	4,114	-	-	0,173	0,296	0,000	-	-	0,161	-	-	0,043	0,961
EST	AM2N	-924,18	1866,36	721,33	10,085	4,131	-	-	0,175	0,262	0,013	-	-	0,160	0,000	-	0,043	0,961
EST	SC2N	-924,19	1866,37	719,87	9,655	4,093	-	-	0,169	0,285	0,000	-	-	0,000	0,161	-	0,035	0,961
EST	AM2N2mG	-924,31	1876,62	359,51	10,781	3,922	0,773	1,047	0,158	7,370	0,000	0,000	0,289	0,065	0,098	0,154	0,088	0,960
EST	SC2N2m	-924,63	1873,26	359,42	7,863	4,033	-	-	0,156	0,546	0,000	0,000	0,813	0,000	0,160	0,948	0,058	0,959
EST	SI2NG	-924,94	1865,89	723,07	32,541	2,415	0,318	2,523	0,193	-	-	0,164	-	-	-	-	0,052	0,964
EST	AM2mG	-925,37	1874,75	719,43	6,293	4,478	2,088	0,011	-	0,134	0,015	0,258	0,015	0,149	0,000	0,060	-	0,956
EST	SI2N	-925,40	1862,80	721,98	10,477	4,096	-	0,038	0,180	-	-	0,157	-	-	-	-	-	0,961
EST	AM2NG	-925,53	1873,07	725,91	8,357	1,973	3,403	3,704	0,205	25,737	0,000	-	-	0,000	0,169	-	0,039	0,967
EST	SCG	-925,97	1869,94	719,09	10,958	4,602	0,007	0,022	-	2,111	0,499	-	-	0,135	0,002	-	-	0,950
EST	IM	-926,81	1865,62	720,27	7,182	3,746	-	-	-	0,218	0,000	-	-	0,154	-	-	-	0,958
EST	AM	-926,82	1867,64	719,72	7,444	3,690	-	-	-	0,242	0,000	-	-	0,153	0,001	-	-	0,958
EST	SC2m	-926,84	1873,68	720,18	7,229	3,754	-	-	-	0,000	0,000	0,212	0,000	0,000	0,154	0,055	-	0,958
EST	SC	-926,92	1867,85	719,77	7,192	3,803	-	-	-	0,201	0,000	-	-	0,021	0,133	-	-	0,959
EST	AM2m	-927,00	1874,01	719,66	7,162	3,768	-	-	-	0,001	0,213	0,125	0,370	0,129	0,024	0,937		0,958
EST	SIG	-927,30	1866,59	722,93	7,112	2,290	1,722	2,363	-	-	-	0,159	-	-	-	-	-	0,962
EST	SI	-927,47	1862,94	721,02	7,842	3,692	-	-	-	-	-	0,152	-	-	-	-	-	0,959
INDIAN	SC2N2mG	-1089,03	2206,07	158,61	0,627	0,293	63,983	6,320	0,200	0,809	1,906	0,155	1,014	4,500	1,307	0,597	0,405	0,967
INDIAN	SC2NG	-1094,92	2211,85	1201,86	0,214	0,196	43,187	4,390	0,185	1,676	5,703	-	-	0,126	0,337	-	0,497	0,954
INDIAN	IM2mG	-1097,36	2216,72	1033,20	0,085	0,203	97,558	2,372	-	0,747	12,454	0,999	2,161	0,378	-	0,643	-	0,963
INDIAN	IM2NG	-1109,26	2238,51	789,68	1,418	0,539	10,748	7,966	0,108	0,959	2,400	-	-	0,331	-	-	0,494	0,953
INDIAN	AM2mG	-1109,85	2243,70	1669,74	0,398	0,197	9,359	30,228	-	0,000	29,521	15,486	0,010	0,420	0,109	0,411	-	0,982
INDIAN	SC2N	-1122,51	2263,02	727,32	5,482	2,295	-	-	0,130	2,220	4,670	-	-	0,214	0,048	-	0,400	0,943
INDIAN	IM2N	-1126,84	2269,67	739,28	5,336	2,056	-	-	0,122	1,051	3,508	-	-	0,313	-	-	0,498	0,941
INDIAN	SC2N2m	-1122,87	2269,74	370,69	3,827	1,134	-	-	0,105	1,980	7,041	0,125	0,000	0,190	0,062	0,621	0,492	0,936
INDIAN	IMG	-1126,89	2269,77	1027,58	0,087	0,088	99,196	4,902		0,900	5,412	-	-	-	0,398	-	-	0,963
INDIAN	AM2N	-1126,60	2271,19	748,55	5,449	1,794	-	-	0,115	1,006	4,186	0,315	-	0,000	-	-	0,499	0,938
INDIAN	AM2N2mG	-1129,46	2286,93	399,77	3,975	0,961	0,351	3,274	0,070	0,446	11,366	1,125	0,013	0,231	0,001	0,127	0,488	0,943
INDIAN	AMG	-1135,33	2288,66	966,36	0,113	0,084	79,583	5,089	-	0,809	5,715	-	-	0,397	0,000	-	-	0,955
INDIAN	AM2N2m	-1133,39	2290,79	391,41	3,757	1,204	-	-	0,072	1,563	7,570	2,764	2,231	0,221	0,000	0,221	0,474	0,941
INDIAN	SC2mG	-1147,39	2318,78	1066,86	0,338	6,184	34,144	0,010	-	0,958	24,798	2,337	0,411	0,296	0,000	0,915	0,968	-
INDIAN	SC2m	-1152,52	2325,04	713,24	3,979	0,492	-	-	-	0,103	56,955	1,354	4,513	0,199	0,024	0,346	-	0,936
INDIAN	SI2NG	-1157,17	2330,34	753,03	3,045	0,761	12,270	94,413	0,195	-	-	-	-	0,217	-	-	0,465	0,971
INDIAN	SCG	-1165,06	2348,12	748,06	0,619	0,282	22,245	27,958	-	0,851	1,145	-	-	0,055	0,228	-	-	0,959
INDIAN	IM2m	-1166,24	2350,47	719,46	5,645	0,428	-	-	-	0,040	43,398	0,002	2,032	0,219	-	0,321	-	0,941
INDIAN	AM2m	-1166,00	2352,00	716,57	5,398	0,435	-	-	-	0,008	43,034	0,091	1,960	0,224	0,000	0,319	-	0,940
INDIAN	AM2NG	-1170,85	2363,70	757,31	3,150	0,923	6,087	25,253	0,265	2,429	0,000	-	-	0,000	0,189	-	0,450	0,962

INDIAN	SI2N	-1184,54	2381,08	749,27	16,206	3,653	-	0,489	0,147	-	-	-	-	0,179	-	-	-	0,953
INDIAN	SC	-1194,61	2403,22	716,88	3,626	0,562	-	-	-	1,152	5,477	-	-	0,173	0,047	-	-	0,935
INDIAN	SIG	-1205,58	2423,15	740,82	2,001	0,246	12,215	98,268	-	-	-	-	-	0,222	-	-	-	0,969
INDIAN	IM	-1210,16	2432,31	721,65	4,207	0,479	-	-	-	0,272	3,566	-	-	0,219	-	-	-	0,937
INDIAN	AM	-1210,17	2434,34	721,73	4,149	0,484	-	-	-	0,281	3,520	-	-	0,218	0,000	-	-	0,936
INDIAN	SI	-1251,12	2510,25	752,25	5,634	0,868	-	-	-	-	-	-	-	0,163	-	-	-	0,946
CLIFF	SC2N2mG	-1006,59	2039,18	123,06	4,120	0,879	24,700	19,600	0,174	0,088	0,024	0,074	0,008	3,120	0,891	0,481	0,400	0,987
CLIFF	SC2mG	-1007,72	2039,43	232,67	7,298	0,966	3,644	10,674		14,629	34,813	0,085	0,006	0,653	0,049	0,055		0,960
CLIFF	IM2mG	-1012,75	2047,50	278,01	23,134	0,304	3,579	94,775	-	4,009	0,117	0,006	0,009	1,143	-	0,198	-	0,988
CLIFF	SI2NG	-1029,34	2074,68	294,85	12,253	1,702	27,521	98,361	0,119	-	-	-	-	0,961	-	-	0,448	0,989
CLIFF	SC2NG	-1028,00	2078,00	272,58	10,968	4,458	20,293	49,353	0,105	0,122	0,000	-	-	0,220	0,892	-	0,482	0,988
CLIFF	IM2NG	-1036,39	2092,77	293,76	5,238	2,869	26,849	46,905	0,101	0,120	0,000	-	-	0,971	-	-	0,437	0,987
CLIFF	AM2mG	-1038,86	2101,73	284,22	4,711	0,486	38,158	37,117	-	0,374	0,002	0,377	4,215	0,139	0,856	0,778	-	0,986
CLIFF	AM2N2mG	-1073,93	2175,86	142,21	21,190	1,004	10,048	31,605	0,146	0,000	0,002	15,277	6,429	0,455	0,669	0,252	0,051	0,988
CLIFF	IMG	-1091,86	2199,73	301,48	5,015	0,233	21,076	99,793	-	0,055	0,094	-	-	0,906	-	-	-	0,987
CLIFF	SI2N	-1095,88	2203,76	271,21	97,226	24,494	-	0,392	0,042	-	-	-	-	1,063	-	-	-	0,985
CLIFF	AM2NG	-1091,63	2205,26	289,56	19,034	0,589	5,755	59,724	0,088	6,293	0,000	-	-	0,066	0,938	-	0,101	0,988
CLIFF	AM2N2m	-1091,48	2206,96	143,19	21,736	13,935	-	-	0,046	0,287	0,099	0,009	0,130	0,912	0,001	0,064	0,492	0,980
CLIFF	SIG	-1100,99	2213,98	300,46	4,248	0,345	72,363	99,879	-	-	-	-	-	0,913	-	-	-	0,987
CLIFF	AMG	-1101,53	2221,07	285,14	34,177	0,366	1,789	99,311	-	0,033	0,033	-	-	0,000	0,978	-	-	0,988
CLIFF	SC2N	-1101,93	2221,86	247,19	15,254	12,250	-	-	0,101	0,404	0,110	-	-	0,828	0,271	-	0,371	0,972
CLIFF	IM2N	-1107,36	2230,72	259,60	21,504	10,229	-	-	0,100	0,172	0,000	-	-	1,020	-	-	0,379	0,975
CLIFF	AM2N	-1111,39	2240,77	281,61	0,011	10,003	-	-	0,100	0,276	0,000	0,894	0,011	0,400	0,972	-	-	-
CLIFF	AM2m	-1111,28	2242,56	334,70	99,474	1,311	-	-	-	25,355	4,612	0,015	0,140	1,799	0,250	0,549	-	0,930
CLIFF	SC2m	-1115,42	2250,84	281,31	93,969	0,859	-	-	-	0,097	29,512	0,206	0,325	1,061	0,091	0,408	-	0,985
CLIFF	IM2m	-1122,31	2262,61	282,45	99,741	0,861	-	-	-	0,000	27,993	0,032	0,106	1,157	-	0,409	-	0,986
CLIFF	SC2N2m	-1119,83	2263,66	128,66	29,411	7,063	-	-	0,110	0,122	0,000	0,000	7,562	0,774	0,229	0,630	0,499	0,978
CLIFF	SC	-1295,69	2605,39	281,52	51,086	1,409	-	-	-	0,135	0,610	-	-	0,895	0,152	-	-	0,982
CLIFF	AM	-1303,73	2621,46	210,36	3,525	85,684	-	-	-	0,013	34,426	-	-	0,211	1,248	-	-	0,970
CLIFF	IM	-1314,32	2640,65	284,34	54,462	1,573	-	-	-	0,043	0,149	-	-	1,026	-	-	-	0,983
CLIFF	SCG	-1319,33	2656,67	292,73	51,314	1,755	6,932	34,177		0,015	0,197	-	-	0,917	0,021	-	-	0,984
CLIFF	SI	-1325,45	2658,89	295,08	50,878	1,828	-	-	-	-	-	-	-	0,907	-	-	-	0,984

Table S2.4: Confidence intervals of contemporary populations following demographic expansion. The confidence interval of the product $N_n x b_n$, was determined for each species for the best model, using the lower and upper bounds of the confidence intervals obtained for parameters N_n and b_n from Supplementary Table 1.

LAKE	MODEL	N1b1	N1b2
Témiscouata	SC2mG	[0;12965.4]	[0;2741.04]
East	SC2N2mG	[0;994.65]	[0;279.48]
Indian	SC2N2mG	[37.26;39.54]	[0;23.22]
Cliff	SC2N2mG	[39.71;182.4]	[6.76;31.22]

2.10 Supplementary figures

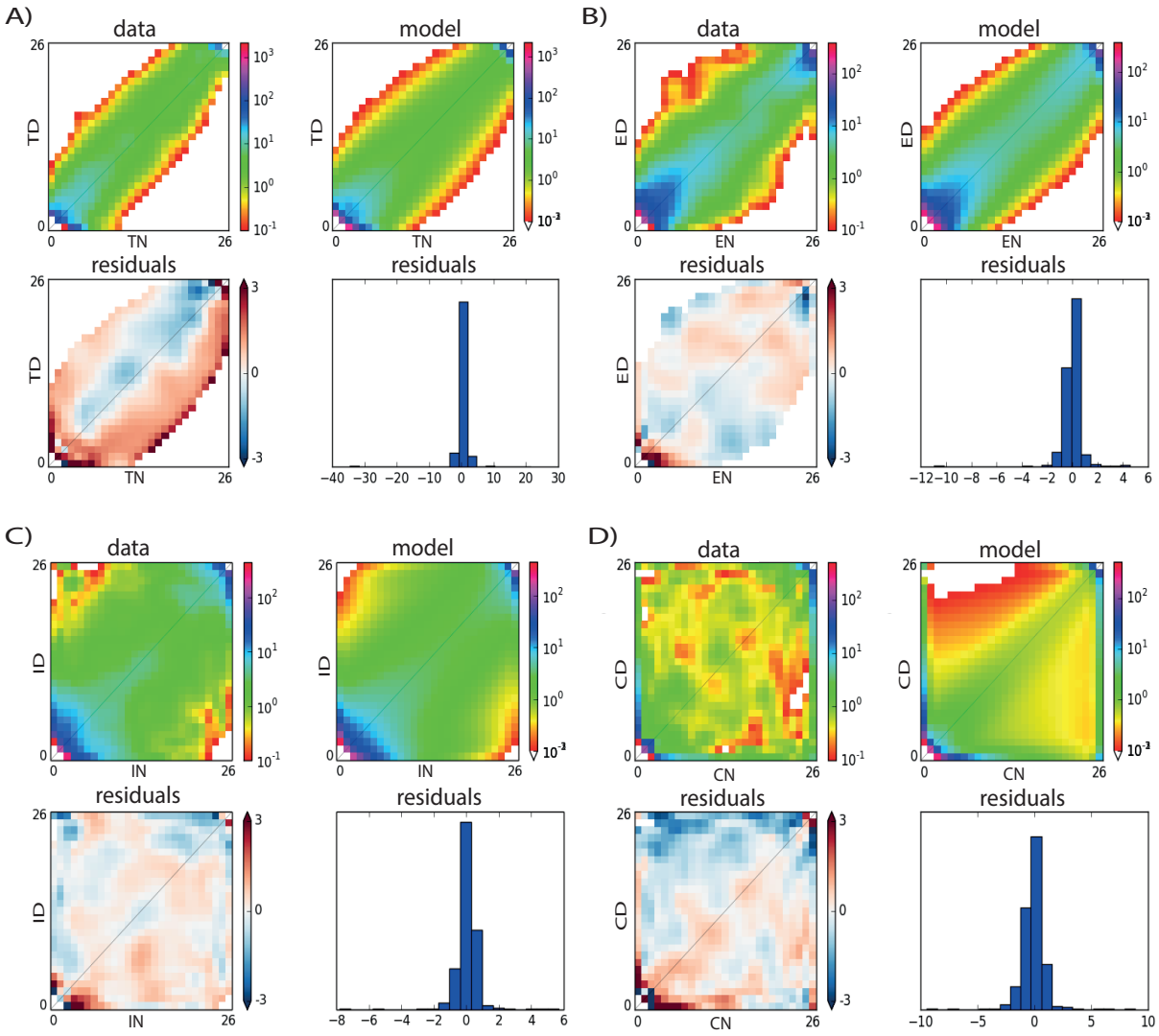


Figure S2.1: The joint allele frequency spectrum (JAFS) obtained from empirical data (upper left panel of each panel) and predicted under the best model (upper right), the Poisson residuals between model and data represented on the spectrum (lower left) and the distribution of residuals (lower right). Blue cells (negative residuals) represent model under-predictions and red cells (positive residuals) model over-predictions. Panels corresponds to A) Témiscouata Lake, B) East Lake, C) Indian Lake and D) Cliff Lake.

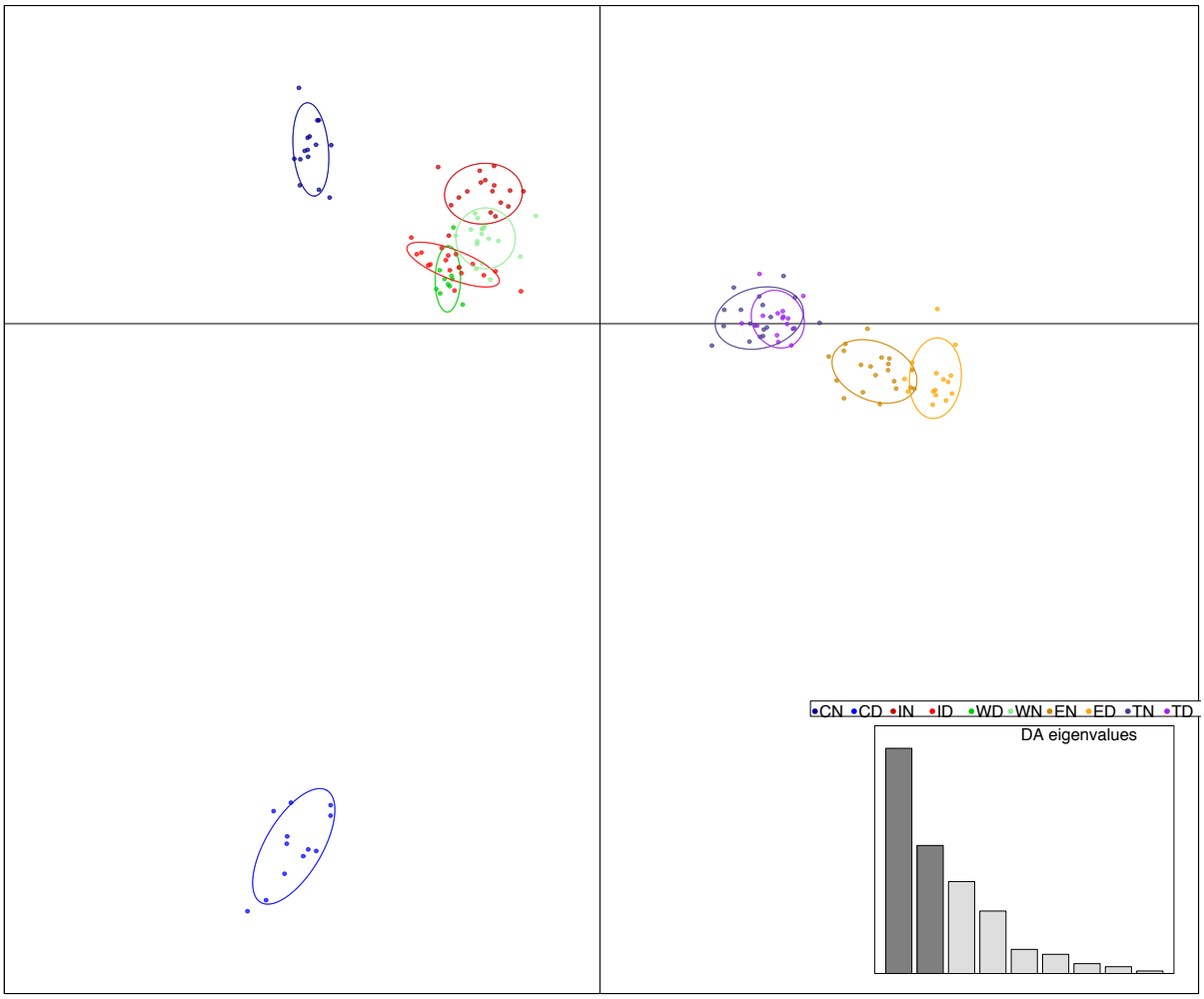


Figure S2.2: Discriminant analysis of principal components (dAPC) of the different lakes (Cliff, Indian, Webster, East, Témiscouata) for either Dwarf or Normal whitefish (D or N), representing relationships between populations. The first (horizontal) axis (LD1-39.5% of the total variance), clustering of populations by lakes, captures the geographical distribution of the lakes. The second (vertical) axis LD2 (22.5%) mostly separates dwarf and normal whitefish populations from Cliff Lake.

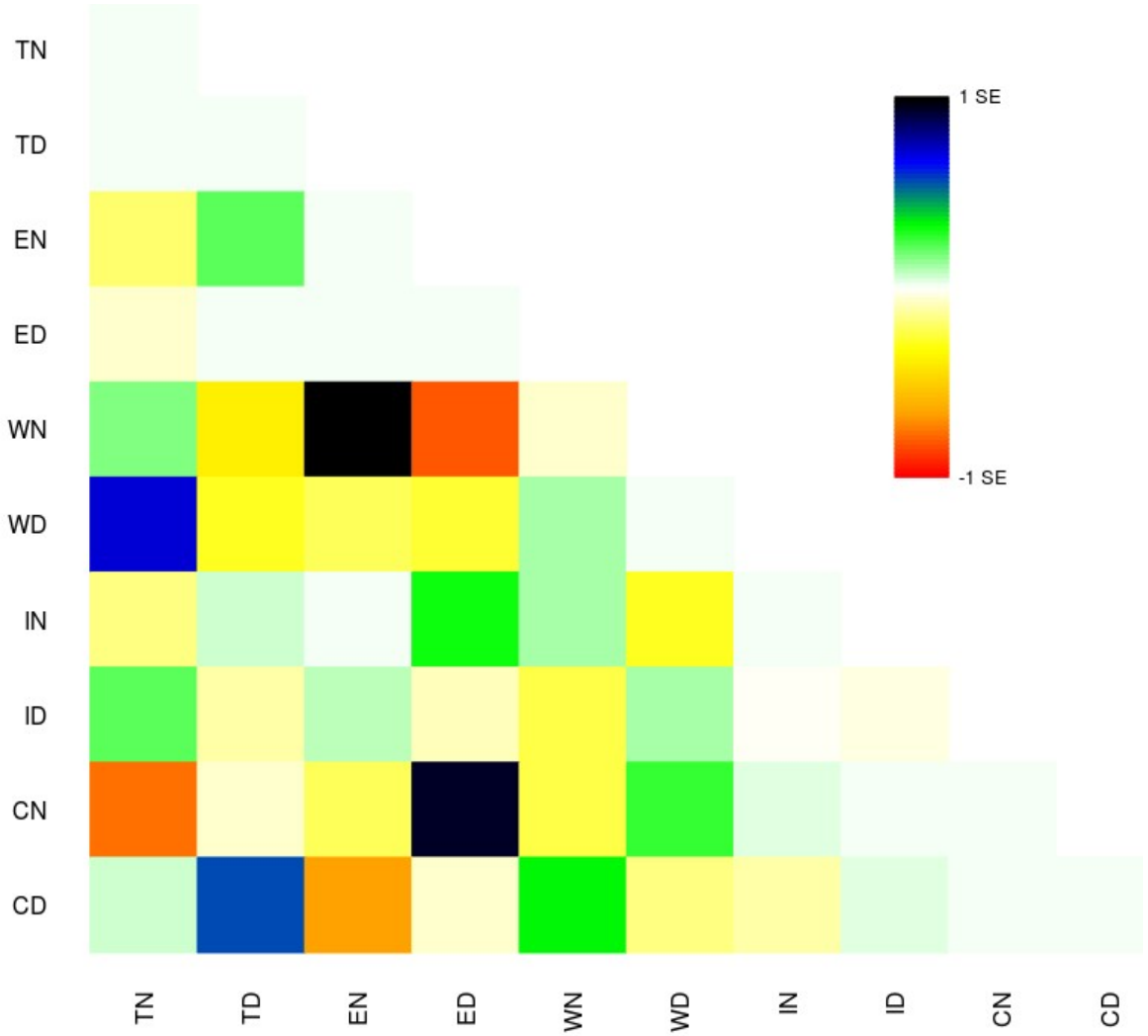


Figure S2.3: Residual fit from the maximum likelihood tree plotted in Figure 7. The residual covariance between each pair of populations was divided by the average standard error across all pairs. Green blue and dark colors indicate population pairs that are more closely related.

Chapitre 2: Differential introgression following postglacial secondary contact in European whitefish species pairs.

3.1 Résumé

La différenciation phénotypique parallèle est généralement associée à la divergence adaptative parallèle en réponse à des changements de conditions environnementales similaires. Ainsi, le parallélisme peut aboutir selon différents processus évolutifs tels que la divergence adaptative parallèle causée par une pression de sélection divergente, ou bien un seul évènement de divergence suivi de flux génique entre populations. Reconstruire l'histoire de l'évolution afin de déchiffrer la contribution relative du flux de gènes et de la pression sélective sur la divergence génomique est fondamental pour comprendre les événements de spéciation. Dans cette étude, nous étudions l'histoire de la divergence de réplicats de paires d'espèces limnétique et benthique de Lavaret (*Coregonus lavaretus*), dans deux lacs Norvégiens et deux lacs Suisses. Les modèles démographiques testés et ajustés au spectre joint déplié des fréquences alléliques, obtenus à partir de de marqueurs génomiques de type SNP génotypés chez six individus par populations. Nos analyses supportent un modèle de contact secondaire post-glaciaire asymétrique entre les lignées glaciaires dans tous les lacs testés, accompagnant la diversification du Lavaret. De plus, nos résultats suggèrent que les patrons génomiques de différenciation hétérogènes ont été façonnés par l'action conjointe de la sélection liée affectant les taux de tri des lignées pendant la phase allopatrique, et par la migration hétérogène érodant la divergence à différentes intensités le long du génome à la suite du contact secondaire. Dans l'ensemble, nos analyses ont mis en lumière l'interaction entre la démographie, la sélection et la contingence historique propre à chaque lac ayant influencé les niveaux de différenciation et affecté des études phylogéographiques précédentes. Cette étude fournit donc de nouvelles informations sur les processus de démographie historique et sélectif qui ont façonné la divergence associée à la spéciation écologique chez le corégone européen.

3.2 Abstract

Parallel phenotypic differentiation is generally attributed to parallel adaptive divergence as response to similar environmental contrasts. Such parallelism may actually originate from several evolutionary scenarios ranging from repeated parallel divergence caused by divergent selection or, a unique divergence event followed by gene flow. Reconstructing the evolutionary history in order to decipher the relative contribution of gene flow and selective pressure on genomic divergence is fundamental to understand speciation events. In this study, we investigate the divergence history of replicate European whitefish (*Coregonus lavaretus*), limnetic and benthic species pairs for two lakes in Norway and two lakes in Switzerland. Demographic models were tested and fitted with the unfolded joint allele frequency spectrum built from genome-wide SNPs of six individuals per populations. We found support for a model of asymmetrical post-glacial secondary contact between glacial lineages in all tested lakes, which accompanied whitefish diversification. Moreover, our results suggest that heterogeneous genomic patterns of differentiation were shaped by the joint action of linked selection affecting rates of lineage sorting during allopatric phase, and heterogeneous migration eroding divergence at different rates along the genome following the secondary contact. All in, our analyses shed the light on the interplay between demography, selection and historical contingency per lake that influenced levels of differentiation and outputs from previous phylogeographic studies. This study thus provides new insights into the historical demographic and selective process that shaped the divergence associated to ecological speciation in European whitefish.

3.3 Introduction

Independent phenotypic divergence among closely related and locally adapted nascent species provides evidence for population phenotypic response to similar constraints stemming from environmental variation (Endler 1986; Losos 2011). Therefore, parallel phenotypic divergence among replicate species pairs provides a valuable framework to understand the molecular basis of species diversification by natural selection (Elmer and Meyer 2011). Repeated phenotypic diversification can arise through similar or different genomic bases as a function of the origin and effects of selected mutations. Different scenarios regarding the origin of adaptive variation are usually distinguished, depending on whether selection has been recruiting *de novo* mutations (Pearce et al. 2009), mutations already present in the standing genetic variation (Yeaman, Hodgins, et al. 2016), or alternatively adaptive alleles transferred by gene flow between populations from similar environments (Schluter and Conte 2009; Welch and Jiggins 2014). These distinctions are important for understanding whether repeated phenotypic divergence is facilitated by the existence of divergent alleles that have evolved in isolation in the past, while being later reused by selection over shorter time scales (Welch and Jiggins 2014; Van Belleghem et al. 2018).

Beyond the similarities of the genomic basis and the origin of the mutations associated with the parallel phenotypic differentiation, this raises the question of the chronology and the mode of divergence (Welch and Jiggins 2014). Indeed, parallelism does not necessarily imply independent evolution, since it could alternatively result from a unique divergence event that is shared among replicates (Bierne et al. 2013). Consequently, it is important to decipher whether phenotypic diversification has evolved in a context of primary or secondary divergence, which are two different evolutionary scenarios with respect to gene flow (Smadja and Butlin 2011). Primary divergence scenario is usually associated to repeated independent divergence events among replicates in response to similar selective pressures. By opposition, secondary divergence may involve a single divergence event between two lineages followed by their spatial redistribution and admixture.

In both secondary and primary divergence scenarios, the evolutionary history of divergence between lineages is jointly influenced by demography and selection, which

modulates the amount of effective gene flow and lineage sorting across the genome (Barton and Bengtsson 1986; Sousa and Hey 2013; Harrison and Larson 2016a).

Demographic inferences and modeling development provides a powerful framework in order to decipher the relative contribution of selective pressures and demography during evolutionary processes underlying parallel phenotypic differentiation (Butlin et al. 2014). In particular, methods that integrate varying introgression rates among loci allow capturing and measuring the barrier effect caused by speciation genes (Sousa et al. 2013; Tine et al. 2014; Roux, Fraïsse, Romiguier, Anciaux, Galtier, and Bierne 2016b; Rougeux et al. 2017), while capturing the effect of linked selection by allowing loci to experience varying rates of genetic drift (Sousa et al. 2013; Roux, Fraïsse, Romiguier, Anciaux, Galtier, and Bierne 2016b; Rougeux et al. 2017). This extended framework provides the opportunity to integrate the effects of heterogeneous selection and gene flow, separately or simultaneously, in order to increase the resolution of demographic inferences, describe origin of patterns of heterogeneous genetic differentiation, and reconstruct the history of speciation (Rougeux et al. 2017).

The European whitefish (*Coregonus lavaretus*), hereafter whitefish, represents a valuable model to study the evolutionary history of repeated parallel phenotypic differentiation. Indeed, the *C. lavaretus* species complex is composed by several whitefish species distributed across the European continents, notably in alpine lakes from Switzerland (Hudson et al. 2010) and lakes from Fennoscandinavia (Douglas et al. 1999; Østbye et al. 2005). Ecological divergence in this system has resulted in phenotypic differentiation associated with the occupation of limnetic and benthic ecological niches (Vonlanthen et al. 2009; Praebel et al. 2013). Phenotypic divergence is believed to have evolved repeatedly across different lakes (Douglas et al. 1999; Hudson et al. 2010). While the sympatric species pairs of the Alpine whitefish radiation have been taxonomically distinguished, this is not the case for Norwegian whitefish. Nevertheless, we here use the terms limnetic and benthic species to refer to these similar morphotypes found in different European locations. The limnetic and benthic whitefish species differ in body size, body shape, and habitat utilisation and resources with a feeding-related morphology (Vonlanthen et al. 2009; Praebel et al. 2013). Such differences between species in the use of habitat and ecological niches likely induce reproductive isolation as a by-product of ecological divergence (Schluter 2000; Woods et al. 2009; Praebel et al. 2013). However, it is still

unclear whether primary or secondary divergence has accompanied whitefish diversification.

Mitochondrial DNA (mtDNA) based phylogeographic studies identified three clades in the whitefish distribution area: the North European, the Siberian and the South European clades (Østbye et al. 2005). Those three clades reflect the occupation of three different glacial refugia during the last glacial period (Østbye et al. 2005). However, haplotypes of each clade are observed in some area suggesting that lineage mixing has resulted from habitat recolonization following glacial retreat (Østbye et al. 2005). Moreover, genetic analysis comparing limnetic and benthic species revealed heterogeneous genetic differentiation along the genome between sympatric species in Alpine lakes (Feulner and Seehausen 2018), suggesting local reduction of gene flow in some genomic regions associated with adaptive divergence. While such heterogeneous genome divergence has not been characterized in studies on Norwegian whitefish, intra-species populations from different lakes showed more genetic similarities between them than between sympatric species, suggesting a common origin for the populations of a given species (Praebel et al. 2013). Although such results suggest that limnetic and benthic whitefish species share a common history of divergence across Europe, this remains to be tested using historical divergence inference methods in order to understand the evolutionary processes associated with their diversification.

Here, we use RNAseq single nucleotide polymorphism (SNP) (De Wit et al. 2015) to document the joint allele frequency spectrum (JAFS) of each sympatric species pairs and infer its demographic divergence history. We test for temporal and genomic variations in the rate of gene flow for each sympatric species pairs. We then characterize genetic relationships and perform historical gene flow analyses among lakes to determine a general parsimonious scenario for whitefish diversification. Finally, we document the relative contribution of selective and demographic parameters in shaping the genomic differentiation between sympatric species pairs.

3.4 Material and Methods

Sampling, libraries preparation and sequencing

Four lakes harbouring sympatric limnetic and benthic species were sampled across the distribution range of *C. lavaretus* in Europe. Six individual per species (i.e., 12 indivudals per lake) were collected in two Scandinavian lakes in Norway (Skrukkebukta and Langfjordvatn lakes) and two lakes from Switzerland (Lucerne and Zurich lakes) (Figure 3.1), for a total of 48 individuals. Liver tissue was taken on each individual and stored in RNAlater before mRNA extractions.

We used the same data as generated in a previous study (Rougeux et al. 2018). However, the analyses realized in the previously paper were on totally different topics. Briefly, total RNA was extracted using the RNAeasy Mini Kit following manufacturer's informations (Qiagen, Hilden, Germany). We first quantified the amount of RNA via a NanoDrop2000 spectrophotometer (Thermo Scientific, Waltham, MA, USA), and assessed RNA quality using the 2100 Bioanalyser (Agilent, Santa Clara, CA, USA). We finally measured RNA concentration with Quant-iT RiboGreen RNA Assay Kit (Invitrogen, Life Technologies, Carlsbad, CA, USA) on samples reaching our quality criterions (RIN value greater than or equal to eight) before library preparation.

Individual libraries were prepared from 2µg of RNA using the TruSeq RNA sample preparation kit V2 (Illumina, San Diego, CA, USA) following the manufacturer's instructions. Library size and concentration were evaluated using DNA High Sensitivity chip on the 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA). Single read sequencing (100bp) was performed on the Illumina HiSeq 2000 platform for the 48 libraries of the present study (form the 72 libraries total including the Lake Whitefish, see Rougeux et al. 2018, eight libraries per lane) at the McGill University and Genome Quebec Innovation Centre (Montreal, Canada).

Genotyping

In order to document the extent of polymorphism within *C. lavaretus* and among divergent sympatric limnetic and benthic species pairs, individuals reads of the 48 individuals of *C. lavaretus* were mapped to a reference transcriptome (Rougeux et al. 2018)

using Bowtie2 v2.1.0 (Li and Durbin 2010). Six benthic individuals of the American Lake Whitefish (*C. clupeaformis*) were also sequenced and aligned to the same reference transcriptome to provide outgroup information for orienting mutations and thus identified the derived alleles within *C. lavaretus*.

BAM files were generated and sorted using Samtools v1.3 (Li et al. 2009), and filtered to remove duplicate reads that were removed with the Picard-tools program v1.119 (http://broadinstitute.github.io/picard/). SNPs calling from mapped reads was realized with Freebayes v0.9.10-3-g47a713eb (Garrison and Marth 2012), considering alleles with at least two reads per sample (default), and kept good-confidence alignments with a minimum quality of five and a minimum coverage five.

Variable sites with a minimum coverage lower than three reads per individuals were filtered out. We then used vcffilter program from vcflib (Garrison and Marth 2012) to process the VCF file generated by Freebayes. We retained biallelic SNPs with a phred scaled quality score above 30, individual genotypes with a phred score higher than 20. Following quality control steps, we removed miscalled and low-quality SNPs for subsequent population genomics analyses from the VCF file using VCFtools (Danecek et al. 2011). SNPs with more than 10% of missing genotypes in at least one *C. lavaretus* population were removed. A lower exclusion threshold of 50% was applied for the *C. clupeaformis* outgroup population. We controlled for Hardy–Weinberg disequilibrium within each population using a P-value exclusion threshold of 0.05. Finally, we merged the filtered data sets of limnetic and benthic populations within each lake together with the Lake Whitefish outgroup, resulting in four lake-outgroup VCF files containing 412,292, 408,999, 410,433 and 408,071 SNPs for Langfjordvatn, Skrukkebulta, Lucerne and Zurich lakes, respectively. Finally, only oriented loci (i.e., monomorphic loci in *C. clupeaformis* and polymorphic in *C. lavaretus*) were kept after removing loci that were polymorphic in the outgroup (Tine et al. 2014; Rougeux et al. 2017). This resulted in 157,516, 136,499, 117,125, 92,678 oriented SNPs for Langfjordvatn, Skrukkebukta, Lucerne and Zurich lakes, respectively, that were used to generate the unfolded JAFS of each lake, by projecting the data to 10 chromosomes (i.e., 5 diploid copies) per population (Figure 3.2).

Divergence history inferences

The demographic and selective histories of the four species pairs was inferred using a custom version of the software ∂a∂i v1.7 (Gutenkunst et al. 2009). On the basis of the four basic models representing alternative modes of divergence: Strict Isolation (SI), Isolation with Migration (IM), Ancient Migration (AM) and Secondary Contact (SC), we considered 13 models extensions that integrate demographic and selective parameters. The demographic part was related to the mode of divergence, using an ancestral population of size N_{ref} that splits into two populations of size N_1 (limnetic) and N_2 (benthic) during T_S (SI and IM), $T_{AM}+T_S$ (AM), or T_S+T_{SC} (SC) generations. Migration events during periods of T_S (SI and IM), T_{AM} (AM), and T_{SC} (SC) generations occured at variable rates m_{e12} from population 2 (benthic) to population 1 (limnetic), and vice-versa. Next, we incorporated extensions to the models (heterogeneous gene flow: -2m) in order to capture the effect of selection inducing reduced gene flow around loci associated with adaptive divergence. These semi-permeability models allowed quantifying the proportion P of loci freely exchanged at rate m_e while the proportion 1-P corresponds to the genomic proportion affected by a reduced effective migration rate m'_{e12} from population 2 (benthic) to population 1 (limnetic), and vice-versa (Tine et al. 2014). We further incorporated the effect of background selection by considering local genomic variation in N_e at linked neutral sites across the genome due to Hill-Robertson effects (Hill and Robertson 1966; Charlesworth et al. 1993), using a scaling factor hrf < 1, multiplying the N_e of each population. These linked-selection models (-2N) allowed quantifying the proportion 1-Q of loci that are affected by linked selection (Rougeux et al. 2017). Then, we quantified the proportion of correct SNPs orientation using parameter O (Tine et al. 2014).

We generated 20 independent optimizations runs for each model in each lake in order to check for model convergence and kept the best fit run of each model in order to perform model comparisons for each lake on the basis of the Akaike information criterion (AIC), in order to account for the number of parameters in each model. We retained models with the smallest AIC ($\Delta AIC_i < 10$) and estimated the weighted AIC (wAIC) to estimate the relative probability for a model to be the best (Rougeux et al. 2017). We then estimated parameter uncertainty for each retained model using the Godambe information matrix method from ∂a∂i v1.7. A nonparametric procedure was used to estimate confidence intervals (CIs) and standard-errors (se) from 1,000 bootstrap resampling. Finally, we

converted parameter value estimates (Table 1) by applying equations from Rougeux et al. (2017), allowing converting parameter estimates from the estimated theta (θ).

Patterns of shared ancestry and admixture

We generated a unique VCF file composed by only polymorphic loci that were retained after filtering in the four studied lakes to search for signals of shared ancestry and gene flow among all four species pairs. Using the 47,556 retained SNPs, we performed a discriminant analysis of principal components (dAPC), in Adegenet v2.0.0 (Jombart et al. 2010), partitioning the variance into a between-group and within-group in order to maximize discrimination between groups. The dAPC characterized the genetic structure and relationships among lakes and species, in the entire system (Figure 3.3A) and in Norway and Switzerland separately (Figure 3.3B and Figure 3.3C).

Finally, we inferred historical relationships among populations using TreeMix v1.12 and f4-statistics (Pickrell and Pritchard 2012). TreeMix allows modeling migration events between populations by discrete mixture events that are added to a bifurcating population tree. Consequently, such events may either reflect gene flow between sympatric populations and/or excess of shared ancestral polymorphism among populations from different lakes that are geographically isolated. We allowed three migration events to be inferred among the branches of the whitefish population tree, in agreement to the maximized variance explained (Figure S3.1).

3.5 Results

A total of $1.19x10^9$ 100bp single raw reads were generated from the 48 individuals of *C. lavaretus* ($1.5x10^9$ when including the 12 *C. clupeaformis* individuals) (Table S3.1**)**. A total of $1.44x10^9$ (including both *C. lavaretus* and *C. clupeaformis*) trimmed and cleaned reads were aligned to the reference transcriptome generated in (Rougeux et al. 2018) resulting in the mapping of $8.3x10^8$ reads.

Divergence models

The JAFS of each the four species pairs showed a high density of SNPs with similar allele frequencies (*i.e.*, located around the diagonal) (Figure 3.2). However, the presence of markers located on the outer frame of the spectra showed that some regions of the genome display strong differences in allele frequencies between sympatric species.

For each lake, the model producing the best fit to the observed JAFS was an extension of the secondary contact (SC) model, in which gene flow occurred following an allopatric phase. The incorporation of selective parameters in the demographic models generally improved AIC values that penalize model likelihood by the number of parameters to avoid overfitting. For all the four lakes, heterogeneous gene flow along the genome was a necessary component of the secondary contact model (SC2m), while Skrukkebukta and Lucerne lakes also integrated in their respective best-fit models local reductions in N_e to account for linked selection (SC2N2m; Table 1 and Table S3.2). All the highest ranked model for each lake received a w_{AIC} = 1.0, thus showing maximal support for the SC2m model in Langfjordvatn and Zurich lakes, and the SC2N2m model in Skrukkebukta and Lucerne lakes.

Inference of model parameters

The inferred proportion of correctly oriented markers in the unfolded JAFS (*O*) ranged from 98% to 99%, indicating that almost all of ancestral allelic states were correctly inferred using *C. clupeaformis* as an outgroup. The conversion of estimated demographic and selective parameters allowed intra- and inter-lake populations comparisons. Here, we considered only the best-fit model for each lake. In lakes from Norway, the effective population size of the limnetic populations (*nu1*) were about three times greater than the

effective populations size of the benthic populations (*nu2*), although wide confidence interval overlapped between of sympatric limnetic and benthic populations were similar in Lucerne Lake but a higher effective population size was observed for the limnetic population in Zurich Lake (Table 1). The amount of gene flow followed the same pattern with a higher rate of gene flow from the limnetic species to the benthic species (m_{e21}), but not in Lucerne Lake where the amount of gene flow were comparable for both directions. Finally, the fraction of the genome showing unconstrained gene flow (*P*) was estimated to 95%, 64%, 50% and 91% for Lakngfjordvatn, Skrukkebukta, Lucerne and Zurich, respectively. These values suggest that limnetic and benthic whitefish from Lucerne and Skrukkebukta lakes are more strongly reproductively isolated than those from Zurich and Lakngfjordvatn lakes.

Consistent with these observations, the highest ranked model for Lucerne and Skrukkebukta lakes also included heterogeneous effective population size across the genome. The fraction (*Q*) of the genome with a locally reduced N_e was estimated to 7% for Skrukkebukta Lake and 2% for Lucerne Lake, and the proportion (*hrf*) of reduction in N_e in those fractions was 13% and 62% in Skrukkebukta and Lucerne lakes, respectively, although this amount was associated with a large uncertainty for Skrukkebukta (Table 1).

Finally, the estimated duration of the allopatric phase (T_s) was variable among lakes, although in the same order of magnitude, with T_s of 121,000; 69,000; 155,000 and 107,000 years for Langfjordvatn, Skrukkebukta, Lucerne and Zurich, respectively. Moreover, the inferred duration of secondary contact (T_{sc}) corresponds roughly with the last glacial retreat, with secondary gene flow being initiated 12,400; 15,600; 24,200 and 28,600 years ago for Langfjordvatn, Skrukkebukta, Lucerne and Zurich, respectively.

A shared history of divergence

Partitioning the genetic variation using a dAPC, either within Norway and Switzerland separately, and in the entire system revealed distinct patterns. In Norway, populations clustered by species along the first axis (LD1 6.4% of the variance) whereas the second axis (LD2 5.3% of the variance) separated the populations according to their lake of origin (Figure 3.3A). A similar pattern emerged from the dAPC on lakes from Switzerland whereby the first axis (LD1 9.9% of the variance) mostly separated the species, while the second axis (LD2 7.4% of the variance) clustered the populations by lake (Figure

3.3B). When projecting the partitioned genetic variation of all four lakes in the same dAPC, the first axis discriminated Norway and Switzerland populations, while the second axis discriminated lakes within each region, albeit to a lesser extent (Figure 3.3C). The third axis separated lakes in Norway, and the fourth axis discriminated limnetic and benthic species from Norway (Figure 3.3D). The fifth axis separated the species across the entire system (Figure 3.3E). We note that within the entire system, despite proximity between sympatric populations for Zurich and Skrukkebukta lakes, populations from the same species show more similarities between them than with sympatric populations of the other species.

The genetic relationships among populations analyzed with TreeMix (Figure 3.4) revealed a clear separation between Norway and Switzerland, with Norwegian populations showing less divergence among them. Within regions this analysis revealed two levels of signal that is for both regions, sympatric populations were grouped by lake, reflecting the effect of gene flow following secondary contact in each lake. Migrations links between allopatric populations of the same species (*i.e.*, from different lakes) illustrated the excess of ancestral shared polymorphism (*i.e.*, shared ancestry). This is particularly the case in Switzerland between the related populations of the limnetic species from Lucerne and Zurich lakes. We observed a more complex pattern in Norway with links between an ancestral population of the Skrukkebukta Lake (shared branch between species from Skrukkebukta Lake) and the benthic species of Langfjordvatn Lake, and reciprocally between an ancestral population of Langfjordvatn Lake and the limnetic species of Skrukkebukta. This suggests an excess of shared polymorphism between related populations of the same species between lakes, considering the inferred population bifurcating tree. This pattern was confirmed with TreeMix inference allowing two migration events, with a link between related populations of the limnetic species (Figure S3.3). Finally, patterns of genetic similarity between allopatric populations of the same species were confirmed by *f*4-statistics, identifying excess of shared polymorphism between populations of the same species between different lakes, with significance test on their respective z-scores ($P<0.001$) (Table S3.3). Those results were observed mainly within Norway and Switzerland analysed separately.

3.6 Discussion

Repeated evolution of phenotypic diversification (*i.e.*, parallel phenotypic divergence) may stem from independent histories of divergence among diverging populations, or have a common basis of divergence with before independent evolution (Bierne et al. 2013; Welch and Jiggins 2014). Therefore, determining the mode of divergence and, by extension, inferring the timing of divergence and gene flow between diverging population is required in order to understand the origin of the divergence leading to parallel phenotypic diversification (Rosenblum et al. 2014; Rougeux et al. 2017; Herman et al. 2018).

Limnetic and benthic species of the *C. lavaretus* complex offer a relevant system to study the origin of parallel phenotypic differentiation. While previous phylogeographic studies aimed to understand and elucidate the evolutionary history of such diversification, they relied on mtDNA (Østbye et al. 2005) or a combination of mtDNA and reduced genomic representation such as AFLP (Hudson et al. 2010). However, recent population genomics study based on RADseq data focused on identification of outlier loci associated with divergent selection between limnetic and benthic species in two lakes (Feulner and Seehausen 2018). Here, our transcriptomic analysis allowed achieving an unprecedented coverage of genome-wide variation. Our data thus provide new insights into the historical demography and mode of divergence between diverging sympatric species pairs of European whitefish, by addressing questions related to (i) the origin of the phenotypic diversification, (ii) the time of divergence between sympatric species, (iii) the amount of gene flow during divergence and (iv) identifying the demographic and selective pressure shaping the genomic differentiation of the diverging species pairs. This study thus provides a complementary framework for understanding the repeated evolution of the limnetic and the benthic species in whitefish, as it was previously realized for the Lake Whitefish (*C. clupeaformis*) in North-America (Rougeux et al. 2017).

Evidences of secondary contact

Firstly, we inferred the divergence history of each species pair separately, using the JAFS as a summary statistics of genome-wide differentiation. The best scenario to explain our data was a secondary contact (SC) for each species pair of the four studied lakes. It is

noteworthy that these data allowed identifying the secondary contact despite low degree of genetic divergence between some species pairs. This corroborates our observation that the allopatric period was relatively long compared to the period of gene flow following secondary contact (Roux, Fraïsse, Romiguier, Anciaux, Galtier, and Bierne 2016b) and that the resulting JAFS showed in each lake an excess of shared intermediate frequency alleles, as expected following the erosion of divergence by gene flow (Alcala et al. 2016).

The secondary contact scenario was clearly supported for each lake. Previous observations from phylogeographic studies (Østbye et al. 2005; Hudson et al. 2010) inferred secondary contact between mitochondrial clades (*i.e.*, glacial lineages), while in some lakes, the origin of sympatric species was assumed to be the result from an adaptive radiation from a single clade due to a lack of polymorphism (Østbye et al. 2005). Such pattern was also observed in the American Lake Whitefish (*C. clupeaformis*) system, in which one of the five studied lakes where sympatric limnetic and benthic species co-occur (East Lake) was fixed for only one mitochondrial haplotype. Consequently, this observation was previously interpreted as a support for sympatric divergence in this lake (Pigeon, Chouinard, and Bernatchez 1997b). However, recent historical demographic analysis based on genome-wide polymorphism data inferred a secondary contact between two glacial lineages also in this lake, where mitochondrial lineage fixation has probably been facilitated by asymmetrical gene flow (Rougeux et al. 2017). Here, estimated parameters from JAFS analysis indicated an important asymmetrical effective population size between sympatric species for most lakes, associated with an asymmetrical gene flow mainly from the limnetic species to the benthic species (except for Lucerne Lake). Interestingly, in Lucerne Lake sympatric species display two different mitochondrial lineages, which is not the case in Zurich Lake where both species are fixed for the 'limnetic haplotype' lineage (Østbye et al. 2005). A similar pattern emerge for Norwegian lakes where mainly only one mitochondrial haplotype is more frequent (Østbye et al. 2005; Kahilainen and Østbye 2006; Østbye et al. 2006; Praebel et al. 2013). Thus, the asymmetrical gene flow from the limnetic to the benthic species occurred mainly due to an asymmetrical effective population size between both sympatric species pairs in each lakes with a higher N_e, but not in Lucerne Lake in which both clades were described (Østbye et al. 2005). Then the asymmetrical gene flow inferred in our analysis is consistent with expected patterns, as gene flow between

hybridizing population is expected to occur from the larger to the smaller population (Beysard et al. 2011).

Repeated evolution of phenotypic diversification is likely to result from divergent selection acting following post-glacial lake colonization and minimizing competitive interactions with other glacial lineage (Østbye et al. 2005; Bernatchez et al. 2010). Moreover, evolution of different morphs in sympatry resulting from a contact between different lineages has been described in several freshwater fishes besides North American and European Lake whitefish, for instance ciscoes *C. artedii* complex (Turgeon and Bernatchez 2003), but also in anadromous fishes such as the rainbow smelt (*Osmerus mordax*) (Bernatchez 1997), as well as between avian species such as the pied flycatcher and the collared flycatcher (Nadachowska-Brzyska et al. 2013).

Our demographic inferences allowed quantifying the genomic proportion affected by selective pressure by portioning reduction of gene flow around speciation genes (*i.e.*, gene involved in reproductive isolation between diverging species), and genomic regions affected by linked selection. This results confirm that improving historical demographic inferences models could allow increasing the resolution of such analysis, and disentangling the relative contribution of the different evolutionary forces (Rougeux et al. 2017). These parameters developed in few studies (Sousa et al. 2013; Roux, Fraïsse, Romiguier, Anciaux, Galtier, and Bierne 2016b; Rougeux et al. 2017) improved demographic inferences, as despite the increased number of parameters those complex models better fitted the data, as observed in the saltmarsh beetle (*Pogonus chalceus*) (Van Belleghem et al. 2018).

Temporal aspects in studies of speciation are fundamental as they can influence the rate of divergence, duration of putative gene flow or complete isolation. Our study brings new evidence supporting previous findings based on mtDNA that the time of divergence matches with the last glacial period (Østbye et al. 2005). In Norway, we found that the divergence time (T_S) between sympatric species was about 121,000 years before present (ybp) for Langfjordvatn Lake and about 69,000 ybp for Skrukkebukta Lake. Similarly, T_S between sympatric species was about 155,000 ybp for Lucerne Lake and about 107,000 ybp for Zurich Lake. Such times of divergence (including confidence intervals) correspond to the late Würm (Alpes) and Weichselian (North Europe) glacial episodes during the Pleistocene (110,000 to 10,000 ybp). Thus, those inferred times and geographic location

correspond to the glacial lineages identified in previous study in which one clade was localized at the North of actual Europe, and a second clade in a refugia located south of the 53°N (Østbye et al. 2005). Consequently, both glacial lineages evolved independently during an allopatric phase, suggesting accumulation of genetic incompatibilities. Furthermore, the time of secondary contact between species was around 12,400; 15,600; 24,000 and 28,000 ybp for Langfjordvatn, Skrukebukkta, Lucerne and Zurich lakes, respectively. Those inferences match with the retreat period of the ice sheet 20,000 ybp in Switzerland (Andersen and Borns 1994) allowing lakes colonization, and with the younger lakes formation in Norway around 10,300 and 12,000 ybp (Björck 1995). Finally, the concordance in the duration of allopatric phase with the distribution of the ice sheet during the last glacial period, and the time of secondary contact with the ice sheet retreat or the lakes formations indicate similarities in the evolutionary history of the lakes both from Norway and Switzerland.

A shared history of divergence

The global and simultaneous analysis of the four species pairs helped clarifying the genetic relationship between the different populations of limnetic and benthic species in order to ask whether species pairs have a common history before the secondary contact.

The genetic similarities revealed by TreeMix produced different levels of grouping in the population tree. The different hierarchical levels were observed through the tree as sympatric species pairs clustered by lake, and lakes clustered per region. This pattern likely reflects the relative importance of gene flow within lakes and genetic drift among lakes. Yet, shorter branches for populations from Norway indicate that those (sympatric) populations are less divergent. However, inferring migration events among populations within the tree enabled us to detect links between populations of the same species but from different lakes. This kind of patterns is more likely to indicate shared genetic variation from a co-ancestry (Rougeux et al. 2017). Namely, inferred links between limetic populations from Zurich and Lucerne lakes support the view that these populations of the same species were genetically similar before being isolated in their respective lakes. Then, a more complex relationship emerged between populations from Norway. Here, the ancestral population from Skrukkebukta Lake was linked to the benthic population of Langfjordvatn Lake, and reciprocally for the ancestral population of Skrukkebukta and Landfjordvatn lakes. This is

consistent with an important genetic similarity between all populations from these lakes due to recent divergence and extensive gene flow. It is, however, important to note that the second parsimonious tree explaining the data (allowing two migration events, Figure S3.3) linked populations of the limnetic species from Skrukkebukta and Langfjordvatn lakes. Finally, the *f*4-statistics analysis supported the results from TreeMix and revealed genetic similarities between benthic populations from Lucerne and Zurich lakes but also between populations of the same species across Norway and Switzerland (Table S3.3), suggesting a common origin of the populations of the same species.

The analysis of diversity patterns performed with the dAPC provided a complementary approach in order to disentangle signals of genetic differentiation between species from genetic differentiation among lakes. At the regional scale, the first axis separated the species (Figure 5.3A and Figure 5.3B), and the second axis separated the lakes. Within the entire system, the fifth axis allowed to identify differentiation between both species for the four species pairs. The projection of limnetic and benthic populations from Lucerne Lake indicated the positions of the two least introgressed (more differentiated) populations of the dataset, in accordance with results obtained from demographic inferences (Table 1) and with previous observations (Østbye et al. 2005). Therefore, these populations were defined as representative of two glacial lineages separated on the 5th axis. Interestingly, Zurich Lake populations as well as limnetic and benthic populations from Norway were separated by the same axis, while Skrukkebukta and Langfjordvatn species pairs showed less differentiation (Figure 5.3C). Thus, variable levels of separation along the 5th axis could mirror variable level of proportion of the ancestry from initial glacial lineages due to variable levels of introgression during the secondary contact.

To summarize, complementarity of the approaches realized in this study suggest that the most parsimonious scenario corresponds to a secondary contact between two main glacial lineages (clades as identified in Østbye *et al.*, 2005) that were isolated for roughly 100,000 (range), with different degrees of introgressive hybridization between two glacial lineages in different lakes resulting from variable demographic contingencies following lake colonization some 10,000-20,000 years ago. Because roads of colonization were described in much details by (Østbye et al. 2005), we focused our analysis on deciphering the timing of historical migration between diverging species, and the selective pressures involved in the genomic landscape of divergence. The retreat of the ice sheet, 20,000 years ago,

corresponds to contact zone of the two glacial lineages from the south and north of Europe (Andersen and Borns 1994; Björck 1995; Østbye et al. 2005). Then, the colonization waves of newly formed post-glacial lakes are likely to have provoked and maintained the secondary contact and the origin of parallel genetic divergence across whitefish species pairs.

All in, we propose that the genetic divergence in whitefish species pairs result from a complex interplay between the genetic divergence between glacial lineages during the allopatric phase (allowing for mutation accumulation and lineage sorting), the reduced introgression around locally adapted loci, but also selective pressures on phenotypes and the demographic contingency of each species pairs. Then, we argue that integrating selective events in demographic models allow disentangling the relative contribution of deterministic factors (*e.g.*, selection) from historical contingency in order to better understand the historical demography of the populations, and then address questions of biological diversification. This is of particular importance for systems in which strict sympatric divergence was assumed without real test for alternative models, as in cichlids (Kautt et al. 2016; Meier, Sousa, et al. 2017) although alternative models were tested in some lakes and revealed secondary contact through several colonisation waves (Martin et al. 2015). Finally, modern demographic inferences shed the light on an increasingly occurrence of secondary contact in the process of divergence observed in sympatry, which was eased by the allopatric phase allowing increase of genetic divergence leading eventually to reproductive isolation.

3.7 Acknowledgement

We thank Kim Praebel, Shripathi Bhat and the Freshwater ecology group at UiT for participating with the sampling in Norway, and Ole Seehausen for samples from Switzerland. This research was supported by a Discovery Research grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to L.B. L.B also holds the Canadian Research Chair in genomics and conservation of aquatic resources, which funded the research infrastructure for this project.

3.8 Tables

Table 3.1: The demographic divergence history of whitefish species pairs inferred parameters. For all lakes, details of statistics and demographic parameter values for the fittest models determined under the threshold of ΔAIC < 10. The table contains in this order the maximum likelihood (MLE) for each model, the value of Akaike information criterion (AIC) and the weighted AIC (wAIC). Then, the inferred demographic parameter values converted with the estimate of THETA: the ancestral effective population size before population split (Nref); the effective population size after split for limnetic (N_1) and benthic (N_2) populations; The Hill-Robertson factor (*hrf*) corresponds to the degree to which the effective population size of diverging populations is locally reduced linked selection effect. Migration parameters (in migrant per generation) include migration rates from benthic population to limnetic population (m_{e12}) and reciprocally (m_{e21}), and a second category of effective migration rates ($m_{e'12}$) and ($m_{e'21}$) applying to a second category of loci. Time parameters include the duration (in years) of the allopatric divergence period (T_s), and the duration of the migration period (T_{SC}). Finally, the table also contains proportion parameters such as the proportion (Q) of the genome with effective population sizes N_1 and N_2. A proportion (P) of the genome is occupied by loci with effective migration rates m_{e12} and m_{e21} (and a second category of loci occupying a fraction 1-P of the genome has effective population sizes m'_{e12} and m'_{e21}). The parameter (O) is the proportion of correct SNP orientation. Numbers into brackets denote 95% confidence intervals obtained using the MLE parameter values ±2 s.e.

Lake	Model	MLE	AIC	THETA	wAIC	Nref	N1	N2	hrf	me12	me21	me'12	me'21	Ts	Tsc	P	Q	O
Lf	**SC2m**	-1003.5	2027.1	2756.0	1.0	3608.5	708835.4	182275.6	-	0.00051	0.00072	0.00004	0.00034	121497.7	12377.1	0.95	-	0.98
							[76535.9;1341021.3]	[0;375859.8]		[0.0002;0.0007]	[0.0007;0.0008]	[0;0.0001]	[0.0002;0.0003]	[53802.5;189192.8]	[4041.5;10608.9]	[0.36;1.53]		[0.98;0.99]
Sk	**SC2N2m**	-803.2	1630.4	3990.2	1.0	6028.9	685784.1	244350.1	0.13	0.00102	0.01149	0.00010	0.00045	68789.4	15614.8	0.64	0.07	0.98
							[481947.9;887088.1]	[50642.5;438479.7]	[0.09;0.17]	[0.0005;0.0015]	[0.008;0.0149]	[0;0.0001]	[0;0.001]	[60348.9;77651.8]	[0;50642.5]	[0.21;1.07]	[0.00;0.19]	[0.96;1.00]
Lu	**SC2N2m**	-887.0	1797.9	3440.0	1.0	6057.4	295539	297235	0.62	0.00079	0.00069	0.00058	0.00001	154833.5	24168.9	0.50	0.02	0.98
							[141621.2;449456.6]	[61482.2;532563.7]	[0.00;1.29]	[0.0003;0.0012]	[0;0.0013]	[0;0.001]	[0;0.00008]	[31801.1;277865.7]	[3392.1;44945.6]	[0.35;0.65]	[0.00;0.07]	[0.96;1.00]
Zu	**SC2m**	-1133.2	2286.3	3988.0	1.0	8874.7	174565.1	79517.2	-	0.00005	0.02647	0.00000	0.00014	106851.2	28576.5	0.91	-	0.99
							[0;765974.1]	[11803.3;147231.0]		[0;0.0013]	[0.0093;0.0436]	[0;0.0142]	[0;0.0004]	[59016.6;154685.7]	[1242.4;54668.0]	[0.33;1.48]		[0.99;0.99]

3.9 Figures

Figure 3.1: Geographic locations of the lakes where sympatric whitefish species pairs were sampled. Two lakes were sampled in Norway: Langfjodrvatn and Skrukkebukta lakes, and two lakes were samples in Switzerland: Zurich and Lucerne lakes.

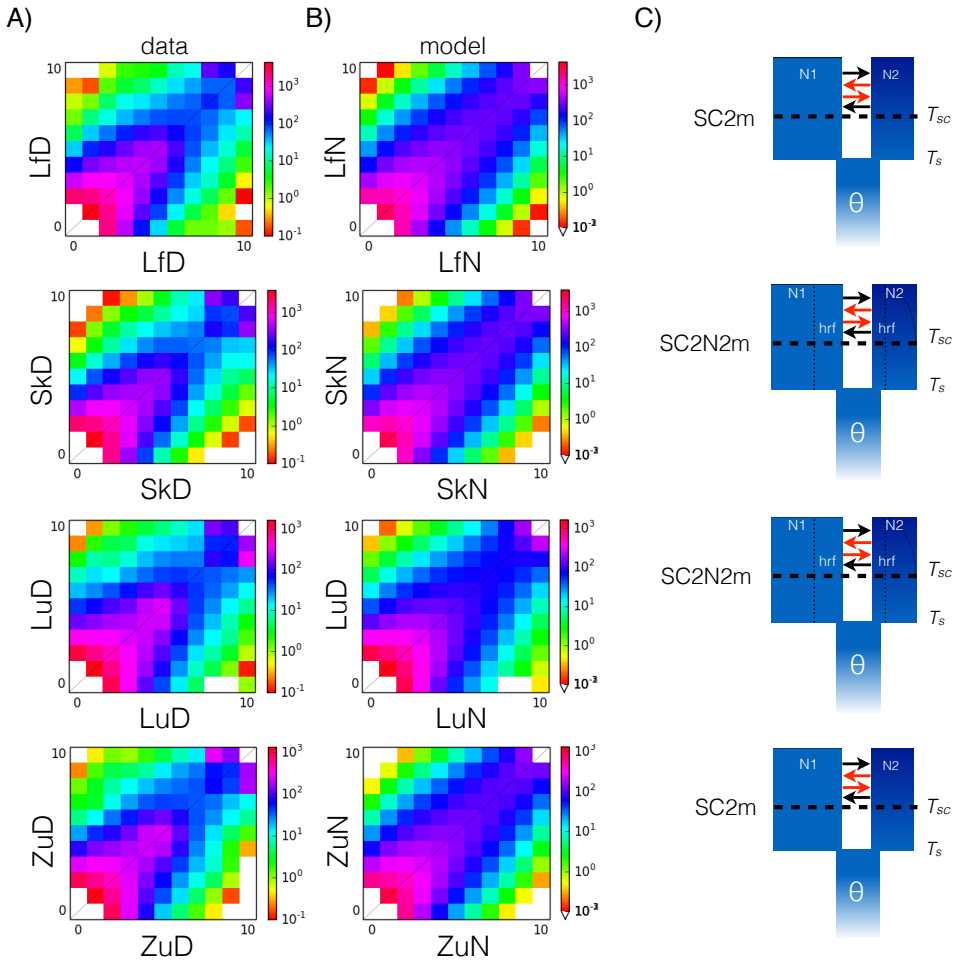


Figure 3.2: Historical demography of whitefish species-pairs. A) Observed joint allele frequency spectrum (JAFS) for benthic (-B; x axis) and limnetic (-L; y axis) populations for each of four lakes (Lf, Langfjordvatn Lake; Sk, Skrukkebukta Lake; Lu, Lucerne Lake and Zu, Zurich lake), obtained by projection of empirical data to 10 chromosomes. B) Predicted JAFS of the fittest model for each lake. For each JAFS, the color scale correspond to the number of SNPs for each bin, defined by the counts of the number of derived allele observed in limnetic and benthic populations. (C) Representation of the fittest model for each lake. Each blue block corresponds to a population; dashed lines indicate the time of secondary contact; black arrows symbolize neutral gene flow, while red arrows symbolize reduced gene flow around locally adapted loci.

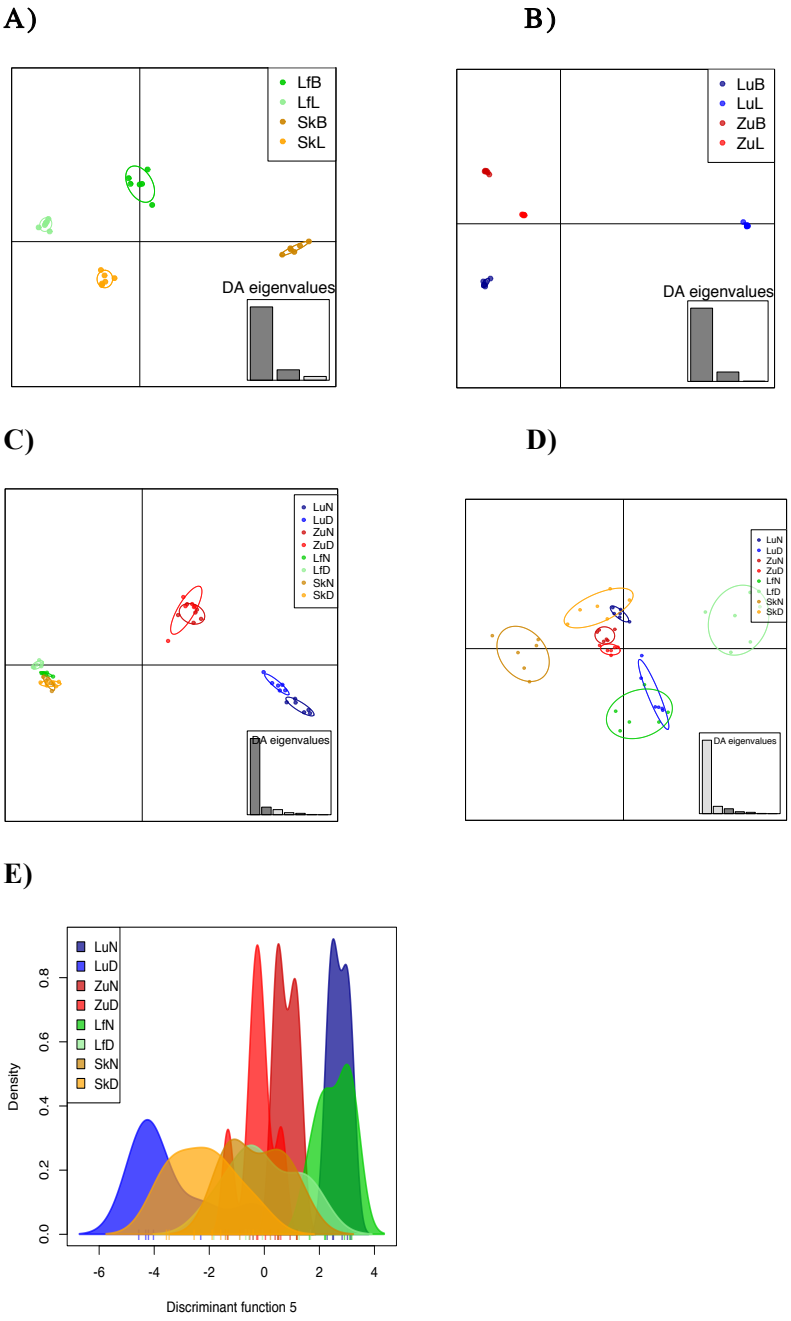


Figure 3.3: Genetic structure and relationships among lakes and species. A) Discriminant analysis of principal components (dAPC) of the lakes in Norway and B), for the lakes in Switzerland. For both regions, the first axis captures the signal of species differentiation and the second axis separates species pairs into their respective lakes. C) dAPC of the different lakes of the study. The first axis separates lakes according to their region. The second axis discriminate mainly lakes from Switzerland. D) dAPC including all lakes, with most of the variation discriminating populations from Norway, capturing the signal of lakes on the axis three and the signal of species

differentiation on the axis four. E) The fifth axis captures the residual genetic proximity of populations of the same species, allowing separating limnetic and benthic sympatric populations.

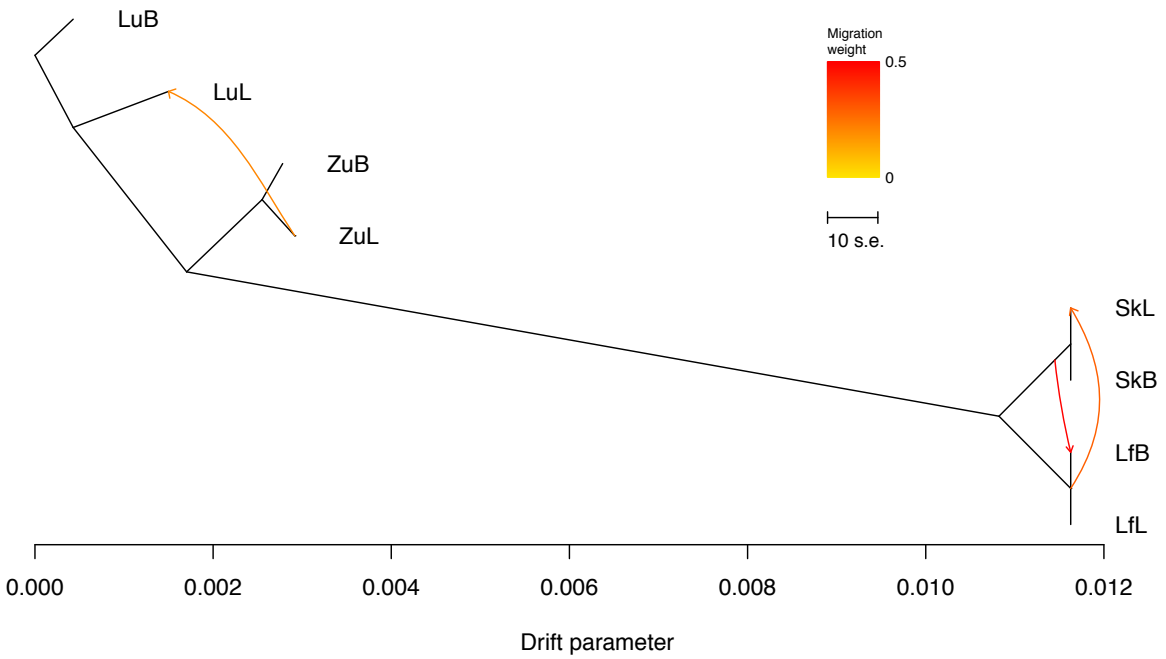


Figure 3.4: Shared ancestral genetic variation between allopatric and admixture events between sympatric species pairs. The TreeMix tree represented here allowed three migration events. Migrations events links limnetic populations from different lakes in Switzerland, and ancestral populations from each lake in Norway with contemporary populations. Horizontal branch lengths are proportional to the amount of genetic drift in each branch, and the scale bar indicates 10 times the average standard error (SE) of the entries in the covariance matrix between pairs of populations. Color scale indicates the weight of inferred migration events.

3.10 Supplementary tables

Table S3.1: Number of raw and filtered reads per individual.

Library	Individual	Species	Raw_reads	Trimmed_reads
LIB1	SK08	Benthic	20582507	19753880
LIB10	Z30	Limnetic	20619799	19688042
LIB11	CN11	Benthic	20169272	19295658
LIB12	LF38	Benthic	24858616	23818507
LIB14	L85	Limnetic	24427813	23359144
LIB15	SK10	Benthic	22828163	21912572
LIB16	CN10	Benthic	25277259	24265292
LIB17	Z23	Limnetic	21296255	20525864
LIB18	LF109'	Limnetic	30904936	29818408
LIB2	LF112	Limnetic	22833090	21850810
LIB20	L56	Limnetic	25402654	24518086
LIB22	Z27	Limnetic	21385377	20566388
LIB23	L46	Limnetic	20070114	19294910
LIB24	SK14	Limnetic	25153932	24251100
LIB26	SK39	Limnetic	25376555	24304127
LIB27	LF41	Benthic	23221581	22394148
LIB28	Z12	Benthic	28935800	27811520
LIB29	Z19	Benthic	18838742	18186632
LIB3	SK09	Benthic	19912485	19111721
LIB32	SK11	Benthic	21830728	21021795
LIB33	LF113	Limnetic	34202220	32821823
LIB34	Z13	Benthic	14858370	14212670
LIB35	L53	Limnetic	19664051	18865826
LIB36	SK03	Limnetic	27947100	26854678

LIB37	CN12	Benthic	26388655	25351805
LIB4	Z15	Benthic	27740869	26602363
LIB40	Z29	Limnetic	31023995	29809348
LIB41	Z11	Benthic	19537598	18931029
LIB42	L68	Limnetic	28025442	27205514
LIB43	CD21	Limnetic	20270143	19650600
LIB45	SK02	Limnetic	25145896	24375925
LIB46	L86'	Benthic	17235327	16676866
LIB47	CD20	Limnetic	25391937	24577489
LIB48	LF37'	Benthic	21936910	21266681
LIB49	Z20	Benthic	22155698	21220544
LIB50	SK71	Limnetic	19642561	18838271
LIB52	CD19	Limnetic	28958463	27755684
LIB53	CD22	Limnetic	27820603	26670427
LIB54	LF35'	Benthic	24449486	23452258
LIB55	L49	Limnetic	30066776	28875128
LIB56	LF39	Benthic	28375763	27242770
LIB57	SK19	Benthic	24999646	23739395
LIB58	CD17	Limnetic	27140147	25803019
LIB59	L83	Limnetic	22364555	21266933
LIB6	CN6	Benthic	27467349	26358596
LIB60	CN5	Benthic	19599732	18676563
LIB61	LF108'	Limnetic	30595792	29159603
LIB62	SK48	Benthic	25208038	24004805
LIB63	LF111	Limnetic	42568708	40665197
LIB64	Z31	Limnetic	20743920	19797211
LIB65	L82	Limnetic	29186773	27884328
LIB66	LF107'	Limnetic	36016311	34438149

LIB67	L81	Limnetic	31202915	29850242
LIB68	CN7	Benthic	32713977	31286322
LIB69	Z37	Limnetic	9701395	9001555
LIB7	L52	Limnetic	29081812	27955687
LIB70	CD18	Limnetic	33866686	32293410
LIB72	SK36	Limnetic	24041196	22983982
LIB8	L58	Limnetic	27575796	26458514
LIB9	LF36'	Benthic	23115781	22169931

Table S3.2: The demographic divergence history of whitefish species pairs inferred raw parameters. For all lakes, details of statistics and demographic parameter values for the fittest models determined under the threshold of ΔAIC < 10. The table contains in this order the maximum likelihood (MLE) for each model, the value of Akaike information criterion (AIC) and the weighted AIC (wAIC). Then, the inferred demographic parameter values converted with the estimate of THETA: the ancestral effective population size before population split (Nref); the effective population size after split for limnetic (N_1) and benthic (N_2) populations; The Hill-Robertson factor (*hrf*) corresponds to the degree to which the effective population size of diverging populations is locally reduced linked selection effect. Migration parameters include migration rates from benthic population to limnetic population (m_{e12}) and reciprocally (m_{e21}), and a second category of effective migration rates ($m_{e'12}$) and ($m_{e'21}$) applying to a second category of loci. Time parameters include the duration of the allopatric divergence period (T_s), and the duration of the migration period (T_{SC}). Finally, the table also contains proportion parameters such as the proportion (Q) of the genome with effective population sizes N_1 and N_2. A proportion (P) of the genome is occupied by loci with effective migration rates m_{e12} and m_{e21} (and a second category of loci occupying a fraction 1-P of the genome has effective population sizes m'_{e12} and m'_{e21}). The parameter (O) is the proportion of correct SNP orientation. Numbers into brackets denote 95% confidence intervals obtained using the MLE parameter values ±2 s.e.

Lake	Model	MLE	AIC	THETA	wAIC	Nref	N1	N2	hrf	me12	me21	me'12	me'21	Ts	Tsc	P	Q	O
Lf	**SC2m**	1003.5	2027.1	2756.0	1.0	3608.5	28.06	7.22	-	1.05	1.49	0.08	0.70	4.81	0.49	0.95	-	0.98
							[0.00;71.80]	[0.00;21.88]		[0.47;1.63]	[1.39;1.58]	[0.00;0.25]	[0.58;0.81]	[2.13;7.49]	[0.16;0.72]	[0.36;1.53]		[0.98;0.99]
Sk	**SC2N2m**	803.2	1630.4	3990.2	1.0	6028.9	16.25	5.79	0.13	3.51	39.57	0.33	1.56	1.63	0.37	0.64	0.07	0.98
							[11.14;21.36]	[0;27.67]	[0.09;0.17]	[1.79;5.22]	[27.68;51.46]	[0.29;0.38]	[0.00;3.54]	[0.00;4.20]	[0.00;1.20]	[0.21;1.07]	[0.00;0.19]	[0.96;1.00]
Lu	**SC2N2m**	887.0	1797.9	3440.0	1.0	6057.4	6.97	7.01	0.62	2.75	2.39	2.01	0.04	3.01	0.57	0.50	0.02	0.98
							[3.34;10.60]	[1.45;12.56]	[0.00;1.29]	[1.34;4.17]	[0.16;4.62]	[0.30;3.72]	[0.00;0.29]	[0.75;5.27]	[0.08;1.06]	[0.35;0.65]	[0.00;0.07]	[0.96;1.00]
Zu	**SC2m**	1133.2	2286.3	3988.0	1.0	8874.7	2.81	1.28	-	0.23	134.24	0.00	0.71	1.72	0.46	0.91	-	0.99
							[0.00;12.33]	[0.19;2.37]		[0.00;6.63]	[47.33;221.15]	[0.00;72.19]	[0.00;2.45]	[0.95;2.49]	[0.02;0.88]	[0.33;1.48]		[0.99;0.99]

Table S3.3: Significant (*P*<0.001) *f*4-statistics inferred between populations.

trees	f4	se	z.score
SkL,ZuL;LuL,LuB	-0.000289398	8.83947e-05	-3.27393
LfB,SkB;SkL,ZuL	-0.000302725	9.11541e-05	-3.32103
LfL,ZuL;LuL,LuB	-0.000314762	8.97307e-05	-3.50785
LfL,LfB;SkB,ZuB	-0.000347533	0.000105049	-3.30831
LfB,SkL;SkB,ZuB	-0.000351642	0.000111237	-3.16118
LfB,SkB;SkL,LuL	-0.000355242	8.92466e-05	-3.98046
LfB,ZuL;LuL,LuB	-0.000360487	9.2397e-05	-3.9015
LfB,SkB;SkL,LuB	-0.000369928	9.64598e-05	-3.83505
LfB,SkL;SkB,LuB	-0.000400086	0.000102122	-3.91772
LuL,ZuB;ZuL,SkB	-0.000543418	9.88115e-05	-5.49955
LfL,SkL;SkB,LuB	-0.000570101	0.000115928	-4.9177
LfL,SkL;SkB,ZuB	-0.000699175	0.000123984	-5.63923
LuL,SkB;ZuL,LuB	-0.000703014	0.000142664	-4.92776
LfL,ZuB;ZuL,LuB	-0.000725595	0.000134019	-5.41411
LfL,ZuL;ZuB,LuB	-0.000792352	0.000111192	-7.12599
SkL,ZuB;ZuL,LuB	-0.000808234	0.000119212	-6.77982
ZuL,LuB;SkB,ZuB	-0.000813327	0.000126141	-6.44777
LfB,ZuB;ZuL,LuB	-0.000880529	0.000129311	-6.80941
SkL,LuB;LuL,ZuB	-0.000884312	0.000117211	-7.54459
LuL,SkB;ZuB,LuB	-0.000898513	0.0001216	-7.38911
LfB,LuB;LuL,ZuB	-0.000906957	0.000110068	-8.23999
SkL,ZuL;ZuB,LuB	-0.000921427	0.000103795	-8.87734
LfB,ZuL;ZuB,LuB	-0.000969871	0.000112437	-8.62594
LuL,ZuB;SkB,LuB	-0.000972923	0.000111574	-8.71997

LfB,LuB;LuL,ZuL	-0.000996299	0.000109071	-9.13442
SkL,LuB;LuL,ZuL	-0.000997505	0.000126554	-7.88205
LfL,LuB;LuL,ZuB	-0.00103875	0.000116779	-8.89505
LuL,ZuL;SkB,LuB	-0.00104882	0.000121565	-8.62759
LfL,LuB;LuL,ZuL	-0.00110551	0.000132905	-8.31803
LuL,ZuB;ZuL,LuB	-0.00151634	0.000108171	-14.018
LuL,ZuL;ZuB,LuB	-0.00178773	0.00011765	-15.1954
LfL,LuB;ZuL,SkB	-0.0089921	0.000207565	-43.3219
LfL,ZuB;LuL,SkB	-0.0091991	0.000211424	-43.5103
LfB,LuB;ZuL,SkB	-0.00931705	0.000228717	-40.7362
LfL,LuL;ZuL,SkB	-0.0093379	0.000233718	-39.9537
LfL,ZuL;LuL,SkB	-0.0093946	0.000233427	-40.2464
LfB,ZuB;LuL,SkB	-0.00941484	0.000221105	-42.5807
LfB,ZuL;LuL,SkB	-0.00961034	0.000250455	-38.3715
SkL,LuB;ZuL,SkB	-0.00964484	0.000230229	-41.8924
LfB,LuL;ZuL,SkB	-0.00966285	0.000248114	-38.9452
SkL,ZuB;LuL,SkB	-0.00974383	0.000222589	-43.7751
LfL,ZuB;ZuL,SkB	-0.00988132	0.000235304	-41.9939
SkL,ZuL;LuL,SkB	-0.00993933	0.000250901	-39.6146
SkL,LuL;ZuL,SkB	-0.00999064	0.000251751	-39.6847
LfL,LuB;LuL,SkB	-0.0100976	0.000237668	-42.4861
LfB,ZuB;ZuL,SkB	-0.0102063	0.00025279	-40.3746
LfB,LuB;LuL,SkB	-0.0103133	0.000245447	-42.0186
SkL,ZuB;ZuL,SkB	-0.0105341	0.00025673	-41.0317
SkL,LuB;LuL,SkB	-0.0106423	0.000247947	-42.9219
LuL,LuB;ZuL,ZuB	0.000271391	6.23069e-05	4.35571
LfL,LfB;ZuL,SkB	0.000324948	0.000108168	3.00412

LfB,SkL;ZuL,SkB	0.000327791	0.000102432	3.20009
LfB,SkL;LuL,SkB	0.000328997	0.000100173	3.28429
LfL,LuL;SkL,SkB	0.000334994	0.000100333	3.33883
LuL,LuB;ZuL,SkB	0.000345801	9.38548e-05	3.68443
LfL,LfB;SkL,SkB	0.000361239	6.41679e-05	5.62958
LfL,ZuL;SkL,SkB	0.000386304	9.21214e-05	4.19342
LfL,LuB;SkL,SkB	0.000391397	9.17034e-05	4.26808
LfL,ZuB;SkL,SkB	0.000423605	9.03778e-05	4.68704
LfL,LuB;LfB,SkB	0.000441017	9.72435e-05	4.53518
LfL,SkB;LfB,SkL	0.000449706	9.40272e-05	4.78272
LfL,LuL;LfB,SkB	0.000455703	0.000101324	4.49746
LfL,ZuL;LuL,ZuB	0.000477591	9.54868e-05	5.00164
LfL,ZuL;LfB,SkB	0.000508219	0.000102753	4.94602
LfL,ZuB;LfB,SkB	0.000521669	9.75046e-05	5.3502
LfL,SkL;LuL,SkB	0.000544737	0.000121826	4.47143
LfB,ZuL;LuL,ZuB	0.000609384	0.000105313	5.78642
SkL,ZuL;LuL,ZuB	0.000632029	0.000101154	6.24817
LfB,LuL;ZuL,LuB	0.000635812	0.000131827	4.82309
LfL,SkL;ZuL,SkB	0.000652739	0.000117379	5.56094
LfL,ZuB;LuL,ZuL	0.000682225	0.000110126	6.19492
SkL,LuL;ZuL,LuB	0.000708108	0.000141867	4.99134
LuL,ZuL;SkB,ZuB	0.000738917	0.0001059	6.97751
SkL,ZuB;LuL,ZuL	0.000790227	0.000108322	7.29518
LfL,LuL;ZuL,LuB	0.000790746	0.000155456	5.08663
LfB,ZuB;LuL,ZuL	0.000791433	0.00010815	7.3179
LfL,SkL;LfB,SkB	0.000810945	0.000100456	8.07267
LfB,LuL;ZuB,LuB	0.000817861	0.000109676	7.45708

SkL,LuL;ZuB,LuB	0.000866305	0.000121894	7.10706
ZuL,SkB;ZuB,LuB	0.000889219	0.000107962	8.23639
LfL,LuL;ZuB,LuB	0.00099538	0.000127395	7.81334
LfL,LuB;SkB,ZuB	0.00905886	0.000215133	42.1083
LfL,ZuL;SkB,LuB	0.00907983	0.0002122	42.7889
LfL,LuL;SkB,ZuB	0.00913327	0.000218959	41.7122
LfL,ZuB;SkB,LuB	0.00915573	0.000207002	44.2301
LfB,ZuL;SkB,LuB	0.00924985	0.000236786	39.0642
LfB,ZuL;SkL,LuB	0.00927491	0.000223507	41.4971
LfB,ZuB;SkB,LuB	0.00932574	0.000227201	41.0462
LfB,LuB;SkL,ZuL	0.00934721	0.00022781	41.0306
LfL,LuB;SkL,ZuL	0.0093835	0.000225621	41.5897
LfB,ZuB;SkL,LuB	0.00938811	0.000228725	41.0455
LfB,LuB;SkB,ZuB	0.00940639	0.000225843	41.6501
LfL,LuB;LfB,ZuL	0.00943312	0.000228196	41.3378
LfB,LuB;SkL,ZuB	0.00943655	0.000229084	41.1926
LfL,LuB;SkL,ZuB	0.00945026	0.000228319	41.3907
LfB,LuL;SkL,ZuB	0.00945456	0.000217455	43.4782
LfL,ZuL;SkL,LuB	0.00946614	0.000213363	44.3664
LfL,LuL;SkL,ZuB	0.00946826	0.000215355	43.9659
LfB,ZuB;SkL,LuL	0.0094772	0.000216377	43.7994
LfB,LuL;SkB,ZuB	0.0094808	0.000224693	42.1944
LfL,LuB;LfB,ZuB	0.00949988	0.000229594	41.3768
LfL,ZuB;SkL,LuB	0.00957933	0.00022084	43.3768
LfL,ZuL;LfB,LuB	0.00958805	0.000221888	43.2113
LfL,LuL;LfB,ZuB	0.00958897	0.000219851	43.6159
LfL,ZuB;SkL,LuL	0.0096227	0.00021965	43.8092

LfB,ZuL;SkL,LuL	0.0096354	0.000234963	41.0082
LfB,LuL;SkL,ZuL	0.00963661	0.000237651	40.5494
SkL,ZuL;SkB,LuB	0.00964993	0.000238656	40.4346
LfL,LuL;SkL,ZuL	0.0096729	0.000238882	40.4923
LfL,ZuB;LfB,LuB	0.0096774	0.000219812	44.0258
LfL,ZuB;LfB,LuL	0.00972077	0.000221873	43.8123
SkL,ZuB;SkB,LuB	0.00972583	0.000234049	41.5546
SkL,LuB;SkB,ZuB	0.00975803	0.000234454	41.6203
LfL,ZuL;SkL,LuL	0.0097809	0.000235335	41.5616
LfL,LuL;LfB,ZuL	0.00979361	0.000243895	40.1551
SkL,LuL;SkB,ZuB	0.00983244	0.000229996	42.7505
LfL,ZuL;SkB,ZuB	0.00987219	0.000239523	41.216
LfL,ZuL;LfB,LuL	0.00990281	0.000242771	40.7908
LfL,LuL;SkB,LuB	0.0101286	0.000237533	42.641
LfB,ZuL;SkB,ZuB	0.0102197	0.000256353	39.8658
LfB,ZuL;SkL,ZuB	0.0102448	0.000245724	41.6922
LfL,ZuL;SkL,ZuB	0.0102585	0.000239012	42.9204
LfB,ZuB;SkL,ZuL	0.0102686	0.000248009	41.4044
LfB,LuL;SkL,LuB	0.0102724	0.000247216	41.5524
LfB,LuL;SkB,LuB	0.0102987	0.000251559	40.9394
LfL,ZuB;SkL,ZuL	0.0103049	0.000246084	41.8757
LfB,LuB;SkL,LuL	0.0103435	0.00024529	42.1685
LfL,ZuL;LfB,ZuB	0.0103804	0.000242023	42.8902
LfL,ZuB;LfB,ZuL	0.010403	0.000243497	42.7234
LfL,LuL;SkL,LuB	0.0104636	0.000240686	43.4743
LfL,LuB;SkL,LuL	0.010489	0.000249228	42.0861
LfL,LuB;LfB,LuL	0.0105386	0.000249335	42.267

SkL,ZuL;SkB,ZuB	0.0105714	0.000262663	40.2469
LfL,LuL;LfB,LuB	0.0105844	0.000238332	44.4101
SkL,LuL;SkB,LuB	0.0106988	0.000254505	42.0375

Chapitre 3: Convergent transcriptomic landscapes accompany intercontinental parallel evolution within a Nearctic Coregonus (Salmonidae) sister-species complex.

4.1 Résumé

Contrairement à la multitude d'études se focalisant sur les bases génomiques de la divergence phénotypique adaptative, le rôle de l'expression génique durant le processus de spéciation a été moins étudié. La convergence de patrons de différentiels d'expression entre paires d'espèces proches peut refléter le rôle de la sélection naturelle durant le processus de spéciation écologique. Dans cette étude, nous testons la convergence inter-continentale de signatures de différentiel d'expression entre paires d'espèces limnétiques et benthique de Grand Corégone et de Lavaret, entre six replicats de paires d'espèces sympatriques, répartis en Amérique du Nord, en Norvège et en Suisse. En se basant sur un assemblage de transcriptome *de novo*, nous avons caractérisé la variation génomique et le différentiel d'expression entre espèces limnétiques et benthiques sympatriques entre régions et continents. Nous avons trouvé des gènes différentiellement exprimés (DEG) entre espèces limnétique et benthiques qui sont enrichis en polymorphisme partagé au sein des lignées soeur Grand Corégone-Lavaret. Nous avons ensuite utilisé des approches de co-variation de génotypiques et d'expression géniques afin d'identifier les cibles de sélection polygéniques. Nous avons identifié des loci differenciés parallèles entre espèces ainsi que des DEG sur-exprimés chez l'espèce limnétique par comparaison à l'espèce benthique. Nous avons aussi découvert un *cis*-eQTL parallèle affectant le niveau d'expression génique de la pyruvate kinase, un gène impliqué dans le métabolisme de la glycolyse. Notre approche de transcriptomique comparative nous a permis d'identifier des fonctions biologiques et des gènes candidats enrichis en polymorphisme partagé et associés à la divergence phénotypique parallèle entre paires d'espèces, au sein des lignées soeurs. Nos résultats indiquent le rôle de la sélection naturelle sur le maintien de la diversité de traits phénotypiques impliqués dans la spéciation entre espèces limnétiques et benthiques de Grand Corégone et de Lavaret.

4.2 Abstract

In contrast to the plethora of studies focusing on the genomic basis of adaptive phenotypic divergence, the role of gene expression during speciation has been much less investigated and consequently, less understood. Yet, the convergence of differential gene expression patterns between closely related species-pairs might reflect the role of natural selection during the process of ecological speciation. Here, we test for intercontinental convergence in differential transcriptional signatures between limnetic and benthic sympatric species-pairs of Lake Whitefish (*Coregonus clupeaformis*) and its sister-lineage, the European Whitefish (*C. lavaretus*), across six replicated sympatric species-pairs (two in North America, two in Norway and two in Switzerland). Following *de novo* transcriptome assembly, we characterized genomic variation and differential gene expression between sympatric limnetic and benthic species across regions and continents. We found differentially expressed genes (DEG) between limnetic and benthic whitefish that were enriched in shared polymorphism among sister-lineages. We then used both genotypes and co-variation in expression in order to infer polygenic selection at gene level. We identified parallel outliers and DEG involving genes primarily over-expressed in limnetic species relative to the benthic species. We also discovered a parallel-shared *cis*-eQTL affecting the pyruvate kinase expression level, a gene involved in glycolysis metabolism. Overall, comparative transcriptomics across continents allowed identifying biological functions and candidate genes enriched in shared polymorphism associated with parallel phenotypic divergence across species-pairs among the sister-lineages. Our results are consistent with a longstanding role of natural selection in maintaining diversity at phenotypic traits involved in ecological speciation between limnetic and benthic whitefishes.

4.3 Introduction

Deciphering the genomic basis of differential adaptations between divergent populations, ultimately leading to ecological speciation, has been of foremost interest over the last decade. The concept of adaptive divergence implies different population phenotypic responses to constraints associated with selective pressures stemming from different environments. This is particularly well illustrated by the occurrence of independent parallel phenotypic evolution among closely related and locally adapted nascent species (Endler 1986; Orr 2005b; Losos 2011). Parallel phenotypic evolution can emerge from repeated divergence of the same genomic regions (Conte, Arnegard, Peichel, and Schluter 2012a) or from different genes involved in similar or different biological pathways (Cohan and Hoffmann 1989; Losos 2011), and has been associated with changes in gene expression during adaptive divergence (Pavey et al. 2010; Manceau et al. 2011; Harrison et al. 2012). Indeed, adaptive genetic changes include both variations in regulation of gene expression resulting in phenotypic variation (Rebeiz et al. 2009) and sequence variation altering proteins structures (Hoekstra et al. 2006). Genetic variation underlying parallel phenotypic changes may originate from parallelism at the molecular level that has arisen from *de novo* mutations affecting the same genes (Manceau et al. 2010; Rockman 2012). However, such mutations are generally associated with loci of large effect controlling the expression of a given phenotypic trait with a mono-/oligo-genic architecture (Manceau et al. 2010). This contrasts with the polygenic architecture of most complex traits, including those thought to be involved in ecological speciation and more generally in adaptation (Savolainen, Lascoux, and Merilä 2013b; Yeaman 2015; Gagnaire and Gaggiotti 2016). For such traits, standing genetic variation is usually seen as an important source of adaptive mutations, although the origin of standing variants may be complex and varied (Welch and Jiggins 2014). Several recent studies showed that standing variation may originate from past admixture events (Roesti et al. 2014; Martin et al. 2015; Meier, Sousa, et al. 2017; Rougeux et al. 2017), suggesting an important role of anciently diverged variants in the process of ecological speciation. Despite the increasing number of studies underlining the fundamental role of standing variation as the main fuel for adaptation (Barrett and Schluter 2008; Schrider and Kern 2017), relatively few have focused on the possible consequences of different levels of standing genetic variation (*e.g.*, because of different historical contingency) across populations on the fate of parallel phenotypic evolution (Nelson and Cresko 2018).

The increased resolution of genomic analyses has allowed revealing the importance of complex polygenic genomic architectures of traits involved in local adaptation, sometimes accompanied by genetic redundancy (Slatkin 1978; Berg and Coop 2014b; Yeaman 2015; Laporte, Pavey, et al. 2016; Yeaman, Hodgins, et al. 2016; Babin et al. 2017; Bay et al. 2017). Despite repeated claims for the importance of investigating the molecular basis of local adaptation and ecological speciation from a polygenic standpoint (Yeaman, Hodgins, et al. 2016; Babin et al. 2017; Bay et al. 2017; Harrisson et al. 2017; Jain and Stephan 2017), such studies are still in their infancy, particularly as pertaining to the investigation of parallel speciation.

In this study, we compare the genomic basis of limnetic-benthic divergence among sympatric species-pairs from two different evolutionary lineages; the North American lake whitefish (*Coregonus clupeaformis* species complex) and the European whitefish (*C. lavaretus* species complex). As limnetic and benthic species-pairs are reproductively isolated, they may therefore fulfill the main criterion of biological species. Therefore in this study, *C. clupeaformis* and *C. lavaretus* refer only to the main phylogeographic lineages from northern America and Europe, respectively (*i.e.*, the upper hierarchical level of differentiation, see below). By contrast, we use the term "species" strictly to refer to limnetic and benthic whitefish independent of their continent of origin. *C. clupeaformis* and *C. lavaretus* lineages became geographically isolated ~500,000 years ago and have evolved independently since then (Bernatchez and Dodson 1991; Bernatchez and Dodson 1994; Jacobsen et al. 2012). This system thus offers a valuable model to study the genomic and transcriptomic underpinnings of parallel differential adaptations leading to ecological speciation in independent lineages. Indeed, both lineages comprise several isolated lakes harbouring partially reproductively isolated sympatric benthic (normal) and limnetic (dwarf) species-pairs. The European whitefish appears to be of admixed origin following contact between glacial sub-lineages (Hudson et al. 2010), resulting in independent events of intra-lacustrine (sympatric) evolution of benthic and limnetic species across Scandinavian and Alpine lakes (Douglas et al. 1999; Østbye et al. 2005; Østbye et al. 2006). The North American lake whitefish sympatric species-pairs are also the result of a post-glacial secondary-contact between two glacial sub-lineages during the late Pleistocene. The allopatric phase that likely lasted about 60,000 years allowed the accumulation of genomic incompatibilities, while character displacement leading to phenotypic and ecological

divergence followed secondary contact in sympatry around 12,000 years ago (Bernatchez and Dodson 1990; Bernatchez and Dodson 1991; Pigeon, Chouinard, and Bernatchez 1997b; Rougeux et al. 2017).

Lake whitefish have been the subject of numerous studies pertaining to the ecological and genomic basis of adaptive divergence between limnetic and benthic species. The limnetic species differ from the benthic species in their use of habitat and trophic resources, with a higher metabolic rate and more active swimming behaviour for foraging and predator avoidance (Bernatchez et al. 1999; Trudel et al. 2001; Rogers et al. 2002), reduced energy allocated to growth relative to benthic whitefish (Trudel et al. 2001; Rogers and Bernatchez 2004) and differences in morphology and life history traits (Rogers and Bernatchez 2007a) which are under polygenic control (Gagnaire, Pavey, et al. 2013; Laporte, Rogers, Dion-Côté, Normandeau, Gagnaire, Dalziel, Chebib, and Bernatchez 2015b). Given the recent divergence between geographical lineages, an important part of the even more recent divergence between species may have relied on divergent selection acting on standing genetic variation. A possible outcome could be that divergent selection between species has been actively maintaining shared polymorphism at selected variants across continents, protecting them from being lost by drift.

The main goal of this study was to investigate the role of ancestral polymorphism on differential transcriptional signatures between limnetic and benthic species. Our aim was to test the general hypothesis that an overlapping polygenic basis underlies the parallel phenotypic divergence observed between sympatric species-pairs from sister lineages living on two different continents. Specifically, i) we first documented the amount and functional role of shared ancestral genetic polymorphism within coding genes among populations of the entire system, ii) we then quantified the extent of differential gene expression between benthic and limnetic species at the local (lake), regional and inter-continental scales, iii) we tested whether genes differentially expressed display an excess of shared polymorphism between sister-lineages, and finally iv) explored associations between polymorphism and variation in expression at the gene level.

4.4 Material and Methods

Sample collection, library preparation and sequencing

C. clupeaformis samples were collected from Indian Lake and Cliff Lake, Maine (USA) (Figure 4.1) in 2010, and correspond to samples used in previous RAD-Seq studies (Gagnaire, Pavey, et al. 2013; Rougeux et al. 2017). These lakes are part of a well-studied lake whitefish system (Bernatchez et al. 2010) and comprise the most divergent species-pairs along the divergence continuum described in previous studies (Renaut et al. 2012; Gagnaire, Pavey, et al. 2013; Rougeux et al. 2017). In parallel, *C. lavaretus* individuals were sampled in two Scandinavian lakes in Norway: Skrukkebukta, Langfjordvatn and two alpine lakes in Switzerland: Zurich and Lucerne (Figure 4.1). We chose these European lakes as they each contained only two sympatric limnetic-benthic populations (i.e., excluding potential genetic interactions with other sympatric whitefish forms that occur in other lakes) consistent with our sampling for *C. clupeaformis*. Also, the sympatric limnetic-benthic populations from Fennoscandinavia have a different evolutionary origin from those from the central Alpine lakes region (Østbye et al. 2005). For each species pair, six benthic and six limnetic individuals were sampled (72 samples in total). Fresh liver biopsies were taken, flash frozen, and stored at -80°C for Lake whitefish, while European whitefish livers were stored directly in RNAlater.

Total RNA was extracted using the RNAeasy Mini Kit following the manufacturer's instructions (Qiagen, Hilden, Germany). RNA quantification was done with a NanoDrop2000 spectrophotometer (Thermo Scientific, Waltham, MA, USA), and quality was assessed using the 2100 Bioanalyser (Agilent, Santa Clara, CA, USA). Only high-quality samples with a RIN value greater than or equal to eight (intact rRNA and no detectable trace of gDNA) were kept for subsequent steps. Final RNA concentration was measured with Quant-iT RiboGreen RNA Assay Kit (Invitrogen, Life Technologies, Carlsbad, CA, USA) before library preparation.

Individual libraries were prepared from 2µg of RNA using the TruSeq RNA sample preparation kit V2 (Illumina, San Diego, CA, USA) following the manufacturer's instructions. Library size and concentration were evaluated using DNA High Sensitivity chip on the 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA). Single read sequencing (100bp) was

performed on the Illumina HiSeq 2000 platform for the 72 libraries (eight libraries per lane for a total of nine lanes) at the McGill University and Genome Quebec Innovation Centre (Montreal, Canada).

De novo transcriptome assembly and annotation

Raw sequencing reads were cleaned to remove adaptor and individual tag sequences, and trimmed using Trimmomatic v0.36 (Bolger et al. 2014). We applied a quality score of 30 across a 10bp sliding window and removed all reads < 60 nucleotides in length after processing. FLASh-merged reads, v1.2.11 with default parameters, (Magoc and Salzberg 2011) from all individuals were used to assemble a de novo reference transcriptome using Trinity v2.2.0 suite (Haas et al. 2013). We aimed to build an orthologous gene composite reference for the two lineages. Contigs lacking an ORF longer than 200bp were discarded such as redundant transcript per ORF in favour of a uniq ORF per transcript using TransDecoder v3.0.1 pipeline (-Transdecoder.LongOrfs) (TransDecoder 2016). In absence of reference genome, only the longest isoform per transcript were kept (Pasquier et al. 2016). Finally, a scaling factor of one Transcript per Million (TPM) was applied to normalize the raw reads count per gene for the gene expression analysis. In parallel, we also assembled a separate reference transcriptome for lake and European whitefish following same procedure as detailed above. From normalized transcripts, we considered reciprocal best hit within each transcriptome, for both *C. clupeaformis* and *C. lavaretus* to identify and discard paralogous genes (Carruthers et al. 2018). We then blasted each reference transcriptome to the common reference and discarded unmapped contigs, avoiding imbalance mapping between species of both lineages. We kept only common transcripts (i.e., found in both lineages) that we refer to as orthologous genes. We found 98,7% and 98.2% of such orthologous hits for *C. lavaretus* and *C. clupeaformis*, respectively. We finally used a BlastX approach against the Swissprot database (http://www.uniprot.org) and the Ensembl database for Danio rerio (Zv9) to annotate the reference transcriptome.

Differential gene expression analysis

Individual reads were mapped to the orthologous gene reference using Bowtie2 v2.1.0 (Langmead and Salzberg 2012) and resulting Bam files were parsed to estimate individual reads counts with eXpress v1.5.1 (Roberts and Pachter 2012). Differential

expression analysis were conducted with the R packages edgeR v3.18.1 (Robinson et al. 2010) and DESeq2 v1.14.1 (Love et al. 2014), but only results from DESeq2 analysis are presented, given to the similarities between outputs of the two tools (results not shown). First, we realized independent gene expression comparisons between species within lakes and identified shared genes among the different hierarchical levels as overlaps. In order to take into account the strong hierarchical structure of the studied populations (lake, region and continent), generalized linear models were built to allow for comparisons between benthic and limnetic species while integrating progressively lakes, regions and continents (hereafter called 'phylogeographic') effects, as covariates on gene expression. Final model for limnetic and benthic comparisons across continents was composed by: '~Species+Lake+Region+Continent+Species*Continent'. We then controlled for presence of DEGs associated with phylogeographic effects. Inference of differentially expressed genes (DEGs) relied on the integration of the size factor per library, to correct for heterogeneity in sample sequencing depth. DEGs were determined based on a q-value < 0.05. Then, GO enrichment analysis was performed with GOATOOLS (Tang, Klopfenstein, et al. 2015), based on Fisher's exact test. For all tested lists of genes, GO enrichment was associated with FDR<0.05 (Benjamini-Hochberg correction) (Benjamini and Hochberg 1995) and a minimum of three genes represented per category.

SNP genotyping and sequence divergence measures

In order to document the extent of polymorphism within and divergence between *C. clupeaformis* and *C. lavaretus* and among divergent sympatric benthic and limnetic species-pairs, individual reads were mapped (71.41% overall alignment mean success rate) to the common reference transcriptome using Bowtie2 v2.1.0 'end-to-end' alignment (Langmead and Salzberg 2012). Resulting SAM files were converted to BAM files and sorted using Samtools v1.3 (Li et al. 2009) and duplicates were removed with the Picard-tools program v1.119 (http://broadinstitute.github.io/picard/). The physical mapping information of reads to the reference was used for calling SNPs with Freebayes v0.9.10-3-g47a713e (Garrison and Marth 2012). Variable sites were filtered for a minimum coverage of three reads per individual in order to consider an allele. Further steps of genotype quality filtering were then applied to the data. We used the vcffilter program from vcflib (https://github. com/ekg/vcflib) to process the Variant Call Format (VCF) file obtained from Freebayes, in order to specifically retain biallelic SNPs with a phred scaled quality score above 30, a genotype quality with a

phred score higher than 20, and an allele depth balance between 0.30 and 0.70 (except for low-frequency variants for which we did not apply an allele depth balance filter). Following these quality control steps, we filtered the resulting VCF file using VCFtools v0.1.12b (Danecek et al. 2011), in order to remove miscalled and low quality SNPs for subsequent population genomics analyses. For each of the 12 populations, we kept loci with less than 10% of missing genotypes and filtered for Hardy-Weinberg disequilibrium using a p-value exclusion threshold of 0.01. Finally, we merged the VCF files from all the 12 populations, resulting in a unique VCF file containing 20,911 SNPs passing all the filters in each population. Since we did not apply any minor allele frequency threshold within populations, the final VCF represents a non-ascertained dataset of genetic variation. Intra-population nucleotide diversity (π) was estimated using a sliding-window of 100bp and a step size of 20bp with VCFtools v0.1.12b (Danecek et al. 2011) on the VCF file of shared variants and invariants sites. Finally, we estimated the between-species nucleotide diversity (Dxy) using a custom perl script, using a 100bp sliding-window with step-size of 20pb (in order to consider a maximum number of window due to the transcriptome specificities), along our reference transcriptome. Here, the absolute measure of divergence (Dxy) was inferred from formula detailed in previous studies (Nei 1987; Cruickshank and Hahn 2014).

Shared polymorphism and historical relationship among lineages and species

The inference of historical relationship among all populations was performed using Treemix v1.12 (Pickrell and Pritchard 2012) applied to the VCF file containing 20,911 polymorphic SNPs. This program uses the covariance structure of allele frequencies between populations and a Gaussian approximation for genetic drift to build a maximum likelihood graph relating populations with the ancestral genetic pool. The number of migration edges was determined empirically to improve the fit to the inferred tree. Migration edges between populations may either reflect gene flow, or the retention of shared ancestral polymorphism among geographically isolated populations.

We then quantified and compared the amount of shared polymorphism retained at the lake, region and continental hierarchical levels. More precisely, from our genotypes matrix we defined the number of polymorphic SNPs that were shared among all populations, for each hierarchical level (i.e., between sympatric species at lake level, among all population of Norway, Switzerland and Maine at region level, and among all

populations across continents at continent level). We also tested for the increased probability of DEGs relative to non-DEGs to display shared polymorphism, which could hint to a possible role for selection in maintaining variation at these genes. We defined the proportions of DEGs with shared polymorphism (DEG_{SP}) relative to the total number of DEGs (DEG_T), and of Non-DEGs with shared polymorphism ($NDEG_{SP}$) relative to the total number of Non-DEGs ($NDEG_T$). From these two proportions we realised a ratio test to compare the relative proportion of shared polymorphism (SPRT: Shared Polymorphism Ratio Test) in each categories of genes (i.e., DEGs and NDEGs):

$$SPRT = \frac{(DEG_{SP}/DEG_T)}{(NDEG_{SP}/NDEG_T)}$$

Confident intervals (CIs) of 95% were determined using1000 bootstrapping iterations per comparisons on the empirical dataset. Obtained ratios and associated CIs were compared to the expected ratio of one (i.e., no difference in the amount of shared polymorphism between DEGs and NDEGs) to test for enrichment of shared polymorphism in DEGs (i.e., SPRT>1).

Detection of adaptive variation

The detection of adaptive variation using F_{ST}-based approaches is challenging in study systems with complex population structures, which can be accounted for by multivariate outlier detection methods (Duforet-Frebourg et al. 2016; Luu et al. 2017). Redundancy Analysis (RDA) (Legendre and Legendre 1998) is an efficient constrained ordination method to detect (adaptive) variation under the effect of divergent selection, especially when the selection gradient is weakly correlated with population structure (Capblancq et al. 2018; Forester et al. 2018). In order to account for the hierarchical population structure, we used a conditioned (partial) Redundancy Analysis (cRDA), as implemented in the vegan v2.4-3 (Jari Oksanen et al. 2018) R package, to identify genes that diverge the most between limnetic and benthic species, independently of population structure. Consequently, the x-axis of the cRDA corresponds to the variance explained by the constrained 'Species' variable and the y-axis corresponds to the PC1 of the PCA analysis (nested into the RDA) and represents the unconstrained variance. Because RDA can be sensitive to missing data, we only kept loci that contained no missing genotypes across populations, representing a total of 9,093 SNPs. Briefly, a RDA allows evaluating the

variation that can be explained by the applied constraints. We conditioned the RDA analysis to remove the effects of continents, regions and lakes to control for the hierarchical genetic structure. We tested the significance level of the cRDA with an analysis of variance (ANOVA), performed with 1,000 permutations. From the conditioned ordination, each SNP was assigned a locus score that corresponds to the coordinates used to ordinate points and vectors. Then, we identified outlier SNPs by putting significance thresholds at +/- 2.6 and 3.0 standard deviations from the mean score of the constrained axis, corresponding to p-value of 0.01 and 0.001, respectively (Forester et al. 2018).

We also applied cRDA to gene expression data (i.e., on the 32,725 orthologous genes) and tested the significance of the constrained ordination model with an ANOVA using 1,000 permutations. We aimed at identifying co-varying DEGs between limnetic and benthic species after correction for hierarchical population structure and a significance threshold on expression scores (p-value of 0.01 and of 0.001) was applied. Moreover, this approach would limit the local effect of highly expressed genes in the identification of outliers/DEGs if they were not associated to the constrained independent variables. Such co-varying DEGs could reflect the effect of polygenic selection acting in parallel between benthic and limnetic whitefish.

Gene subnetworks analysis

In order to test for selection acting on sub-networks of genes involved in common biological pathways, we performed a gene network analysis designed specifically to detect polygenic selection (Gouy et al. 2017). The level of differential expression between limnetic and benthic whitefish captured by individual locus scores in the expression cRDA analysis was scaled to a z-score, such that individual locus scores have a mean of 0 and a standard deviation of 1. We obtained Kegg Ontology (KO) for each transcript with an Entrez gene ID annotation from the KASS (Kegg Automatic Annotation Server, http://www.genome.jp/tools/kaas/). Polygenic selection was tested using the R package signet (Gouy et al. 2017) on the Danio rerio and Homo sapiens KEGG databases (we present only the results obtained from the human database because of lack of power using the smaller D. rerio database). This package defines sub-networks of genes that interact with each other and present similar patterns attributed to selection, such as co-variation in expression level for genes involved in the same biological pathway. A null distribution of

sub-network scores was generated by random sampling to create 10,000 sub-networks of variable sizes. Each pathway of the KEGG database was parsed to identify gene sub-networks with a high score using 10,000 iterations of simulated annealing. Finally, the p-value of the sub-networks showing parallelism in gene expression was inferred based on the distribution of 10,000 permuted scores from the randomly generated sub-networks. We then tested for similarities among limnetic species for differential gene expression against benthic species across continents (sub-networks $P<0.05$).

eQTL analysis

We related differential gene expression with sequence divergence to identify eQTLs. We thus generated a new VCF file containing shared loci among all the populations that showed polymorphism across continents (i.e., trans-continental polymorphisms shared between *C. clupeaformis* and *C. lavaretus*), which corresponded to 2,240 SNPs. We extracted the 1,272 associated annotated genes and their expression level and tested for correlations between genotype and expression level (eQTL), using different models. First we applied a linear model testing for gene expression variation in response to allele frequency variations by controlling for the lake and continent covariates as environmental effects to correct for population genetic structure (expression~genotype+covariates). We compared the tested linear model to a theoretical simulated dataset, as suggested by the author of the package (Shabalin 2012). Second, we realized independent comparisons with one covariate at a time, and compare outputs from both approaches in order to exclude remaining cis-eQTL associated with environmental effect. This analysis was run with the R package MatrixEQTL v2.1.1 (Shabalin 2012). We identified significant ($FDR<0.05$) cis-eQTL by focusing on SNPs affecting the expression level of the gene to which they were physically linked, from the output of the linear model including all covariates.

4.5 Results

Reference transcriptome assembly

A total of 1.74 x 10^9 100bp single-end reads were generated from 72 individuals (Table S4.1). Filtered libraries (1.69 x 10^9 reads) were used to de novo assemble a composite reference transcriptome composed of 54,514 contigs (mean length = 1,121bp ; N50 = 1,672bp). This liver-specific assembly is consistent with the number of transcripts assembled using several organs separately in one *C. clupeaformis* male and one *C. lavaretus* female (range: 66,996 - 74,701, respectively (Pasquier et al. 2016)). Comparing transcriptomes separately assembled in *C. clupeaformis* and *C. lavaretus* to identify orthologous genes and filtering out paralogous genes (i.e., self-mapped transcripts hits, 15%), we ended up with a reference transcriptome of 32,725 annotated contigs (comparable to (Carruthers et al. 2018)) that were used for downstream analyses of gene expression and sequence divergence. The reference transcriptome N50 was 1,797bp with a contig size distribution ranging from 297bp to 13,427bp and a mean contig size of 1,185bp.

Genetic relationship among populations

The analysis of genetic relationships among the studied populations with TreeMix revealed the presence of shared polymorphisms maintained across the entire hierarchical genetic structure (Figure 4.1 and Figure S4.1). The different hierarchical levels were composed by limnetic-benthic species-pairs (hereafter, Species-pair level) in both North America and Europe. Similarities in branch length at the Species-pair level reflected the similar degrees of genetic differentiation among species-pairs from different regions consistent with a relatively similar timing of divergence and postglacial admixture among species-pairs from both continents. The second hierarchical level was composed by intra-continent regional divergence (hereafter, Region level), which was represented by the two European regions; Central alpine (Switzerland) and Fennoscandinavia (Norway). The highest hierarchical level of divergence was between the two sister-lineages *C. clupeaformis* and *C. lavaretus*. The sharing of ancestral polymorphism between geographically isolated taxa across continents was captured by the inferred migration link connecting two limnetic populations from Maine and Norway (Figure 4.1). This link indicates an excess of shared ancestral polymorphism after accounting for drift along the population

tree, which could indicate the presence of balanced polymorphisms across continents or past admixture events.

Trans-continental polymorphism quantification

Given the evidence for shared polymorphisms maintained among species from different continents, we documented the overall extent of trans-continental polymorphism. Trans-continental polymorphism corresponds to ancestral variation shared among all populations of limnetic and benthic species from North America (*C. clupeaformis*) and Europe (*C. lavaretus*). In other words, loci that are polymorphic on both continents. Among the 20,911 SNPs initially obtained after genotyping and filtering steps, we identified 2,241 SNPs (10.7%) distributed among 1,251 genes (3.8%) that met our criteria of trans-continental polymorphic loci. The genes containing trans-continental polymorphisms showed a significantly higher level of nucleotide diversity (π) within species (Wilcoxon signed-rank test, $P<0.001$, Figure S4.2), and a higher level of absolute divergence (Dxy, Nei 1987) between limnetic/benthic species (Wilcoxon signed-rank test, $P<0.001$, Figure S4.3) compared to genes with no trans-continental SNPs. These results indicate that evolutionary processes act differentially between these two categories of genes.

Parallel genetic differentiation

We used conditioned ordination to test whether divergence at the limnetic/benthic species-pair level involves parallel changes in allele frequency across sister-lineages from different continents (*C. clupeaformis-C. lavaretus*). We conditioned the ordination to account for the hierarchical genetic structure among populations (Lake, Region and Continent). The cRDA thus allowed the identification of variants associated with limnetic-benthic species divergence, explaining 2.8% of the total genetic variance (ANOVA, $F=1.259$, $P=0.001$), after controlling for the variance explained by regional and continental population structure (Figure 4.2A and Figure S4.4). The distribution of individual locus scores on the first cRDA axis discriminating all limnetic and benthic samples allowed identifying 348 outlier markers ($P<0.001$, 3.0 s.d.) showing parallel allele frequency differences between limnetic and benthic species across both continents. These 348 SNPs, which represent 15% of the 2,241 of the trans-continental polymorphic loci may be interpreted as representing shared genetic bases of limnetic-benthic species divergence across continents. Gene ontology analysis of transcripts associated with parallel outlier

SNPs revealed significant enrichment (P<0.001) in metabolic process (i.e., catabolism), immune system process and developmental process, for example (TableS1).

Differential gene expression between Limnetic and Benthic Species

Using the benthic whitefish populations as the reference level, we quantified differentially expressed genes (DEG) between limnetic and benthic whitefish at the Lake and Region levels using multivariate generalized linear model (glm). Our expectation was a decreasing number of shared DEGs in higher comparisons levels, due to a reduced fraction of shared regulatory variants. North American lakes, Cliff and Indian, showed 3,175 (9.7%) and 238 (0.7%) of significantly differentially expressed genes (False discovery rate; FDR<0.05) between sympatric limnetic and benthic species, respectively (Figure 4.3). In both Cliff Lake and Indian Lake, approximately twice as many genes were up-regulated in the limnetic species compared to the benthic species (Cliff Lake: 2,001 vs. 1,174, χ^2 test, P<0.001; Indian Lake: 159 vs. 79, χ2 test, P<0.001) (Figure S4.5). The lower level of DEGs identified in Indian Lake was associated to a higher inter-individuals variance (Figure S4.5). In addition, we observed that for a higher significance threshold (q-value<0.1) 1,926 DEGs were identified between species in Indian Lake, and 318 of those were shared with DEGs from Cliff Lake. In Langfjordvatn and Skukkebukta lakes from Norway, 276 and 112 significant DEGs were identified between limnetic and benthic whitefish, respectively. Limnetic whitefish showed a higher magnitude gene expression repression relative to benthic whitefish (Figure S4.6), as opposite to North America, more genes were down-regulated in limnetic populations (45 vs. 231, χ^2 test, P<0.001 in Langfjordvatn; 44 vs. 68, χ^2 test, P=0.023, in Skrukkebukta). In Swiss lakes, 3,727 and 1,392 genes showed a significantly different expression level between limnetic and benthic populations in Lake Lucerne and Lake Zurich, respectively. In contrast with North America and Norway however, a similar number of genes were down- and up-regulated in the limnetic species compared to benthic species in both lakes (1,870 vs. 1,857, χ^2 test, P=0.831 in Lake Lucerne and 691 vs. 701, χ^2 test, P=0.787, in Lake Zurich).

To document the extant of parallelism at the Region level, we considered the proportion of DEGs between limnetic and benthic whitefish while integrating the lakes as covariates in a multivariate glm. Thus, the degree of parallelism in DEG was the largest in Swiss lakes whereby 1,439 DEGs between limnetic and whitefish species among lakes

from Switzerland. These parallel genes showed a higher proportion of down-regulated genes in limnetic species (625 vs. 814, χ^2test, P<0.001) (Figure S4.7). Similarly in Maine, 126 DEGs between limnetic and benthic species were found among Indian and Cliff lake. These genes showed a higher gene expression in the limnetic species than in the benthic species (818 vs. 541, χ^2test, P<0.001) (Figure S4.7). Finally, in Norway, 126 DEGs were identified between limnetic and benthic whitefish among lakes from Norway. Here, however, parallel DEGs comprised a significantly higher proportion of down-regulated genes in the limnetic species (53 vs. 239, χ^2 test, P<0.001) (Figure S4.8).

Gene ontology (GO) enrichment analysis at the Lake and Region levels provided evidence for parallelism in biological functions. Indeed, DEGs were significantly enriched (FDR<0.05, see Table S4.2) in limnetic species for immune system response, detoxification and antioxidant activity in both North America and Europe. Moreover, we found enrichment in genes associated to growth and development at Lake (Indian, Lucerne, Skrukkebukta and Langfjordvatn lakes) and Region levels (Maine) in benthic whitefish. DEGs were also enriched in limnetic species for metabolic processes, electron carrier activity and catabolic processes, which are associated with differences in the metabolic rate between limnetic and benthic species (Laporte, Dalziel, Martin, and Bernatchez 2016a; Dalziel, Laporte, Guderley, et al. 2017; Dalziel, Laporte, Rougeux, et al. 2017).

At the Continent level (integrating Region and Continent as covariates), we found 156 parallel DEGs between limnetic and benthic species between both continents. Again, these 156 genes showed similar proportions of up- and down-regulated genes in all limnetic whitefish compared to all benthic whitefish (72 vs. 84, χ^2 test, P<0.299; Figure S4.8). From enriched biological functions, some were were associated to trade-off associated to the limnetic niche colonization; metabolic process (P=0.016) and antigen binding (P=0.021), which are associated with immune response (e.g., Immunoglobulin domain) and cellular metabolic process (P=0.002) (e.g., SPRY-associated domain). DEGs were enriched for oxido-reductase activity genes were involved in metabolic activity (e.g., TSTA3, a gene able to activate fructose and mannose metabolism via an oxydo-reductase step, involved in Glycolysis; Hsp90, a gene responding to environmental stress with effects on growth).

Enrichment in trans-continental polymorphism in DEGs

The excess of ancestral polymorphism shared among limnetic and benthic species across both continents suggests the existence of a mechanism responsible for the maintenance of balanced ancestral variation against the stochastic effect of drift in each lineage. In order to further test if the retention of ancestral polymorphism could be linked to differential selection on adaptive traits between limnetic and benthic species, we tested if cis-regulating regions (i.e., regions physically linked to transcripts) of DEGs show an increased probability of having shared polymorphisms. The shared polymorphism ratio test (SPRT), which compares the proportions of shared polymorphism in DEGs to non-differentially expressed genes (NDEGs), revealed an enrichment of shared polymorphism in DEGs at the three hierarchical levels (Figure 4.4).

Identification of gene sub-networks showing patterns of parallel gene expression

A conditioned RDA (cRDA) performed on the expression data of the 32,725 orthologous contigs revealed that 2.5% of total variance in expression was explained by net differences between limnetic and benthic species across continents (ANOVA, $F=2.516$, $P=0.006$, Figure 4.2B), while accounting for the hierarchical population structure. The z-scored distribution of the gene expression on the first RDA axis constrained for divergence between limnetic and benthic whitefish ranged from [-4.14; 3.99] (Figure S4.9). Applying two different significance thresholds ($P<0.01$ and $P<0.001$) allowed identifying 272 ($P<0.01$) and 66 ($P<0.001$) putative outliers DEGs. These were significantly enriched for biological regulation ($P<0.001$) and metabolic process ($P<0.001$) in both sets of genes, and growth ($P=0.026$) for the subset of 272 genes (Table S4.2). 22 out of the 156 DEGs (14%; more than expected by chance, hypergeometric test, $P<0.001$) identified with the glm analysis overlapped with the 272 DEGs from the cRDA analysis on gene expression, including the previously mentioned genes TSTA3 and Hsp90.

We then addressed the polygenic basis of transcriptomic differences using DEGs found in the ordination analysis. The z-scored transformation of cRDA's genes expression scores was used as a quantitative measure for assessing the extent of parallel expression between limnetic and benthic whitefish across continents. A total of 22,188 out of the 32,725 genes that were successfully annotated with an Entrez gene ID were analysed with signet based on information from KEGG databases. In signet, gene sub-networks were

identified for each pathway and we considered the significance of sub-networks (P<0.05) in the analysis against the Homo sapiens KEGG database.

Ten metabolic pathways with significant sub-networks of genes were identified (Table S4.3). Five of these pathways shared genes and were therefore merged together to identify genes showing convergent patterns among the significant sub-networks (Figure 4.5), mainly for peripheral genes (Figure S4.10). From the ten identified metabolic pathways composed of 73 parallel DEGs, three categories of metabolic functions were represented. The first category corresponded to energetic metabolism (e.g., pentose phosphate pathway, glycerolipid metabolism, nicotinate and nicotinamide metabolism, FoxO signalling pathways) which is involved in regulation of glycolysis and energy production (from ATP to NADH). The second category was the detoxification metabolism and immune system (e.g., CytP450 and glutathione metabolism), which is mainly associated to detoxification and oxidative stress, maintaining cell integrity by preventing damage due to reactive oxygen species (ROS). The third category was the cell cycle metabolism and control (e.g., FoxO signaling pathways, purine and pyrimidine metabolism, cAMP signalling pathway, PI3K-Akt signaling pathway). These pathways play critical roles in regulating diverse cellular functions including metabolism, growth, proliferation, survival, transcription and protein synthesis.

cis-eQTL markers associated with expression

We finally tested whether the level of expression was associated with SNPs located within a given gene. We found 451 SNPs associated with the level of expression of the gene to which they belong (cis-eQTL) with a significance threshold of P=0.05. That is, variable level of expression was associated with the three possible genotypes at a given SNP. Controlling for multiple tests, we retained 134 significant (FDR<0.05) cis-eQTL across continents associated to limnetic and benthic species variation.

We identified SNPs and genes showing overlap between the different analyses. Indeed, two significant cis-eQTL overlapped the 20 outliers SNPs from the cRDA (hypergeometric test, P<0.001). They were physically linked to the complement factor H (CFH, Entrez 3075; Figure 4.6B) and Protein Kinase AMP-Activated Non-Catalytic Subunit Beta 1 (Prkab1, Entrez 19079; Figure 4.6C). These genes are respectively involved in immune response and in regulation of the cellular energy metabolism. Moreover, among the

134 significant cis-eQTL, 19 genes were shared with genes identified in the polygenic subnetwork analysis (hypergeometric test, P=0.006). However, only the pyruvate kinase gene (PKM, isoform M2; Entrez 5315) remained significant in both (sub-networks and eQTL) analyses (Figure 4.4). This gene encodes a protein involved in glycolysis, which generates ATP and pyruvate. The level of expression of this gene was higher in heterozygous and homozygous individuals for the minor allele (Figure 4.6A and Figure S4.11, linear model, P=0.001). Finally, we inferred the gene structure (i.e., identification of 5' and 3' UTR, Exonic and Intronic regions) of our de novo assembled transcriptome and more particularly the PKM gene. We localised the variant affecting the level of expression of the PKM gene in its 3'UTR region, which could impact the regulation of the transcription of this gene.

4.6 Discussion

C. clupeaformis (North America) and *C. lavaretus* (Eurasia) sister lineages became geographically isolated and diverged independently since at least 500,000 years ago (Bernatchez and Dodson 1991; Bernatchez and Dodson 1994; Jacobsen et al. 2012). Yet, they maintained similar habitat preferences in cold freshwater lakes (Bernatchez and Dodson 1991; Douglas et al. 1999; Østbye et al. 2005; Østbye et al. 2006), with a frequent occurrence of sympatric species pairs being respectively associated to benthic and limnetic ecological and trophic niches (Lu and Bernatchez 1999; Amundsen, Bøhn, and Våga 2004b; Kahilainen and Østbye 2006; Landry et al. 2007; Häkli et al. 2018). Both *C. clupeaformis* and *C. lavaretus* were exposed to climatic oscillations during the late Pleistocene, initiating diversification trough allopatric divergence followed by secondary contacts at the regional scale (Rougeux et al. 2017), and lacustrine sympatric ecological specialisation to limnetic and benthic habitats in each region. Here, the analysis of gene sequence divergence and differential expression in limnetic-benthic species has the potential to provide new insights into the genomic basis of parallel adaptation and parallel ecological speciation.

A salient result from our analysis is that pairs of incipient species from divergence events that occurred on both continents exhibit significant parallelism in differential gene expression associated with repeated divergent adaptation to limnetic and benthic ecological niches. The identification of significant DEGs involved in energetic metabolism, immune response, cell cycle and growth is congruent with previous transcriptomic analysis conducted in *C. clupeaformis* on the same organ tissue, highlighting life history trade-offs between growth and energetic costs associated with limnetic niche colonization, albeit focusing on reduced transcriptome representation (low resolution microarrays developed for another species) and candidate genes (St-Cyr et al. 2008; Jeukens et al. 2009). Our results thus confirm that parallel patterns of transcriptional responses at the gene, gene network and biological function levels accompany parallel phenotypic divergence among independently evolved species-pairs. Our results also show that seeking to detect parallel DEGs based on a single gene approach may lack the power to detect polygenic changes in expression levels, as it could be expected for the complex phenotypic traits involved in the divergence of these species-pairs. Indeed, a gene-by-gene approach may be too

conservative and not well adapted to capture subtle parallel DEGs at genes involved in the same or closely related biological pathways under selection. Our second statistical approach based on transcript abundance co-variation allowed us to integrate this level of information. As predicted by theory (Slatkin 1978; Yeaman 2015), we found a greater number of DEGs associated with species divergence after correcting for hierarchical population structure than with a generalized linear model approach. Consistent with results from the negative binomial generalized linear model, the redundancy analysis allowed identifying a parallel genetic basis of phenotypic and ecological divergence by revealing parallel DEGs between limnetic and benthic species. Moreover, we found that these DEGs are involved in several metabolic pathways belonging to energetic, growth, cell cycle metabolisms and transcription factor, regulating genes associated with energetic metabolism, as observed in (Jacobs et al. 2018). This approach also allowed detecting congruent expression signals at the integrated pathway scale, where the same effect on the selected phenotype can be achieved via regulation of different genes, because of the complexity and redundancy of the multi-genic regulatory systems (Yeaman 2015). The accumulated results coupled with previous analyses on this system thus highlight the repeated action of natural selection on similar polygenic bases of phenotypic traits (St-Cyr et al. 2008; Jeukens et al. 2009; Jeukens and Bernatchez 2011; Gagnaire, Normandeau, et al. 2013; Gagnaire, Pavey, et al. 2013), mainly on genes pertaining to network periphery in the identified module.

In addition to trans-continental parallelism in interspecific divergence of transcript abundance during ecological speciation, we found parallel differentiation between limnetic and benthic species between continents also at the gene sequence level. Parallel outliers represented 15% of shared outliers between replicate species-pairs, consistent with what has been found in other systems (ranging from 6% to 28%; (Deagle et al. 2012; Westram et al. 2014; Ravinet et al. 2015; Le Moan et al. 2016; Meier et al. 2018)). We also identified a substantial number of genes exhibiting shared (*i.e.*, trans-continental) polymorphism across the complex of *Coregonus* lineages and radiations. As observed in the tree species *Populus tremula*, genes in the network's periphery (*i.e.*, with lower level of connectivity) are more likely to be enriched in polymorphism due to less evolutionary constraints (Mähler et al. 2017). For these genes, patterns of genetic diversity and DEGs enrichment between limnetic and benthic species suggest the action of divergent selection in the presence of

gene flow (Charlesworth et al. 1997) globally maintaining alleles associated with different expression levels between sympatric species. The fact that absolute genetic divergence between sympatric limnetic and benthic species in genes with trans-continental polymorphism was elevated compared to other genes may reflect the active maintenance of these alleles over the long term, since the lineages divergence, by the interaction of divergent selection and admixture between sympatric species-pairs (Han et al. 2017; Ma et al. 2017). The action of gene flow between limnetic and benthic whitefish may indeed further contribute to maintain the alleles favoured in each species in a balanced state within each lake, thus protecting polymorphism from being lost at a global scale even across continents. This could be eased by the apparently highly polygenic nature of the traits under divergent selection, meaning that the intensity of selection acting on each underlying locus could be weak (Le Corre and Kremer 2012). Therefore, even a modest amount of gene flow could possibly maintain a balanced polymorphism within each species-pair. This however, remains to be investigated more formally, since strong divergent selection without migration within lakes would also help the maintenance of polymorphism across continents. Moreover, such genomic patterns could be directly caused by the sorting of sieved ancestral balanced polymorphism, which genomic signatures would mimic patterns built by other evolutionary process (*e.g.*, gene flow, as discussed above) (Guerrero and Hahn 2017).

The identification of orthologous genes with trans-continental polymorphism associated with differential expression between benthic and limnetic species supports the existence of *cis*-acting SNPs on transcripts abundance. Moreover, the characterisation of DEGs enriched in shared polymorphism across continents suggests the long-term action of some form of balancing selection, maintaining ancestral polymorphisms that predate the onset of regional and continental divergence of the different limnetic-benthic species-pairs. Consistent with theory and empirical studies (Zheng et al. 2011), our analysis of orthologous genes supports a role of polymorphism originating from standing genetic variation both in protein coding sequences (CDS) and regulatory motifs (*e.g.*, untranslated regions UTRs) in the process of adaptive divergence between limnetic-benthic whitefish sister species (Zheng et al. 2011). For instance, we found a parallel *cis*-eQTL in the 3'UTR of the pyruvate kinase gene (*PKM*), affecting the relative expression level of this gene between species. The *PKM* isoform M2 corresponds to a glycolytic enzyme (iso-enzyme)

that is expressed in liver tissue. Given the importance of 3'UTRs in regulating the transcription process and transcripts abundance (Merritt et al. 2008; Wittkopp and Kalay 2011), this 3'UTR SNP could be under divergent selection between limnetic and benthic species and therefore protected from being lost by drift within populations over the long term within any given limnetic-benthic pair as hypothesized above. Consequently, it is likely that such a *cis*-eQTL could have been recruited from standing genetic variation by natural selection, increasing in frequency in limnetic whitefish on both continents, while modifying the level of expression of a central gene in energetic metabolism.

The inferred module of gene co-expression form liver tissue allowed quantifying a partial view of the gene co-expression network associated with species phenotypic differentiation. Interactions between nearby genes within the module could results in *cis*-regulation affecting the level of expression of other genes and ultimately, affect the activity of genes farther in the genome (Boyle et al. 2017). It is noteworthy that genes interactions between metabolic pathways are conserved among sister-lineages. Indeed, it has been showed that co-expression modules are maintained through evolutionary times despite variation in the set of regulatory genes that activate them (Tanay et al. 2005). Moreover, those patterns of gene expression changes, maintained across the system, could be associated to the variational adaptation, by which co-evolving traits are integrated into the same module (Wagner et al. 2007). Despite the partial portrait of the polygenic basis of phenotypic differentiation between species, we stress that no gene of main effect (*i.e.*, hub gene) was identified in the module. Such patterns would suggest that modularity in gene expression and genes interaction into a module can recruit less constrained genes without affecting highly constrained central genes (Wagner 1996), while peripheral genes could be directly and indirectly involved in the co-expression network via a 'hub-gene' effect in the initial metabolic pathways of the recruited gene. However, further investigations on quantifying the gene-gene interactions or protein-protein interactions (PPIs), in order to infer individual gene constraint levels or position in the module, should be realized in a more formal framework on several tissues. Thus, this could allow identifying a most complete picture (qualitatively and quantitatively) of the gene co-expression network associated with phenotypic differentiation between limnetic and benthic species. Finally, it could be interesting to measure the independent relative effect size of genes from the co-expressed module, despite the complex phenotypic traits associated with ecological niche adaptation.

In conclusion, our study provides a quantitative assessment of DEGs and gene sequence divergence based on an extensive transcriptomic dataset, enabling to infer the effects of polygenic divergent selection acting on complex traits that diverge between sympatric benthic and limnetic species, within both the *C. clupeaformis* and *C. lavaretus* species radiations (Gagnaire, Pavey, et al. 2013; Laporte, Rogers, Dion-Côté, Normandeau, Gagnaire, Dalziel, Chebib, and Bernatchez 2015b). Our results confirm previous studies of differential gene expression between sympatric limnetic and benthic species (St-Cyr et al. 2008; Jeukens et al. 2009; Jeukens and Bernatchez 2011). Moreover, they extend previous findings by revealing patterns of parallelism between species on two continents, derived from two evolutionary lineages that diverged at least half a million years ago on separate continents. Furthermore, they show the effects of polygenic selection on genes associated with fundamental and constrained metabolic pathways, such as functions associated with energetic metabolism (Dalziel, Laporte, Guderley, et al. 2017). Due to the additive effects of multiple genes in controlling the expression of polygenic phenotypic traits, the probability of identifying a shared genetic basis from standing genetic variation (likely to increase with the number of genes involved) is higher compared to the alternative *de novo* mutation to generate local polymorphism. This suggests an important contribution of ancestral polymorphism in the repeated evolution of sympatric species pairs. This was illustrated by the identification of a genetic variant in the UTR gene region associated with phenotypic differences between species, as previously reported in other taxa (Schluter et al. 2004; Wittkopp et al. 2004; Jones et al. 2012; Uebbing et al. 2016; Verta et al. 2016; Wang et al. 2017). The resolution of future studies could be enhanced using a comparative whole-genome resequencing approach to provide a more detailed understanding of the genomic architecture of phenotypic differences between species, and the role of old standing variants in ecological speciation. Moreover, given the apparently important role of divergent gene expression in the adaptive radiation of the whitefish species complex, we are currently investigating variation in regulatory mechanisms (*e.g.*, epigenetic variation) involved in controlling levels of gene expression.

4.8 Acknowledgements

Jérémy Le Luyer, Ben J. G. Sutherland provided valuable advices and discussions on transcriptomic analysis, Martin Laporte for inspiring and stimulating discussions about the lake whitefish system, thanks guys. We thank Alexandre Gouy for support with *signet* in the early times of the package, as well as Kyle W. Wellband for commenting on an earlier version of the manuscript. We thank Shripathi Bhat and the Freshwater ecology group at UiT for participating with the sampling in Norway. Finally, we are grateful to Associate Editors as well as anonymous referees for their very constructive and valuable inputs on an earlier version of the manuscript.

4.9 Figures

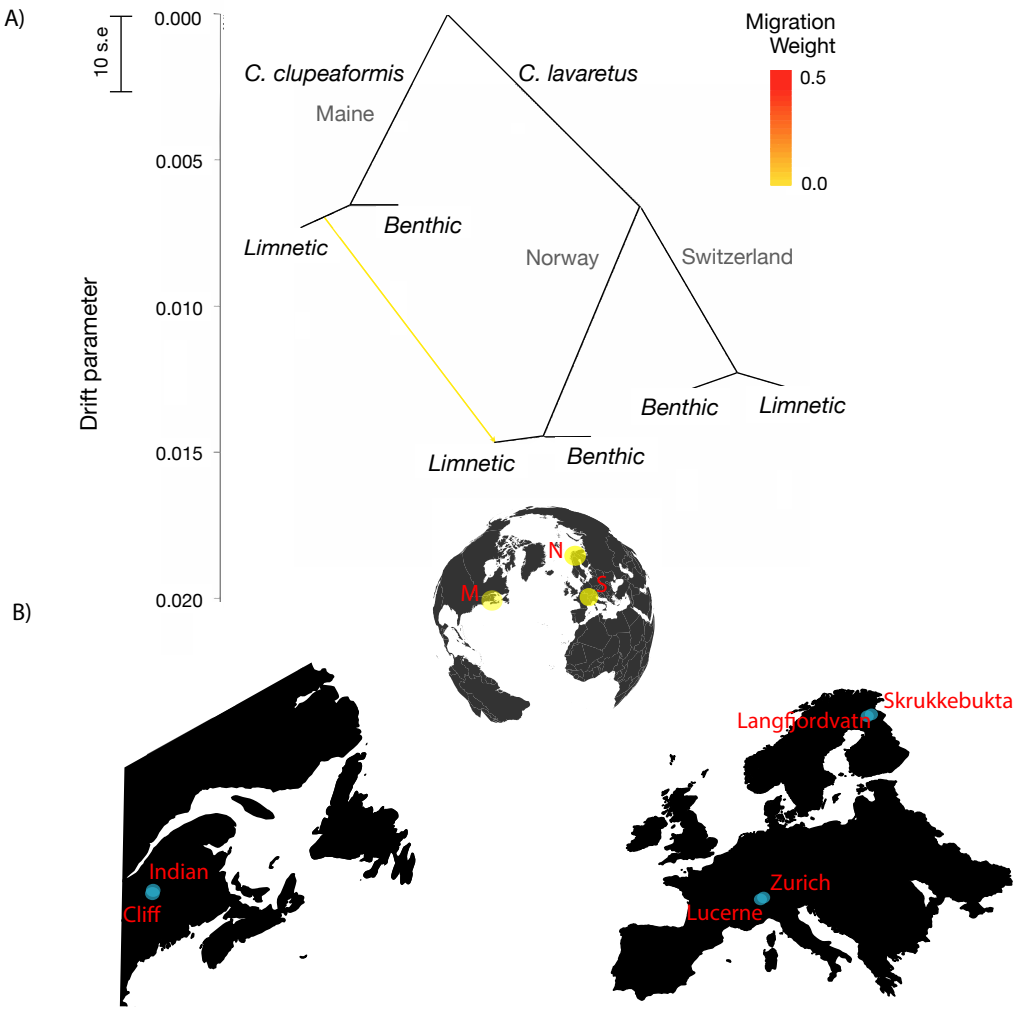


Figure 4.1: Details about the whitefish study system. A) Treemix analysis illustrating independent differentiation between sympatric Benthic and Limnetic species, from the closely related sister lineages *C. clupeaformis* in North America and *C. lavaretus* in Europe. Vertical branch lengths are proportional to the amount of genetic drift in each branch, the scale bar indicates 10 times the average standard error (s.e) of the entries in the covariance matrix between pairs of populations and the color scale indicates the weight of inferred migration events or shared ancestral polymorphism in absence of possible gene flow across continents. B) Locations of the three sampled regions (yellow circles) M: Maine for *C. clupeaformis*, N: Norway and S: Switzerland for *C. lavaretus*. Two lakes (blue circles) containing sympatric species-pairs were sampled per region.

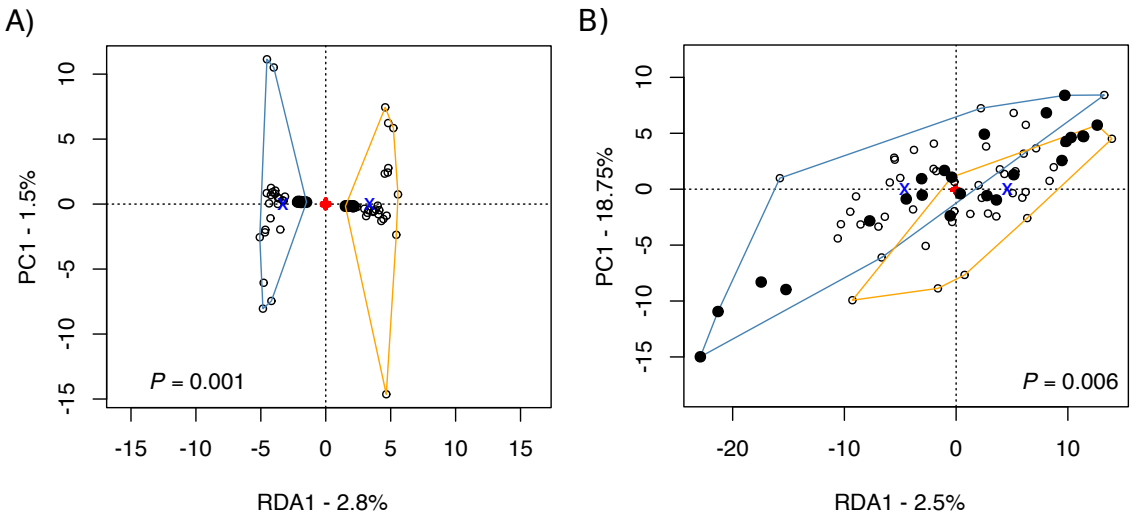


Figure 4.2: Conditioned redundancy analysis (cRDA) clustering individuals per species (all limnetic vs. all benthic from both continents). The cRDA analysis was used to capture divergence between limnetic and benthic species (along RDA axis 1) while correcting for three hierarchical levels of population structure (lake, region and continent) based on A) genotypes and B) gene expression levels. Orange and blue clusters correspond to benthic and limnetic species, respectively. Circles represent individuals. Open and filled circles represent the *C. lavaretus* and *C. clupeaformis* lineages, respectively.

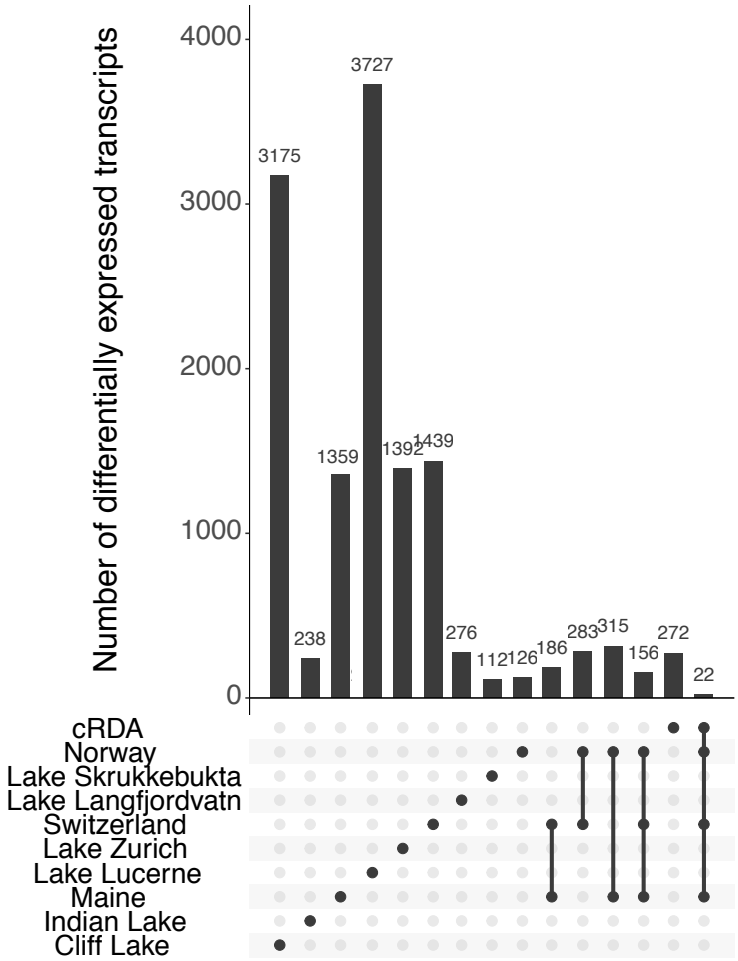


Figure 4.3: Frequency of shared differentially expressed transcripts between species across hierarchical levels. Limnetic/Benthic comparisons are indicated by the dots (intra-lake and intra-region) or linked dots (inter-regions and inter-continents). The number of significant differentially expressed genes (DEGs) associated to species divergence (FDR<0.05) per comparison is indicated on top of each bar, as determined from the multivariates tests, from the cRDA analysis for all limnetic to all benthic comparison., as well as the overlap of DEGs between cRDA and DESeq2.

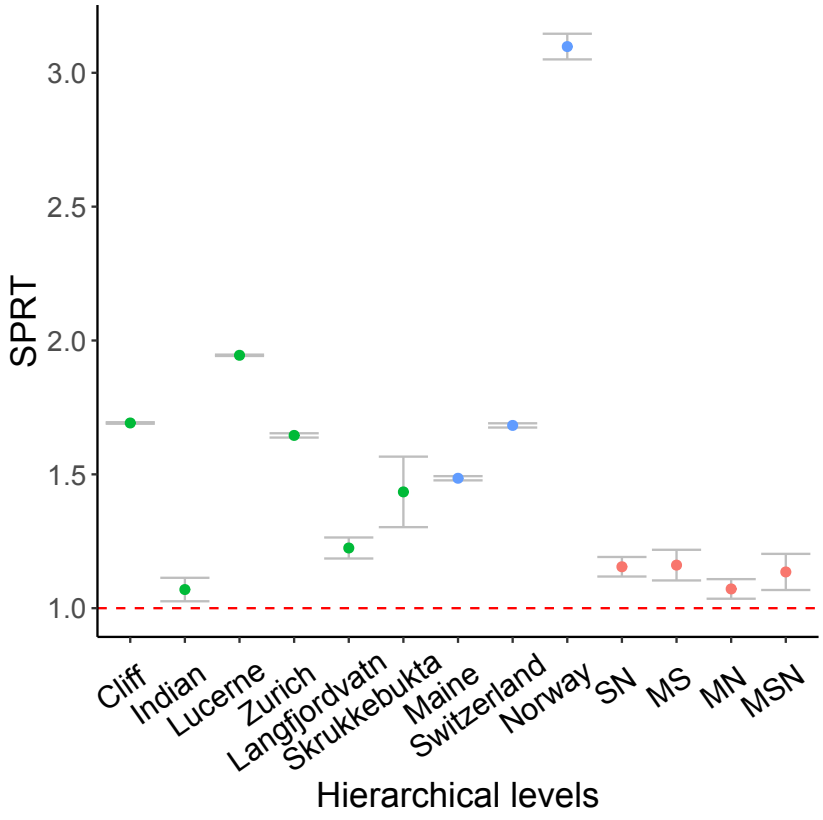


Figure 4.4: Shared polymorphism enrichment among DEGs compared to Non-DEGs. Test ratio (SPRT) of the proportion of shared trans-continental polymorphism in DEGs compared to non-DEGs, for intra-lake (green dots), inter-lake within regions (blue dots) and inter-regions (SN: Switzerland/Norway, MS: Maine/Switzerland, MN: Maine/Norway and MSN: Maine/Switzerland/Norway; red dots) comparisons. All ratios are above the red-dashed line (y=1), which illustrates the general enrichment of shared polymorphism in DEGs compared to non-DEGs. Grey bars indicate the 95% confidence interval associated with the observed value obtained using 1000 bootstrap resampling of the dataset.

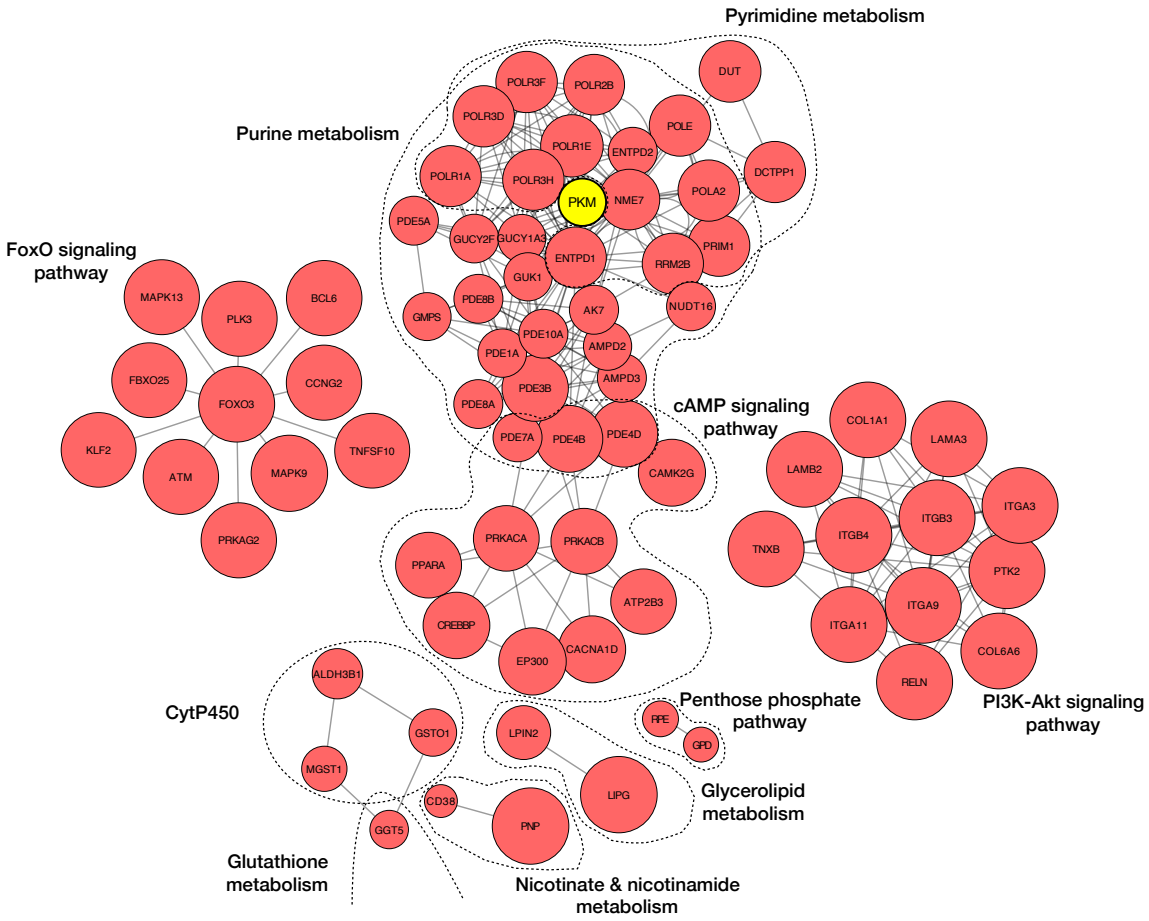


Figure 4.5: Merged significant subnetworks for Limnetic/Benthic species comparisons. Detailed subset of significant differentially expressed genes (DEGs) between species displayed among corresponding metabolic pathways. Pyruvate kinase (PKM) gene expression (in yellow) is associated with a *cis*-eQTL in its 3'UTR. Gene annotation is based on the Ensembl nomenclature. For each node represented by a gene, the relative size is proportional to the contribution score of the associated gene to the significance of the metabolic pathway. The score for each gene corresponds to a probability of convergent adaptation between individuals of the same species.

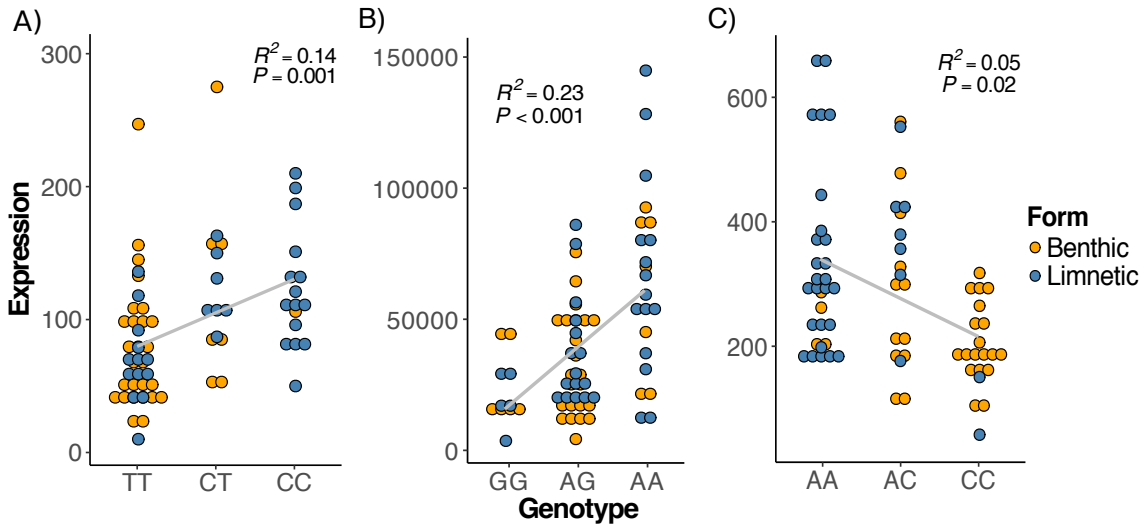


Figure 4.6: Associations between significant *cis*-eQTL genotypes and the level of expression of three genes in limnetic and benthic whitefish species, independently of their geographic origin. Three examples of transcripts abundance per individual (circles) varying with genotypes for a *cis*-eQTL located in 3'UTR of A) pyruvate kinase (PKM), B) complement factor H (CFH) and C) Protein Kinase AMP-Activated Non-Catalytic Subunit Beta 1 (Prkab1) genes. The grey line corresponds to the linear model fitted to the data and associated statistics (coefficient of determination: R^2 and *p*-value: *P*) detailed in each panel. Individuals from benthic and limnetic sister-species are represented in orange and blue, respectively.

4.10 Supplementary tables

Table S4.1: Number of raw and filtered reads obtained per individual.

Number of raw reads	Number of filtered reads	Library ID	Individual ID	Species
20582507	19753880	LIB1	SK08	Benthic
22833090	21850810	LIB2	LF112	Limnetic
19912485	19111721	LIB3	SK09	Benthic
27740869	26602363	LIB4	Zu15	Benthic
17582083	16896683	LIB5	ID2	Limnetic
27467349	26358596	LIB6	CN6	Benthic
29081812	27955687	LIB7	L52	Limnetic
27575796	26458514	LIB8	L58	Limnetic
23115781	22169931	LIB9	LF36	Benthic
20619799	19688042	LIB10	Z30	Limnetic
20169272	19295658	LIB11	CN11	Benthic
24858616	23818507	LIB12	LF38	Benthic
27739704	26628586	LIB13	ID7	Limnetic
24427813	23359144	LIB14	L85	Limnetic
22828163	21912572	LIB15	SK10	Benthic
25277259	24265292	LIB16	CN10	Benthic
21296255	20525864	LIB17	Z23	Limnetic
30904936	29818408	LIB18	LF109	Limnetic
23034903	22170784	LIB19	ID3	Limnetic
25402654	24518086	LIB20	L56	Limnetic
17533268	16715432	LIB21	IN8	Benthic
21385377	20566388	LIB22	Z27	Limnetic
20070114	19294910	LIB23	L46	Limnetic
25153932	24251100	LIB24	SK14	Limnetic
21627171	20817338	LIB25	IN5	Benthic
25376555	24304127	LIB26	SK39	Limnetic
23221581	22394148	LIB27	LF41	Benthic

28935800	27811520	LIB28	Z12	Benthic
18838742	18186632	LIB29	Z19	Benthic
4822124	25833236	LIB30	ID1	Limnetic
24259963	23292157	LIB31	IN9	Benthic
21830728	21021795	LIB32	SK11	Benthic
34202220	32821823	LIB33	LF113	Limnetic
14858370	14212670	LIB34	Z13	Benthic
19664051	18865826	LIB35	L53	Limnetic
27947100	26854678	LIB36	SK03	Limnetic
26388655	25351805	LIB37	CN12	Benthic
21151309	20110167	LIB38	IN6	Benthic
25185594	24150716	LIB39	ID4	Limnetic
31023995	29809348	LIB40	Z29	Limnetic
19537598	18931029	LIB41	Z11	Benthic
28025442	27205514	LIB42	Lu68	Limnetic
20270143	19650600	LIB43	CD21	Limnetic
15707040	15224135	LIB44	IN7	Benthic
25145896	24375925	LIB45	SK02	Limnetic
17235327	16676866	LIB46	Lu86	Benthic
25391937	24577489	LIB47	CD20	Limnetic
21936910	21266681	LIB48	LF37	Benthic
22155698	21220544	LIB49	Z20	Benthic
19642561	18838271	LIB50	SK71	Limnetic
22853137	21907482	LIB51	ID9	Limnetic
28958463	27755684	LIB52	CD19	Limnetic
27820603	26670427	LIB53	CD22	Limnetic
24449486	23452258	LIB54	LF35	Benthic
30066776	28875128	LIB55	L49	Limnetic
28375763	27242770	LIB56	LF39	Benthic
24999646	23739395	LIB57	SK19	Benthic
27140147	25803019	LIB58	CD17	Limnetic
22364555	21266933	LIB59	L83	Limnetic

19599732	18676563	LIB60	CN5	Benthic
30595792	29159603	LIB61	LF108	Limnetic
25208038	24004805	LIB62	SK48	Benthic
42568708	40665197	LIB63	LF111	Limnetic
20743920	19797211	LIB64	Z31	Limnetic
29186773	27884328	LIB65	L82	Limnetic
36016311	34438149	LIB66	LF107	Limnetic
31202915	29850242	LIB67	L81	Limnetic
32713977	31286322	LIB68	CN7	Benthic
9701395	9001555	LIB69	Z37	Limnetic
33866686	32293410	LIB70	CD18	Limnetic
20858977	19875693	LIB71	IN10	Benthic
24041196	22983982	LIB72	SK36	Limnetic

Table S4.2: Gene ontology analysis results for significant enrichment. Significant GO term (*P*<0.05) from different analyses. From expression data, with identification of DEGs based on DEseq2 (FDR<0.05) and cRDA by applying different thresholds (P<0.01 and P<0.001, respectively), and from genotype data from cRDA analysis with a significance threshold of P<0.01.

Method (significant threshold)	Hierachichal level	GO_number	Biological Function	FDR or *p*-value
DESeq2 (FDR<0.05)	Cliff	GO:0009058	biosynthetic process	7.54e-10
DESeq2 (FDR<0.05)	Cliff	GO:0006807	nitrogen compound metabolic process	1.88e-05
DESeq2 (FDR<0.05)	Cliff	GO:0008152	metabolic process	0.000162
DESeq2 (FDR<0.05)	Cliff	GO:0016043	cellular component organization	0.00019
DESeq2 (FDR<0.05)	Cliff	GO:0007155	cell adhesion	0.000253
DESeq2 (FDR<0.05)	Cliff	GO:0071840	cellular component organization or biogenesis	0.000346
DESeq2 (FDR<0.05)	Cliff	GO:0022610	biological adhesion	0.000459
DESeq2 (FDR<0.05)	Cliff	GO:0050789	regulation of biological process	0.00139
DESeq2 (FDR<0.05)	Cliff	GO:0071704	organic substance metabolic process	0.00172
DESeq2 (FDR<0.05)	Cliff	GO:0044237	cellular metabolic process	0.00188
DESeq2 (FDR<0.05)	Cliff	GO:0098754	detoxification	0.00192
DESeq2 (FDR<0.05)	Cliff	GO:0065007	biological regulation	0.00252
DESeq2 (FDR<0.05)	Cliff	GO:0007165	signal transduction	0.00377
DESeq2 (FDR<0.05)	Cliff	GO:0098727	maintenance of cell number	0.0039
DESeq2 (FDR<0.05)	Cliff	GO:0042330	taxis	0.00545
DESeq2 (FDR<0.05)	Cliff	GO:1990748	cellular detoxification	0.00567
DESeq2 (FDR<0.05)	Cliff	GO:0044703	multi-organism reproductive process	0.0083
DESeq2 (FDR<0.05)	Cliff	GO:0072376	protein activation cascade	0.00965
DESeq2 (FDR<0.05)	Cliff	GO:0044710	single-organism metabolic process	0.014
DESeq2 (FDR<0.05)	Cliff	GO:0061687	detoxification of inorganic compound	0.0146
DESeq2 (FDR<0.05)	Cliff	GO:0044706	multi-multicellular organism process	0.0149
DESeq2 (FDR<0.05)	Cliff	GO:0051716	cellular response to stimulus	0.0169
DESeq2 (FDR<0.05)	Cliff	GO:0044419	interspecies interaction between organisms	0.0172
DESeq2 (FDR<0.05)	Cliff	GO:0022402	cell cycle process	0.0204
DESeq2 (FDR<0.05)	Cliff	GO:0045321	leukocyte activation	0.0226
DESeq2 (FDR<0.05)	Cliff	GO:0044238	primary metabolic process	0.0287
DESeq2 (FDR<0.05)	Cliff	GO:0044764	multi-organism cellular process	0.0493

DESeq2 (FDR<0.05)	Cliff	GO:1990904	ribonucleoprotein complex	2.9e-10
DESeq2 (FDR<0.05)	Cliff	GO:0044279	other organism membrane	3.17e-05
DESeq2 (FDR<0.05)	Cliff	GO:0005623	cell	0.00025
DESeq2 (FDR<0.05)	Cliff	GO:0033643	host cell part	0.000906
DESeq2 (FDR<0.05)	Cliff	GO:0043234	protein complex	0.00136
DESeq2 (FDR<0.05)	Cliff	GO:0044423	virion part	0.0032
DESeq2 (FDR<0.05)	Cliff	GO:1902494	catalytic complex	0.00376
DESeq2 (FDR<0.05)	Cliff	GO:0019013	viral nucleocapsid	0.00401
DESeq2 (FDR<0.05)	Cliff	GO:0044464	cell part	0.00852
DESeq2 (FDR<0.05)	Cliff	GO:0034358	plasma lipoprotein particle	0.00918
DESeq2 (FDR<0.05)	Cliff	GO:0032994	protein-lipid complex	0.0101
DESeq2 (FDR<0.05)	Cliff	GO:0042995	cell projection	0.0102
DESeq2 (FDR<0.05)	Cliff	GO:0055036	virion membrane	0.0146
DESeq2 (FDR<0.05)	Cliff	GO:0051286	cell tip	0.015
DESeq2 (FDR<0.05)	Cliff	GO:0072562	blood microparticle	0.0182
DESeq2 (FDR<0.05)	Cliff	GO:0044217	other organism part	0.0241
DESeq2 (FDR<0.05)	Cliff	GO:0044424	intracellular part	0.0244
DESeq2 (FDR<0.05)	Cliff	GO:0097458	neuron part	0.0263
DESeq2 (FDR<0.05)	Cliff	GO:0044462	external encapsulating structure part	0.0264
DESeq2 (FDR<0.05)	Cliff	GO:0044456	synapse part	0.0268
DESeq2 (FDR<0.05)	Cliff	GO:0003735	structural constituent of ribosome	2.13e-10
DESeq2 (FDR<0.05)	Cliff	GO:0005198	structural molecule activity	2.31e-10
DESeq2 (FDR<0.05)	Cliff	GO:0005085	guanyl-nucleotide exchange factor activity	0.000336
DESeq2 (FDR<0.05)	Cliff	GO:0098772	molecular function regulator	0.000378
DESeq2 (FDR<0.05)	Cliff	GO:1905361	L-serine transporter activity	0.00142
DESeq2 (FDR<0.05)	Cliff	GO:0005057	signal transducer activity, downstream of receptor	0.00149
DESeq2 (FDR<0.05)	Cliff	GO:0004871	signal transducer activity	0.00235
DESeq2 (FDR<0.05)	Cliff	GO:0039660	structural constituent of virion	0.00567
DESeq2 (FDR<0.05)	Cliff	GO:0038023	signaling receptor activity	0.00766
DESeq2 (FDR<0.05)	Cliff	GO:0019808	polyamine binding	0.00942
DESeq2 (FDR<0.05)	Cliff	GO:0016740	transferase activity	0.011
DESeq2 (FDR<0.05)	Cliff	GO:0016491	oxidoreductase activity	0.0121
DESeq2 (FDR<0.05)	Cliff	GO:0016209	antioxidant activity	0.0229
DESeq2 (FDR<0.05)	Cliff	GO:0042056	chemoattractant activity	0.0251
DESeq2 (FDR<0.05)	Cliff	GO:0004872	receptor activity	0.0253
DESeq2 (FDR<0.05)	Cliff	GO:0060089	molecular transducer activity	0.0253
DESeq2 (FDR<0.05)	Cliff	GO:0097159	organic cyclic compound binding	0.0328
DESeq2 (FDR<0.05)	Cliff	GO:0016531	copper chaperone activity	0.0352

DESeq2 (FDR<0.05)	Cliff	GO:0004601	peroxidase activity	0.0362
DESeq2 (FDR<0.05)	Cliff	GO:0016787	hydrolase activity	0.0368
DESeq2 (FDR<0.05)	Cliff	GO:0019825	oxygen binding	0.0401
DESeq2 (FDR<0.05)	Cliff	GO:1901363	heterocyclic compound binding	0.0459
DESeq2 (FDR<0.05)	Cliff	GO:0001871	pattern binding	0.0462
DESeq2 (FDR<0.05)	Cliff	GO:0033218	amide binding	0.0481
DESeq2 (FDR<0.05)	Indian	GO:0050789	regulation of biological process	0.00182
DESeq2 (FDR<0.05)	Indian	GO:0065007	biological regulation	0.0025
DESeq2 (FDR<0.05)	Indian	GO:0009987	cellular process	0.00317
DESeq2 (FDR<0.05)	Indian	GO:0051606	detection of stimulus	0.00363
DESeq2 (FDR<0.05)	Indian	GO:0044767	single-organism developmental process	0.00481
DESeq2 (FDR<0.05)	Indian	GO:0032502	developmental process	0.00815
DESeq2 (FDR<0.05)	Indian	GO:0009058	biosynthetic process	0.00817
DESeq2 (FDR<0.05)	Indian	GO:0048646	anatomical structure formation involved in morphogenesis	0.00832
DESeq2 (FDR<0.05)	Indian	GO:0044699	single-organism process	0.00989
DESeq2 (FDR<0.05)	Indian	GO:0071840	cellular component organization or biogenesis	0.0128
DESeq2 (FDR<0.05)	Indian	GO:0044763	single-organism cellular process	0.0134
DESeq2 (FDR<0.05)	Indian	GO:0016043	cellular component organization	0.0156
DESeq2 (FDR<0.05)	Indian	GO:0044237	cellular metabolic process	0.0207
DESeq2 (FDR<0.05)	Indian	GO:0044238	primary metabolic process	0.0251
DESeq2 (FDR<0.05)	Indian	GO:0008152	metabolic process	0.0272
DESeq2 (FDR<0.05)	Indian	GO:0061687	detoxification of inorganic compound	0.043
DESeq2 (FDR<0.05)	Indian	GO:0007165	signal transduction	0.0434
DESeq2 (FDR<0.05)	Indian	GO:0071704	organic substance metabolic process	0.0435
DESeq2 (FDR<0.05)	Indian	GO:0032991	macromolecular complex	0.000154
DESeq2 (FDR<0.05)	Indian	GO:0044424	intracellular part	0.00113
DESeq2 (FDR<0.05)	Indian	GO:0012505	endomembrane system	0.00354
DESeq2 (FDR<0.05)	Indian	GO:0044464	cell part	0.00409
DESeq2 (FDR<0.05)	Indian	GO:0044422	organelle part	0.00503
DESeq2 (FDR<0.05)	Indian	GO:0043226	organelle	0.00599
DESeq2 (FDR<0.05)	Indian	GO:1902494	catalytic complex	0.00822
DESeq2 (FDR<0.05)	Indian	GO:0044446	intracellular organelle part	0.00905
DESeq2 (FDR<0.05)	Indian	GO:1990904	ribonucleoprotein complex	0.00983
DESeq2 (FDR<0.05)	Indian	GO:0043234	protein complex	0.0115
DESeq2 (FDR<0.05)	Indian	GO:0043229	intracellular organelle	0.012
DESeq2 (FDR<0.05)	Indian	GO:0016020	membrane	0.0125

DESeq2 (FDR<0.05)	Indian	GO:0043227	membrane-bounded organelle	0.0281
DESeq2 (FDR<0.05)	Indian	GO:0031252	cell leading edge	0.0326
DESeq2 (FDR<0.05)	Indian	GO:0005576	extracellular region	0.0334
DESeq2 (FDR<0.05)	Indian	GO:0098948	intrinsic component of postsynaptic specialization membrane	0.043
DESeq2 (FDR<0.05)	Indian	GO:0030246	carbohydrate binding	0.000553
DESeq2 (FDR<0.05)	Indian	GO:0003700	transcription factor activity, sequence-specific DNA binding	0.000843
DESeq2 (FDR<0.05)	Indian	GO:0001071	nucleic acid binding transcription factor activity	0.000844
DESeq2 (FDR<0.05)	Indian	GO:0033218	amide binding	0.00158
DESeq2 (FDR<0.05)	Indian	GO:0016530	metallochaperone activity	0.0023
DESeq2 (FDR<0.05)	Indian	GO:0016015	morphogen activity	0.0073
DESeq2 (FDR<0.05)	Indian	GO:0097159	organic cyclic compound binding	0.0235
DESeq2 (FDR<0.05)	Indian	GO:0044877	macromolecular complex binding	0.0236
DESeq2 (FDR<0.05)	Indian	GO:1901363	heterocyclic compound binding	0.0279
DESeq2 (FDR<0.05)	Indian	GO:0001871	pattern binding	0.0299
DESeq2 (FDR<0.05)	Indian	GO:0004133	glycogen debranching enzyme activity	0.036
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0009607	response to biotic stimulus	0.000326
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0009605	response to external stimulus	0.000505
DESeq2 (FDR<0.05)	Langfjordvatn	GO:1902578	single-organism localization	0.000719
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0002376	immune system process	0.0012
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0042221	response to chemical	0.00655
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0033036	macromolecule localization	0.00931
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0048589	developmental growth	0.0201
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0002252	immune effector process	0.0267
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0045321	leukocyte activation	0.0273
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0051641	cellular localization	0.035
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0006955	immune response	0.0377
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0065007	biological regulation	0.038
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0009987	cellular process	0.0386
DESeq2 (FDR<0.05)	Langfjordvatn	GO:1902728	positive regulation of growth factor dependent skeletal muscle satellite cell proliferation	0.0497
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0061687	detoxification of inorganic compound	0.0497
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0040007	growth	0.0498
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0043226	organelle	0.00102
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0044421	extracellular region part	0.00109

DESeq2 (FDR<0.05)	Langfjordvatn	GO:0044446	intracellular organelle part	0.00114
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0044422	organelle part	0.00148
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0043227	membrane-bounded organelle	0.0024
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0044424	intracellular part	0.0025
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0043229	intracellular organelle	0.00567
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0005577	fibrinogen complex	0.00893
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0031224	intrinsic component of membrane	0.0123
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0044459	plasma membrane part	0.0124
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0043230	extracellular organelle	0.0152
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0098890	extrinsic component of postsynaptic membrane	0.0169
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0098888	extrinsic component of presynaptic membrane	0.0169
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0044463	cell projection part	0.021
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0031012	extracellular matrix	0.0246
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0005886	plasma membrane	0.032
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0042995	cell projection	0.0481
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0098948	intrinsic component of postsynaptic specialization membrane	0.0497
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0055036	virion membrane	0.0497
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0005488	binding	2.8e-05
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0097159	organic cyclic compound binding	0.00247
DESeq2 (FDR<0.05)	Langfjordvatn	GO:1901363	heterocyclic compound binding	0.00368
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0022892	substrate-specific transporter activity	0.00442
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0051920	peroxiredoxin activity	0.00446
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0043167	ion binding	0.00488
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0032542	sulfiredoxin activity	0.00846
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0022857	transmembrane transporter activity	0.0114
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0005215	transporter activity	0.0145
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0038023	signaling receptor activity	0.019
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0060090	binding, bridging	0.0191
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0005326	neurotransmitter transporter activity	0.0317
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0005515	protein binding	0.0333
DESeq2 (FDR<0.05)	Langfjordvatn	GO:0004871	signal transducer activity	0.038
DESeq2 (FDR<0.05)	Langfjordvatn	GO:1901505	carbohydrate derivative transporter activity	0.0451
DESeq2 (FDR<0.05)	Skrukkebukta	GO:0033036	macromolecule localization	0.000368
DESeq2 (FDR<0.05)	Skrukkebukta	GO:0009653	anatomical structure morphogenesis	0.00321
DESeq2 (FDR<0.05)	Skrukkebukta	GO:0051235	maintenance of location	0.00679

DESeq2 (FDR<0.05)	Skrukkebukta	GO:0048511	rhythmic process	0.0136
DESeq2 (FDR<0.05)	Skrukkebukta	GO:0032991	macromolecular complex	0.00712
DESeq2 (FDR<0.05)	Skrukkebukta	GO:0005789	endoplasmic reticulum membrane	0.0129
DESeq2 (FDR<0.05)	Skrukkebukta	GO:1990904	ribonucleoprotein complex	0.0234
DESeq2 (FDR<0.05)	Skrukkebukta	GO:0031974	membrane-enclosed lumen	0.0235
DESeq2 (FDR<0.05)	Skrukkebukta	GO:0043233	organelle lumen	0.0235
DESeq2 (FDR<0.05)	Skrukkebukta	GO:0016491	oxidoreductase activity	0.0218
DESeq2 (FDR<0.05)	Skrukkebukta	GO:0039660	structural constituent of virion	0.044
DESeq2 (FDR<0.05)	Lucerne	GO:0008152	metabolic process	1.43e-08
DESeq2 (FDR<0.05)	Lucerne	GO:0071704	organic substance metabolic process	4.58e-08
DESeq2 (FDR<0.05)	Lucerne	GO:0044237	cellular metabolic process	2.5e-07
DESeq2 (FDR<0.05)	Lucerne	GO:0006807	nitrogen compound metabolic process	9.79e-07
DESeq2 (FDR<0.05)	Lucerne	GO:0044238	primary metabolic process	3.43e-06
DESeq2 (FDR<0.05)	Lucerne	GO:0007165	signal transduction	5.6e-06
DESeq2 (FDR<0.05)	Lucerne	GO:0006457	protein folding	1.47e-05
DESeq2 (FDR<0.05)	Lucerne	GO:0009056	catabolic process	3.83e-05
DESeq2 (FDR<0.05)	Lucerne	GO:0009653	anatomical structure morphogenesis	0.000148
DESeq2 (FDR<0.05)	Lucerne	GO:0003006	developmental process involved in reproduction	0.00017
DESeq2 (FDR<0.05)	Lucerne	GO:0051234	establishment of localization	0.000375
DESeq2 (FDR<0.05)	Lucerne	GO:0044767	single-organism developmental process	0.000445
DESeq2 (FDR<0.05)	Lucerne	GO:0009987	cellular process	0.000921
DESeq2 (FDR<0.05)	Lucerne	GO:0048856	anatomical structure development	0.000928
DESeq2 (FDR<0.05)	Lucerne	GO:0007155	cell adhesion	0.00107
DESeq2 (FDR<0.05)	Lucerne	GO:0022610	biological adhesion	0.0011
DESeq2 (FDR<0.05)	Lucerne	GO:0044085	cellular component biogenesis	0.00131
DESeq2 (FDR<0.05)	Lucerne	GO:0051179	localization	0.00142
DESeq2 (FDR<0.05)	Lucerne	GO:0044710	single-organism metabolic process	0.00152
DESeq2 (FDR<0.05)	Lucerne	GO:0032502	developmental process	0.00194
DESeq2 (FDR<0.05)	Lucerne	GO:0001776	leukocyte homeostasis	0.00339
DESeq2 (FDR<0.05)	Lucerne	GO:0048646	anatomical structure formation involved in morphogenesis	0.00414
DESeq2 (FDR<0.05)	Lucerne	GO:0098754	detoxification	0.00437
DESeq2 (FDR<0.05)	Lucerne	GO:0048589	developmental growth	0.00442
DESeq2 (FDR<0.05)	Lucerne	GO:0044707	single-multicellular organism process	0.00594
DESeq2 (FDR<0.05)	Lucerne	GO:0044702	single organism	0.00606

			reproductive process	
DESeq2 (FDR<0.05)	Lucerne	GO:0044700	single organism signaling	0.00638
DESeq2 (FDR<0.05)	Lucerne	GO:0007154	cell communication	0.00955
DESeq2 (FDR<0.05)	Lucerne	GO:1990748	cellular detoxification	0.0112
DESeq2 (FDR<0.05)	Lucerne	GO:0040007	growth	0.0133
DESeq2 (FDR<0.05)	Lucerne	GO:0006955	immune response	0.0139
DESeq2 (FDR<0.05)	Lucerne	GO:0032501	multicellular organismal process	0.0144
DESeq2 (FDR<0.05)	Lucerne	GO:0022414	reproductive process	0.0201
DESeq2 (FDR<0.05)	Lucerne	GO:0023052	signaling	0.0212
DESeq2 (FDR<0.05)	Lucerne	GO:1902728	positive regulation of growth factor dependent skeletal muscle satellite cell proliferation	0.0226
DESeq2 (FDR<0.05)	Lucerne	GO:0000920	cell separation after cytokinesis	0.0226
DESeq2 (FDR<0.05)	Lucerne	GO:0061687	detoxification of inorganic compound	0.0226
DESeq2 (FDR<0.05)	Lucerne	GO:0072376	protein activation cascade	0.0249
DESeq2 (FDR<0.05)	Lucerne	GO:0065009	regulation of molecular function	0.0263
DESeq2 (FDR<0.05)	Lucerne	GO:1902578	single-organism localization	0.0418
DESeq2 (FDR<0.05)	Lucerne	GO:0043500	muscle adaptation	0.0471
DESeq2 (FDR<0.05)	Lucerne	GO:1990904	ribonucleoprotein complex	3.41e-10
DESeq2 (FDR<0.05)	Lucerne	GO:0044424	intracellular part	3.55e-09
DESeq2 (FDR<0.05)	Lucerne	GO:0031974	membrane-enclosed lumen	3.71e-09
DESeq2 (FDR<0.05)	Lucerne	GO:0043233	organelle lumen	3.71e-09
DESeq2 (FDR<0.05)	Lucerne	GO:0044464	cell part	1.92e-08
DESeq2 (FDR<0.05)	Lucerne	GO:0044446	intracellular organelle part	6.54e-08
DESeq2 (FDR<0.05)	Lucerne	GO:0044422	organelle part	4.42e-07
DESeq2 (FDR<0.05)	Lucerne	GO:0044421	extracellular region part	1.36e-06
DESeq2 (FDR<0.05)	Lucerne	GO:0043226	organelle	1.6e-06
DESeq2 (FDR<0.05)	Lucerne	GO:0043229	intracellular organelle	8.65e-06
DESeq2 (FDR<0.05)	Lucerne	GO:0005576	extracellular region	5.02e-05
DESeq2 (FDR<0.05)	Lucerne	GO:0072562	blood microparticle	6.83e-05
DESeq2 (FDR<0.05)	Lucerne	GO:0030054	cell junction	8.81e-05
DESeq2 (FDR<0.05)	Lucerne	GO:0043228	non-membrane-bounded organelle	0.000193
DESeq2 (FDR<0.05)	Lucerne	GO:0034358	plasma lipoprotein particle	0.000203
DESeq2 (FDR<0.05)	Lucerne	GO:0032991	macromolecular complex	0.000209
DESeq2 (FDR<0.05)	Lucerne	GO:0032994	protein-lipid complex	0.000252
DESeq2 (FDR<0.05)	Lucerne	GO:0017053	transcriptional repressor complex	0.000309
DESeq2 (FDR<0.05)	Lucerne	GO:0005886	plasma membrane	0.000396
DESeq2 (FDR<0.05)	Lucerne	GO:0043235	receptor complex	0.000415
DESeq2 (FDR<0.05)	Lucerne	GO:0043227	membrane-bounded	0.000757

			organelle	
DESeq2 (FDR<0.05)	Lucerne	GO:0008250	oligosaccharyltransferase complex	0.000999
DESeq2 (FDR<0.05)	Lucerne	GO:0043230	extracellular organelle	0.0015
DESeq2 (FDR<0.05)	Lucerne	GO:0005615	extracellular space	0.00174
DESeq2 (FDR<0.05)	Lucerne	GO:0097470	ribbon synapse	0.00206
DESeq2 (FDR<0.05)	Lucerne	GO:0005911	cell-cell junction	0.00297
DESeq2 (FDR<0.05)	Lucerne	GO:0005789	endoplasmic reticulum membrane	0.00405
DESeq2 (FDR<0.05)	Lucerne	GO:0098590	plasma membrane region	0.00872
DESeq2 (FDR<0.05)	Lucerne	GO:0034518	RNA cap binding complex	0.00896
DESeq2 (FDR<0.05)	Lucerne	GO:0044279	other organism membrane	0.0107
DESeq2 (FDR<0.05)	Lucerne	GO:0032420	stereocilium	0.0124
DESeq2 (FDR<0.05)	Lucerne	GO:0042571	immunoglobulin complex, circulating	0.013
DESeq2 (FDR<0.05)	Lucerne	GO:0000785	chromatin	0.0148
DESeq2 (FDR<0.05)	Lucerne	GO:0070161	anchoring junction	0.0155
DESeq2 (FDR<0.05)	Lucerne	GO:0090543	Flemming body	0.0158
DESeq2 (FDR<0.05)	Lucerne	GO:0005623	cell	0.0166
DESeq2 (FDR<0.05)	Lucerne	GO:0042995	cell projection	0.0185
DESeq2 (FDR<0.05)	Lucerne	GO:0005905	clathrin-coated pit	0.0292
DESeq2 (FDR<0.05)	Lucerne	GO:0044420	extracellular matrix component	0.0312
DESeq2 (FDR<0.05)	Lucerne	GO:0099572	postsynaptic specialization	0.0373
DESeq2 (FDR<0.05)	Lucerne	GO:0097458	neuron part	0.0386
DESeq2 (FDR<0.05)	Lucerne	GO:0071944	cell periphery	0.0431
DESeq2 (FDR<0.05)	Lucerne	GO:0044297	cell body	0.0482
DESeq2 (FDR<0.05)	Lucerne	GO:0003824	catalytic activity	9.81e-07
DESeq2 (FDR<0.05)	Lucerne	GO:0016853	isomerase activity	2.74e-06
DESeq2 (FDR<0.05)	Lucerne	GO:0016874	ligase activity	1.89e-05
DESeq2 (FDR<0.05)	Lucerne	GO:0016787	hydrolase activity	3.58e-05
DESeq2 (FDR<0.05)	Lucerne	GO:0003700	transcription factor activity, sequence-specific DNA binding	0.000318
DESeq2 (FDR<0.05)	Lucerne	GO:0001071	nucleic acid binding transcription factor activity	0.000318
DESeq2 (FDR<0.05)	Lucerne	GO:0097159	organic cyclic compound binding	0.000343
DESeq2 (FDR<0.05)	Lucerne	GO:0005215	transporter activity	0.000454
DESeq2 (FDR<0.05)	Lucerne	GO:1901363	heterocyclic compound binding	0.000834
DESeq2 (FDR<0.05)	Lucerne	GO:0005488	binding	0.000986
DESeq2 (FDR<0.05)	Lucerne	GO:1901505	carbohydrate derivative transporter activity	0.00122
DESeq2 (FDR<0.05)	Lucerne	GO:0000988	transcription factor activity, protein binding	0.00144
DESeq2 (FDR<0.05)	Lucerne	GO:0000989	transcription factor activity,	0.00197

			transcription factor binding	
DESeq2 (FDR<0.05)	Lucerne	GO:0022892	substrate-specific transporter activity	0.00267
DESeq2 (FDR<0.05)	Lucerne	GO:0017056	structural constituent of nuclear pore	0.0036
DESeq2 (FDR<0.05)	Lucerne	GO:0022857	transmembrane transporter activity	0.00563
DESeq2 (FDR<0.05)	Lucerne	GO:0005085	guanyl-nucleotide exchange factor activity	0.00831
DESeq2 (FDR<0.05)	Lucerne	GO:0044877	macromolecular complex binding	0.00854
DESeq2 (FDR<0.05)	Lucerne	GO:0098772	molecular function regulator	0.0112
DESeq2 (FDR<0.05)	Lucerne	GO:0060090	binding, bridging	0.0118
DESeq2 (FDR<0.05)	Lucerne	GO:0051184	cofactor transporter activity	0.0118
DESeq2 (FDR<0.05)	Lucerne	GO:0004871	signal transducer activity	0.0126
DESeq2 (FDR<0.05)	Lucerne	GO:0016829	lyase activity	0.0204
DESeq2 (FDR<0.05)	Lucerne	GO:0032451	demethylase activity	0.0386
DESeq2 (FDR<0.05)	Lucerne	GO:0036094	small molecule binding	0.0419
DESeq2 (FDR<0.05)	Zurich	GO:0002376	immune system process	5.93e-10
DESeq2 (FDR<0.05)	Zurich	GO:0044710	single-organism metabolic process	6.05e-10
DESeq2 (FDR<0.05)	Zurich	GO:0006955	immune response	9.67e-09
DESeq2 (FDR<0.05)	Zurich	GO:0050896	response to stimulus	6.81e-07
DESeq2 (FDR<0.05)	Zurich	GO:0044699	single-organism process	1.42e-06
DESeq2 (FDR<0.05)	Zurich	GO:0009607	response to biotic stimulus	1.43e-06
DESeq2 (FDR<0.05)	Zurich	GO:0009605	response to external stimulus	4.94e-06
DESeq2 (FDR<0.05)	Zurich	GO:0002253	activation of immune response	4.31e-05
DESeq2 (FDR<0.05)	Zurich	GO:0050900	leukocyte migration	0.000127
DESeq2 (FDR<0.05)	Zurich	GO:0042221	response to chemical	0.000155
DESeq2 (FDR<0.05)	Zurich	GO:0006807	nitrogen compound metabolic process	0.000209
DESeq2 (FDR<0.05)	Zurich	GO:0006950	response to stress	0.000237
DESeq2 (FDR<0.05)	Zurich	GO:0032259	methylation	0.000609
DESeq2 (FDR<0.05)	Zurich	GO:1902578	single-organism localization	0.00077
DESeq2 (FDR<0.05)	Zurich	GO:0072376	protein activation cascade	0.00105
DESeq2 (FDR<0.05)	Zurich	GO:0098602	single organism cell adhesion	0.00105
DESeq2 (FDR<0.05)	Zurich	GO:0009719	response to endogenous stimulus	0.0011
DESeq2 (FDR<0.05)	Zurich	GO:0051179	localization	0.00156
DESeq2 (FDR<0.05)	Zurich	GO:0044763	single-organism cellular process	0.00161
DESeq2 (FDR<0.05)	Zurich	GO:1990748	cellular detoxification	0.00172
DESeq2 (FDR<0.05)	Zurich	GO:0051234	establishment of localization	0.00173

DESeq2 (FDR<0.05)	Zurich	GO:0016043	cellular component organization	0.00198
DESeq2 (FDR<0.05)	Zurich	GO:0071840	cellular component organization or biogenesis	0.00258
DESeq2 (FDR<0.05)	Zurich	GO:0098754	detoxification	0.00306
DESeq2 (FDR<0.05)	Zurich	GO:0051235	maintenance of location	0.0032
DESeq2 (FDR<0.05)	Zurich	GO:0002507	tolerance induction	0.00367
DESeq2 (FDR<0.05)	Zurich	GO:0009628	response to abiotic stimulus	0.00517
DESeq2 (FDR<0.05)	Zurich	GO:0051641	cellular localization	0.00683
DESeq2 (FDR<0.05)	Zurich	GO:0002252	immune effector process	0.00754
DESeq2 (FDR<0.05)	Zurich	GO:0022402	cell cycle process	0.0089
DESeq2 (FDR<0.05)	Zurich	GO:0007165	signal transduction	0.00965
DESeq2 (FDR<0.05)	Zurich	GO:1903046	meiotic cell cycle process	0.0105
DESeq2 (FDR<0.05)	Zurich	GO:0044700	single organism signaling	0.0112
DESeq2 (FDR<0.05)	Zurich	GO:0023052	signaling	0.0115
DESeq2 (FDR<0.05)	Zurich	GO:0051606	detection of stimulus	0.0182
DESeq2 (FDR<0.05)	Zurich	GO:0061687	detoxification of inorganic compound	0.0242
DESeq2 (FDR<0.05)	Zurich	GO:0009056	catabolic process	0.0253
DESeq2 (FDR<0.05)	Zurich	GO:0031294	lymphocyte costimulation	0.0292
DESeq2 (FDR<0.05)	Zurich	GO:0019835	cytolysis	0.039
DESeq2 (FDR<0.05)	Zurich	GO:0022601	menstrual cycle phase	0.0426
DESeq2 (FDR<0.05)	Zurich	GO:0044848	biological phase	0.0426
DESeq2 (FDR<0.05)	Zurich	GO:0031224	intrinsic component of membrane	4.19e-10
DESeq2 (FDR<0.05)	Zurich	GO:0044425	membrane part	4.81e-10
DESeq2 (FDR<0.05)	Zurich	GO:0005615	extracellular space	1.85e-07
DESeq2 (FDR<0.05)	Zurich	GO:0044459	plasma membrane part	3.12e-07
DESeq2 (FDR<0.05)	Zurich	GO:0044421	extracellular region part	4.39e-05
DESeq2 (FDR<0.05)	Zurich	GO:1902494	catalytic complex	5.6e-05
DESeq2 (FDR<0.05)	Zurich	GO:0009986	cell surface	0.000136
DESeq2 (FDR<0.05)	Zurich	GO:0000785	chromatin	0.000262
DESeq2 (FDR<0.05)	Zurich	GO:0016020	membrane	0.000418
DESeq2 (FDR<0.05)	Zurich	GO:0044464	cell part	0.000751
DESeq2 (FDR<0.05)	Zurich	GO:0031974	membrane-enclosed lumen	0.00108
DESeq2 (FDR<0.05)	Zurich	GO:0043233	organelle lumen	0.00108
DESeq2 (FDR<0.05)	Zurich	GO:0005576	extracellular region	0.00118
DESeq2 (FDR<0.05)	Zurich	GO:0032991	macromolecular complex	0.00459
DESeq2 (FDR<0.05)	Zurich	GO:0005886	plasma membrane	0.00533
DESeq2 (FDR<0.05)	Zurich	GO:0005789	endoplasmic reticulum membrane	0.0057
DESeq2 (FDR<0.05)	Zurich	GO:0098590	plasma membrane region	0.00831
DESeq2 (FDR<0.05)	Zurich	GO:0043234	protein complex	0.0142
DESeq2 (FDR<0.05)	Zurich	GO:0098552	side of membrane	0.0386
DESeq2 (FDR<0.05)	Zurich	GO:1902773	GTPase activator complex	0.0426

DESeq2 (FDR<0.05)	Zurich	GO:0043218	compact myelin	0.0426
DESeq2 (FDR<0.05)	Zurich	GO:0005215	transporter activity	1.68e-07
DESeq2 (FDR<0.05)	Zurich	GO:0022892	substrate-specific transporter activity	2.75e-06
DESeq2 (FDR<0.05)	Zurich	GO:0022857	transmembrane transporter activity	5.5e-06
DESeq2 (FDR<0.05)	Zurich	GO:0004872	receptor activity	0.000249
DESeq2 (FDR<0.05)	Zurich	GO:0060089	molecular transducer activity	0.000249
DESeq2 (FDR<0.05)	Zurich	GO:0038023	signaling receptor activity	0.000589
DESeq2 (FDR<0.05)	Zurich	GO:0016829	lyase activity	0.000726
DESeq2 (FDR<0.05)	Zurich	GO:0045735	nutrient reservoir activity	0.00181
DESeq2 (FDR<0.05)	Zurich	GO:1901476	carbohydrate transporter activity	0.00273
DESeq2 (FDR<0.05)	Zurich	GO:0051184	cofactor transporter activity	0.00284
DESeq2 (FDR<0.05)	Zurich	GO:0016491	oxidoreductase activity	0.00298
DESeq2 (FDR<0.05)	Zurich	GO:1901363	heterocyclic compound binding	0.00496
DESeq2 (FDR<0.05)	Zurich	GO:0015607	fatty-acyl-CoA transporter activity	0.00528
DESeq2 (FDR<0.05)	Zurich	GO:0003824	catalytic activity	0.00593
DESeq2 (FDR<0.05)	Zurich	GO:0097159	organic cyclic compound binding	0.00607
DESeq2 (FDR<0.05)	Zurich	GO:0042056	chemoattractant activity	0.0062
DESeq2 (FDR<0.05)	Zurich	GO:0004879	RNA polymerase II transcription factor activity, ligand-activated sequence-specific DNA binding	0.00622
DESeq2 (FDR<0.05)	Zurich	GO:1905361	L-serine transporter activity	0.00737
DESeq2 (FDR<0.05)	Zurich	GO:0005200	structural constituent of cytoskeleton	0.00874
DESeq2 (FDR<0.05)	Zurich	GO:0061783	peptidoglycan muralytic activity	0.00982
DESeq2 (FDR<0.05)	Zurich	GO:0016874	ligase activity	0.0118
DESeq2 (FDR<0.05)	Zurich	GO:0005326	neurotransmitter transporter activity	0.0122
DESeq2 (FDR<0.05)	Zurich	GO:0019825	oxygen binding	0.0197
DESeq2 (FDR<0.05)	Zurich	GO:0030246	carbohydrate binding	0.0235
DESeq2 (FDR<0.05)	Zurich	GO:0060090	binding, bridging	0.0254
DESeq2 (FDR<0.05)	Zurich	GO:0005515	protein binding	0.0317
DESeq2 (FDR<0.05)	Zurich	GO:0016853	isomerase activity	0.032
DESeq2 (FDR<0.05)	Zurich	GO:0016531	copper chaperone activity	0.0427
DESeq2 (FDR<0.05)	Maine	GO:0009058	biosynthetic process	1.58e-07
DESeq2 (FDR<0.05)	Maine	GO:0007165	signal transduction	9.95e-06
DESeq2 (FDR<0.05)	Maine	GO:0050789	regulation of biological process	0.000464
DESeq2 (FDR<0.05)	Maine	GO:0071840	cellular component organization or biogenesis	0.00101

DESeq2 (FDR<0.05)	Maine	GO:0016043	cellular component organization	0.00123
DESeq2 (FDR<0.05)	Maine	GO:0065009	regulation of molecular function	0.00137
DESeq2 (FDR<0.05)	Maine	GO:0007155	cell adhesion	0.00145
DESeq2 (FDR<0.05)	Maine	GO:0051641	cellular localization	0.00154
DESeq2 (FDR<0.05)	Maine	GO:0065007	biological regulation	0.00171
DESeq2 (FDR<0.05)	Maine	GO:0048870	cell motility	0.00172
DESeq2 (FDR<0.05)	Maine	GO:0040011	locomotion	0.00178
DESeq2 (FDR<0.05)	Maine	GO:0051716	cellular response to stimulus	0.00241
DESeq2 (FDR<0.05)	Maine	GO:0032259	methylation	0.00254
DESeq2 (FDR<0.05)	Maine	GO:0006950	response to stress	0.00449
DESeq2 (FDR<0.05)	Maine	GO:0007154	cell communication	0.00473
DESeq2 (FDR<0.05)	Maine	GO:0022610	biological adhesion	0.00479
DESeq2 (FDR<0.05)	Maine	GO:0098602	single organism cell adhesion	0.00485
DESeq2 (FDR<0.05)	Maine	GO:0045321	leukocyte activation	0.00711
DESeq2 (FDR<0.05)	Maine	GO:0022402	cell cycle process	0.00813
DESeq2 (FDR<0.05)	Maine	GO:0033036	macromolecule localization	0.00962
DESeq2 (FDR<0.05)	Maine	GO:0042330	taxis	0.0118
DESeq2 (FDR<0.05)	Maine	GO:0002376	immune system process	0.0193
DESeq2 (FDR<0.05)	Maine	GO:0048589	developmental growth	0.0201
DESeq2 (FDR<0.05)	Maine	GO:0006807	nitrogen compound metabolic process	0.0233
DESeq2 (FDR<0.05)	Maine	GO:0044406	adhesion of symbiont to host	0.0315
DESeq2 (FDR<0.05)	Maine	GO:0044706	multi-multicellular organism process	0.0318
DESeq2 (FDR<0.05)	Maine	GO:0007610	behavior	0.034
DESeq2 (FDR<0.05)	Maine	GO:0019835	cytolysis	0.0353
DESeq2 (FDR<0.05)	Maine	GO:0065008	regulation of biological quality	0.0401
DESeq2 (FDR<0.05)	Maine	GO:0051410	detoxification of nitrogen compound	0.0416
DESeq2 (FDR<0.05)	Maine	GO:1990904	ribonucleoprotein complex	2.23e-05
DESeq2 (FDR<0.05)	Maine	GO:0044279	other organism membrane	0.000873
DESeq2 (FDR<0.05)	Maine	GO:0016020	membrane	0.00266
DESeq2 (FDR<0.05)	Maine	GO:0044423	virion part	0.00463
DESeq2 (FDR<0.05)	Maine	GO:0033643	host cell part	0.00513
DESeq2 (FDR<0.05)	Maine	GO:0005615	extracellular space	0.0055
DESeq2 (FDR<0.05)	Maine	GO:0031975	envelope	0.00698
DESeq2 (FDR<0.05)	Maine	GO:0098796	membrane protein complex	0.0136
DESeq2 (FDR<0.05)	Maine	GO:0043226	organelle	0.0142
DESeq2 (FDR<0.05)	Maine	GO:0044326	dendritic spine neck	0.0159
DESeq2 (FDR<0.05)	Maine	GO:0044459	plasma membrane part	0.0172

DESeq2 (FDR<0.05)	Maine	GO:0042995	cell projection	0.0181
DESeq2 (FDR<0.05)	Maine	GO:0043234	protein complex	0.0185
DESeq2 (FDR<0.05)	Maine	GO:0043229	intracellular organelle	0.0223
DESeq2 (FDR<0.05)	Maine	GO:0055036	virion membrane	0.0232
DESeq2 (FDR<0.05)	Maine	GO:0043227	membrane-bounded organelle	0.0232
DESeq2 (FDR<0.05)	Maine	GO:0019013	viral nucleocapsid	0.0315
DESeq2 (FDR<0.05)	Maine	GO:0000785	chromatin	0.0346
DESeq2 (FDR<0.05)	Maine	GO:0044425	membrane part	0.0384
DESeq2 (FDR<0.05)	Maine	GO:1902773	GTPase activator complex	0.0416
DESeq2 (FDR<0.05)	Maine	GO:1990015	ensheathing process	0.0416
DESeq2 (FDR<0.05)	Maine	GO:0005623	cell	0.0454
DESeq2 (FDR<0.05)	Maine	GO:0034358	plasma lipoprotein particle	0.0495
DESeq2 (FDR<0.05)	Maine	GO:0003735	structural constituent of ribosome	9.98e-11
DESeq2 (FDR<0.05)	Maine	GO:0005198	structural molecule activity	1.06e-06
DESeq2 (FDR<0.05)	Maine	GO:0016491	oxidoreductase activity	3.25e-06
DESeq2 (FDR<0.05)	Maine	GO:0005085	guanyl-nucleotide exchange factor activity	4.37e-06
DESeq2 (FDR<0.05)	Maine	GO:0098772	molecular function regulator	5.05e-05
DESeq2 (FDR<0.05)	Maine	GO:0019808	polyamine binding	0.00173
DESeq2 (FDR<0.05)	Maine	GO:0032451	demethylase activity	0.00326
DESeq2 (FDR<0.05)	Maine	GO:0048037	cofactor binding	0.00333
DESeq2 (FDR<0.05)	Maine	GO:0016740	transferase activity	0.005
DESeq2 (FDR<0.05)	Maine	GO:0005515	protein binding	0.00613
DESeq2 (FDR<0.05)	Maine	GO:0032947	protein complex scaffold	0.0074
DESeq2 (FDR<0.05)	Maine	GO:0030234	enzyme regulator activity	0.0112
DESeq2 (FDR<0.05)	Maine	GO:0039660	structural constituent of virion	0.015
DESeq2 (FDR<0.05)	Maine	GO:0004871	signal transducer activity	0.0184
DESeq2 (FDR<0.05)	Maine	GO:0009055	electron carrier activity	0.0234
DESeq2 (FDR<0.05)	Maine	GO:0016209	antioxidant activity	0.026
DESeq2 (FDR<0.05)	Maine	GO:0005057	signal transducer activity, downstream of receptor	0.0282
DESeq2 (FDR<0.05)	Maine	GO:0090484	drug transporter activity	0.0344
DESeq2 (FDR<0.05)	Maine	GO:0016787	hydrolase activity	0.0375
DESeq2 (FDR<0.05)	Maine	GO:0000386	second spliceosomal transesterification activity	0.0416
DESeq2 (FDR<0.05)	Maine	GO:0045174	glutathione dehydrogenase (ascorbate) activity	0.0416
DESeq2 (FDR<0.05)	Maine	GO:0005326	neurotransmitter transporter activity	0.0467
DESeq2 (FDR<0.05)	Norway	GO:0009605	response to external stimulus	0.000425
DESeq2 (FDR<0.05)	Norway	GO:0002376	immune system process	0.000934
DESeq2 (FDR<0.05)	Norway	GO:0009607	response to biotic stimulus	0.00167

DESeq2 (FDR<0.05)	Norway	GO:0006955	immune response	0.0042
DESeq2 (FDR<0.05)	Norway	GO:0033036	macromolecule localization	0.0114
DESeq2 (FDR<0.05)	Norway	GO:0044238	primary metabolic process	0.0166
DESeq2 (FDR<0.05)	Norway	GO:0071704	organic substance metabolic process	0.0208
DESeq2 (FDR<0.05)	Norway	GO:0006807	nitrogen compound metabolic process	0.0213
DESeq2 (FDR<0.05)	Norway	GO:0042221	response to chemical	0.0267
DESeq2 (FDR<0.05)	Norway	GO:0008152	metabolic process	0.0294
DESeq2 (FDR<0.05)	Norway	GO:0050900	leukocyte migration	0.0339
DESeq2 (FDR<0.05)	Norway	GO:0065007	biological regulation	0.0377
DESeq2 (FDR<0.05)	Norway	GO:0051641	cellular localization	0.0401
DESeq2 (FDR<0.05)	Norway	GO:0044237	cellular metabolic process	0.0493
DESeq2 (FDR<0.05)	Norway	GO:0043227	membrane-bounded organelle	0.00115
DESeq2 (FDR<0.05)	Norway	GO:0044459	plasma membrane part	0.00308
DESeq2 (FDR<0.05)	Norway	GO:0043226	organelle	0.0031
DESeq2 (FDR<0.05)	Norway	GO:0043229	intracellular organelle	0.00597
DESeq2 (FDR<0.05)	Norway	GO:0005577	fibrinogen complex	0.00994
DESeq2 (FDR<0.05)	Norway	GO:0044424	intracellular part	0.0117
DESeq2 (FDR<0.05)	Norway	GO:0098890	extrinsic component of postsynaptic membrane	0.0178
DESeq2 (FDR<0.05)	Norway	GO:0098888	extrinsic component of presynaptic membrane	0.0178
DESeq2 (FDR<0.05)	Norway	GO:0005576	extracellular region	0.0221
DESeq2 (FDR<0.05)	Norway	GO:0097458	neuron part	0.0244
DESeq2 (FDR<0.05)	Norway	GO:0098589	membrane region	0.0248
DESeq2 (FDR<0.05)	Norway	GO:0098552	side of membrane	0.0472
DESeq2 (FDR<0.05)	Norway	GO:0022857	transmembrane transporter activity	0.000608
DESeq2 (FDR<0.05)	Norway	GO:0005488	binding	0.000966
DESeq2 (FDR<0.05)	Norway	GO:0005215	transporter activity	0.00189
DESeq2 (FDR<0.05)	Norway	GO:0022892	substrate-specific transporter activity	0.00204
DESeq2 (FDR<0.05)	Norway	GO:0097159	organic cyclic compound binding	0.0126
DESeq2 (FDR<0.05)	Norway	GO:1901363	heterocyclic compound binding	0.0149
DESeq2 (FDR<0.05)	Norway	GO:0043167	ion binding	0.0187
DESeq2 (FDR<0.05)	Norway	GO:0005200	structural constituent of cytoskeleton	0.0259
DESeq2 (FDR<0.05)	Norway	GO:0019534	toxin transporter activity	0.0266
DESeq2 (FDR<0.05)	Norway	GO:0004871	signal transducer activity	0.0311
DESeq2 (FDR<0.05)	Norway	GO:0038023	signaling receptor activity	0.0339
DESeq2 (FDR<0.05)	Norway	GO:0005326	neurotransmitter transporter activity	0.0352
DESeq2 (FDR<0.05)	Norway	GO:1901505	carbohydrate derivative	0.0499

			transporter activity	
DESeq2 (FDR<0.05)	Switzerland	GO:0006955	immune response	4.94e-10
DESeq2 (FDR<0.05)	Switzerland	GO:0044710	single-organism metabolic process	6.61e-10
DESeq2 (FDR<0.05)	Switzerland	GO:0002376	immune system process	1.04e-08
DESeq2 (FDR<0.05)	Switzerland	GO:0044699	single-organism process	2.71e-06
DESeq2 (FDR<0.05)	Switzerland	GO:0009607	response to biotic stimulus	1.46e-05
DESeq2 (FDR<0.05)	Switzerland	GO:0050896	response to stimulus	1.69e-05
DESeq2 (FDR<0.05)	Switzerland	GO:0002253	activation of immune response	3.3e-05
DESeq2 (FDR<0.05)	Switzerland	GO:0009605	response to external stimulus	8.43e-05
DESeq2 (FDR<0.05)	Switzerland	GO:0002507	tolerance induction	0.000227
DESeq2 (FDR<0.05)	Switzerland	GO:0002252	immune effector process	0.000296
DESeq2 (FDR<0.05)	Switzerland	GO:0072376	protein activation cascade	0.000378
DESeq2 (FDR<0.05)	Switzerland	GO:0042221	response to chemical	0.000505
DESeq2 (FDR<0.05)	Switzerland	GO:0044763	single-organism cellular process	0.000862
DESeq2 (FDR<0.05)	Switzerland	GO:0051641	cellular localization	0.001
DESeq2 (FDR<0.05)	Switzerland	GO:0006950	response to stress	0.00113
DESeq2 (FDR<0.05)	Switzerland	GO:0031294	lymphocyte costimulation	0.0019
DESeq2 (FDR<0.05)	Switzerland	GO:0016043	cellular component organization	0.00323
DESeq2 (FDR<0.05)	Switzerland	GO:0009056	catabolic process	0.00398
DESeq2 (FDR<0.05)	Switzerland	GO:0050900	leukocyte migration	0.00444
DESeq2 (FDR<0.05)	Switzerland	GO:0009719	response to endogenous stimulus	0.0055
DESeq2 (FDR<0.05)	Switzerland	GO:0032259	methylation	0.00844
DESeq2 (FDR<0.05)	Switzerland	GO:0009628	response to abiotic stimulus	0.00903
DESeq2 (FDR<0.05)	Switzerland	GO:0071840	cellular component organization or biogenesis	0.00934
DESeq2 (FDR<0.05)	Switzerland	GO:0007165	signal transduction	0.00976
DESeq2 (FDR<0.05)	Switzerland	GO:1903046	meiotic cell cycle process	0.0112
DESeq2 (FDR<0.05)	Switzerland	GO:0022402	cell cycle process	0.0127
DESeq2 (FDR<0.05)	Switzerland	GO:0044700	single organism signaling	0.018
DESeq2 (FDR<0.05)	Switzerland	GO:0023052	signaling	0.0183
DESeq2 (FDR<0.05)	Switzerland	GO:0044085	cellular component biogenesis	0.0201
DESeq2 (FDR<0.05)	Switzerland	GO:0098602	single organism cell adhesion	0.0234
DESeq2 (FDR<0.05)	Switzerland	GO:0001776	leukocyte homeostasis	0.0239
DESeq2 (FDR<0.05)	Switzerland	GO:0000075	cell cycle checkpoint	0.0262
DESeq2 (FDR<0.05)	Switzerland	GO:1902578	single-organism localization	0.0353
DESeq2 (FDR<0.05)	Switzerland	GO:0044848	biological phase	0.044
DESeq2 (FDR<0.05)	Switzerland	GO:0022601	menstrual cycle phase	0.044
DESeq2 (FDR<0.05)	Switzerland	GO:0009987	cellular process	0.0475

DESeq2 (FDR<0.05)	Switzerland	GO:0048511	rhythmic process	0.0478
DESeq2 (FDR<0.05)	Switzerland	GO:0031224	intrinsic component of membrane	3.83e-10
DESeq2 (FDR<0.05)	Switzerland	GO:0044425	membrane part	4.84e-10
DESeq2 (FDR<0.05)	Switzerland	GO:0044421	extracellular region part	6.88e-05
DESeq2 (FDR<0.05)	Switzerland	GO:0044464	cell part	0.000115
DESeq2 (FDR<0.05)	Switzerland	GO:0005615	extracellular space	0.000153
DESeq2 (FDR<0.05)	Switzerland	GO:0044459	plasma membrane part	0.000725
DESeq2 (FDR<0.05)	Switzerland	GO:0009986	cell surface	0.00341
DESeq2 (FDR<0.05)	Switzerland	GO:0031974	membrane-enclosed lumen	0.00551
DESeq2 (FDR<0.05)	Switzerland	GO:0043233	organelle lumen	0.00551
DESeq2 (FDR<0.05)	Switzerland	GO:0005576	extracellular region	0.00868
DESeq2 (FDR<0.05)	Switzerland	GO:1902494	catalytic complex	0.0116
DESeq2 (FDR<0.05)	Switzerland	GO:0043230	extracellular organelle	0.0121
DESeq2 (FDR<0.05)	Switzerland	GO:0045178	basal part of cell	0.0169
DESeq2 (FDR<0.05)	Switzerland	GO:0000785	chromatin	0.0307
DESeq2 (FDR<0.05)	Switzerland	GO:0098552	side of membrane	0.0435
DESeq2 (FDR<0.05)	Switzerland	GO:1902773	GTPase activator complex	0.044
DESeq2 (FDR<0.05)	Switzerland	GO:0043218	compact myelin	0.044
DESeq2 (FDR<0.05)	Switzerland	GO:0004872	receptor activity	2.04e-05
DESeq2 (FDR<0.05)	Switzerland	GO:0060089	molecular transducer activity	2.04e-05
DESeq2 (FDR<0.05)	Switzerland	GO:0038023	signaling receptor activity	0.000107
DESeq2 (FDR<0.05)	Switzerland	GO:0003735	structural constituent of ribosome	0.000222
DESeq2 (FDR<0.05)	Switzerland	GO:0016874	ligase activity	0.000315
DESeq2 (FDR<0.05)	Switzerland	GO:0005215	transporter activity	0.0011
DESeq2 (FDR<0.05)	Switzerland	GO:0022892	substrate-specific transporter activity	0.00136
DESeq2 (FDR<0.05)	Switzerland	GO:0022857	transmembrane transporter activity	0.00182
DESeq2 (FDR<0.05)	Switzerland	GO:0045735	nutrient reservoir activity	0.00193
DESeq2 (FDR<0.05)	Switzerland	GO:0051184	cofactor transporter activity	0.00335
DESeq2 (FDR<0.05)	Switzerland	GO:0015607	fatty-acyl-CoA transporter activity	0.00563
DESeq2 (FDR<0.05)	Switzerland	GO:1905361	L-serine transporter activity	0.00808
DESeq2 (FDR<0.05)	Switzerland	GO:1901476	carbohydrate transporter activity	0.0122
DESeq2 (FDR<0.05)	Switzerland	GO:0016829	lyase activity	0.0135
DESeq2 (FDR<0.05)	Switzerland	GO:0004879	RNA polymerase II transcription factor activity, ligand-activated sequence-specific DNA binding	0.0213
DESeq2 (FDR<0.05)	Switzerland	GO:0016491	oxidoreductase activity	0.0222
DESeq2 (FDR<0.05)	Switzerland	GO:0003824	catalytic activity	0.0327
DESeq2 (FDR<0.05)	Switzerland	GO:0004871	signal transducer activity	0.0421

DESeq2 (FDR<0.05)	Switzerland	GO:0042056	chemoattractant activity	0.0423
DESeq2 (FDR<0.05)	Maine-Norway	GO:0032259	methylation	0.000507
DESeq2 (FDR<0.05)	Maine-Norway	GO:0019835	cytolysis	0.0195
DESeq2 (FDR<0.05)	Maine-Norway	GO:0065007	biological regulation	0.0257
DESeq2 (FDR<0.05)	Maine-Norway	GO:0007165	signal transduction	0.031
DESeq2 (FDR<0.05)	Maine-Norway	GO:0009653	anatomical structure morphogenesis	0.037
DESeq2 (FDR<0.05)	Maine-Norway	GO:0050789	regulation of biological process	0.0437
DESeq2 (FDR<0.05)	Maine-Norway	GO:0005576	extracellular region	0.000634
DESeq2 (FDR<0.05)	Maine-Norway	GO:0070161	anchoring junction	0.00417
DESeq2 (FDR<0.05)	Maine-Norway	GO:0032991	macromolecular complex	0.00525
DESeq2 (FDR<0.05)	Maine-Norway	GO:0043234	protein complex	0.0112
DESeq2 (FDR<0.05)	Maine-Norway	GO:0044421	extracellular region part	0.0195
DESeq2 (FDR<0.05)	Maine-Norway	GO:0030672	synaptic vesicle membrane	0.0213
DESeq2 (FDR<0.05)	Maine-Norway	GO:0034358	plasma lipoprotein particle	0.0241
DESeq2 (FDR<0.05)	Maine-Norway	GO:0032994	protein-lipid complex	0.0251
DESeq2 (FDR<0.05)	Maine-Norway	GO:0098576	lumenal side of membrane	0.0287
DESeq2 (FDR<0.05)	Maine-Norway	GO:1990904	ribonucleoprotein complex	0.0368
DESeq2 (FDR<0.05)	Maine-Norway	GO:0016942	insulin-like growth factor binding protein complex	0.0474
DESeq2 (FDR<0.05)	Maine-Norway	GO:0044326	dendritic spine neck	0.0474
DESeq2 (FDR<0.05)	Maine-Norway	GO:0016491	oxidoreductase activity	0.000168
DESeq2 (FDR<0.05)	Maine-Norway	GO:0048037	cofactor binding	0.00557
DESeq2 (FDR<0.05)	Maine-Norway	GO:0009055	electron carrier activity	0.00899
DESeq2 (FDR<0.05)	Maine-Norway	GO:0045174	glutathione dehydrogenase (ascorbate) activity	0.00966
DESeq2 (FDR<0.05)	Maine-Norway	GO:0016209	antioxidant activity	0.0304
DESeq2 (FDR<0.05)	Maine-Norway	GO:0098772	molecular function regulator	0.0313
DESeq2 (FDR<0.05)	Maine-Norway	GO:0001871	pattern binding	0.0496
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0007165	signal transduction	0.000587
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0032259	methylation	0.00621
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0044238	primary metabolic process	0.0213
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0050789	regulation of biological process	0.0302
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0065007	biological regulation	0.0313
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0044237	cellular metabolic process	0.0379
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0019835	cytolysis	0.0449
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0051716	cellular response to stimulus	0.0469
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0032991	macromolecular complex	6.08e-05
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0043234	protein complex	0.0029
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0097223	sperm part	0.00569
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:1990015	ensheathing process	0.00571

DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0070161	anchoring junction	0.0109
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0043229	intracellular organelle	0.0135
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0098576	lumenal side of membrane	0.017
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0098948	intrinsic component of postsynaptic specialization membrane	0.0338
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0043228	non-membrane-bounded organelle	0.04
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0043230	extracellular organelle	0.0446
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0030672	synaptic vesicle membrane	0.0477
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0098772	molecular function regulator	0.00755
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0036094	small molecule binding	0.0242
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0030234	enzyme regulator activity	0.0271
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0003823	antigen binding	0.0297
DESeq2 (FDR<0.05)	Maine-Switzerland	GO:0004791	thioredoxin-disulfide reductase activity	0.0393
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0016043	cellular component organization	0.00759
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0009987	cellular process	0.00819
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0071840	cellular component organization or biogenesis	0.00972
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0001776	leukocyte homeostasis	0.00991
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0051037	regulation of transcription involved in meiotic cell cycle	0.0173
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0048609	multicellular organismal reproductive process	0.0232
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0051641	cellular localization	0.0278
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0009605	response to external stimulus	0.0312
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0051606	detection of stimulus	0.0333
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0065007	biological regulation	0.0405
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0051235	maintenance of location	0.0408
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0002376	immune system process	0.0465
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0008283	cell proliferation	0.0468
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0008021	synaptic vesicle	0.0156
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0098890	extrinsic component of postsynaptic membrane	0.0173
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0098888	extrinsic component of presynaptic membrane	0.0173
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0032991	macromolecular complex	0.0186
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0098794	postsynapse	0.025
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:1990904	ribonucleoprotein complex	0.0284
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0030054	cell junction	0.0321
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0009986	cell surface	0.0399
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0060090	binding, bridging	0.00604

DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0005488	binding	0.0102
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0022857	transmembrane transporter activity	0.0129
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0005215	transporter activity	0.0224
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0019534	toxin transporter activity	0.0258
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0097159	organic cyclic compound binding	0.032
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:1901363	heterocyclic compound binding	0.0373
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0030234	enzyme regulator activity	0.0373
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:0022892	substrate-specific transporter activity	0.0446
DESeq2 (FDR<0.05)	Norway-Switzerland	GO:1901505	carbohydrate derivative transporter activity	0.0472
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0006807	nitrogen compound metabolic process	0.000597
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0007165	signal transduction	0.00151
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0044237	cellular metabolic process	0.00189
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0044238	primary metabolic process	0.00563
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0071704	organic substance metabolic process	0.00792
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0065007	biological regulation	0.00962
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0050789	regulation of biological process	0.0115
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0009058	biosynthetic process	0.0116
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0032259	methylation	0.0125
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0008152	metabolic process	0.0167
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0051716	cellular response to stimulus	0.0436
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0032991	macromolecular complex	0.000102
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0043234	protein complex	0.000716
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0099080	supramolecular complex	0.00204
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0099081	supramolecular polymer	0.00204
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:1990015	ensheathing process	0.0048
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0098576	lumenal side of membrane	0.0143
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0097223	sperm part	0.022
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0098948	intrinsic component of postsynaptic specialization membrane	0.0284
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0030672	synaptic vesicle membrane	0.0348
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0035749	myelin sheath adaxonal region	0.0424
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0098936	intrinsic component of postsynaptic membrane	0.047
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0016491	oxidoreductase activity	0.0031
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0003823	antigen binding	0.0215
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0004791	thioredoxin-disulfide reductase activity	0.0331

DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0048037	cofactor binding	0.043
DESeq2 (FDR<0.05)	Maine-Norway-Switzerland	GO:0016209	antioxidant activity	0.0447
RDAexp (P<0.01)	Limnetic-Benthic	GO:0065007	biological_regulation	0.000446
RDAexp (P<0.01)	Limnetic-Benthic	GO:0008152	metabolic_process	0.000547
RDAexp (P<0.01)	Limnetic-Benthic	GO:0040007	growth	0.0259
RDAexp (P<0.01)	Limnetic-Benthic	GO:0032991	macromolecular_complex	0.00126
RDAexp (P<0.01)	Limnetic-Benthic	GO:0044422	organelle_part	0.00398
RDAexp (P<0.01)	Limnetic-Benthic	GO:0044464	cell_part	0.00716
RDAexp (P<0.01)	Limnetic-Benthic	GO:0043226	organelle	0.0375
RDAexp (P<0.01)	Limnetic-Benthic	GO:0005576	extracellular_region	0.0463
RDAexp (P<0.01)	Limnetic-Benthic	GO:0005488	binding	0.00877
RDAexp (P<0.001)	Limnetic-Benthic	GO:0008152	metabolic_process	0.0168
RDAexp (P<0.001)	Limnetic-Benthic	GO:0065007	biological_regulation	0.0186
RDAexp (P<0.001)	Limnetic-Benthic	GO:0009987	cellular_process	0.0194
RDAexp (P<0.001)	Limnetic-Benthic	GO:0050896	response_to_stimulus	0.0472
RDAexp (P<0.001)	Limnetic-Benthic	GO:0044464	cell_part	0.0067
RDAexp (P<0.001)	Limnetic-Benthic	GO:0044422	organelle_part	0.0392
RDAexp (P<0.001)	Limnetic-Benthic	GO:0005488	binding	0.00298
RDAgeno (P<0.01)	Limnetic-Benthic	GO:0044421	extracellular region part	0.0148
RDAgeno (P<0.01)	Limnetic-Benthic	GO:0044425	membrane_part	0.0216
RDAgeno (P<0.01)	Limnetic-Benthic	GO:0043226	organelle	0.0392
RDAgeno (P<0.001)	Limnetic-Benthic	GO:0008152	metabolic process	1.07e-05
RDAgeno (P<0.001)	Limnetic-Benthic	GO:0009987	cellular process	0.00025
RDAgeno (P<0.001)	Limnetic-Benthic	GO:0002376	immune system process	0.00076
RDAgeno (P<0.001)	Limnetic-Benthic	GO:0044699	single-organism process	0.00157
RDAgeno (P<0.001)	Limnetic-Benthic	GO:0032502	developmental process	0.0104
RDAgeno (P<0.001)	Limnetic-Benthic	GO:0031974	membrane-enclosed lumen	9.67e-07
RDAgeno (P<0.001)	Limnetic-Benthic	GO:0005576	extracellular region	5.97e-06
RDAgeno (P<0.001)	Limnetic-Benthic	GO:0044421	extracellular region part	1.55e-05
RDAgeno (P<0.001)	Limnetic-Benthic	GO:0044464	cell part	2.92e-05
RDAgeno (P<0.001)	Limnetic-Benthic	GO:0044422	organelle part	0.00011

Table S4.3: Significant subnetworks from the KEGG database. The significant subnets (P<0.05), the number of genes composing the subnet and their respective pathways are indicated. The subnet composition is detailed by the Entrez ID of genes. Aiming to control for any bias in the significance threshold caused by highly expressed genes, we normalized the distribution of the counts.

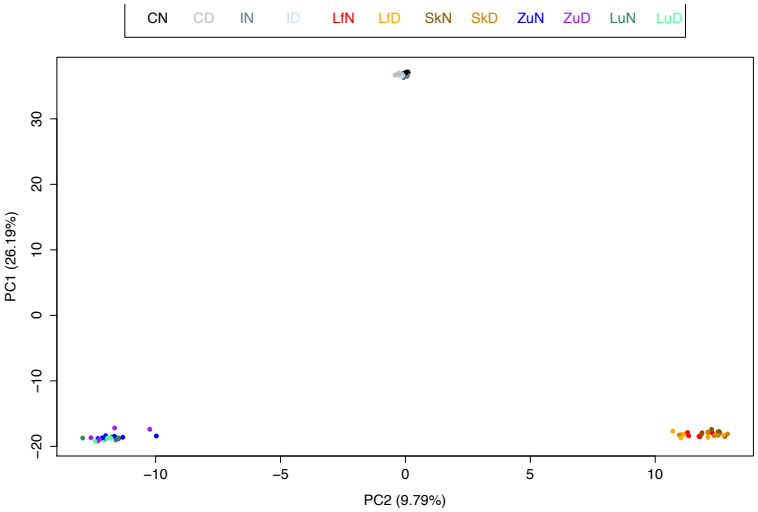
Database	Pathway	Pathway size	Subnet size	Subnet score	p-value	Subnet genes (Entrez ID)
KEGG	Pentose phosphate pathway	10	2	0.788	0.024	5226; 6120
KEGG	Purine metabolism	49	37	0.432	0.031	108; 112; 271; 272; 318; 661; 953; 954; 2982; 2986; 2987; 5136; 5140 ;5142; 5144; 5150; 5151; 5315; 5426; 5431 5439; 5557; 8622; 8654; 8833; 10621; 10846; 23649; 25885; 29922; 50484; 51728; 64425; 87178; 122481; 131870; 171568

						661; 953; 1854; 5426; 5431; 5557; 10621;
KEGG	Pyrimidine metabolism	29	14	0.393	0.035	23649; 25885; 29922; 50484; 64425; 79077; 171568
KEGG	Glycerolipid metabolism	18	2	0.362	0.036	9388; 9663
KEGG	Drug metabolism - cytochrome P450	8	3	0.341	0.036	221; 4257; 9446
KEGG	Glutathione metabolism	12	3	0.337	0.036	2687; 4257; 9446
KEGG	Nicotinate and nicotinamide metabolism	8	2	0.301	0.040	952; 4860
KEGG	cAMP signaling pathway	50	16	0.255	0.047	492; 572; 776; 818; 1387; 2033; 5140; 5142; 5144; 5465; 5566; 5567; 5594; 5595; 9586; 90993
KEGG	FoxO signaling pathway	44	18	0.251	0.047	208; 472; 604; 901; 1026; 1263; 2309; 5594; 5595; 5601; 5603; 5934; 8743; 10000; 10110; 10365;

						26260; 51422 1277; 3675; 3680; 3690; 3691; 3909; 3913; 5649; 5747; 7148; 22801; 131873
KEGG	PI3K-Akt signaling pathway	83	12	0.249	0.047	

4.11 Supplementary figures

A)



B)

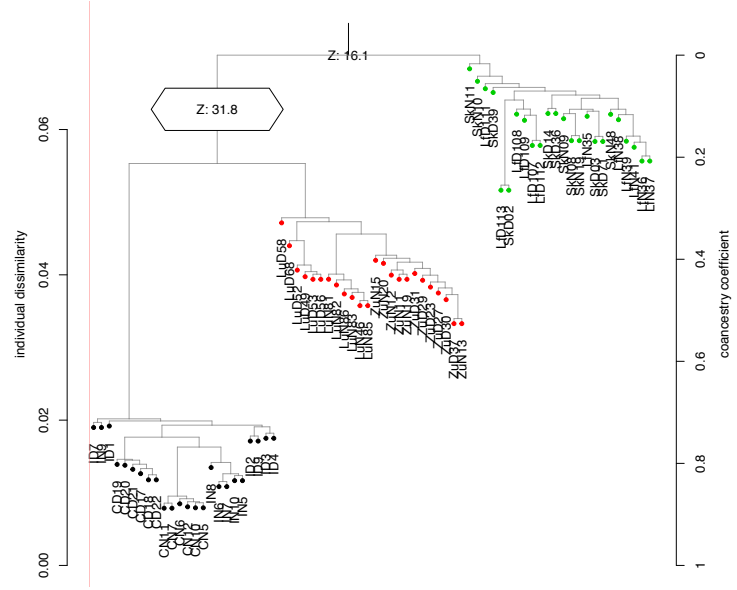


Figure S4.1: Genetic clustering of limnetic and benthic species' populations. A) Principal component analysis (PCA) generated from 20,911 SNPs. Individuals from Cliff (C), Indian (I), Langfjordvatn (LF), Skrukkebukta (SK), Zurich (Zu) and Lucerne (Lu) lakes for limnetic (-D) and benthic (-N) species are projected along PC1 discriminated populations of *C. clupeaformis* form *C. lavaretus*, and PC2 lakes from Norway and Switzerland. B) Similarity tree decomposing the individual genetic distance across the system. Less differentiated species-pairs clustered together within region (*e.g.*, Norway in green and Indian Lake in Maine region, balck).

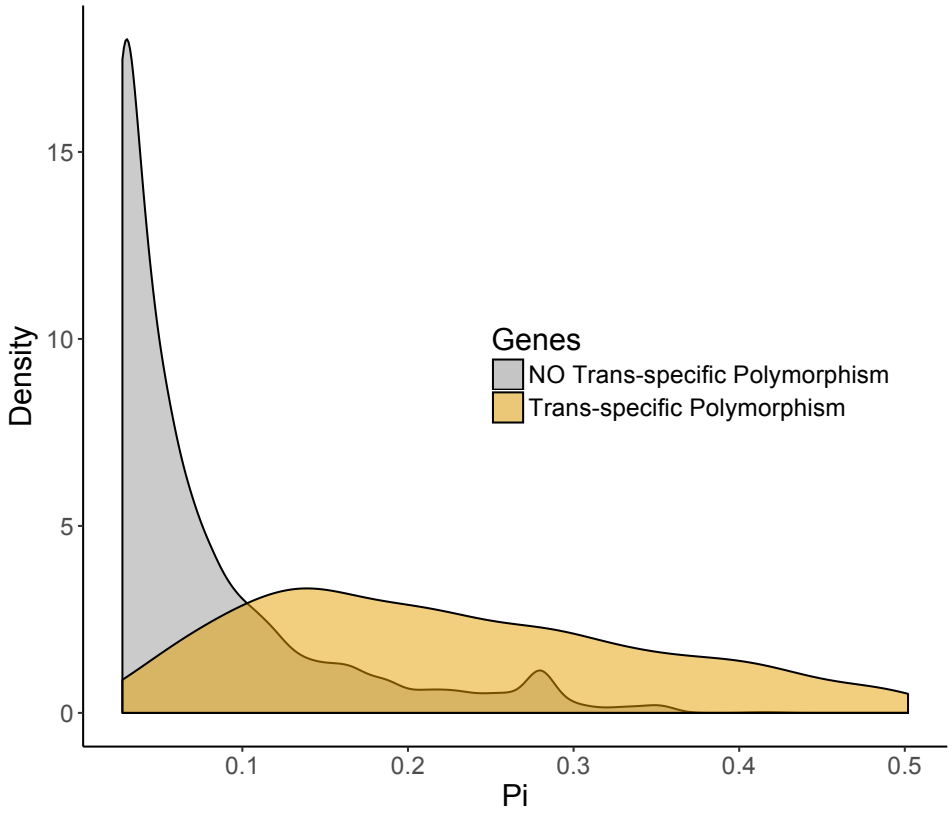


Figure S4.2: Distribution of genetic diversity (P) for two categories of genes. Genes with trans-specific polymorphism (yellow) show higher genetic diversity among species than genes without trans-specific polymorphism (Student's *t*-test, *p*-value<0.001).

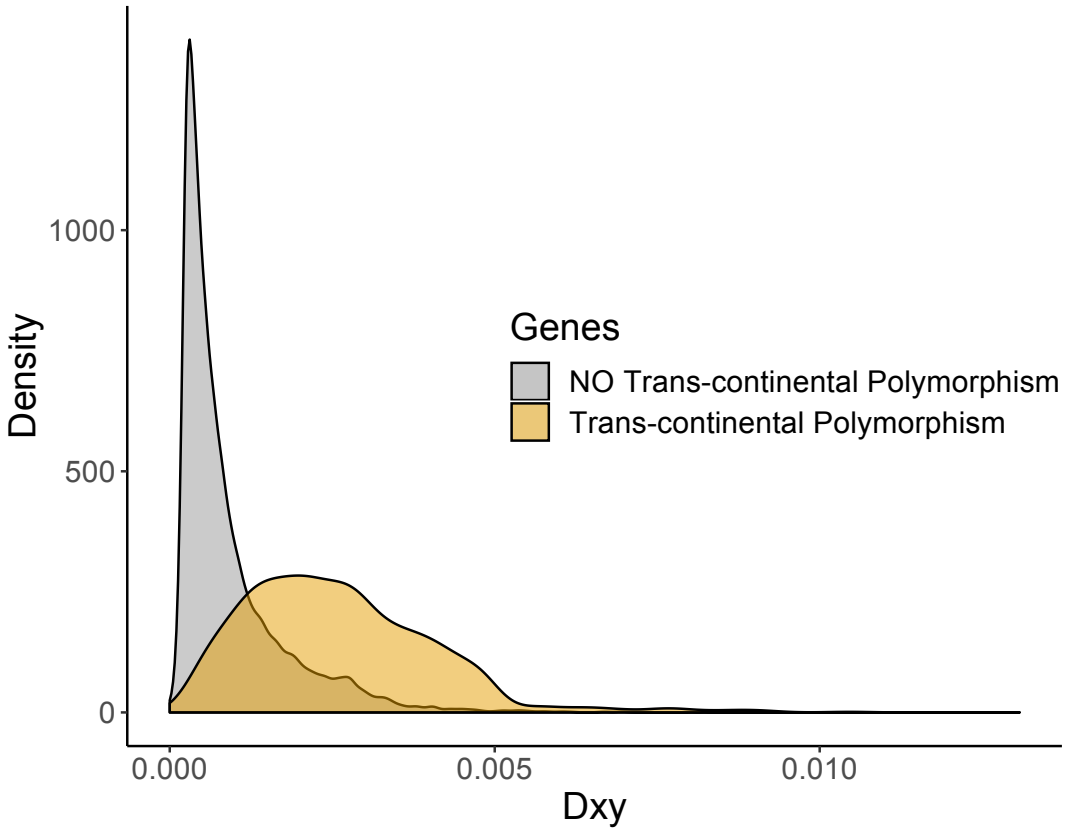


Figure S4.3: Distribution of absolute genetic divergence (Dxy) for two categories of genes. Genes with trans-specific polymorphism (yellow) showed higher genetic divergence between species than genes without trans-specific polymorphism (Student's *t*-test, *p*-value<0.001).

Figure S4.4: Transcripts z-score distribution from genotypic data. Based on allele frequencies, each locus score obtained from cRDA analysis was transformed in z-score. Full and dashed red lines correspond to significance threshold of *p*-value of 0.01 and 0.001, respectively.

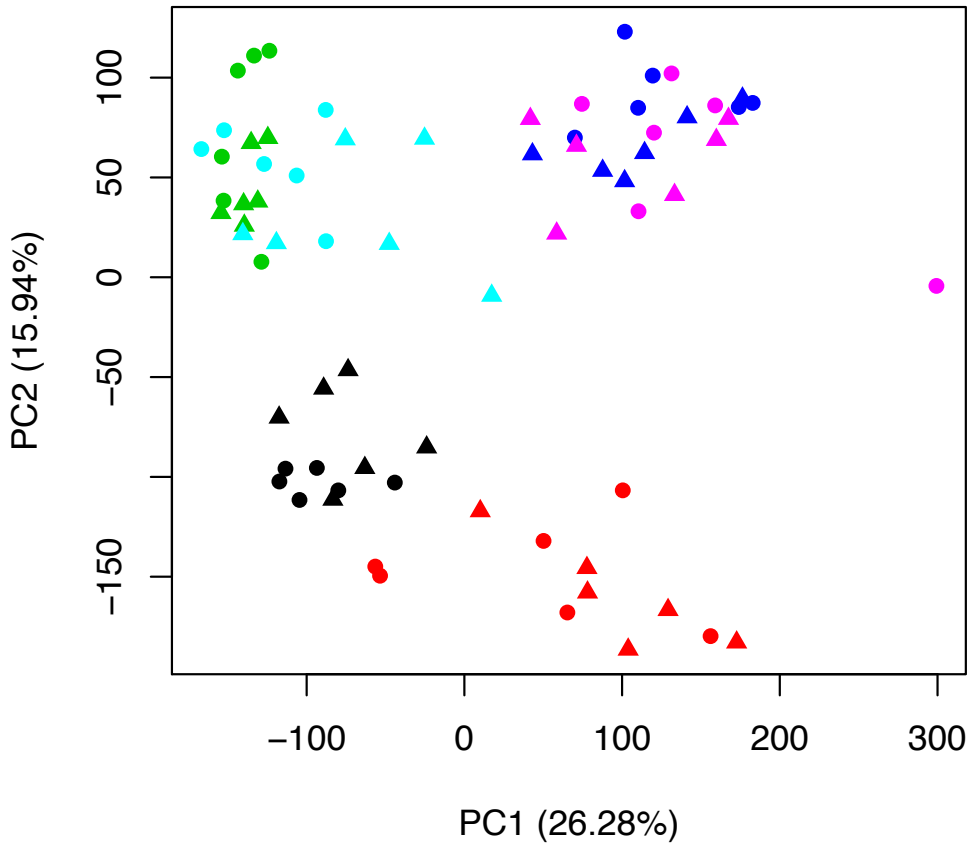


Figure S4.5: Principal component analysis on gene expression level between individuals. The PCA indicates the transcriptomic relationships between individuals of Cliff (black), Indian (red), Langfjordvatn (green), Skrukkebukta (lightblue), Zurich (pink) and Lucerne (blue) lakes for limnetic (circles) and benthic (triangles) species are projected along PC1 and PC2. PC1 separated lakes from Norway and Switzerland, but also C.liff and Indian Lakes. PC2 allowed discriminating populations of *C. clupeaformis* form *C. lavaretus*. Indian, Skrukkebukta and Zurich (less differentiated lakes per regions) showed more inter-indivual gene expression level variation.

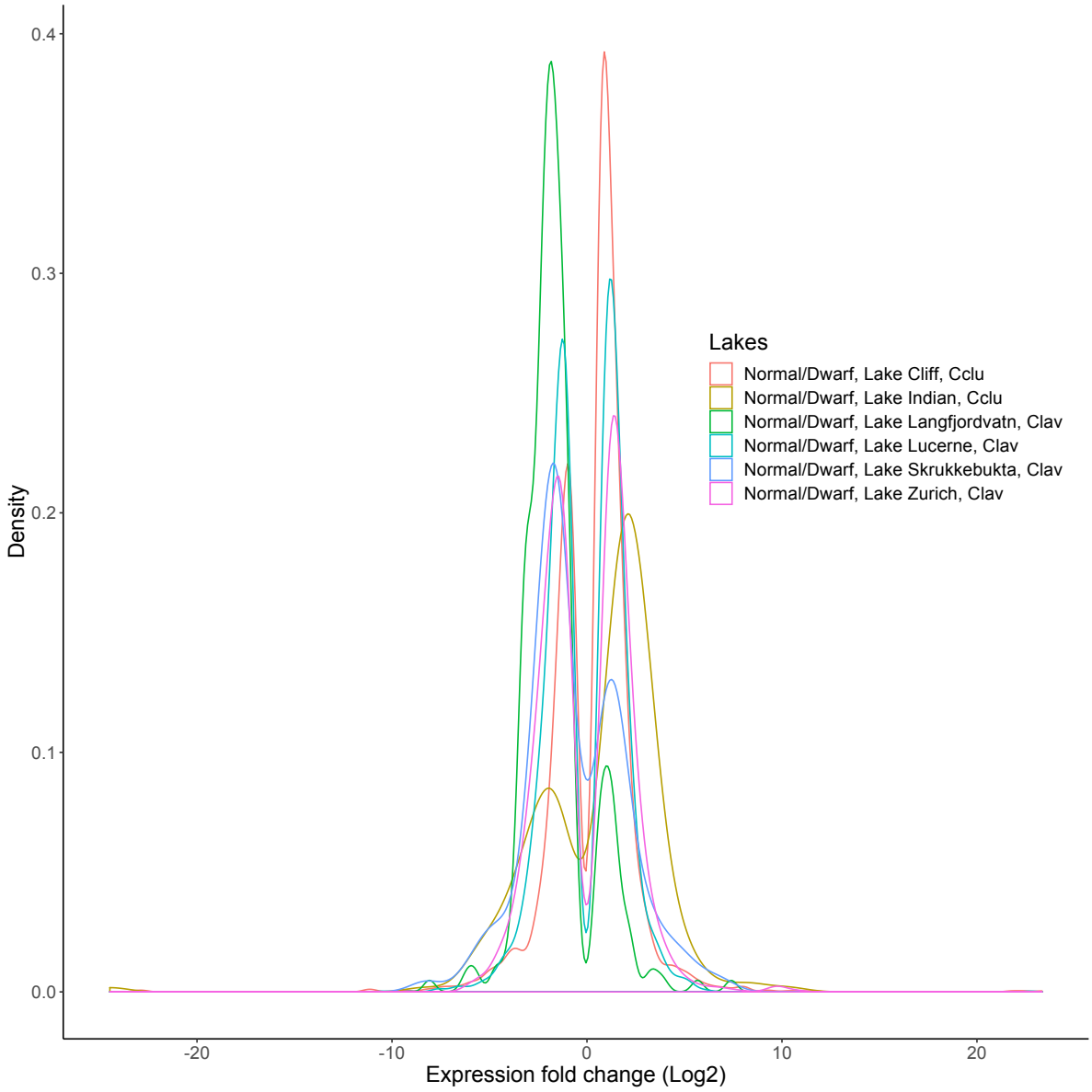


Figure S4.6: Expression fold changes for differentially expressed genes for intra-lake Limnetic/Benthic species comparisons. Each density plot is associated to a lake with the corresponding color. Positive and negative fold changes (log2) are associated to an over-expression and an under-expression, respectively, for genes in limnetic species relative to the benthic species.

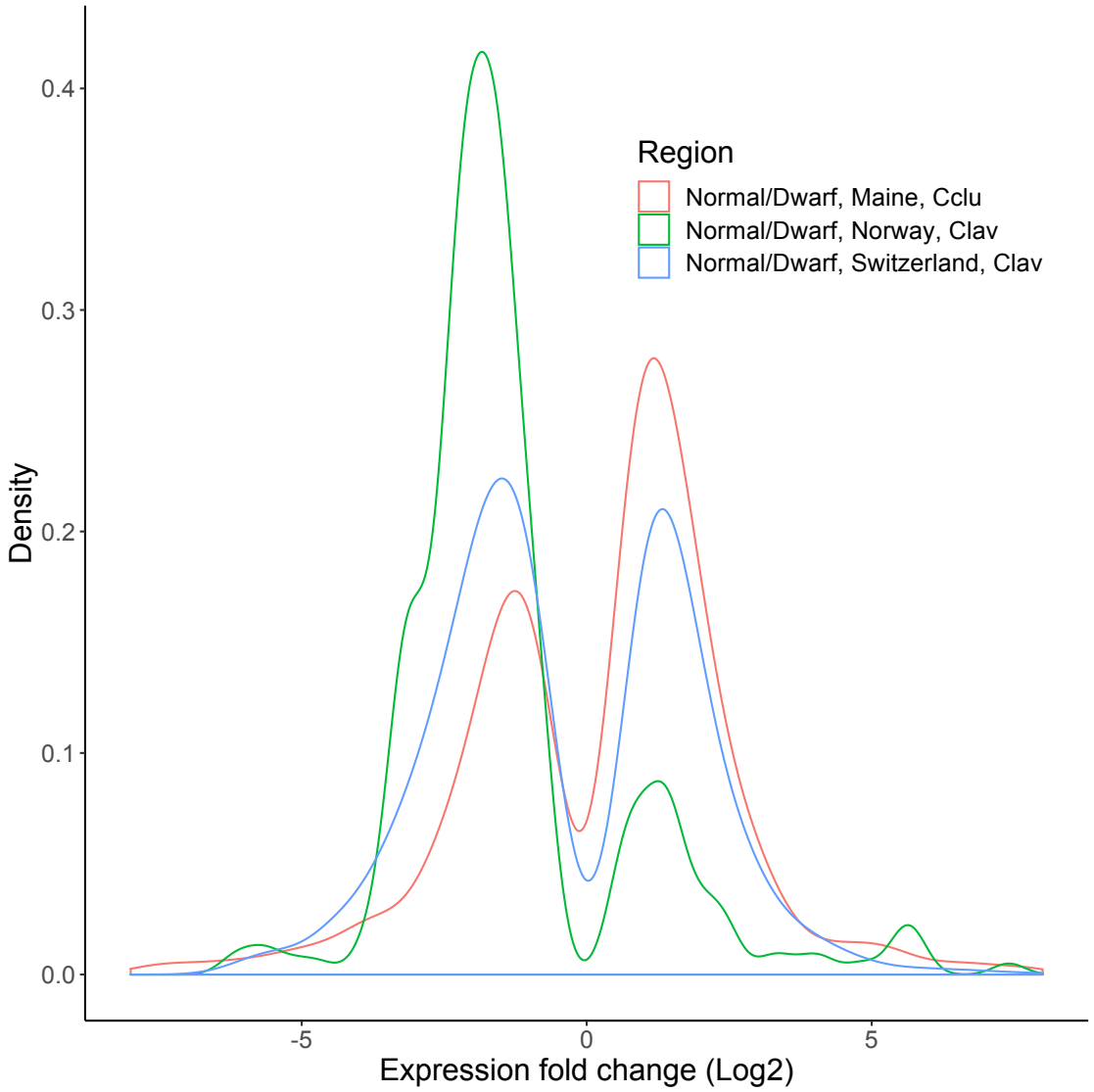


Figure S4.7: Expression fold changes for differentially expressed genes for intra-region Limnetic/Benthic species comparisons. Each density plot is associated to a region and composed by overlap between DEGs within region with the corresponding color. Positive and negative fold changes (log2) are associated to an over-expression and an under-expression, respectively, for genes in limnetic species relative to the benthic species.

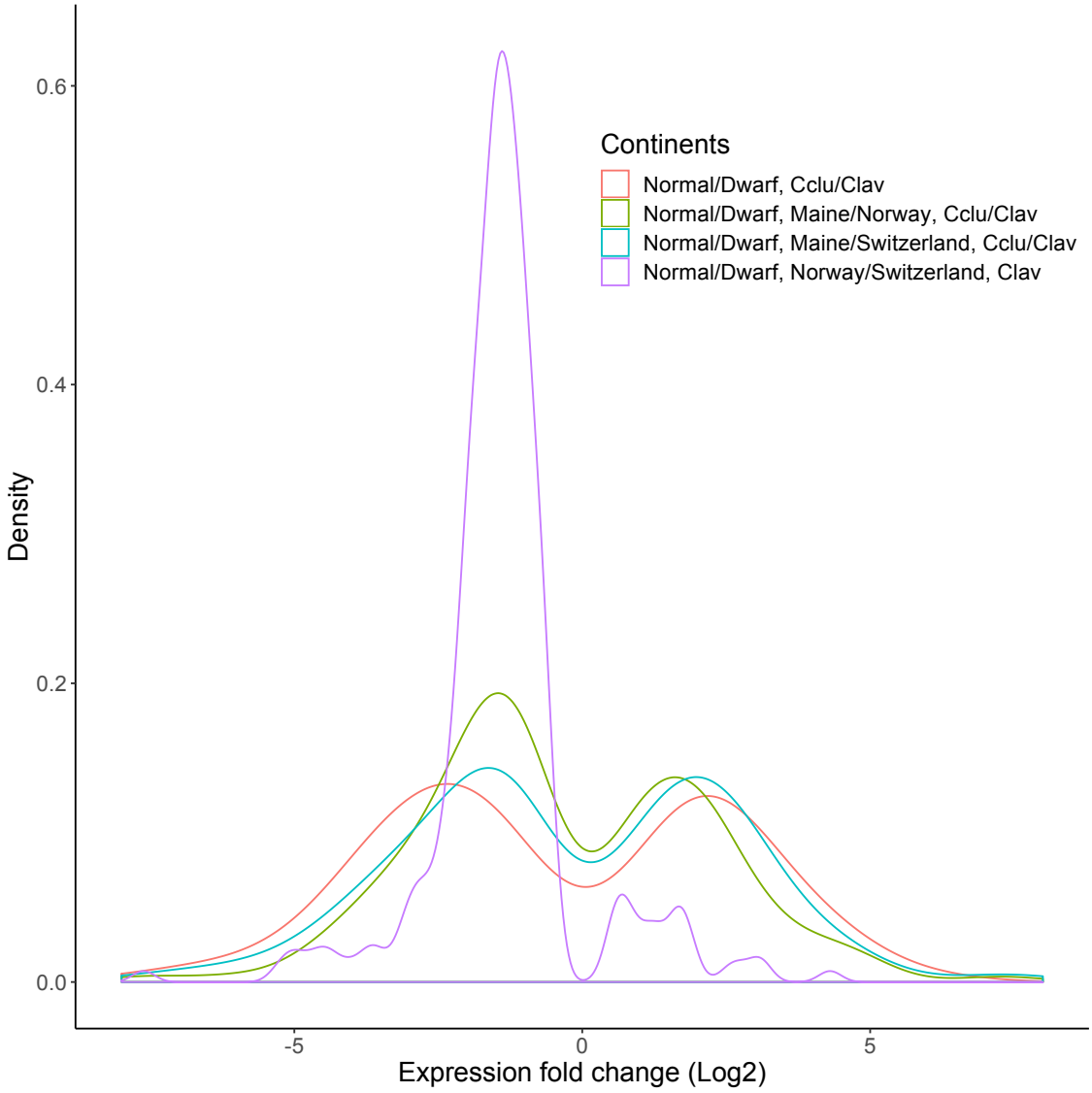


Figure S4.8: Expression fold changes for differentially expressed genes for across continents Limnetic/Benthic species comparisons. Each density plot is associated to a continental comparison and composed by overlap between DEGs among regions with the corresponding color. Positive and negative fold changes (log2) are associated to an over-expression and an under-expression, respectively, of genes in limnetic species relative to the benthic species.

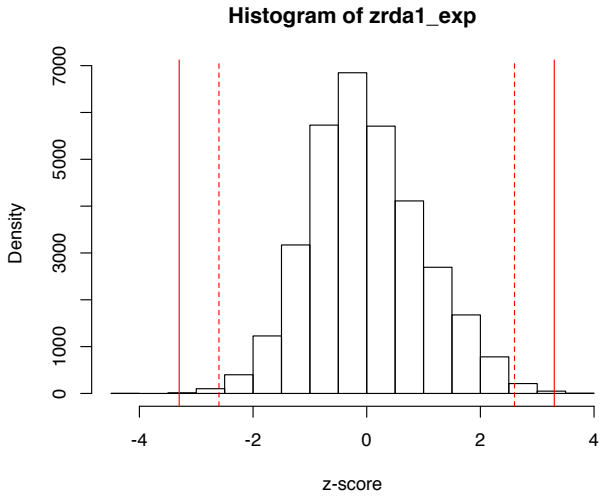


Figure S4.9: Transcripts z-score distribution from abundance transcripts data. Each transcript score obtained from cRDA analysis was transformed in z-score. Full and dashed red lines correspond to significance threshold of *p*-value of 0.01 and 0.001, respectively.

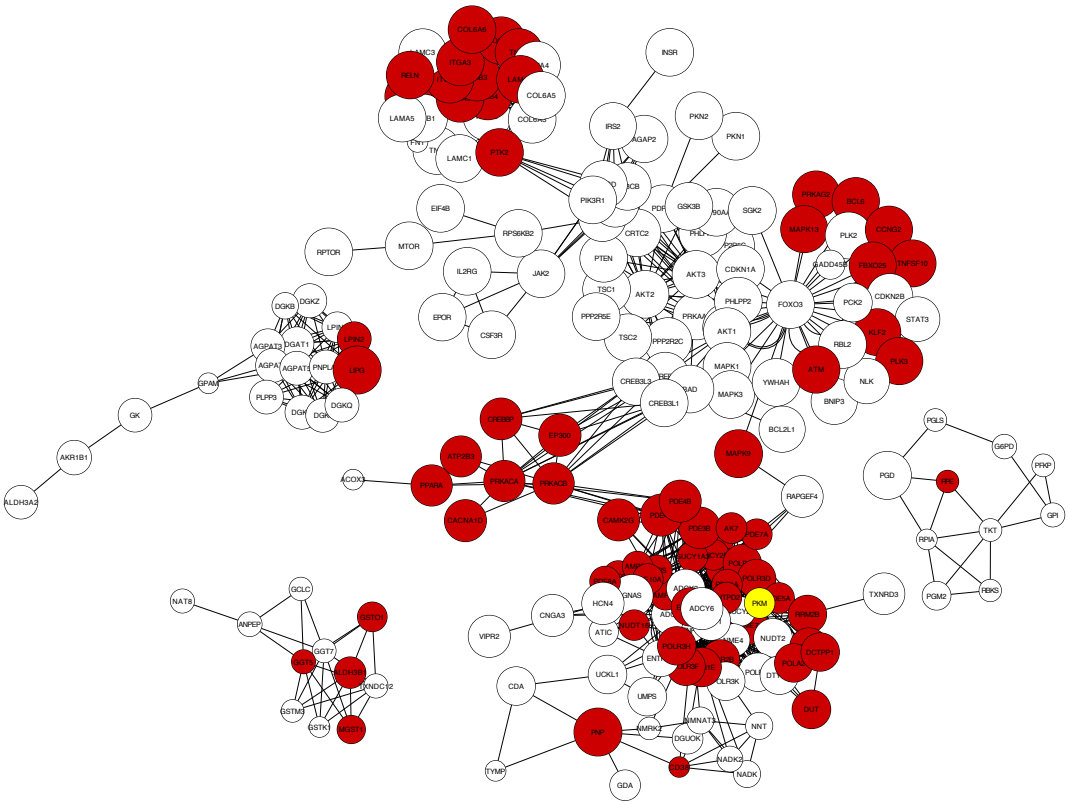


Figure S4.10: Significant subnetworks for Limnetic/Benthic species comparisons. The highest scoring genes within each metabolic pathway are shown in red, composing a subset of significant differentially expressed genes (DEGs), between species. Pyruvate kinase (PKM) gene expression (in yellow) is associated with a *cis*-eQTL in its 3'UTR. Gene identification is based on the Ensembl nomenclature. For each node, the relative size is proportional to the contribution score of the associated gene to the significance of the metabolic pathway. The score for each gene corresponds to a probability of convergent adaptation.

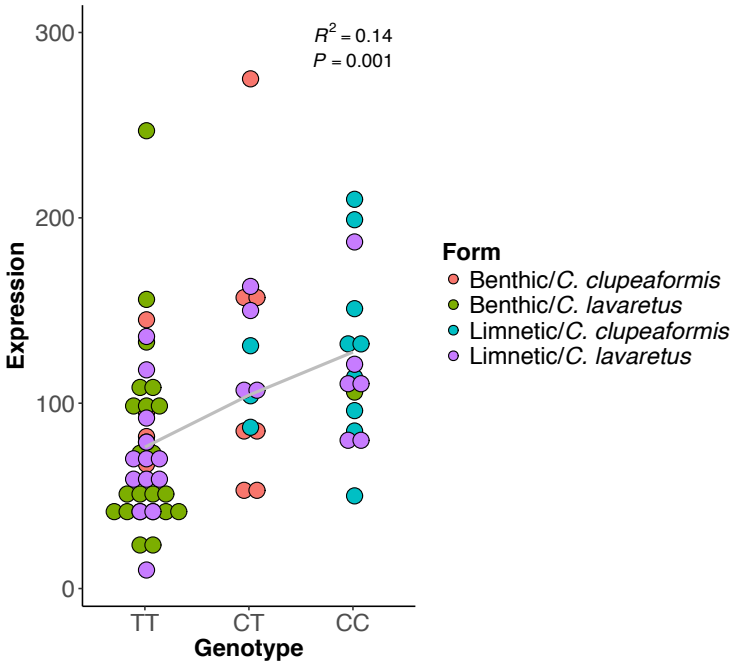


Figure S4.11: Association between a *cis*-eQTL genotypes and level of expression of the PKM gene between species. Transcripts abundance per individual (circles) as a function of genotypes for the *cis*-eQTL locus. The grey line corresponds to the linear model fitted to the data and showing significant differential expression between genotypes (*P*=0.001) represented species for each continent.

Chapitre 4: Genomic, epigenomic and transcriptomic differentiation in inter-continental parallel speciation of whitefish (*Coregonus sp*.) sympatric pairs.

5.1 Résumé

Des contrastes environnementaux similaires sont souvent à l'origine de la divergence adaptative parallèle dans les réplicats de la différenciation phénotypique parallèle. Un tel contexte est propice à l'identification de bases moléculaires impliquées dans des traits phénotypiques divergents. Outre les associations génotype-phénotype impliquées dans la divergence adaptative, la méthylation peut également affecter le niveau phénotypique via des modifications de l'expression génique. Dans cette étude, nous avons combiné les données génomiques, épigénomiques et transcriptomiques de 64 individus afin de quantifier le rôle relatif de la variation génétique et de la méthylation dans l'évolution parallèle de paires d'espèces de corégones limnétique et benthique, sur quatre paires sympatriques. Dans un premier temps, nous avons identifié 149 régions différentiellement méthylées parallèles (DMR) entre les espèces du système limnétique et benthique, qui étaient corrélés aux niveaux d'expression génique. 92 DMRs hyperméthylés chez les espèces limnétiques ont induit une répression de l'expression du gène auquel ils appartenaient, comparativement aux espèces benthiques, tandis que 57 DMRs hypométhylés ont coïncidé avec des gènes surexprimés. En outre, 108 variants génétiques parallèles (eQTL) ont affecté le niveau d'expression des gènes entre les espèces. Des analyses multivariées ont permis de dissocier la contribution relative de la génomique et de l'épigénomique sur la variation de l'expression génique entre espèces limnétiques et benthiques. La majeure partie de la variation de l'expression génique était expliquée par le génomique (4,1%) et l'intéraction entre les variations génomiques et épigénomiques (épigénomique «facilitée») (46,7%), tandis que l'épigénomique seulement (méthylation «pure») expliquait 2,3% de la variation de l'expression génique dans le système. Dans l'ensemble, les approches comparatives de plusieurs niveaux biologiques ont permis d'identifier les DMRs parallèles entre espèces, l'origine de la variation de la méthylation et de quantifier l'effet relatif de l'épigénétique sur la variation de l'expression des gènes. Cette étude fournit donc une rare documentation qualitative et quantitative sur le rôle de la méthylation de l'ADN au cours du processus de spéciation écologique.

5.2 Abstract

Similar environmental contrasts are often the source of parallel adaptive divergence in replicates of parallel phenotypic differentiation. Such context is propitious to identify molecular bases involved in diverging phenotypic traits. Besides genotype-phenotype associations involved in adaptive divergence, epigenetics may also affect phenotypic level via changes in gene expression. Here, we combined genomic, epigenomic and transcriptomic data from 64 individuals in order to quantify the relative role of genomic and methylation variation in the parallel evolution of limnetic-benthic whitefish species pairs across four sympatric pairs. We first found evidence for 149 parallel differentially methylated regions (DMRs) between species across the system, which modulated gene expression levels. 92 hypermethylated DMRs in limnetic species induced a repression of expression of the gene to which they belonged, relatively to benthic species, while 57 hypomethylated DMRs coincided with overexpressed genes. Furthermore, 108 parallel genetic variants (eQTL) affected gene expression between species. Multivariate analyses were used to partition the contribution of genomic and epigenomic to gene expression variation between species. Most of the gene expression variation was explained by genomic (4.1%) and interaction between genomic and epigenomic variations ('facilitated' epigenomic) (46.7%), while epigenomic only ('pure' methylation) explained 2.3% of the gene expression variation across the system. Overall, comparative approaches from several biological levels allowed identifying parallel DMRs between species, the origin of methylation variation and quantifying the relative effect epigenetic on gene expression variation. This study thus provides a rare qualitative and quantitative documentation of the role of DNA methylation during the process of ecological speciation.

5.3 Introduction

Understanding the origin and the processes associated with organisms' diversification remains a major goal of evolutionary biology. Populations' diversification is generally a consequence of stochastic events affecting independently the genetic pool of both diverging populations (*e.g.*, genetic drift inducing random genetic divergence through the time) (Song *et al.* 2006), and may include response to differential adaptation (Van Belleghem *et al.* 2018), eventually leading to ecological speciation. Indeed, adaptive divergence results in phenotypic responses caused by selective pressures stemming from different environments (Schluter 2001). Particular contexts allow identification of similar selection pressure in different populations, notably when independent parallel phenotypic evolution occurs in closely related and locally adapted nascent species (Orr 2005c; Elmer and Meyer 2011; Losos 2011; Conte, Arnegard, Peichel, and Schluter 2012b). Consequently, the emergence of new phenotypes due to changes in selective pressures, between parallel diverging populations, relies on the genomic bases, the environmental conditions and their interaction on the considered traits. Genetic variation inducing parallel phenotypic diversification may i) originates from independent *de novo* mutation (Pearce et al. 2009; Manceau et al. 2011; Conte, Arnegard, Peichel, and Schluter 2012b), ii) have a single origin and spread by gene flow (Schluter and Conte 2009; Abbott et al. 2013; Bierne et al. 2013), or iii) be independently recruited from standing genetic variation (Colosimo et al. 2004; Yeaman, Hodgins, et al. 2016; Rougeux et al. 2018). However, identical *de novo* mutations and gene flow are unlikely to occur between geographically isolated populations, and maintaining standing genetic variation have been show to increase the probability of genetic parallelism, particularly in the evolution of complex traits underlying a polygenic selection (Barrett and Schluter 2008; Schrider and Kern 2017). Moreover, phenotypic parallelism can be environmentally induced through phenotypic plasticity (*i.e.*, capacity of a genotype to produce several phenotypes in response to different environmental conditions; (Pigliucci *et al.* 2006), which may also allow diverging populations to cross valley of a changing adaptive landscape in order to reach a new phenotypic optimum (Pfennig *et al.* 2010). It is generally infered that such molecular mechanism underlying the modulation of phenotypic variation under changing environmental conditions is the results of epigenetic changes (Bossdorf *et al.* 2008).

Epigenetics refers to changes in gene function that are not associated to changes in gene sequence (Youngson and Whitelaw 2008). As such, epigenetics can modify the individual phenotype during development and/or in response to a changing environment without altering the DNA sequence (Weaver *et al.* 2004). Epigenetic variation is considered as i) 'obligatory' when relying completely on genetic variation, ii) 'facilitated' when indirectly potentiated by the genotype, and iii) as 'pure' when independent of genotypes (Richards 2006). Although all types of epigenetic variation can influence the level of gene expression, 'pure' epigenetic variants could allow exploring the adaptive landscape in changing conditions more quickly due to their elevated mutation rate and their response to environmental changes (Klironomos *et al.* 2013).

One of the most commonly studied epigenetic variation at the population level is DNA methylation, a biochemical modification of the DNA sequence adding a methyl group, generally to a cytosine within a CpG dinucleotides (Jones 2012). Level of DNA methylation in regulatory regions can influence the level of gene expression, with generally a negative correlation between DNA methylation level and gene expression (Jones 2012), while opposite pattern has been documented with an increase methylation in core genes (Feng *et al.* 2010). Consequently, if genes affected by methylation are associated to a phenotype under a selective pressure, methylation could contribute to generate an adaptive phenotypic response (Klironomos *et al.* 2013). Such fast plastic response may allow organisms to adjust their phenotype in a changing or a new environment. On the other hand, the loss of phenotypic plasticity may occur through genetic assimilation (*i.e.*, phenomenon by which an environmentally induced phenotypic character becomes acquired, with underlying genotypic bases, through the process of selection and eliminating the environmental component of this trait;(Waddington 1961). Since genetic assimilation ensures the stability of a phenotype, theory predict that the proportion of 'pure' epigenetic variation explaining phenotypic variation will decrease during the speciation process (Pfennig *et al.* 2010; de Villemereuil *et al.* 2018; Kawecki). Moreover, considering that environmental conditions may have a strong effect on methylation variation, parallel patterns of methylation should also be observed in independent related species inhabiting environments with similar selective pressure.

Here, we compared two sister species complexes of whitefish (Lake whitefish: *Coregonus clupeaformis* and European whitefish: *C. lavaretus*), both comprising benthic

and limnetic specialists (Lu & Bernatchez 1999; Praebel *et al.* 2013). The Lake whitefish and the European whitefish (hereafter, lineages) evolved independently on both continents since they became geographically isolated ~500,000 years ago (Bernatchez and Dodson 1991; Bernatchez and Dodson 1994; Jacobsen et al. 2012). The limnetic-benthic species complex evolved independently in several post-glacial lakes in the two lineages during the last Pleistocene. (Bernatchez and Dodson 1990; Bernatchez and Dodson 1991; Pigeon, Chouinard, and Bernatchez 1997b; Østbye et al. 2005; Rougeux et al. 2017). Limnetic and benthic species are partially reproductively isolated (Rogers and Bernatchez 2006). The limnetic species colonized the free limnetic ecological niche through an adaptive divergence with an associated phenotypic diversification, such as slower growth, slender body, higher metabolic rate and more active swimming behaviour (Bernatchez et al. 1999; Trudel et al. 2001; Rogers et al. 2002; Dalziel et al. 2015; Laporte, Rogers, Dion-Côté, Normandeau, Gagnaire, Dalziel, Chebib, and Bernatchez 2015a; Laporte, Dalziel, Martin, and Bernatchez 2016b). Finally, the phenotypic diversification is associated to polygenic control (Gagnaire, Pavey, et al. 2013; Laporte, Rogers, Dion-Côté, Normandeau, Gagnaire, Dalziel, Chebib, and Bernatchez 2015a; Rougeux et al. 2018). As such, this whitefish species complex offers a valuable model to study the molecular bases associated with parallel phenotypic differentiation.

The main goal of this study was to investigate the role of epigenetics in phenotypic diversification during independent speciation events. We first tested for non-random pattern of DNA methylation by assessing parallelism between the limnetic and benthic species from both North America and Europe. Combining transcriptomes and whole epigenomes resequencing data, we then investigated the effect of DNA methylation differentiation and eQTL on patterns of gene expression. Finally, we tested for the relative contribution of genetics and epigenetics components on patterns of gene expression in order to estimate the proportion of i) ‘pure’ genetic, ii) ‘obligatory’ and ‘facilitated’ epigenetic, and iii) ‘pure’ epigenetic explaining gene expression variation, in both species complexes.

5.4 Material and Methods

Sample collection

Our sample design involved the sampling of two lakes in North America (Cliff Lake and Indian Lake) and two lakes in Europe from Norway (Langfjordvatn Lake) and Switzerland (Zurich Lake), each comprising sympatric limnetic-benthic species pairs (Figure 5.1). Eight individuals per species per lake were sampled, for a total of 64 individuals DNA and RNA were extracted from liver tissue of each sample, stored in -80°C and in RNA later (All samples from Europe were stored in RNA later). Whole genomic DNA was isolated using the DNeasy Tissue Kit (Qiagen, Valencia, CA). DNA integrity was checked on an agarose gel (1%), and quantified with optic density measure (NanoDrop^{TM}2000, Thermo Scientific, Waltham, MA) before a quantitative-PCR measures. Individual libraries were built at the McGill University and Genome Quebec Innovation Centre (Montreal, Canada), for Illumina paired-end 150bp whole-genome bi-sulfite sequencing (WGBS) on the Illumina HiSeq4000 and HiSeqX, respectively. The 64 libraries for the WGBS were pooled on 16 lanes (two flow-cells, four individuals pooled per lane). From the eight individuals per population, we extracted the total RNA from liver tissue of six individuals (48 individuals total) as detailed in (Rougeux et al. 2018). Briefly, we choose the liver tissue for its homogeneous tissue characteristics and its multiples functions such as growth regulation in Salmonids (Rise et al. 2006), but also in energy metabolism and protein function which show heritable divergence in limnetic and benthic species (St-Cyr et al. 2008). RNA was extracted using the RNeasy Mini Kit following the manufacturer's instructions (Qiagen, Hilden, Germany). RNA quantity and quality were assessed using the NanoDrop^{TM}2000 (Thermo Scientific, Waltham, MA), and the 2100 Bioanalyser (Agilent, Santa Clara, CA, USA). Single read sequencing (100bp) was performed on the Illumina HiSeq 2000 platform for the 72 libraries (eight libraries per lane for a total of nine lanes), which included other populations used in (Rougeux et al. 2018), at the McGill University and Genome Quebec Innovation Centre (Montreal, Canada).

Gene expression analysis

Gene expression analysis was realized as in a previous study (Rougeux et al. 2018). However, we realized an original gene expression comparison based on a log-likelihood

ratio test, as detailed below. Raw sequences were cleaned from adaptor and tag sequences, and trimmed using Trimmomatic v0.36 (Bolger et al. 2014). Individual reads were mapped to the reference transcriptome using Bowtie2 v2.1.0 (Langmead and Salzberg 2012). eXpress v1.5.1 (Roberts and Pachter 2012) was used to estimate individual reads counts from BAM files. Then, differential expression analysis was realized with DESeq2 v1.14.1 (Love et al. 2014). Counts matrix was normalized using size factors and was log2-transformed. Considering the hierarchical structure of the studied system, we used a log-likelihood ratio test (LRT) approach in order to test for the effect of species (benthic species determined as reference), by building a full general linear model allowing species comparisons while integrating continents effect as covariate, and controlling for structuration with a reduced model. A significant threshold of a FDR < 0.1 was applied to determine differentially expressed genes (DEGs). Supplementary models testing for comparisons between continents and lakes only were realized in order to control for DEGs associated with environmental effects (e.g., lakes or continent).

SNP calling

Raw data from the 48 transcriptomes were cleaned and trimmed using cutadapt (v1.10) (Martin 2011). Cleaned reads were aligned to the indexed to the reference transcriptome (Rougeux et al. 2018) with Bowtie2 v2.1.0 (Langmead and Salzberg 2012). SAM files obtained were converted to BAM files, sorted using Samtools v1.3 (Li et al. 2009) and cleared from duplicates with the Picard-tools v1.119 program (http://broadinstitute.github.io/picard/). Reads alignments information to the reference transcriptome was used for genotyping SNPs with a minimum of quality of alignment of four and a minimum number of five reads to call SNPs with Freebayes v0.9.10-3-g47a713e (Garrison and Marth 2012). Vcffilter program from vcflib (Garrison2012) was used in order to keep variable sites with a minimum coverage of three reads per individual, biallelic SNPs with a phred scaled quality score above 30, a genotype quality with a phred score higher than 20. Then, we filtered the resulting VCF file using VCFtools (Danecek et al. 2011), in order to remove miscalled and low quality SNPs for subsequent population genomics analyses. For each of the eight populations, we kept loci with less than 10% of missing genotypes and filtered for Hardy-Weinberg disequilibrium using a p-value exclusion threshold of 0.05. Finally, we merged the VCF files from all eight populations, resulting in a unique VCF file containing 161,675 SNPs passing all the filters, in order to correct WGBS data (see below),

and a VCF file of 9,093 SNPs shared among all populations (trans-species polymorphism) and with no missing data for subsequent analysis.

Methylation calling and differential methylation analyses

Raw sequencing reads were trimmed for quality (≥25), error rate (threshold of 0.15) and adaptor sequence using trim_galore v0.4.5 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Trimmed sequences were aligned to the reference transcriptome using BSseeker2 v2.1.5 (Guo et al. 2013) with Bowtie2 v2.1.0 (Langmead and Salzberg 2012) in the end-to-end alignment mode. We cleaned the BAM files by removing duplicates with the Picard-tools v1.119 program (http://broadinstitute.github.io/picard/), before determining methylation levels for each site by using the BSseeker2 methylation call step. The raw methylation file was filtered by removing C-T DNA polymorphism identified from 'SNP calling' step, in order to avoid 'false' epigenetic variation at those positions (Le Luyer et al. 2017). We used the CGmapTools suite v0.1.1 (Guo et al. 2017) to extract only CpGs sites determined as CG context (avoiding CHH and CHG contexts), and with a minimum of 10X coverage and a maximum of 100X in order to avoid noise from repetitive elements and paralogs genes. Generalized linear model, with the hierarchical population structure as covariate, was used to identify differentially methylated loci (DMLs) and differentially methylated regions (DMRs) with the DSS R package (Wu et al. 2013), using a smoothing strategy on 500bp. We also controlled for interaction between terms allowing direct comparisons between species. Then, we compared directly all limnetic individuals to all benthic individuals across continents. DMLs were defined when showing at least 20% of difference between species and a significant threshold of $P < 0.05$. DMRs were retained when at least five CpGs occurred in a minimum sequence of 50bp, when CpGs showed a minimum of 10% of methylation difference between species and a significant threshold of $P < 0.05$. DMRs distant by 50bp or less were merged together to define a same DMR.

Gene ontology analyses

Gene ontology (GO) enrichment analyses were performed on gene expression and methylation with GOATOOLS (Tang, Lyons, et al. 2015). We tested significant DEGs and DMRs using Fisher's exact tests and GO enrichment were associated with FDR < 0.05

(Benjamini-Hochberg correction) and we kept GO categories represented by at least three genes.

Effects of genetic and DNA methylation on gene expression

Gene expression level was related to sequence polymorphism to identify eQTLs. We used the R package MatrixEQTL v2.1.1 (Shabalin 2012) to perform association mapping for local eQTL affecting the expression level of the transcripts to which they were directly physically linked (cis-eQTL). From the 9,093 shared SNPs among all eight populations, we retained loci showing polymorphism across continents, which corresponded to 5,424 SNPs. eQTL were identified through linear model in order to identify differential expression between species, while considering lakes and continents as covariates. A Bonferroni correction was applied to correct for multiple testing, and significance of identified cis-eQTL was accepted with a FDR<0.01.

Then, we report the gene expression level of genes affected by a parallel DMR. Considering parallel DMRs, only one DMR per gene were identified. In the case of identified DMRs, we directly associated the level of differential methylation between species to the gene expression differentiation level (Log2 Fold Change) to which the DMR belongs. This approach allowed defining genes with and without DMRs and cis-eQTL.

Variance of gene expression explained by genomic and epigenomic variation

In order to We used redundancy analyses (RDAs) to estimate the percentage of gene expression variation explained by genomic variation alone, ‘pure’ epigenomic variation and both molecular components combined within ‘obligatory’ and ‘facilitated’ epigenomic. Then, partial RDAs (pRDAs) were applied to quantifying the proportion of variance in gene expression explained by genomic variation, when controlling for epigenomic variation, and vice versa. Both RDAs and pRDAs were performed using the function ‘rda’ available in the Vegan R package (Jari Oksanen et al. 2018). Previous to RDAs and pRDAs analysis, principal components analyses (PCA) on gene expression, genomic and epigenomic matrix were performed, with Vegan. Principal component (PCs) factors from the PCA were used as the measure of gene expression, genomic and epigenomic variation. Only factors explaining at least 2.0% of the variation were kept. Selection of the best genomic and epigenomic PCs explaining gene expression variation was performed with backward

selection, using the function ‘ordistep’ of Vegan. Finally, variation partitioning was computed from RDAs and pRDAs with the function ‘varpart’ (same R package). Thus, variation of gene expression explained by i) genomic variation after controlling for epigenomic was consider as ‘pure’ genetic effect, ii) epigenomic variation after controlling for genomic as ‘pure’ epigenetic effect and iii) both genomic and epigenomic variation as a combination of ‘obligatory’ and ‘facilitated’ epigenetic effect. This procedure was repeated for comparisons involving limnetic and benthic species across continents and between limnetic and benthic species within each continents.

5.5 Results

Parallel differential gene expression between benthic and limnetic species

Liver tissue RNAseq generated a total of 1.15 x 10^9 100bp raw single-end reads from 48 individuals (Table S5.1). Filtered libraries (1.13 x 10^9 reads) were aligned to the transcriptome reference to quantify the amount of reads per transcript per individuals. Using the benthic species populations as reference in conditions (i.e., benthic and limnetic) comparisons, we quantified differentially expressed genes (DEG) between conditions across continents using a generalized linear model, including lake, continent and species information as covariates. We found 156 parallel DEGs between limnetic and benthic species between both continents. These parallel genes showed equal proportions of down- and up-regulated genes in limnetic species relatively to the benthic species (72 vs. 84; χ^2 = 1.07, df = 1, P=0.29).

Gene ontology enrichment analysis revealed an overrepresentation of modules associated with metabolic processes (GO:0006807, nitrogen compound metabolic process; GO:0044237, cellular metabolic process; GO:0044238, primary metabolic process; GO:0071704, organic substance metabolic process; GO:0008152, metabolic process), immune system response and antioxidant activity (GO:0016209, antioxidant activity; GO:0003823, antigen binding) and methylation (GO:0032259, methylation) in limnetic species (Table S5.2).

Differential methylation between benthic and limnetic species pairs

A total of 10.9 x 10^9 150bp paired-end reads were generated from 64 individuals for the WGBS. Such sequencing efficiency corresponds to an average of 13.6X coverage (s.d. 1.6X) per individual (Table S5.3).

From the methylation call filtered for C-T genetic polymorphism and CpGs corresponding to CGs context, the umber of CpGs per population were as follows for limnetic: Cliff (864,143 s.d. 399,951, due to one individual with low sequence coverage), Indian (1,001,147 s.d. 95,639), Langfjordvatn (914,369 s.d. 78,472), Zurich (952,032 s.d. 75,774) and for benthic: Cliff (995,605 s.d. 64,301), Indian (1,012,432 s.d. 39,255), Langfjordvatn (883,485 s.d. 89648), Zurich (981,415 s.d. 81,630).

The number of differentially methylated loci (DMLs) between species averaged at 17,921 DMLs. More precisely, we found 21,843 and 15,878 DMLs between limnetic and benthic species from Cliff Lake and Indian Lake, and 16,593 and 17,370 DMLs between species from Langfjordvatn Lake and Zurich Lake. The CpGs, comparisons between limnetic and benthic species for each lake allowed identifying a mean number of 619 differentially methylated regions (DMRs) across all lakes. Thus, we found 748, 532, 575 and 619 DMRs in Cliff, Indian, langfjordvatn and Zurich lakes, respectively.

Testing for parallel methylation between limnetic and benthic species

We then tested for the occurrence of parallel differential methylation between all limnetic and all benthic fish across both continents. Generalized linear model identified 9,819 significant common DMLs between all populations of the limnetic species and all populations of the benthic species. We also found 149 parallel DMRs distributed among 147 genes between all populations of the limnetic species and all populations of the benthic species. From the 149 DMRs, approximately twice as many DMRs were hypermethylated in the limnetic species relatively to the benthic species. Indeed, 92 DMRs showed patterns of hypermethylation, while 57 DMRs were hypomethylated (57 vs. 92; $\chi^2 = 8.22$, df = 1, P=0.004). Such observations suggest a non-random epigenetic pattern underlying phenotypic differentiation in those species complexes.

We tested for gene ontology enrichment analysis in genes associated to the 149 parallel DMRs (Table S5.4). We found that the 92 hypermethylated DMRs in the limnetic species were associated with an overrepresentation of biological process linked to growth and developmental process through several functions (GO:0048856, anatomical structure development; GO:0032502, developmental process; GO:0044699, single-organism process; GO:0009987, cellular process). We also observed an enrichment in genes associated to variable rate of cell cycle process (GO:0022402 BP, cell cycle process) in the 57 hypermethylated DMRs in the benthic species.

DMRs are associated with gene expression repression

The 57 hypomethylated DMRs in the limnetic species were associated with a positive log2 Fold change, corresponding to an overexpression of genes in limnetic species relative to the benthic species (Figure 5.2A). Moreover, hypermethylated genes in limnetic

species were associated with a repression of the gene expression in limnetic species relative to the benthic species (Figure 5.2A). This difference in gene expression between hyper- vs. hypomethylated DMRs was highly significant (Wilcoxon test, $W = 4577$, $P < 0.001$). Finally, we found that the association between DNA methylation and transcriptional repression in general was stronger for more significant DMRs (Figure S3.1).

eQTL effect on differential gene expression between species

We contrasted the variance in gene expression between species in genes to which a cis-eQTL was associated and in genes without cis-eQTL. We found 108 significant (FDR < 0.01) cis-eQTL between all limnetic and all benthic whitefish from both continents and no DMR was observed in those 108 genes with cis-eQTL, which therefore represents a measure on genetic effects only. We noted that the genes without cis-eQTL showed a pronounced variance in gene expression compared to those with a cis- eQTL. We also observed a highly significant difference between both genes categories (cis- vs. no cis-QTL) for genes that were repressed in the limnetic species relative to the benthic species (i.e., negative Log2 fold change) (F test, $F = 0.37232$, num df = 49, $P < 0.001$), whereas this was not the case for genes that were overexpressed in limnetic whitefish (F test, $F = 0.79361$, num df = 57, $P = 0.26$) (Figure 5.2B). However, the difference in levels of gene expression between limnetic and benthic species was more pronounced in genes with eQTL compared to genes without eQTL, both in repressed (Wilcoxon test, $W = 323420$, $P < 0.001$), and overexpressed (Wilcoxon test, $W = 647930$, $P < 0.001$) genes in limnetic whitefish.

Relative genetic and epigenomics effects on patterns of gene expression

The comparison of all limnetic and all benthic species across the two continents ('All dataset') allowed identifying seven PCs axes which individually explained more than 2.0% of the variation, and together explaining 64.7% of the total observed variance in gene expression. Following backward selections on PCs foactors explaining more than 2.0% of genomic and epigenomic variation, PCs foactors 1, 2, 11 and 14 were selected for genomic data (representing a total of 40.9% of the genomic variance), while PCs axes 1 and 2 were selected for epigenomic data (representing a total of 25.6% of epigenomic variance). Interestingly, gene expression, genomic and epigenomic PCAs allowed separating individuals per continents and clustering lake populations at global scale ('All dataset'

column), but also at continental scale ('America' and 'Europe' columns) (Figure 5.3 – panels of rows 2 to 4). Variance partitioning for limnetic and benthic species across continents showed that RDA including genomic and epigenomic factors explained 53.1% of gene expression variation ($P < 0.001$; Figure 5.3 – top left panel). Decomposing the factors, genomic factors alone explained 50.8% of genes expression ($P < 0.001$) while epigenomic factors alone explained 49.0% ($P < 0.001$). Genomic factors explained 4.1% of genes expression ($P = 0.042$) when controlling for epigenomic ('pure' genetic control), while epigenomic factors explained 2.3% of gene expression ($P = 0.063$) when controlling for genomic ('pure' epigenetic control), and thus, 46.7% of genes expression variation was explained by the combined effect of both genomic and epigenomic factors ('obligatory' and 'facilitated' epigenetic control).

The same analyses were also performed on each continent separately. In North America, 16 PCs axes individually explaining more than 2.0% of the variation were retained for gene expression data (representing a total of 94.6% of the variance in gene expression). PCs factors 1, 2 and 3 for genomic data (representing a total of 33.0% of the genomic variance), and PCs factors 1, 2, 4, 9 and 19 for epigenomic data (representing a total of 29.6% of the epigenomic variance) explained more than 2.0% of the variation and consequently were selected by backward selection. Partitioning the variance in North American limnetic and benthic whitefish allowed quantifying the variance in gene expression (31.3%, $P < 0.001$) explained by both genomic and epigenomic factors (Figure 5.3 – top central panel), whereas genomic and epigenomic factors alone explained 29.5% ($P < 0.001$) and 23.6% ($P < 0.001$), respectively. After controlling for epigenomic and genomic on each other, genomic factors ('pure' genetic control) explained 7.8% ($P = 0.023$), while epigenomic factors ('pure' epigenetic control) did not explain a significant proportion variance in gene expression. However, 21.8% of the variance in gene expression was explained by the interacting effect of genomic and epigenomic factors ('obligatory' and 'facilitated' epigenetic control). In Europe, we retained 12 PCs axes that explained more than 2.0% variance in gene expression (explaining together 83.4% of the variance in gene expression). PCs axes 1, 10, 13 and 22 for genomic data (29.3% of the genomic variance), and PCs axes 1, 17 and 19 for epigenomic data (16.1% of epigenomic variance) were identified by backward selection. Variance in the European limnetic and benthic whitefish revealed that RDA including genomic and epigenomic factors explained 51.4% of the

variance in gene expression ($P < 0.001$; Figure 5.3 – top right panel), while genomic and epigenomic factors alone explained 51.2% ($P < 0.001$) and 43.3% ($P < 0.001$) of the gene expression variance, respectively. After controlling for epigenomic and genomic on each other, genomic factors ('pure' genetic control) explained 8.1% ($P = 0.023$), while epigenomic factors ('pure' epigenetic control) did not explain a significant proportion variance in gene expression. However, 43.1% of the variance in gene expression was explained by the interacting effect of genomic and epigenomic factors ('obligatory' and 'facilitated' epigenetic control).

5.6 Discussion

Our study focus on two sister lineages of whitefish that diverged and evolved independently for about ~500,000 years (Bernatchez and Dodson 1994) and have postglacially colonised cold freshwater lakes of Eurasia and North America since then (Bernatchez and Dodson 1991; Douglas et al. 1999; Østbye et al. 2005; Hudson et al. 2010). In both lineages, diverging limnetic and benthic sympatric species pairs are associated with their respective trophic niches (Kahilainen and Østbye 2006; Landry and Bernatchez 2010; Häkli et al. 2018). The sympatric parallel phenotypic diversification occurred in several lakes from the secondary contact of diverged glacial lineages (*i.e.*, sub-lineages evolving independently during the last glaciation event (Rougeux *et al.* 2017; Rougeux *et al.* unpubl.). In this study, whole transcriptome and epigenome resequenced data offered the opportunity to analyze both the levels of DNA methylation and gene sequence divergence and combined these to the analysis of differential gene expression on representative sympatric pairs of limnetic-benthic whitefish species on the two continents. This allowed questions about the relative importance of different molecular mechanisms associated with the phenotypic differentiation in a context of parallel ecological speciation.

Origin of parallel differential methylation

The main result stemming from the comparisons of DNA methylation levels between limnetic and benthic species is that all sympatric species pairs exhibit significant differentially methylated regions (DMRs) when found in same environment (lake). Then comparing populations belonging to the limnetic and benthic species all analysed with the same tissue (liver) allowed identifying, i) several identical differentially methylated loci (DMLs) between limnetic and benthic whitefish from different sympatric species pairs within continents; ii) 9,819 shared DMLs between all populations of the limnetic species compared to the benthic species across continents, and iii) DMRs between limnetic and benthic whitefish form different sympatric species pairs and also 149 shared DMRs across continents between all limnetic *vs.* all benthic whitefish (parallel DMRs). Indeed, the amount of DMLs between sympatric limnetic and benthic whitefish was comparable among lakes, although a greater absolute number of DMLs was observed in Cliff Lake (North America; 21,843 DMLs) and Zurich Lake (Switzerland; 17,370 DMLs), where sympatric species pairs are the most genetically and phenotypically differentiated of those comparisons (Rougeux

et al. 2018). In addition, the distribution of DMLs across the transcriptome allowed reconstituting DMRs. Again, these two most differentiated lakes showed a higher number of DMRs than the other lakes (Cliff Lake: 748 DMRs and Zurich lake: 619 DMRs). Our previous detailed transcriptomic study on these same lakes also showed that the number of differentially expressed genes between limnetic and benthic whitefish was pronounced in these two lakes as well (Rougeux *et al.* 2018). Moreover, the DMLs we found between limnetic and benthic whitefish of different sympatric species pairs allowed identifying parallel DMLs and parallel DMRs between all limnetic and benthic whitefish across both continents. Together, this suggests that epigenetic differentiation is at least partly influenced by a selective pressures associated with the use of different ecological niches by limnetic and benthic whitefish (Schluter 2001; Nosil 2012). Such selective pressure could be: i) related to ensure the genome integrity by silencing transposable elements (TEs) through constitutive methylation (Rey *et al.* 2016), ii) directly associated to the modulation of the gene expression by genetically induced methylation (Richards 2006), and iii) a direct response to local ecological conditions (Duncan et al. 2014; Metzger and Schulte 2017). Despite the fact that our dataset could not distinguish between those potential selective pressures, the parallel pattern observed in this study suggests a role of epigenetic in the process of adaptive divergence, and ultimately ecological speciation of limnetic and benthic whitefish.

Functional effects of DMRs

In addition to the extant of parallelism we observed among DMRs, we were interested to test whether DMRs had and effect on the level of gene transcription and if so, what biological functions were involved. We identified DEGs significantly enriched in metabolic processes and immune system. This corroborates previous studies on North American whitefish species pairs in which the authors proposed that the differences in both metabolic and immune functions between limnetic and benthic whitefish may reflect life history trade-offs whereby a more active swimming in order to avoid predation, increase foraging efficiency would induce a more parasites load and consequently, energetic costs that translate into slower growth for example (Lu and Bernatchez 1999; St-Cyr et al. 2008; Landry and Bernatchez 2010; Dalziel et al. 2015; Laporte, Dalziel, Martin, and Bernatchez 2016b; Rougeux et al. 2018). DEGs were also enriched in genes associated with methylation regulation, suggesting a genetic role of DNA methylation. More precisely, the

enrichment for methylation regulation was associated to the process of rRNA 2'-O-methylation (rRNA2'-O-me), a highly complex and specific posttranscriptional modification present in functionally important domains of the ribosome (Krogh et al. 2016). Such mechanism within the ribosome is particularly important in the regulation of the gene expression (Schwanhäusser et al. 2011). Moreover, some functional domains of ribosomes has been showed to be targeted by the rRNA2'-O-me, inducing a modulation of the translation of mRNAs through plastic response, corresponding to a new level of regulation of gene expression (Erales et al. 2017).

We then investigated the possible association between differential patterns of gene expression with the extant of methylation for those genes that showed parallel DMRs between all limnetic vs. all benthic whitefish. Most of the parallel hypomethylated DMRs identified were associated to overexpressed genes in the limnetic species compared to the benthic species. Moreover, parallel hypermethylated DMRs in the limnetic species were associated with a repression of the gene expression. These results support the correlative effect of the DNA methylation level on the regulation of the transcriptional activity of linked genes, where hypermethylation is typically associated with gene expression repression (Jones 1999; Jones 2012). In addition, the most significant DMRs were directly associated with a higher proportion of repressed genes with a repressed transcriptional activity. Such observations underlie the direct effect of the DMRs size (i.e., length), composition (i.e., number of DMLs) and the amount of methylation differences along the DMR between conditions. Identical results were observed during identification of DMRs using statistical correction (Korthauer and Irizarry 2018), instead of effect-size cut-offs which could diminished the resolution of the study (Ford et al. 2017). Then, integrating the individual characteristics of CpGs allows identifying robust DMRs in methylation studies (Korthauer and Irizarry 2018).

The gene ontology analysis of genes with parallel DMRs revealed an enrichment of gene modules involved in growth and developmental processes, and more so for hypermethylated DMRs in the limnetic species, but also genes module regulating cells cycle process for hypomethylated DMRs in the limnetic species. These results suggest a repression of gene expression of genes related to growth and developmental process in the limnetic species while genes associated with a higher metabolism and shorter cell cycle life are over-expressed in the limnetic species relatively to the benthic species. Again, these

observation at the epigenetic level corroborates those of previous studies at the transcriptomic and physiological levels and supported the hypothesis of life history trade-offs between survival (limnetic) and growth (benthic) functions (Derome et al. 2006; St-Cyr et al. 2008; Renaut et al. 2009; Jacobs et al. 2018; Rougeux et al. 2018). On the other hand, none of the parallel DMRs belong to a DEG, suggesting that methylation differentiation is not a major mechanism involved in the higher differential of gene expression in this species divergence. Similar observations have been made in the stiff brome (Brachypodium distachyon), to which CpGs and DMRs and gene expression association were not systematically associated across tissues comparisons (Roessler et al. 2016), as well as in hepatocellular carcinoma cells were differentially expressed genes were not correlated to higher DNA methylation differences (Sun et al. 2018). In addition, it has been showed that the evaluation of the relationship between methylation level and gene expression, considering the log fold change is a more relevant signal of change in gene expression than the absolute differences (Goentoro and Kirschner 2009).

Decoupling epigenetic from genetic control on gene expression

We used multivariate statistical framework to disentangle the relative role of genomic and epigenomic variation on patterns of gene expression. Combined data set (*i.e.*, genomic and epigenomic) explained 53.1% of the gene expression variation between species across continents, with 4.1% attributable to 'pure' genetic, 2.3% to pure epigenomic and 46.7% to the interaction between both levels. Thus, the observed variance in gene expression variation was influenced by higher variation at the genomic than the epigenomic level. Similar observations were realised in a human study, interested in the inheritance of CpG methylation sites among family cohort. They quantified about 3% of CpGs to which the genomic bases were not identified suggesting environmental effect on this CpGs, or missing correspondence with genomic bases (Zaghlool *et al.* 2016). This suggests that the origin and the maintenance of epigenomic variation could itself have a genetic basis and consequently be defined as 'obligatory' or 'facilitated' methylation (Richards 2006). Theory predicts that pure epigenetic variation could allow selection on newly induced phenotypes (Klironomos *et al.* 2013). Yet, our study, along with previous empirical ones mainly pointed to a marginal contribution of 'pure' epigenetic compared to the contribution of genetic variants. Indeed, testing for environmentally-induced methylation in *Arabidopsis thaliana* coupled with and extensive genome wide association study (GWAS) revealed that most of

the methylation variation was directly associated to changes in locally adapted allele frequencies, and so no 'pure' epigenetic was observed (Dubin *et al.* 2015). In addition, similar approach to our was realized on human populations and revealed strong methylation modifications between populations, with a strong genetic basis and a lower contribution of genetic and epigenetic variation on gene expression (Moen *et al.* 2013). Nevertheless, 46.9% of gene expression was not associated to methylation and/or genomic variation, which suggest that other mechanisms associated to genes expression regulation such as histone structure, which may be induced before methylation change, that serve as more stable epigenetic modifications (Park *et al.* 2008), or small RNAs (si- and piRNAs) that are mainly associated to chromatin remodelling and so maintaining genes and/or TEs in a poised state (Fagegaltier *et al.* 2009), are still to be investigate to better understand the role of epigenetic in speciation.

The comparison between limnetic and benthic whitefish within each continent showed similar tendencies as those observed across both continents. That is most of the variation was explained by 'pure' genomic variants (significant for both continents) and interaction between genomics and epigenomics, whereas 'pure' epigenomic only did not explain a significant part of the variance in gene. These results also support the hypothesis that mainly 'obligatory' and/or 'facilitated' epigenetic variants are associated with controlled the level of gene expression. Finally, admitting the exploration of the adaptive landscape at the early stage of phenotypic differentiation has been associated to 'pure' epigenomic variation in an amphibian species (*Scaphiopus couchii*), which evolved a novel ectomorph that likely arisen through a 'plasticity-first' scenario (Ledon-Rettig et al. 2008; Ledon-Rettig et al. 2010). However, such mechanism would be eroded through genetic assimilation (Ehrenreich and Pfennig 2016; Nishikawa and Kinjo 2018), in nascent species such as the whitefish complex.

Another salient observation supporting the prevailing role of genetic over epigenetic variation is that differences in gene expression between limnetic and benthic whitefish was more important for genes for which we identified a *cis*-eQTL than those without *cis*-eQTL. This pattern emerged for overexpressed and repressed genes in the limnetic species relative to the benthic species, and both gene categories with *cis*-eQTL showed les gene expression variance than those without *cis*-eQTL. This observation also suggests that selective pressures act on standing genetic variation and that favoured alleles could

increase in frequency. This was previously describe in the whitefish system (Rougeux *et al.* 2018), but also in several systems. For example, *cis*-variation in a gene promoter could alter the binding site of a regulatory protein and induce newly derived phenotype, as observed in the morphological evolution of two *Drosophila* species (Nagy *et al.* 2018), in the threespine stickleback with a genetic incidence in the EDA gene associated with armor plates (O'Brown *et al.* 2015), or in corals under changing climatic conditions were an important number of favoured alleles frequencies co-vary with gene expression level (Rose *et al.* 2018). Moreover, genes affected by *cis*-eQTL showed more gene expression differentiation between limnetic and benthic species than genes with parallel DMRs between limnetic and benthic whitefish, and more so for overexpressed genes in the limnetic species relatively to the benthic whitefish. Furthermore, despite the effect of methylation on their expression, genes associated with parallel DMRs are not associated with most differentially expressed genes whereas those associated with *cis*- genetic variants are (Jeukens and Bernatchez 2011; Rougeux et al. 2018). While this suggest that variation in gene expression could result from direct changes in the level of DNA methylation to respond to environmental conditions (*e.g.*, 'pure' epigenetic variation) (Richards 2006), such mechanism is not clearly supported by our results whereby we observed that most of the methylation variants are 'obligatory' or 'facilitated' epigenetic modifications. Hypothetically, DMRs could be in linkage disequilibrium with *trans*-acting genetic variant affecting the DNA methylation level of the gene where the DMR is localized, although our data does not allow us to test for this possibility (Taudt *et al.* 2016).

Epigenetics and genomic architecture of adaptive traits

As mentioned above, our analyses revealed that most of the methylation variants associated with variation in gene expression originates from 'obligatory' and 'facilitated' methylation. Moreover, it is noteworthy that previous experimental in controlled conditions along with GWAS studies suggested that phenotypic differences between limnetic and benthic whitefish has a strong genetic basis with little evidence for phenotypic plasticity (Rogers and Bernatchez 2007a; Laporte, Rogers, Dion-Côté, Normandeau, Gagnaire, Dalziel, Chebib, and Bernatchez 2015a; Laporte, Dalziel, Martin, and Bernatchez 2016b). Then, our observations suggest i) that 'pure' epigenetic has never been involved in the process of phenotypic diversification between limnetic and benthic whitefish, or ii) that 'pure' epigenetic was originally present, and initially involved in the process of phenotypic

diversification before being progressively genetically assimilated. It is important to note that both hypotheses cannot be supported with our present results. However, according to the genetic assimilation theory, novel phenotypes could results from phenotypic plasticity (Pfennig *et al.* 2010), and ultimately contribute to speciation if plastic phenotypes are constantly induced by stable environmental selective pressure over the time, facilitating ecologically driven phenotypic diversification and eventually reproductive isolation (West-Eberhard 1989; West-Eberhard 2005), particularly when a magic traits (*e.g.*, mate choice) is involved (Irwin and Price 1999). Of course, it is implied that the potential for phenotypic plasticity of the population allows such response over the time, and that not all individuals have the same potential to avoid and evolutionary dead-end (Robinson 2013). Some empirical and experimental studies ask whether such mechanism could be involved i) in colonization of empty ecological niche such as for the threespine stickleback where comparisons of marine and freshwater populations revealed a loss of plasticity through putative genetic assimilation (McCairns and Bernatchez 2010), and in some birds to which plumage coloration can be diet-dependent and may finally get fixed by genetic assimilation as response to selection (Badyaev *et al.* 2017); ii) in phenotypic diversification as observed in cichlids. Indeed, experimental works on cichlids showed that the initial stages of the adaptive radiation of lineages of East African cichlid fishes could be eased by phenotypic plasticity (Muschick *et al.* 2011; Gunter *et al.* 2017). Moreover, comparisons between several lineages reared in and expose to the same experimental conditions found flat reaction norm in derived lineages compared to lineages identified as ancestral, suggesting a loss of phenotypic plasticity and phenotype canalization through genetic assimilation (Gunter *et al.* 2017). Finally, those observations suggest that both hypotheses are not exclusive but require a well-designed system, such as the ancestral population and the possibility to reared individuals in common garden experiments, in order to answer such complex questions in the context of speciation.

Potential caveats of the study

Our study allowed making several important inferences regarding the relative adaptive role of epigenetic *vs.* genetic variation on patterns of gene expression between limnetic and benthic whitefish from separate continents. Admittedly however, our study design comes with several caveats that must be considered for cautiously interpreting our results. First, fish that were analysed came directly from the wild, as we could not rear them

all in similar controlled conditions. This may have increased inter-individual variation in the level of DNA methylation (both within and between limnetic and benthic whitefish) and therefore reduced the statistical power of detecting CpGs potentially associated with DMLs or DMRs, as well as DEGs. Moreover, given the expensive cost of performing WGBS, we chose to focus on liver tissue only. It is therefore possible that different patterns of association between DNA methylation and gene expression would have been depicted in other tissues. On the other hand, focussing on liver allowed making direct comparisons with previous transcriptomic studies on limnetic and benthic whitefish (St-Cyr *et al.* 2008; Jeukens *et al.* 2010; Rougeux *et al.* 2018). Also, liver is particularly interesting because more genes are expressed in this organ than most other tissues or organs studied in fishes (Derome *et al.* 2006), but also in human where transcriptomic analyses revealed that almost 60% of the human proteins are expressed in the liver and that associated genes are highly expressed (Kampf *et al.* 2014; Uhlen *et al.* 2015). In particular, most genes involved in energetic and metabolic functions are expressed in liver (Uhlen *et al.* 2015), which are highly relevant for investigating the molecular mechanisms involved in the adaptive diversification of species to different ecological niches (Bernatchez *et al.* 2010). Finally, this study focused on the interaction of genotypes, genes expression level and DNA methylation in order to describe the interplay between those different biological levels. Consequently, whole transcriptome and whole genome bisulphite sequencing data were used on a detailed reference transcriptome of orthologous genes to realise all analyses and inferences described above, with an interest for coding regions. Next steps would lead our analysis by characterizing the methylome landscape between all limnetic and all benthic whitefish by using whole genome resequencing of same individuals on the ongoing Lake whitefish genome assembly.

5.7 Acknowledgements

We would like to thank Anne-Marie Dion-Côté for insightful discussions around epigenomics in whitefish, as well as Jérémy Le Luyer for advices and exchanges about methodological approaches. We thank Kim Praebel, Shripathi Bhat and the Freshwater ecology group at UiT for participating with the sampling in Norway, and Ole Seehausen for samples from Switzerland. This research was supported by a Discovery Research grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to L.B. L.B also holds the Canadian Research Chair in genomics and conservation of aquatic resources, which funded the research infrastructure for this project.

5.8 Figures

Figure 5.1: Geographic location of sampling sites corresponding to lakes where sympatric limnetic and benthic species were sampled. Two lakes (blue circles) were sampled in North America: Cliff Lake and Indian Lake in the region of Maine (M) (yellow circles), and two lakes were sampled in Europe: Langfjordvatn Lake in Norway (N) and Zurich Lake in Switzerland (S).

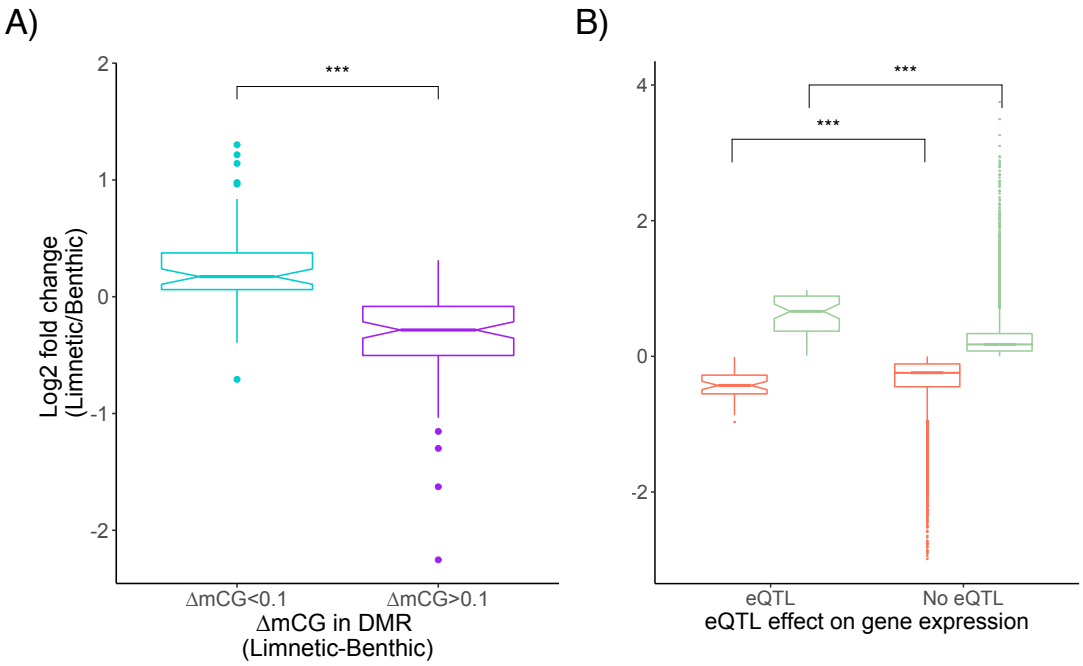


Figure 5.2: Differential gene expression between limnetic and benthic whitefish induced by DNA methylation and cis-eQTL. A) Boxplot showing the level of gene expression differentiation (Log2 fold change) as a function of the difference in methylation level between the limnetic and benthic species (ΔmCG Limnetic-Benthic) for genes associated with a parallel DMR. The light-blue box corresponds to genes with hypomethylated DMRs in limnetic species, whereas the purple box corresponds to genes with hypermethylated DMRs in the limnetic species. Both categories are respectively associated with an overexpression and a repression of gene expression ($P < 0.001$). B) Boxplot showing the level of gene expression differentiation (Log2 fold change) as a function of genes for which we found a presence (eQTL) or absence (No eQTL) of parallel *cis*-eQTL. The red box corresponds to genes with *cis*-eQTL, whereas the light-green box corresponds to genes without *cis*-eQTL. Overexpressed genes in limnetic species (Log2 fold change > 0) are more differentiated when affected by a *cis*-eQTL ($P < 0.001$), as well as repressed genes (Log2 fold change < 0) ($P < 0.001$).

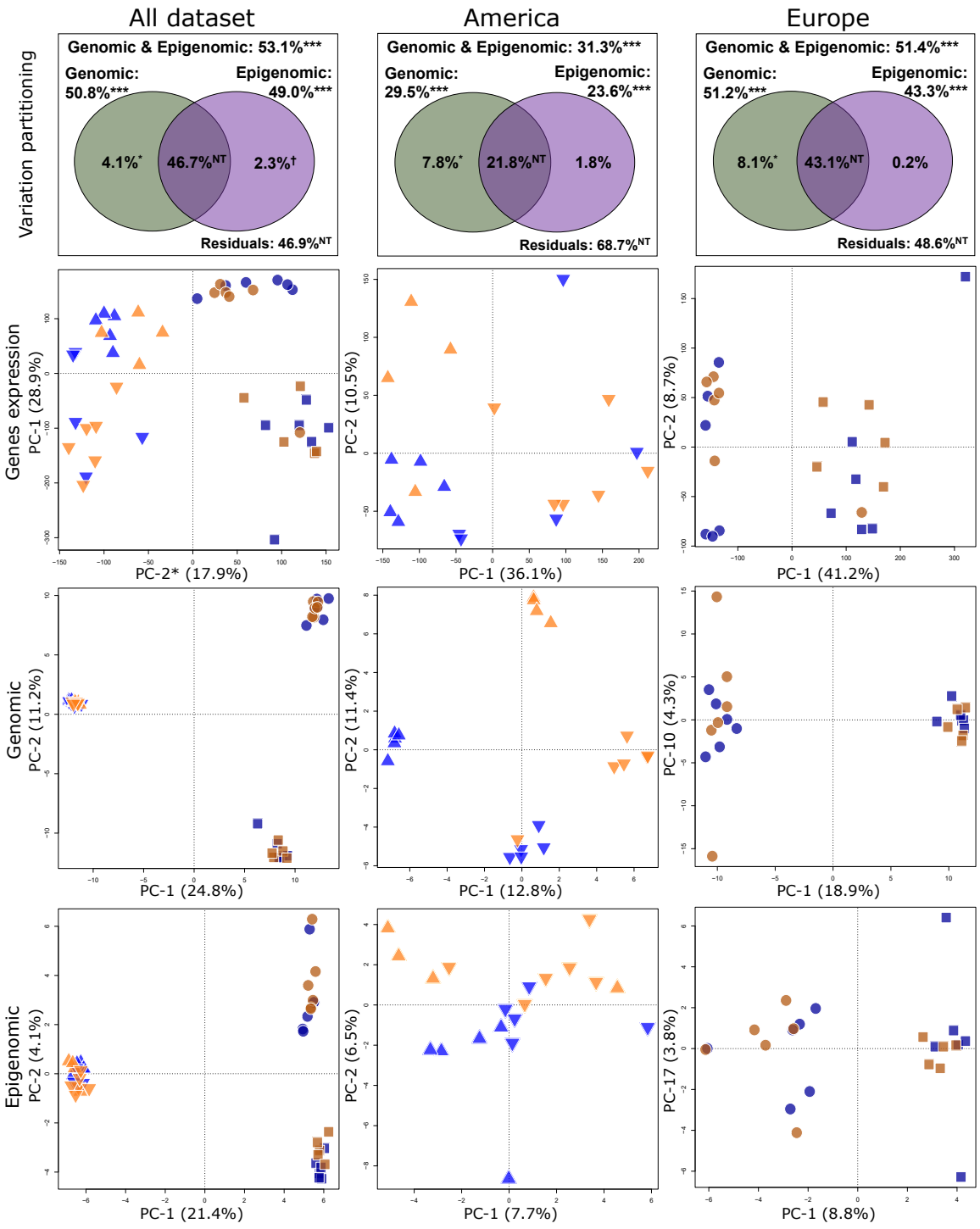


Figure 5.3: Variation partitioning of gene expression level by genomic and epigenomic factors. The figure summarizes the variance partitioning of gene expression associated to genomic and epigenomic data, on panels of the first row, for limnetic-benthic species comparisons across continents (All dataset, first column), and for limnetic-benthic species comparisons in *C. clupeaformis* (North America, second column) and *C. lavaretus* (Europe, third

column). Each panel of the variation partitioning decomposes the variance in gene expression measured in each population of the limnetic and benthic species, at both continental and lakes scale. The total amount of variance explained by the data corresponds to the 'Genomic & Epigenomic' category, while the remaining part is associated to the statistically non-testable (NT) 'Residuals'. Then, the explained variation is decomposed by genomic factor alone ('Genomic') and the epigenomic factor alone ('Epigenomic'). The proportion of variance associated to 'Genomic' data (darkgreen), 'Epigenomic' data (purple) and their intersection (darkpurple, middle) are decomposed in the Venn diagram. Others panels are PCAs illustrating the variance between individuals within populations of both species by means of PCAs, for the three biological levels ('Gene expression', second row; 'Genomic', third row and 'Epigenomic', fourth row). For each PCA, the plotted axes tied in to axes identified by backward selection, and each point corresponds to an individual. Blue and orange points represent limnetic and benthic species, respectively. Lighter tones for both colors are for *C. clupeaformis* (North America) individuals and darker tones for *C. lavaretus* (Europe) individuals, while symbols represent lakes. *Note that on the gene expression PCA for the 'All dataset', we inverted the axes (PC-1 and PC-2) during the projection for readability and comparisons of patterns with other PCAs.

5.9 Supplementary tables

Table S5.1: Number of raw and filtered reads obtained per individual.

Number of raw reads	Number of filtered reads	Library ID	Individual ID	Species
27140147	25803019	LIB58	CD17	Limnetic
33866686	32293410	LIB70	CD18	Limnetic
28958463	27755684	LIB52	CD19	Limnetic
25391937	24577489	LIB47	CD20	Limnetic
20270143	19650600	LIB43	CD21	Limnetic
27820603	26670427	LIB53	CD22	Limnetic
25277259	24265292	LIB16	CN10	Benthic
20169272	19295658	LIB11	CN11	Benthic
26388655	25351805	LIB37	CN12	Benthic
19599732	18676563	LIB60	CN5	Benthic
27467349	26358596	LIB6	CN6	Benthic
32713977	31286322	LIB68	CN7	Benthic
4822124	25833236	LIB30	ID1	Limnetic
17582083	16896683	LIB5	ID2	Limnetic
23034903	22170784	LIB19	ID3	Limnetic
25185594	24150716	LIB39	ID4	Limnetic
27739704	26628586	LIB13	ID7	Limnetic
22853137	21907482	LIB51	ID9	Limnetic
20858977	19875693	LIB71	IN10	Benthic
21627171	20817338	LIB25	IN5	Benthic
21151309	20110167	LIB38	IN6	Benthic
15707040	15224135	LIB44	IN7	Benthic
17533268	16715432	LIB21	IN8	Benthic
24259963	23292157	LIB31	IN9	Benthic
36016311	34438149	LIB66	LF107	Limnetic
30595792	29159603	LIB61	LF108	Limnetic
30904936	29818408	LIB18	LF109	Limnetic
42568708	40665197	LIB63	LF111	Limnetic
22833090	21850810	LIB2	LF112	Limnetic
34202220	32821823	LIB33	LF113	Limnetic
24449486	23452258	LIB54	LF35	Benthic
23115781	22169931	LIB9	LF36	Benthic
21936910	21266681	LIB48	LF37	Benthic
24858616	23818507	LIB12	LF38	Benthic
28375763	27242770	LIB56	LF39	Benthic
23221581	22394148	LIB27	LF41	Benthic
19537598	18931029	LIB41	Z11	Benthic

28935800	27811520	LIB28	Z12	Benthic
14858370	14212670	LIB34	Z13	Benthic
18838742	18186632	LIB29	Z19	Benthic
22155698	21220544	LIB49	Z20	Benthic
21296255	20525864	LIB17	Z23	Limnetic
21385377	20566388	LIB22	Z27	Limnetic
31023995	29809348	LIB40	Z29	Limnetic
20619799	19688042	LIB10	Z30	Limnetic
20743920	19797211	LIB64	Z31	Limnetic
9701395	9001555	LIB69	Z37	Limnetic
27740869	26602363	LIB4	Z15	Benthic

Table S5.2: Gene ontology analysis results for significant enrichment. Significant GO term (P<0.05) from differential gene expression analyses between limnetic and benthic species.

GO:0006807	BP	biological_process	p	nitrogen compound metabolic process	0.000597
GO:0007165	BP	biological_process	p	signal transduction	0.00151
GO:0044237	BP	biological_process	p	cellular metabolic process	0.00189
GO:0044238	BP	biological_process	p	primary metabolic process	0.00563
GO:0071704	BP	biological_process	p	organic substance metabolic process	0.00792
GO:0065007	BP	biological_process	p	biological regulation	0.00962
GO:0050789	BP	biological_process	p	regulation of biological process	0.0115
GO:0009058	BP	biological_process	p	biosynthetic process	0.0116
GO:0032259	BP	biological_process	e	methylation	0.0125
GO:0008152	BP	biological_process	p	metabolic process	0.0167
GO:0051716	BP	biological_process	p	cellular response to stimulus	0.0436
GO:0032991	CC	cellular_component	p	macromolecular complex	0.000102
GO:0043234	CC	cellular_component	p	protein complex	0.000716
GO:0099080	CC	cellular_component	e	supramolecular complex	0.00204
GO:0099081	CC	cellular_component	e	supramolecular polymer	0.00204
GO:1990015	CC	cellular_component	e	ensheathing process	0.0048
GO:0098576	CC	cellular_component	e	lumenal side of membrane	0.0143
GO:0097223	CC	cellular_component	e	sperm part	0.022
GO:0098948	CC	cellular_component	e	intrinsic component of postsynaptic specialization membrane	0.0284
GO:0030672	CC	cellular_component	e	synaptic vesicle membrane	0.0348
GO:0035749	CC	cellular_component	e	myelin sheath adaxonal region	0.0424
GO:0098936	CC	cellular_component	e	intrinsic component of postsynaptic membrane	0.047
GO:0016491	MF	molecular_function	e	oxidoreductase activity	0.0031
GO:0003823	MF	molecular_function	e	antigen binding	0.0215
GO:0004791	MF	molecular_function	e	thioredoxin-disulfide reductase activity	0.0331
GO:0048037	MF	molecular_function	e	cofactor binding	0.043
GO:0016209	MF	molecular_function	e	antioxidant activity	0.0447

Table S5.3: Sequencing efficiency and coverage per WGBS library. The number of reads generated per library (nReads), the number of nucleotides associated (nNucleotides) and the individual coverage associated.

nReads	nNucleotides	Coverage (X)
73428106	22028431800	8.81137272
128872591	38661777300	15.46471092
120073762	36022128600	14.40885144
114373473	34312041900	13.72481676
126932751	38079825300	15.23193012
128312915	38493874500	15.3975498
89843255	26952976500	10.7811906
101163627	30349088100	12.13963524
125040522	37512156600	15.00486264
110977507	33293252100	13.31730084
110086107	33025832100	13.21033284
136320804	40896241200	16.35849648
138579146	41573743800	16.62949752
111383287	33414986100	13.36599444
95107625	28532287500	11.412915
107098869	32129660700	12.85186428
103559153	31067745900	12.42709836
85605173	25681551900	10.27262076
101460525	30438157500	12.175263
100842607	30252782100	12.10111284
83510961	25053288300	10.02131532
106055635	31816690500	12.7266762
117124294	35137288200	14.05491528
119580386	35874115800	14.34964632
113790085	34137025500	13.6548102
124151997	37245599100	14.89823964
110672336	33201700800	13.28068032
116574416	34972324800	13.98892992
132113031	39633909300	15.85356372
105026245	31507873500	12.6031494
110189564	33056869200	13.22274768
114091761	34227528300	13.69101132
121404423	36421326900	14.56853076
122307506	36692251800	14.67690072
127378901	38213670300	15.28546812
113703088	34110926400	13.64437056
85282287	25584686100	10.23387444
117595494	35278648200	14.11145928

120073762	36022128600	14.40885144
114373473	34312041900	13.72481676
126932751	38079825300	15.23193012
128312915	38493874500	15.3975498
117595494	35278648200	14.11145928
120073762	36022128600	14.40885144
114373473	34312041900	13.72481676
126932751	38079825300	15.23193012
128312915	38493874500	15.3975498
113790085	34137025500	13.6548102

Table S5.4: Gene ontology analysis results for significant enrichment. Significant GO term (P<0.05) from differentially methylated regions (DMRs) between all limnetic and all benthic whitefish across continents.

GO:0048856	BP	e	anatomical structure development	0.00161
GO:0032502	BP	e	developmental process	0.00301
GO:0044699	BP	e	single-organism process	0.00543
GO:0009987	BP	e	cellular process	0.00588
GO:0044767	BP	e	single-organism developmental process	0.00665
GO:0022402	BP	p	cell cycle process	0.0159
GO:0044464	CC	e	cell part	3.1e-06
GO:0044424	CC	e	intracellular part	0.000128
GO:0044422	CC	e	organelle part	0.000341
GO:0044446	CC	e	intracellular organelle part	0.00216
GO:0031090	CC	e	organelle membrane	0.00222
GO:0044425	CC	e	membrane part	0.00929
GO:0005622	CC	e	intracellular	0.0183
GO:0044326	CC	e	dendritic spine neck	0.0223
GO:1990794	CC	e	basolateral part of cell	0.0267
GO:0098590	CC	e	plasma membrane region	0.0303
GO:0044456	CC	e	synapse part	0.0318
GO:0043226	CC	e	organelle	0.0374
GO:0097458	CC	e	neuron part	0.0378
GO:0060171	CC	e	stereocilium membrane	0.0397
GO:0045281	CC	e	succinate dehydrogenase complex	0.0397
GO:0044420	CC	e	extracellular matrix component	0.0446
GO:0005488	MF	e	binding	2.98e-05
GO:0043167	MF	e	ion binding	0.000644
GO:0016491	MF	e	oxidoreductase activity	0.00179
GO:0048037	MF	e	cofactor binding	0.00928
GO:0051540	MF	e	metal cluster binding	0.011
GO:0005515	MF	e	protein binding	0.0143
GO:0050436	MF	e	microfibril binding	0.0178
GO:0003824	MF	e	catalytic activity	0.0206
GO:0005201	MF	e	extracellular matrix structural constituent	0.0356

Conclusion

6.1 Conclusions générales

L'objectif général des travaux présentés dans cette thèse consistait à contribuer à la compréhension de la diversification phénotypique dans un contexte de spéciation, en élucidant les bases moléculaires de la divergence adaptative. Plus spécifiquement, notre approche était de combiner différentes approches moléculaires tels que la génomique des populations, la transcriptomique et l'épigénomique afin de relier l'architecture de la différenciation génomique aux les bases moléculaires impliquées dans la diversification phénotypique durant le processus de spéciation. Nous avons identifié l'origine de la divergence chez des réplicats de paires d'espèces montrant une divergence phénotypique parallèle, décrit une partie des bases polygéniques sous-jacentes à cette différenciation phénotypique, et documenté des mécanismes moléculaires associés à cette différenciation entre les espèces en cours de spéciation.

Le système du Grand Corégone présente des caractéristiques propices à l'étude du processus de spéciation, notamment pour répondre à l'objectif de cette thèse. En effet, le système du Grand Corégone a l'avantage de présenter des réplicats naturels de paires d'espèces limnétiques et benthiques observées dans des lacs indépendants, et dont la divergence est relativement récente. De plus, le Grand Corégone possède une espèce sœur: le Lavaret, elle aussi composée d'un complexe de paires d'espèces limnétiques et benthiques observées dans de nombreux lacs isolés les uns des autres en Europe. En outre, ces populations lacustres ont fait l'objet de nombreuses études documentant les différences génétiques et écologiques qui résultent de la divergence adaptative aux niches écologiques benthique et limnétique, ainsi qu'aux aspects physiologiques, morphologiques et phénotypiques associés. Le parallélisme de diversification phénotypique entre localités géographiques indépendantes a ainsi permis des analyses de génomique comparative et rendu possible l'identification des différences entre espèces limnétiques et benthiques qui sont conservées dans l'ensemble des paires sympatriques. Ces différences se sont révélées être potentiellement impliquées dans la diversification phénotypique.

Dans le premier chapitre de cette thèse (chapitre 2), nous avons élucidé l'histoire démographique de la divergence entre espèces limnétiques et benthiques chez *C.*

clupeaformis afin de comprendre les variations de différenciation génomique et phénotypique entre paires d'espèces, dans les lacs étudiés. Nous avons ainsi associé des données génomiques empiriques à des modèles visant à reconstruire les processus démographiques et sélectifs modelant les paysages génomiques de la divergence. Le développement de modèles originaux dans cette étude a permis d'établir un cadre analytique nouveau pour l'inférence de la démographie historique, permettent de tester successivement l'influence de différents facteurs démo-sélectifs lors de la spéciation. De plus, nous avons mis en évidence que la divergence génomique pour chacune des paires d'espèces sympatriques résultait d'un contact secondaire entre deux lignées glaciaires ayant divergées en allopatrie. Cette période allopatrique fut propice à l'accumulation de différences génomiques (*i.e.*, par tri du polymorphisme ancestral et accumulation de nouvelles mutations indépendemment dans chaque lignées). Ces analyses ont révélé l'existence d'un contact secondaire similaire dans tous les lacs, y compris celui un temps suspecté abriter une population unique du fait d'une faible diversité mitochondriale (Bernatchez and Dodson 1991). Outre la robustesse de telles approches, nous avons pu identifier et quantifier les processus sélectifs afin de les dissocier des processus démographiques. Les paysages de différenciation génomique entre paires d'espèces limnétiques et benthiques sympatriques ont ainsi été modélisés en intégrant l'effet joint de la sélection en liaison et du flux génique hétérogène, en plus des effets démographiques classiques. Nos résultats ont montré que le gradient de différenciation génomique et phénotypique observé entre lacs résulte d'interactions complexes entre les forces démographiques comme le taux de migration et la taille efficace, le degré de semi-perméabilité des génomes et l'empreinte de la sélection en liaison. Il est intéressant de noter que de tels patrons génomiques sont observés dans d'autres systèmes empiriques chez qui une méthodologie similaire a été déployée (Van Belleghem et al. 2018). Ces résultats sont également cohérents avec des modèles théoriques visant à illustrer les effets de différents modes de sélection sur le paysage génomique de la divergence au fil du temps (Cruickshank and Hahn 2014; Burri 2017a; Burri 2017b).

Par la suite, nous avons réalisé une étude parallèle d'inférence de l'histoire démo-sélective chez le complexe d'espèces limnétiques et benthiques européennes des lacs post-glaciaires de Suisse et de Norvège (Chapitre 3). Initialement décrites comme des cas de spéciation sympatrique, notamment dans les lacs Suisse, ces paires d'espèces étaient

néanmoins mentionnées par certaines études comme résultant potentiellement de contacts entre différentes lignées. Afin d'interroger l'origine de la diversification phénotypique parallèle observée entre ces paires d'espèces sympatriques limnétiques et benthiques, nous avons mis en œuvre une méthodologie similaire à celle développée chez le Grand Corégone américain. Nous avons mis en évidence des cas de contacts secondaires entre lignées glaciaires ayant évolué en allopatrie durant la dernière glaciation. Ce contact a probablement résulté d'une recolonisation (vers le Nord et vers le Sud) des deux lignées ayant trouvé refuge au Sud (Suisse) et au Nord (Scandinavie) de l'Europe, suite au retrait de la calotte glaciaire et à l'ouverture de voies de colonisations. Bien que les paysages de différenciation génomique entre les espèces limnétiques et benthiques ne soient pas caractérisés pour *C. lavaretus*, nous avons également pu démontrer que des processus évolutifs similaires à ceux révélés chez le Grand Corégone ont influencé la divergence adaptative en Europe.

Ces deux chapitres complémentaires ont permis de connaître l'histoire évolutive des espèces limnétiques et benthiques observées en Amérique du nord et en Europe. Suite à la séparation des lignées américaine et européenne il y a environ 500 000 ans, la dernière période glaciaire a induit des subdivisions indépendantes de différentes lignées glaciaires dans chaque continent, qui sont ensuite entrées en contact dans la région des Apalaches (pour *C. clupeaformis*) et la région centro Européenne (pour *C. lavaretus*) suite au dernier retrait glaciaire. Ces études ont donc mis en évidence des mécanismes évolutifs identiques entre les paires d'espèces des deux continents, mettant en avant l'importance de l'existence d'une période d'isolement géographique préalabe à la divergence phénotypique en sympatrie. Les flux géniques initiés lors des contacts secondaires répétés ont permis de mettre en évidence la nature poreuse des barrières génétiques entre espèces (Barton and Bengtsson 1986), mais aussi l'effet de la sélection aux sites liés induisant une accélération du tri de polymorphisme ancestral lors des épisode d'isolement géographique (Cruickshank and Hahn 2014).

Lors du troisième chapitre (chapitre 4), nous avons comparé des paires d'espèces limnétiques et benthiques provenant d'Amérique du Nord et d'Europe. Cela nous a permis de caractériser les bases génomiques associées au parallélisme de diversification

phénotypique observé entre les continents. Nous avons notamment identifié une partie des bases polygéniques potentiellement sous sélection qui sont associées à la diversification phénotypique. Nous avons également constaté que les gènes différentiellement exprimés sont enrichis en polymorphisme partagé entre les espèces américaine et européenne, soit en polymorphisme ancestral qui aurait été maintenu depuis le premier évènement de divergence entre *C. clupeaformis* et *C. lavaretus*. Finalement, certains gènes différentiellement exprimés entre les espèces limnétiques et benthiques sont associés à des variants génétiques dont la variation de fréquence est corrélée à la variation du niveau d'expression génique. Bien qu'il serait intéressant et pertinent d'élargir le nombre de comparaisons grâce à un échantillonnage plus large (nombre de lacs et d'individus), les résultats suggèrent que la sélection divergente entre espèces limnétiques et benthiques affecte en partie les mêmes gènes, participe ainsi au maintien du polymorphisme aux mutations cis-régulatrices contrôlant le niveau d'expression de ces gènes et induit des changements subtiles de fréquences alléliques chez l'espèce limnétique. Une telle architecture composée d'un nombre élevé de gènes et de changement subtils de fréquences alléliques correspond à l'adaptation polygénique (Yeaman 2015). Finalement, les gènes identifiés comme cibles de la sélection divergente sont associés à du métabolisme énergétique et à la croissance, tel qu'observés dans des études précédentes (Derome et al. 2006; St-Cyr et al. 2008).

Finalement, lors du dernier chapitre (chapitre 5), nous avons caractérisé les patrons de méthylation entre espèces limnétiques et benthiques au sein et entre les continents, dans une approche comparative similaire à celle réalisée dans le quatrième chapitre de cette thèse. La volonté de déterminer les mécanismes moléculaires associés aux réponses phénotypiques causées par des pressions environnementales (potentiellement sélectives), intègre des niveaux biologiques supplémentaires tels que l'épigénétique. Via une approche d'épigénomique écologique, nous avons observé des patrons de méthylation différents entre espèces limnétiques et benthiques qui sont en partie maintenus entre les paires de comparaison et entre continents. Ces différences de méthylation entre espèces, partagées dans l'ensemble du système étudié, sont directement associées à des bases génétiques (*i.e.*, méthylation obligatoire). Les différents niveaux de méthylation affectent l'activité transcriptionnelle des gènes affectés, avec notamment une corrélation négative entre le

niveau de méthylation et le niveau d'expression. De plus, l'amplitude de la différenciation d'expression génique est plus importante lorsque les gènes sont associés à des variants génétiques (*cis*-eQTL) que par des DMRs. Bien que nos analyses manquent de résolution en terme de définition de structures géniques et de séquences flanquantes, nous avons pu identifier l'origine génétique d'une importante proportion de la composante épigénétique correspondant à la méthylation de l'ADN. La méthylation obligatoire correspond à une proportion importante de la méthylation, si ce n'est l'intégralité chez les plantes (Meng et al. 2016) mais aussi chez l'humain (Zaghlool et al. 2016). Cette étude nous permet de proposer un mécanisme de modulation de l'expression génique des gènes affectés par le seul effet des différentiels de méthylation (*i.e.*, en l'absence d'une régulation via des variants génétiques), lorsque ces gènes sont impliqués dans des traits phénotypiques directement (*e.g.*, traits polygéniques) ou indirectement (*e.g.*, gènes pléiotropiques) affectés par des pressions sélectives. Ainsi, les niveaux transcriptionnels des gènes, sous-jacents à des traits phénotypiques corrélés ou anti-corrélés, seraient ajustés en réponse aux pressions sélectives et permettraient d'explorer le paysage adaptatif plus rapidement en générant éventuellement des phénotypes alternatifs.

En conclusion, cette thèse de doctorat contribue à l'approfondissement de nos connaissances concernant les bases génomiques et l'architecture génomique de la divergence adaptative dans des systèmes empiriques, notamment dans les systèmes où des cas de répétition de diversification phénotypique causée par l'adaptation à des conditions environnementales similaires sont observés. Au delà de l'identification de barrières potentielles à la reproduction (*i.e.*, gènes de spéciation) chez des espèces naissantes en cours de divergence, nous montrons ici que les barrières peuvent être multiples du fait de l'architecture des bases génomiques des traits phénotypiques complexes impliqués dans la divergence adaptative, tels que chez le Grand Corégone et le Lavaret. Parmi ces barrières génomiques, celles identifiées et maintenues dans l'ensemble du système indiquent que les voies évolutives sont contraintes par les possibilités de réponses imposées par l'architecture des génomes (*e.g.*, contraintes évolutives sur les gènes, les voies métaboliques), pour les pressions de sélection divergente, ici entre les espèces limnétiques et benthiques. Ces contraintes s'observent par la répétition de l'action

de la sélection divergente, sur des gènes orthologues, notamment sur des variants génétiques partagés issus du polymorphisme ancestral.

6.2 Perspectives

Les résultats de mes travaux suggèrent un rôle majeur du polymorphisme génétique ancestral dans les cas d'adaptations répétées à des conditions environnementales similaires, et plus précisément de diversification phénotypique parallèle dans un contexte de spéciation écologique. L'originalité de cette thèse réside dans la combinaison des approches d'inférences démographiques de populations divergentes, et de la caractérisation de l'origine des gènes associés à la divergence adaptative dans les contextes de spéciation en présence de flux géniques, que très peu d'études à date ont su réaliser. Ainsi, cette thèse apporte une complémentarité à la littérature spécialisée. Par ailleurs, malgré la croissance de la résolution des analyses génomiques (*via* le développement du séquençage), les questions relatives à l'histoire démographique de la divergence ne sont pas traitées dans tous les systèmes d'étude, et certaines hypothèses alternatives à l'histoire suposée ne sont pas testées (Elmer et al. 2012; Meier, Marques, et al. 2017; Meier, Sousa, et al. 2017). Pourtant, l'éventail grandissant des possibilités analytiques permettant d'accroitre la qualité des inférences de sélection en lien avec la diversification phénotypique et laisse place à un champ inexploré de nouvelles interprétations évolutives.

L'étude du parallélisme de diversification phénotypique, dans certains systèmes caractéristiques, permet d'aborder les questions liées à l'origine des bases génomiques de l'adaptation locale, de la divergence adaptative et éventuellement à la spéciation, notamment lorsque l'histoire évolutive des populations est connue. Des études récentes se sont focalisées sur ces aspects pour identifier des bases génétiques issues d'une histoire commune de divergence (Le Moan et al. 2016; Van Belleghem et al. 2018). Outre l'origine des bases génomiques de la diversification phénotypique, dans un contexte de spéciation, certaines études documentent l'architecture de traits phénotypiques oligogéniques (Soria-Carrasco et al. 2014; Zhao and Begun 2017), ou monogéniques (Xie et al. 2019) relativement simples, alors que peu d'études s'intéressent à des traits phénotypiques complexes polygéniques (Van Belleghem et al. 2018). Par conséquences, le Grand Corégone offre des caractéristiques intéressantes pour enrichir nos connaissances sur l'architecture génomique de la diversification phénotypiques dans un contexte de spéciation écologique.

Le Grand Corégone est une espèce non modèle, et de ce fait, dépourvue de référence génomique. Cependant, le séquencage et l'assemblage du génome du Grand Corégone sont en cours de réalisation. Avec le développement des technologies de séquencage et des algorithmes d'assemblage, acquérir une référence génomique n'a jamais semblé aussi réalisable, même pour des espèces dont les caractéristiques génomiques affichent une importante complexité. En effet, durant cette thèse de doctorat j'ai contribué à la supervision de l'assemblage du génome du Grand Corégone, me heurtant à la taille relativement élevée du génome (~3Gb), à sa forte proportion en éléments répétés (~60%), et aux régions génomiques pseudo-tétraploides caractéristiques des Salmonidés, complexifiant les tentatives d'assemblages. Toutefois, le séquencage de longues séquences pourrait permettre d'affiner nos assemblages en chevauchant ces régions à forte complexité. Ensuite, la phase d'annotation, notamment à l'aide des données transcriptomiques produites durant ce doctorat et de régférence génomique chez d'autres espèces de Salmonidés, permettra d'identifier et de caractériser la structure des gènes.

Dès que cette référence génomique sera acquise, des approches de génomique des populations sur l'ensemble du génome pourront être réalisées sur les mêmes individus. En effet, j'ai généré durant ma thèse le matériel nécessaire au reséquencage de huit génomes par population des espèces limnétiques et benthiques en Amérique du Nord et en Europe. Ces données génomiques, pourront nous permettre de mesurer, qualitativement et quantitativement, le parallélisme génomique associé à la diversification phénotypique entre niches limnétique et benthique. Mais au delà d'accroître la résolution, il serait interessant de déterminer un indice permettant de quantifier la probabilité de la contribution de chaque gène au parallélisme de l'adaptation. En d'autres termes, il serait important de quantifier la probabilité d'implication d'un gène dans la répétabilité d'adaptation locale à des contraintes environnementales similaires. Ce type d'approche a été abordé de façon théorique chez l'épinoche à trois épines (Conte, Arnegard, Peichel, and Schluter 2012b), de façon empirique chez des espèces de conifères (Yeaman, Hodgins, et al. 2016), et fait l'objet de projets de recherches qui tendent à homogénéiser les approches méthodologiques et statistiques, et à développer des métriques de quantification de parallélisme au niveau génomique dans des études d'associations génotypes-phénotypes-environnement. Le Grand Corégone pourrait de part les différents niveaux temporels de divergence et le

parallélisme de différenciation phénotypique, contribuer à cette composante de la littérature scientifique s'intéressant à l'évolution de l'adaptation locale.

En outre, les données de reséquencage de génomes entiers pourraient être couplées aux données de transcriptomique et d'épigénomique. En effet, la possibilité de quantifier le niveau d'expression génique par individu pourrait permettre d'associer des niveaux de méthylation génomique à la variation d'expression génique. Cela pourrait être fait en dissociant les effets potentiels de méthylation dans les régions promotrices de gènes, dans les régions UTR, dans les exons et dans les introns, afin de mesurer les effets individuels de ces régions. De plus, il serait possible de quantifier les niveaux de méthylation des éléments transposables dans les populations naturelles. Des études en jardin commun de croisements entre lignées parentales limnétiques et benthiques ont mis en évidence une réactivation des éléments transposables (TEs) et des ARNs non-codants chez les hybrides, induisant des malformations et une mortalité accrue, associée à une instabilité génomique (Dion-Côté et al. 2014). Les TEs sont également capables de générer du polymorphisme phénotypique, de par leur réactivation, en affectant l'expression de gènes localisés à proximité (Hosaka and Kakutani 2018). Il serait donc intéressant d'identifier les différences de méthylations entre paires d'espèces limnétiques et benthiques dans les TEs, et le potentiel effet sur les variations transcriptionnelles des gènes et le maintien de barrières reproductives dans les populations naturelles.

Finalement, il serait pertinent de mener des études comparatives entre paires d'espèces de façon à identifier et mesurer les différences entre les différents lacs et entre continents. En effet, la quantification de différents polymorphismes entre paires d'espèces de différents lacs pourrait compléter la compréhension, notamment en Amérique du Nord, des différents niveaux de différenciation phénotypiques et génomiques. Est-ce que la plus grande différenciation phénotypique est causée par une variance d'activité transcriptionnelle réduite, en lien à de plus forts niveaux de différences de méthylation entre espèces? Est-ce que les régions génomiques hautement différenciées entre espèces sont systématiquement associées à des DMRs, et réciproquement? En d'autres termes, malgré des bases épigénomiques essentiellement obligatoires, pouvons-nous découpler des variants génétiques à forte différenciation des variants épigénétiques? Par extension, il serait intéressant d'intégrer les caractéristiques propres à *C. lavaretus*. Le complexe d'espèces de *C. lavaretus* peut se composer de plus de deux espèces dans certains lacs. Des études

phylogéographiques ont mis en évidence des différences de diversité génétique au niveau de l'ADN mitochondrial (Bernatchez et al. 1989; Bernatchez and Dodson 1994). J'ai observé des patrons similaires à l'échelle de l'ADN nucléaire, et cette réduction de diversité génétique en Amérique du Nord serait la conséquence d'un plus fort impact de la dernière période glaciaire, causant un plus fort goulot d'étranglement sur les populations. Il est alors possible d'émettre l'hypothèse que la plus grande amplitude de radiations adaptatives en Europe est associée à un potentiel évolutif supérieur. Bien que les lacs peuvent fournir davantage d'habitats et de niches écologiques différentes, du fait de leurs tailles supérieures, il est possible de tester cette hypothèse en comparant les différences de polymorphisme génétique, de variance d'expression génique et de méthylation dans les régions génomiques d'intérêts (*i.e.*, associées à la différenciation entre espèces limnétiques et benthiques). L'ensemble des perspectives proposées içi viendrait largement étayer et compléter notre connaissance des mécanismes de la spéciation, qui plus est chez une espèce non-modèle qui pourrait alors prétendre au titre d'espèce modèle.

Bibliographie

Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJE, Bierne N, Boughman J, Brelsford A, Buerkle CA, Buggs R, et al. 2013. Hybridization and speciation. J. Evol. Biol. 26:229–246.

Abbott RJ, Barton NH, Good JM. 2016. Genomics of hybridization and its evolutionary consequences. Molecular Ecology.

Alcala N, Jensen JD, Telenti A. 2016. The genomic signature of population reconnection following isolation: from theory to HIV. G3: Genes 6:107–120.

Ambrose SH. 1998. Late Pleistocene human population bottlenecks, volcanic winter, and differentiation of modern humans. Journal of Human Evolution 34:623–651.

Amundsen P-A, Bøhn T, Våga GH. 2004a. Gill raker morphology and feeding ecology of two sympatric morphs of European whitefish (Coregonus lavaretus). Annales Zoologici Fennici 41:291–300.

Amundsen P-A, Bøhn T, Våga GH. 2004b. Gill raker morphology and feeding ecology of two sympatric morphs of European whitefish (*Coregonus lavaretus*). Annales Zoologici Fennici 41:291–300.

Andersen KG, Borns HJ. 1994. *The Ice Age World*, 1st edn. (Scandinavian University Press, editor.).

Aoki K, Kato M, Murakami N. 2008. Glacial bottleneck and postglacial recolonization of a seed parasitic weevil, Curculio hilgendorfi, inferred from mitochondrial DNA variation. Molecular Ecology 17:3276–3289.

April J, Hanner RH, Dion-Côté A-M, Bernatchez L. 2013. Glacial cycles as an allopatric speciation pump in north-eastern American freshwater fishes. Molecular Ecology 22:409–422.

Aspinwall N. 1974. GENETIC ANALYSIS OF NORTH AMERICAN POPULATIONS OF THE PINK SALMON, ONCORHYNCHUS GORBUSCHA, POSSIBLE EVIDENCE FOR THE NEUTRAL MUTATION-RANDOM DRIFT HYPOTHESIS. Evolution 28:295–305.

Avise JC. 2000. Phylogeography: the history and formation of species.

Ågren J, Oakley CG, McKay JK, Lovell JT, Schemske DW. 2013. Genetic mapping of adaptation reveals fitness tradeoffs in Arabidopsis thaliana. Proceedings of the National Academy of Sciences 110:21077–21082.

Babin C, Gagnaire P-A, Pavey SA, Bernatchez L. 2017. RAD-Seq Reveals Patterns of Additive Polygenic Variation Caused by Spatially-Varying Selection in the American Eel (*Anguilla rostrata*). Genome Biology and Evolution 9:2974–2986.

Badyaev AV, Potticary AL, Morrison ES. 2017. Most Colorful Example of Genetic Assimilation? Exploring the Evolutionary Destiny of Recurrent Phenotypic Accommodation. Am Nat 190:266–280.

Bailey SF, Rodrigue N, Kassen R. 2015. The effect of selection environment on the probability of parallel evolution. Molecular Biology and Evolution 32:1436–1448.

Barbash DA, Siino DF, Tarone AM, Roote J. 2003. A rapidly evolving MYB-related protein causes species isolation in Drosophila. Proceedings of the National Academy of Sciences 100:5302–5307.

Barnard AA, Fincke OM, McPeek MA, Masly JP. 2017. Mechanical and tactile incompatibilities cause reproductive isolation between two young damselfly species. Evolution 71:2410–2427.

Barrett LW, Fletcher S, Wilton SD. 2012. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. Cell. Mol. Life Sci. 69:3613–3634.

Barrett RDH, M'Gonigle LK, Otto SP. 2006. The distribution of beneficial mutant effects under strong selection. Genetics 174:2071–2079.

Barrett RDH, Schluter D. 2008. Adaptation from standing genetic variation. Trends in Ecology & Evolution 23:38–44.

Barton NH, Bengtsson BO. 1986. The barrier to genetic exchange between hybridising populations. Heredity 57:357–376.

Barton NH, de Cara MAR. 2009. The evolution of strong reproductive isolation. Evolution 63:1171–1190.

Barton NH. 1983. MULTILOCUS CLINES. Evolution 37:454–471.

Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, Haudenschild CD, Beckman KB, Shi J, Mei R, et al. 2014. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Research 24:14–24.

Bay RA, Arnegard ME, Conte GL, Best J, Bedford NL, McCann SR, Dubin ME, Chan YF, Jones FC, Kingsley DM, et al. 2017. Genetic Coupling of Female Mate Choice with Polygenic Ecological Divergence Facilitates Stickleback Speciation. Current Biology 27:3344–3349.e3344.

Becquet C, Przeworski M. 2007. A new approach to estimate parameters of speciation models with application to apes. Genome Research 17:1505–1519.

Bengtsson BO. 1994. Hybrid zones and the evolutionary process. Trends in Ecology & Evolution 9:350.

Benjamini Y, Hochberg YHJOTRSSS. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. JSTOR

1:289–300.

Berg JJ, Coop G. 2014a. A Population Genetic Signal of Polygenic Adaptation.W Feldman M, editor. PLoS Genet 10:e1004412–e1004425.

Berg JJ, Coop G. 2014b. A Population Genetic Signal of Polygenic Adaptation.W Feldman M, editor. PLoS Genet 10:e1004412–e1004425.

Bernatchez L, Chouinard A, Lu G. 1999. Integrating molecular genetics and ecology in studies of

adaptive radiation: whitefish, *Coregonus* sp., as a case study. Biological Journal of the Linnean Society 68:173–194.

Bernatchez L, Dodson JJ, Boivin S. 1989. Population bottlenecks: influence on mitochondrial DNA diversity and its effect in coregonine stock discrimination. Journal of Fish Biology 35:233–244.

Bernatchez L, Dodson JJ. 1990. Allopatric origin of sympatric populations of lake whitefish (*Coregonus clupeaformis*) as revealed by mitochondrial-DNA restriction analysis. Evolution 24:890–908.

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, Dodson JJ. 1994. Phylogenetic relationships among Palearctic and Nearctic whitefish (*Coregonus* sp.) populations as revealed by mitochondrial DNA variation. Canadian Journal of Fisheries and Aquaculture.

Bernatchez L, Renaut S, Whiteley AR, Derome N, Jeukens J, Landry L, Lu G, Nolte AW, Ostbye K, Rogers SM, et al. 2010. On the origin of species: insights from the ecological genomics of lake whitefish. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 365:1783–1800.

Bernatchez L, Wilson CC. 1998. Comparative phylogeography of Nearctic and Palearctic fishes. Molecular Ecology 7:431–452.

Bernatchez L. 1997. Mitochondrial DNA analysis confirms the existence of two glacial races of rainbow smelt Osmerus mordaxand their reproductive isolation in the St Lawrence River estuary (Quebec, Canada). Molecular Ecology 6:73–83.

Beysard M, Perrin N, Jaarola M, Heckel G, Vogel P. 2011. Asymmetric and differential gene introgression at a contact zone between two highly divergent lineages of field voles (Microtus agrestis). J. Evol. Biol. 25:400–408.

Bierne N, Gagnaire PA, David P. 2013. The geography of introgression in a patchy environment and the thorn in the...: EBSCOhost. Current Zoology.

Bierne N, Welch J, Loire E, Bonhomme F, David P. 2011. The coupling hypothesis: why genome scans may fail to map local adaptation genes. Molecular Ecology 20:2044–2072.

Björck S. 1995. A review of the history of the Baltic Sea, 13.0-8.0 ka BP. Quaternary International 27:19–40.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Bolnick DI, Fitzpatrick BM. 2007. Sympatric Speciation: Models and Empirical Evidence. Annu. Rev. Ecol. Evol. Syst. 38:459–487.

Bossdorf O, Richards CL, Pigliucci M. 2008. Epigenetics for ecologists. Ecology Letters 11:106–115.

Boyle EA, Li YI, Pritchard JK. 2017. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 169:1177–1186.

Bradshaw HD, Schemske DW. 2003. Allele substitution at a flower colour locus produces a pollinator shift in monkeyflowers. Nature 426:176–178.

Burnham KP, Anderson DR. 2002. Model selection and multimodel inference: a practical information-theoretic approach.

Burri R, Nater A, Kawakami T, Mugal CF, Olason PI, Smeds L, Suh A, Dutoit L, Bureš S, Garamszegi LZ, et al. 2015. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. Genome Research 25:1656–1665.

Burri R. 2017a. Linked selection, demography and the evolution of correlated genomic landscapes in birds and beyond. Molecular Ecology 26:3853–3856.

Burri R. 2017b. Interpreting differentiation landscapes in the light of long-term linked selection. Evolution Letters 112:2109–2114.

Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL. 2002. The cost of inbreeding in Arabidopsis. Nature 416:531–534.

Butlin RK, Saura M, Charrier G, Jackson B, André C, Caballero A, Coyne JA, Galindo J, Grahame JW, Hollander J, et al. 2014. Parallel evolution of local adaptation and reproductive isolation in the face of gene flow. Evolution 68:935–949.

Capblancq T, Luu K, Blum MGB, Bazin É. 2018. Evaluation of redundancy analysis to identify signatures of local adaptation. Mol Ecol Resour 18:1223–1233.

Carruthers M, Yurchenko AA, Augley JJ, Adams CE, Herzyk P, Elmer KR. 2018. De novo transcriptome assembly, annotation and comparison of four ecological and evolutionary model salmonid fish species. BMC Genomics 19:1–17.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: an analysis tool set for population genomics. Molecular Ecology 22:3124–3140.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289–1303.

Charlesworth B, Nordborg M, Charlesworth D. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet. Res. 70:155–174.

Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet 2:e64.

Chevin L-M. 2013. Genetic constraints on adaptation to a changing environment. Evolution 67:708–721.

Cohan FM, Hoffmann AA. 1989. Uniform Selection as a Diversifying Force in Evolution: Evidence from Drosophila. American Naturalist 134:613–637.

Collins S, de Meaux J. 2009. ADAPTATION TO DIFFERENT RATES OF ENVIRONMENTAL CHANGE IN CHLAMYDOMONAS. Evolution 63:2952–2965.

Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G, 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.

Colosimo PF, Peichel CL, Nereng K, Blackman BK, Shapiro MD, Schluter D, Kingsley DM. 2004. The Genetic Architecture of Parallel Armor Plate Reduction in Threespine Sticklebacks.Nipam Patel, editor. PLoS Biol 2:e109.

Comeault AA, Ferreira C, Dennis S, Soria-Carrasco V, Nosil P. 2016. Color phenotypes are under similar genetic control in two distantly related species of Timema stick insect. Evolution 70:1283–1296.

Conte GL, Arnegard ME, Peichel CL, Schluter D. 2012a. The probability of genetic parallelism and convergence in natural populations. Proc. Biol. Sci. 279:5039–5047.

Conte GL, Arnegard ME, Peichel CL, Schluter D. 2012b. The probability of genetic parallelism and convergence in natural populations. Proc. Biol. Sci. 279:5039–5047.

Conte GL, Schluter D. 2013. Experimental confirmation that body size determines mate preference via phenotype matching in a stickleback species pair. Evolution 67:1477–1484.

Cooper TF, Rozen DE, Lenski RE. 2003. Parallel changes in gene expression after 20,000 generations of evolution in Escherichiacoli. Proceedings of the National Academy of Sciences 100:1072–1077.

Coyne JA, Orr HA. 2004. Speciation. Sunderland MA Sinauer Associates Inc

Cruickshank TE, Hahn MW. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Molecular Ecology 23:3133–3157.

Curry RA. 2007. Late glacial impacts on dispersal and colonization of Atlantic Canada and Maine by freshwater fishes. Quaternary Research 67:225–233.

Dalziel AC, Laporte M, Guderley H, Bernatchez L. 2017. Do differences in the activities of carbohydrate metabolism enzymes between Lake Whitefish ecotypes match predictions from transcriptomic studies? Comparative Biochemistry and Physiology, Part B:1–12.

Dalziel AC, Laporte M, Rougeux C, Guderley H, Bernatchez L. 2017. Convergence in organ size but not energy metabolism enzyme activities among wild Lake Whitefish (Coregonus clupeaformis) species pairs.Rogers SM, Xu S, Schluter PM, editors. Molecular Ecology 26:225–244.

Dalziel AC, Martin N, Laporte M, Guderley H, Bernatchez L. 2015. Adaptation and acclimation of aerobic exercise physiology in Lake Whitefish ecotypes (Coregonus clupeaformis). Evolution

69:2167–2186.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

de Queiroz K. 1998. The general lineage concept of species, species criteria, and the process of speciation: A conceptual unification and terminological recommendations. In: Howard DJ, Berlocher SH, editors. *Endless Forms: Species and Speciation.*

. Oxford Univ. Press.:57–75.

de Queiroz K. 2007. Species Concepts and Species Delimitation. Systematic Biol. 56:879–886.

de Villemereuil P, Mouterde M, Gaggiotti OE, Till-Bottraud I. 2018. Patterns of phenotypic plasticity and local adaptation in the wide elevation range of the alpine plant Arabis alpina.Jacquemyn H, editor. J Ecol 106:1952–1971.

De Wit P, Pespeni MH, Palumbi SR. 2015. SNP genotyping and population genomics from expressed sequences - current advances and future possibilities. Molecular Ecology 24:2310–2323.

Deagle BE, Jones FC, Chan YF, Absher DM, Kingsley DM, Reimchen TE. 2012. Population genomics of parallel phenotypic evolution in stickleback across stream-lake ecological transitions. Proc. Biol. Sci. 279:1277–1286.

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

Dion-Côté A-M, Barbash DA. 2017. Beyond speciation genes: an overview of genome stability in evolution and speciation. Current Opinion in Genetics & Development 47:17–23.

Dion-Côté A-M, Renaut S, Normandeau E, Bernatchez L. 2014. RNA-seq reveals transcriptomic shock involving transposable elements reactivation in hybrids of young lake whitefish species. Molecular Biology and Evolution 31:1188–1199.

Dion-Côté A-M, Symonová R, Ráb P, Bernatchez L. 2015. Reproductive isolation in a nascent species pair is associated with aneuploidy in hybrid offspring. Proceedings of the Royal Society B: Biological Sciences 282:20142862–20142862.

Dittmar EL, Oakley CG, Conner JK, Gould BA, Schemske DW. 2016. Factors influencing the effect size distribution of adaptive substitutions. Proc. Biol. Sci. 283:20153065–20153068.

Dobzhansky T. 1937. Genetics and the origin of species.

. New York: Columbia University Press.

Douglas MR, Brunner PC, Bernatchez L. 1999. Do assemblages of *Coregonus* (Teleostei: Salmoniformes) in the Central Alpine region of Europe represent species flocks? Molecular Ecology 8:589–603.

Dubin MJ, Zhang P, Meng D, Remigereau M-S, Osborne EJ, Paolo Casale F, Drewe P, Kahles A, Jean

G, Vilhjálmsson B, et al. 2015. DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. eLife 4:e05255.

Duforet-Frebourg N, Luu K, Laval G, Bazin É, Blum MGB. 2016. Detecting Genomic Signatures of Natural Selection with Principal Component Analysis: Application to the 1000 Genomes Data. Molecular Biology and Evolution 33:1082–1093.

Duncan EJ, Gluckman PD, Dearden PK. 2014. Epigenetics, plasticity, and evolution: How do we link epigenetic change to phenotype? J. Exp. Zool. B Mol. Dev. Evol. 322:208–220.

Duranton M, Allal F, Fraïsse C, Bierne N, Bonhomme F, Gagnaire P-A. 2018. The origin and remolding of genomic islands of differentiation in the European sea bass. Nature Communications 9:2518.

Edmands S. 1999. Heterosis and Outbreeding Depression in Interpopulation Crosses Spanning a Wide Range of Divergence. Evolution 53:1757.

Ehrenreich IM, Pfennig DW. 2016. Genetic assimilation: a review of its potential proximate causes and evolutionary consequences. Ann Bot 117:769–779.

Ellegren H, Smeds L, Burri R, Olason PI, Backström N, Kawakami T, Künstner A, Mäkinen H, Nadachowska-Brzyska K, Qvarnström A, et al. 2012. The genomic landscape of species divergence in Ficedula flycatchers. Nature 491:756–760.

Ellison CK, Burton RS. 2008. INTERPOPULATION HYBRID BREAKDOWN MAPS TO THE MITOCHONDRIAL GENOME. Evolution 62:631–638.

Ellison CK, Wiley C, Shaw KL. 2011. The genetics of speciation: genes of small effect underlie sexual isolation in the Hawaiian cricket Laupala. J. Evol. Biol. 24:1110–1119.

Elmer KR, Lehtonen TK, Kautt AF, Harrod C, Meyer A. 2012. Rapid sympatric ecological differentiation of crater lake cichlid fishes within historic times. BMC Biology 10:70.

Elmer KR, Meyer A. 2011. Adaptation in the age of ecological genomics: insights from parallelism and convergence. Trends in Ecology & Evolution 26:298–306.

Endler JA. 1986. Natural selection in the wild. Princeton Univ. Press: Princeton.

Erales J, Marchand V, Panthu B, Gillot S, Belin S, Ghayad SE, Garcia M, Laforêts F, Marcel V, Baudin-Baillieu A, et al. 2017. Evidence for rRNA 2′-O-methylation plasticity: Control of intrinsic translational capabilities of human ribosomes. Proceedings of the National Academy of Sciences 114:12934–12939.

Evans ML, Bernatchez L. 2012. Oxidative phosphorylation gene transcription in whitefish species pairs reveals patterns of parallel and nonparallel physiological divergence. J. Evol. Biol. 25:1823–1834.

Excoffier L, Dupanloup I, Huerta-SÃ nchez E, Sousa VC, Foll M. 2013. Robust Demographic Inference from Genomic and SNP Data.Akey JM, editor. PLoS Genet 9:e1003905–e1003917.

Eyre-Walker A. 2006. The Distribution of Fitness Effects of New Deleterious Amino Acid Mutations in Humans. Genetics 173:891–900.

Fagegaltier D, Bougé A-L, Berry B, Poisot E, Sismeiro O, Coppée J-Y, Théodore L, Voinnet O, Antoniewski C. 2009. The endogenous siRNA pathway is involved in heterochromatin formation in Drosophila. Proc. Natl. Acad. Sci. U.S.A. 106:21258–21263.

Faria R, Navarro A. 2010. Chromosomal speciation revisited: rearranging theory with pieces of evidence. Trends in Ecology & Evolution 25:660–669.

Fay JC, Wittkopp PJ. 2007. Evaluating the role of natural selection in the evolution of gene regulation. Heredity 100:191–199.

Feder JL, Egan SP, Nosil P. 2012a. The genomics of speciation-with-gene-flow. Trends in Genetics 28:342–350.

Feder JL, Egan SP, Nosil P. 2012b. The genomics of speciation-with-gene-flow. Trends in Genetics 28:342–350.

Feder JL, Nosil P. 2010. The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. Evolution 64:1729–1747.

Feder JL, Xie X, Rull J, Velez S, Forbes A, Leung B, Dambroski H, Filchak KE, Aluja M. 2005. Mayr, Dobzhansky, and Bush and the complexities of sympatric speciation in Rhagoletis. Proceedings of the National Academy of Sciences 102 Suppl 1:6573–6580.

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

Felsenstein J. 1981. Skepticism Towards Santa Rosalia, or Why are There so Few Kinds of Animals? Evolution 35:124.

Feng S, Cokus SJ, Zhang X, Chen P-Y, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, et al. 2010. Conservation and divergence of methylation patterning in plants and animals. Proc. Natl. Acad. Sci. U.S.A. 107:8689–8694.

Feulner PGD, Seehausen O. 2018. Genomic insights into the vulnerability of sympatric whitefish species flocks. Molecular Ecology:mec.14977.

Fisher RA. 1930. The Genetical Theory of Natural Selection. Oxford University Press.

Fishman L, Kelly AJ, Willis JH. 2002. Minor quantitative trait loci underlie floral traits associated with mating system divergence in Mimulus. Evolution 56:2138–2155.

Fishman L, Willis JH. 2006. A cytonuclear incompatibility causes anther sterility in Mimulus hybrids. Evolution 60:1372–1381.

Ford EE, Grimmer MR, Stolzenburg S, Bogdanović O, de Mendoza A, Farnham PJ, Blancafort P, Lister R. 2017. Frequent lack of repressive capacity of promoter DNA methylation identified through genome-wide epigenomic manipulation. :1–43.

Forester BR, Lasky JR, Wagner HH, Urban DL. 2018. Comparing methods for detecting multilocus adaptation with multivariate genotype-environment associations. Molecular Ecology 27:2215–2233.

Gagnaire P-A, Gaggiotti OE. 2016. Detecting polygenic selection in marine populations by combining population genomics and quantitative genetics approaches. Current Zoology 62:603–616.

Gagnaire P-A, Normandeau E, Bernatchez L. 2012. Comparative genomics reveals adaptive protein evolution and a possible cytonuclear incompatibility between European and American Eels. Molecular Biology and Evolution 29:2909–2919.

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

Gagnaire P-A, Pavey SA, Normandeau E, Bernatchez L. 2013. The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by RAD sequencing. Evolution 67:2483–2497.

Gagnon MC, Turgeon J. 2011. Sexual conflict in Gerris gillettei (Insecta: Hemiptera): intraspecific intersexual correlated morphology and experimental assessment of behaviour and fitness. J. Evol. Biol. 24:1505–1516.

Garrison E, 2012. Vcflib: A C++ library for parsing and manipulating VCF files.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. arXiv

arXiv:1207.3907.

Gibson JD, Niehuis O, Verrelli BC, Gadau J. 2010. Contrasting patterns of selective constraints in nuclear-encoded genes of the oxidative phosphorylation pathway in holometabolous insects and their possible role in hybrid breakdown in Nasonia. Heredity 104:310–317.

Glassberg EC, Gao Z, Harpak A, Lan X, Pritchard JK. 2018. Measurement of selective constraint on human gene expression. :1–18.

Goentoro L, Kirschner MW. 2009. Evidence that Fold-Change, and Not Absolute Level, of β-Catenin Dictates Wnt Signaling. Molecular Cell 36:872–884.

Gompert Z, Lucas LK, Nice CC, Fordyce JA, Forister ML, Buerkle CA. 2012. GENOMIC REGIONS WITH A HISTORY OF DIVERGENT SELECTION AFFECT FITNESS OF HYBRIDS BETWEEN TWO BUTTERFLY SPECIES. Evolution 66:2167–2181.

Gouy A, Daub JT, Excoffier L. 2017. Detecting gene subnetworks under selection in biological pathways. Nucl. Acids Res.:1–11.

Greenberg DA, Mooers AØ. 2017. Linking speciation to extinction: Diversification raises contemporary extinction risk in amphibians. Evolution Letters 1:40–48.

Griswold CK. 2006. Gene flow's effect on the genetic architecture of a local adaptation and its

consequences for QTL analyses. Heredity 96:445–453.

Guerrero RF, Hahn MW. 2017. Speciation as a sieve for ancestral polymorphism. Molecular Ecology 26:5362–5368.

Gunter HM, Schneider RF, Karner I, Sturmbauer C, Meyer A. 2017. Molecular investigation of genetic assimilation during the rapid adaptive radiations of East African cichlid fishes. Molecular Ecology 26:6634–6653.

Guo W, Fiziev P, Yan W, Cokus S, Sun X, Zhang MQ, Chen P-Y, Pellegrini M. 2013. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics 14:774–778.

Guo W, Zhu P, Pellegrini M, Zhang MQ, Wang X, Ni Z. 2017. CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data.Birol I, editor. Bioinformatics 34:381–387.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data.McVean G, editor. PLoS Genet 5:e1000695.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494–1512.

Haldane JBS. 1922. Sex ratio and unisexual sterility in hybrid animals. Journal of Genetics 12:101–109.

Haldane JBS. 1927. A Mathematical Theory of Natural and Artificial Selection, Part V: Selection and Mutation. Mathematical Proceedings of the Cambridge Philosophical Society 23:838.

Haller BC, De Léon LF, Rolshausen G, Gotanda KM, Hendry AP. 2012. Magic traits: distinguishing the important from the trivial. Trends in Ecology & Evolution 27:4–5–authorreply5–6.

Han F, Lamichhaney S, Grant BR, Grant PR, Andersson L, Webster MT. 2017. Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. Genome Research 27:1004–1015.

Harr B. 2006. Genomic islands of differentiation between house mouse subspecies. Genome Research 16:730–737.

Harrison PW, Wright AE, Mank JE. 2012. The evolution of gene expression and the transcriptome–phenotype relationship. Seminars in Cell and Developmental Biology 23:222–229.

Harrison RG, Larson EL. 2014. Hybridization, introgression, and the nature of species boundaries. Journal of Heredity 105 Suppl 1:795–809.

Harrison RG, Larson EL. 2016a. Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones.Abbott RJ, Barton NH, Good JM, editors. Molecular Ecology 25:2454–2466.

Harrison RG, Larson EL. 2016b. Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. Molecular Ecology:n/a–n/a.

Harrison RG. 1990. Hybrid zones: windows on evolutionary process. Oxford Surveys in Evolutionary Biology 7:69–128.

Harrisson KA, Amish SJ, Pavlova A, Narum S, Telonis-Scott M, Rourke ML, Lyon J, Tonkin Z, Gilligan DM, Ingram BA, et al. 2017. Signatures of polygenic adaptation associated with climate across the range of a threatened fish species with high genetic connectivity. Molecular Ecology.

Häkli K, Ostbye K, Kahilainen KK, Amundsen P-A, Praebel K. 2018. Diversifying selection drives parallel evolution of gill raker number and body size along the speciation continuum of European whitefish. Ecology and Evolution 8:2617–2631.

Hereford J. 2009. A quantitative survey of local adaptation and fitness trade-offs. Am Nat 173:579–588.

Herman A, Brandvain Y, Weagley J, Jeffery WR, Keene AC, Kono TJY, Bilandzija H, Borowsky R, Espinasa L, O'Quin K, et al. 2018. The role of gene flow in rapid and repeated evolution of cave-related traits in Mexican tetra, Astyanax mexicanus. Molecular Ecology 27:4397–4416.

Hermisson J, Pennings PS. 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169:2335–2352.

Hermisson J, Wagner GP. 2004. The population genetic theory of hidden variation and genetic robustness. Genetics 168:2271–2284.

Hewitt G. 2000. The genetic legacy of the Quaternary ice ages. Nature 405:907–913.

Hewitt GM. 1996. Some genetic consequences of ice ages, and their role in divergence and speciation. Biological Journal of the Linnean Society 58:247–276.

Hewitt GM. 2001. Speciation, hybrid zones and phylogeography—or seeing genes in space and time. Molecular Ecology 10:537–549.

Hewitt GM. 2004. Genetic consequences of climatic oscillations in the Quaternary. Philosophical Transactions of the Royal Society of London B: Biological Sciences 359:183–95–discussion195.

Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. Genetics 167:747–760.

Hey J, Nielsen R. 2007. Integration within the Felsenstein Equation for Improved Markov Chain Monte Carlo Methods in Population Genetics. Proc. Natl. Acad. Sci. U.S.A. 104:2785–2790.

Hietpas RT, Bank C, Jensen JD, Bolon DNA. 2013. SHIFTING FITNESS LANDSCAPES IN RESPONSE TO ALTERED ENVIRONMENTS. Evolution 67:3512–3522.

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. Genet. Res. 8:269–294.

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.

Hosaka A, Kakutani T. 2018. Transposable elements, genome evolution and transgenerational epigenetic variation. Current Opinion in Genetics & Development 49:43–48.

Hudson AG, Vonlanthen P, Seehausen O. 2010. Rapid parallel adaptive radiations from a single hybridogenic ancestral population. Proc. Biol. Sci. 278:58–66.

Irwin D, Price T. 1999. Sexual imprinting, learning and speciation. Heredity 82 Pt 4:347–354.

Jacobs A, Carruthers M, Yurchenko A, Gordeeva N, Alekseyev S, Hooker O, Leong J, Rondeau E, Koop B, Adams C, et al. 2018. Convergence in form and function overcomes non-parallel evolutionary histories in Arctic charr. bioRxiv:1–84.

Jacobsen MW, Hansen MM, Orlando L, Bekkevold D, Bernatchez L, Willerslev E, Gilbert MTP. 2012. Mitogenome sequencing reveals shallow evolutionary histories and recent divergence time between morphologically and ecologically distinct European whitefish (Coregonus spp.). Molecular Ecology 21:2727–2742.

Jain K, Stephan W. 2017. Modes of Rapid Polygenic Adaptation. Molecular Biology and Evolution:1–7.

Jari Oksanen F, Blanchet G, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, OHara RB, Simpson GL, Solymos P, et al. 2018. vegan: Community Ecology Package. R package version 2.5-3. https://CRAN.R-project.org/package=vegan.

Jeukens J, Bernatchez L. 2011. Regulatory versus coding signatures of natural selection in a candidate gene involved in the adaptive divergence of whitefish species pairs (Coregonusspp.). Ecology and Evolution 2:258–271.

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

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

Jiggins CD, Naisbit RE, Coe RL, Mallet J. 2001. Reproductive isolation caused by colour pattern mimicry. Nature 411:302–305.

Jiggins CD. 2006. Sympatric speciation: why the controversy? Current Biology 16:R333–R334.

Johnson NA, Porter AH. 2000. Rapid speciation via parallel, directional selection on regulatory genetic pathways. Journal of Theoretical Biology 205:527–542.

Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11:94.

Jones CD. 2002. The genetic basis of larval resistance to a host plant toxin in Drosophila sechellia. Genet. Res. 78:225–233.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484:55–61.

Jones PA. 1999. The DNA methylation paradox. Trends in Genetics 15:34–37.

Jones PA. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature Publishing Group 13:484–492.

Kahilainen K, Østbye K. 2006. Morphological differentiation and resource polymorphism in three sympatric whitefish Coregonus lavaretus(L.) forms in a subarctic lake. Journal of Fish Biology 68:63–79.

Kampf C, Mardinoglu A, Fagerberg L, Hallström BM, Edlund K, Lundberg E, Pontén F, Nielsen J, Uhlen M. 2014. The human liver-specific proteome defined by transcriptomics and antibody-based profiling. The FASEB Journal 28:2901–2914.

Kassen R, Bataillon T. 2006. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. Nat Genet 38:484–488.

Kautt AF, Machado-Schiaffino G, Meyer A. 2016. Multispecies Outcomes of Sympatric Speciation after Admixture with the Source Population in Two Radiations of Nicaraguan Crater Lake Cichlids.Payseur BA, editor. PLoS Genet 12:e1006157–33.

Kawecki TJ. Conceptual issues in local adaptation. Ecology Letters 7:1225–1241.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge University Press. Cambridge, UK

Kirkpatrick M, Ravigné V. 2002. Speciation by Natural and Sexual Selection: Models and Experiments. Am Nat 159:S22–S35.

Klironomos FD, Berg J, Collins S. 2013. How epigenetic mutations can affect genetic evolution: model and mechanism. Bioessays 35:571–578.

Kopp M, Hermisson J. 2009. The Genetic Basis of Phenotypic Adaptation II: The Distribution of Adaptive Substitutions in the Moving Optimum Model. Genetics 183:1453–1476.

Korthauer K, Irizarry RA. 2018. Genome-wide repressive capacity of promoter DNA methylation is revealed through epigenomic manipulation. :1–27.

Krogh N, Jansson MD, Häfner SJ, Tehler D, Birkedal U, Christensen-Dalsgaard M, Lund AH, Nielsen H. 2016. Profiling of 2'- O-Me in human rRNA reveals a subset of fractionally modified positions and provides evidence for ribosome heterogeneity. Nucl. Acids Res. 44:7884–7895.

Kruuk LE, Baird SJ, Gale KS, Barton NH. 1999. A comparison of multilocus clines maintained by environmental adaptation or by selection against hybrids. Genetics 153:1959–1971.

Labruna MB, Naranjo V, Mangold AJ, Thompson C, Estrada-Peña A, Guglielmone AA, Jongejan F, la Fuente de J. 2009. Allopatric speciation in ticks: genetic and reproductive divergence between geographic strains of Rhipicephalus (Boophilus) microplus. BMC Evol Biol 9:46–12.

Landry L, Bernatchez L. 2010. Role of epibenthic resource opportunities in the parallel evolution of lake whitefish species pairs (Coregonus sp.). J. Evol. Biol. 23:2602–2613.

Landry L, Vincent WF, Bernatchez L. 2007. Parallel evolution of lake whitefish dwarf ecotypes in association with limnological features of their adaptive landscape. J. Evol. Biol. 20:971–984.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Meth 9:357–359.

Laporte M, Dalziel AC, Martin N, Bernatchez L. 2016a. Adaptation and acclimation of traits associated with swimming capacity in Lake Whitefish (*Coregonus clupeaformis*) ecotypes. BMC Evol Biol 16:1–13.

Laporte M, Dalziel AC, Martin N, Bernatchez L. 2016b. Adaptation and acclimation of traits associated with swimming capacity in Lake Whitefish (coregonus clupeaformis) ecotypes. BMC Evol Biol 16:160.

Laporte M, Pavey SA, Rougeux C, Pierron F, Lauzent M, Budzinski H, Labadie P, Geneste E, Couture P, Baudrimont M, et al. 2016. RAD sequencing reveals within-generation polygenic selection in response to anthropogenic organic and metal contamination in North Atlantic Eels. Molecular Ecology 25:219–237.

Laporte M, Rogers SM, Dion-Côté A-M, Normandeau E, Gagnaire P-A, Dalziel AC, Chebib J, Bernatchez L. 2015a. RAD-QTL Mapping Reveals Both Genome-Level Parallelism and Different Genetic Architecture Underlying the Evolution of Body Shape in Lake Whitefish (Coregonus clupeaformis) Species Pairs. G3 (Bethesda) 5:1481–1491.

Laporte M, Rogers SM, Dion-Côté A-M, Normandeau E, Gagnaire P-A, Dalziel AC, Chebib J, Bernatchez L. 2015b. RAD-QTL Mapping Reveals Both Genome-Level Parallelism and Different Genetic Architecture Underlying the Evolution of Body Shape in Lake Whitefish (*Coregonus clupeaformis*) Species Pairs. G3 (Bethesda) 5:1481–1491.

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

Le Luyer J, Laporte M, Beacham TD, Kaukinen KH, Withler RE, Leong JS, Rondeau EB, Koop BF, Bernatchez L. 2017. Parallel epigenetic modifications induced by hatchery rearing in a Pacific salmon. Proceedings of the National Academy of Sciences 1:201711229–6.

Le Moan A, Gagnaire PA, Bonhomme F. 2016. Parallel genetic divergence among coastal-marine ecotype pairs of European anchovy explained by differential introgression after secondary contact. Molecular Ecology 25:3187–3202.

Le Rouzic A, Carlborg O. 2008. Evolutionary potential of hidden genetic variation. Trends in Ecology & Evolution 23:33–37.

Ledon-Rettig CC, Pfennig DW, Crespi EJ. 2010. Diet and hormonal manipulation reveal cryptic genetic variation: implications for the evolution of novel feeding strategies. Proc. Biol. Sci. 277:3569–3578.

Ledon-Rettig CC, Pfennig DW, Nascone-Yoder N. 2008. Ancestral variation and the potential for genetic accommodation in larval amphibians: implications for the evolution of novel feeding strategies. Evolution & Development 10:316–325.

Legendre L, Legendre L. 1998. Numerical ecology. Elsevier. Amsterdam

Lenormand T. 2002. Gene flow and the limits to natural selection. Trends in Ecology & Evolution 17:183–189.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Lohmueller KE. 2014. The impact of population demography and selection on the genetic architecture of complex traits.Williams SM, editor. PLoS Genet 10:e1004379.

Losos JB. 2011. Convergence, adaptation, and constraint. Evolution 65:1827–1840.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15:31–21.

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.

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.

Luikart G, Allendorf FW, Cornuet JM, Sherwin WB. 1998. Distortion of allele frequency distributions provides a test for recent population bottlenecks. Journal of Heredity 89:238–247.

Luu K, Bazin É, Blum MGB. 2017. pcadapt: an R package to perform genome scans for selection based on principal component analysis. Mol Ecol Resour 17:67–77.

Ma T, Wang K, Hu Q, Xi Z, Wan D, Wang Q, Feng J, Jiang D, Ahani H, Abbott RJ, et al. 2017. Ancient polymorphisms and divergence hitchhiking contribute to genomic islands of divergence within a poplar species complex. Proceedings of the National Academy of Sciences 5:201713288–201713288.

MacPherson A, Hohenlohe PA, Nuismer SL. 2015. Trait dimensionality explains widespread variation in local adaptation. Proc. Biol. Sci. 282:20141570–20141570.

Magoc T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27:2957–2963.

Maheshwari S, Barbash DA. 2011. The Genetics of Hybrid Incompatibilities. Annu. Rev. Genet. 45:331–355.

Mallet B, Martos F, Blambert L, Pailler T, Humeau L. 2014. Evidence for Isolation-by-Habitat among Populations of an Epiphytic Orchid Species on a Small Oceanic Island.Canestrelli D, editor. PLoS ONE 9:e87469–12.

Manceau M, Domingues VS, Linnen CR, Rosenblum EB, Hoekstra HE. 2010. Convergence in pigmentation at multiple levels: mutations, genes and function. Philosophical Transactions of the Royal Society of London B: Biological Sciences 365:2439–2450.

Manceau M, Domingues VS, Mallarino R, Hoekstra HE. 2011. The developmental role of Agouti in color pattern evolution. Science 331:1062–1065.

Marie Curie SPECIATION Network, Butlin R, Debelle A, Kerth C, Snook RR, Beukeboom LW, Castillo Cajas RF, Diao W, Maan ME, Paolucci S, et al. 2012. What do we need to know about speciation? Trends in Ecology & Evolution 27:27–39.

Martin CH, Cutler JS, Friel JP, Touokong CD, Coop G, Wainwright PC. 2015. Complex histories of repeated gene flow in Cameroon crater lake cichlids cast doubt on one of the clearest examples of sympatric speciation. Evolution 69:1406–1422.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal 17:10.

Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, Blaxter M, Manica A, Mallet J, Jiggins CD. 2013. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Research 23:1817–1828.

Matuszewski S, Hermisson J, Kopp M. 2014. FISHER'S GEOMETRIC MODEL WITH A MOVING OPTIMUM. Evolution 68:2571–2588.

Matuszewski S, Hermisson J, Kopp M. 2015. Catch Me if You Can: Adaptation from Standing Genetic Variation to a Moving Phenotypic Optimum. Genetics 200:1255–1274.

Mayr E. 1942. Systematics and the Origin of Species from the Viewpoint of a Zoologist. (New York Columbia University Press, editor.).

Mayr E. 1954. Geographic Speciation in Tropical Echinoids. Evolution 8:1.

Mähler N, Wang J, Terebieniec BK, Ingvarsson PK, Street NR, Hvidsten TR. 2017. Gene co-expression network connectivity is an important determinant of selective constraint.Springer NM, editor. PLoS Genet 13:e1006402–e1006433.

McCairns RJS, 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 JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652–654.

McGee LW, Sackman AM, Morrison AJ, Pierce J, Anisman J, Rokyta DR. 2016. Synergistic Pleiotropy Overrides the Costs of Complexity in Viral Adaptation. Genetics 202:285–295.

McGirr JA, Martin CH. 2018. Parallel evolution of gene expression between trophic specialists despite divergent genotypes and morphologies. Evolution Letters 5:40–14.

McKinnon JS, Mori S, Blackman BK, David L, Kingsley DM, Jamieson L, Chou J, Schluter D. 2004. Evidence for ecology's role in speciation. Nature 429:294–298.

McPhail JD. 1992. Ecology and evolution of sympatric sticklebacks (Gasterosteus): evidence for a species-pair in Paxton Lake, Texada Island, British Columbia. Canadian Journal of Zoology 70:361–369.

Meier JI, Marques DA, Mwaiko S, Wagner CE, Excoffier L, Seehausen O. 2017. Ancient hybridization fuels rapid cichlid fish adaptive radiations. Nature Communications 8:14363.

Meier JI, Marques DA, Wagner CE, Excoffier L, Seehausen O. 2018. Genomics of parallel ecological speciation in Lake Victoria cichlids. Molecular Biology and Evolution:1–37.

Meier JI, Sousa VC, Marques DA, Selz OM, Wagner CE, Excoffier L, Seehausen O. 2016. Demographic modeling with whole genome data reveals parallel origin of similar Pundamilia cichlid species after hybridization. Molecular Ecology.

Meier JI, Sousa VC, Marques DA, Selz OM, Wagner CE, Excoffier L, Seehausen O. 2017. Demographic modelling with whole-genome data reveals parallel origin of similar *Pundamilia* cichlid species after hybridization. Molecular Ecology 26:123–141.

Meng D, Dubin M, Zhang P, Osborne EJ, Stegle O, Clark RM, Nordborg M. 2016. Limited Contribution of DNA Methylation Variation to Expression Regulation in Arabidopsis thaliana. PLoS Genet 12:e1006141.

Merrill RM, Dasmahapatra KK, Davey JW, Dell'Aglio DD, Hanly JJ, Huber B, Jiggins CD, Joron M, Kozak KM, Llaurens V, et al. 2015. The diversification of Heliconius butterflies: what have we learned in 150 years? J. Evol. Biol. 28:1417–1438.

Merritt C, Rasoloson D, Ko D, Seydoux G. 2008. 3' UTRs are the primary regulators of gene expression in the C. elegans germline. Current Biology 18:1476–1482.

Metzger DCH, Schulte PM. 2017. Persistent and plastic effects of temperature on DNA methylation across the genome of threespine stickleback (Gasterosteus aculeatus). Proc. Biol. Sci. 284:20171667–20171667.

Mérot C, Salazar C, Merrill RM, Jiggins CD, Joron M. 2017. What shapes the continuum of reproductive isolation? Lessons from Heliconiusbutterflies. Proc. Biol. Sci. 284:20170335–10.

Mihola O, Trachtulec Z, Vlcek C, Schimenti JC, Forejt J. 2009. A mouse speciation gene encodes a meiotic histone H3 methyltransferase. Science 323:373–375.

Moen EL, Zhang X, Mu W, Delaney SM, Wing C, McQuade J, Myers J, Godley LA, Dolan ME, Zhang W. 2013. Genome-wide variation of cytosine modifications between European and African populations and the implications for complex traits. Genetics 194:987–996.

Muller M. 1942. Isolating mechanisms, evolution, and temperature.

. Biol Symp. 6:71–125.

Muschick M, Barluenga M, Salzburger W, Meyer A. 2011. Adaptive phenotypic plasticity in the Midas cichlid fish pharyngeal jaw and its relevance in adaptive radiation. BMC Evol Biol 11:116.

Nadachowska-Brzyska K, Burri R, Olason PI, Kawakami T, Smeds LA, Ellegren H. 2013. Demographic Divergence History of Pied Flycatcher and Collared Flycatcher Inferred from Whole-Genome Re-sequencing Data.Payseur BA, editor. PLoS Genet 9:e1003942.

Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, Baxter SW, Quail MA, Joron M, ffrench-Constant RH, Blaxter ML, et al. 2012. Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 367:343–353.

Nagel L, Schluter D. 1998. BODY SIZE, NATURAL SELECTION, AND SPECIATION IN STICKLEBACKS. Evolution 52:209–218.

Nagy O, Nuez I, Savisaar R, Peluffo AE, Yassin A, Lang M, Stern DL, Matute D, David JR, Courtier-Orgogozo V. 2018. Correlated Evolution of two Sensory Organs via a Single Cis-Regulatory Nucleotide Change. :1–61.

Naisbit RE, Jiggins CD, Mallet J. 2001. Disruptive sexual selection against hybrids contributes to speciation between Heliconius cydno and Heliconius melpomene. Proc. Biol. Sci. 268:1849–1854.

Nei M. 1972. Genetic Distance between Populations. Am Nat 106:283–292.

Nei M. 1987. Molecular Evolutionary Genetics. Columbia University Press

Nelson TC, Cresko WA. 2018. Ancient genomic variation underlies repeated ecological adaptation in young stickleback populations. Evolution Letters 114:7061–13.

Nishikawa K, Kinjo AR. 2018. Mechanism of evolution by genetic assimilation : Equivalence and independence of genetic mutation and epigenetic modulation in phenotypic expression. Biophys Rev 10:667–676.

Noor MA, Grams KL, Bertucci LA, Reiland J. 2001. Chromosomal inversions and the reproductive isolation of species. Proceedings of the National Academy of Sciences 98:12084–12088.

Noor MAF, Bennett SM. 2010. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. Heredity 104:418–418.

Nosil P, Harmon LJ, Seehausen O. 2009. Ecological explanations for (incomplete) speciation. Trends in Ecology & Evolution 24:145–156.

Nosil P, Schluter D. 2011. The genes underlying the process of speciation. Trends in Ecology & Evolution 26:160–167.

Nosil P. 2008. Speciation with gene flow could be common. Molecular Ecology 17:2103–2106.

Nosil P. 2012. Ecological speciation. Oxford University Press.

O'Brown NM, Summers BR, Jones FC, Brady SD, Kingsley DM. 2015. A recurrent regulatory change underlying altered expression and Wnt response of the stickleback armor plates gene EDA. eLife 4:e05290.

Oppenheim SJ, Gould F, Hopper KR. 2012. The genetic architecture of a complex ecological trait: host plant use in the specialist moth, Heliothis subflexa. Evolution 66:3336–3351.

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

Orr HA. 1998. THE POPULATION GENETICS OF ADAPTATION: THE DISTRIBUTION OF FACTORS FIXED DURING ADAPTIVE EVOLUTION. Evolution 52:935–949.

Orr HA. 2005a. THE PROBABILITY OF PARALLEL EVOLUTION. Evolution 59:216–220.

Orr HA. 2005b. The probability of parallel evolution. Evolution 59:216–220.

Orr HA. 2005c. The genetic theory of adaptation: a brief history. Nature Reviews Genetics.

Paaby AB, Rockman MV. 2013. The many faces of pleiotropy. Trends in Genetics 29:66–73.

Pai AA, Pritchard JK, Gilad Y. 2015. The Genetic and Mechanistic Basis for Variation in Gene Regulation. PLoS Genet 11:e1004857.

Park JH, Stoffers DA, Nicholls RD, Simmons RA. 2008. Development of type 2 diabetes following intrauterine growth retardation in rats is associated with progressive epigenetic silencing of Pdx1. J. Clin. Invest. 118:2316–2324.

Pasquier J, Cabau C, Nguyen T, Jouanno E, Severac D, Braasch I, Journot L, Pontarotti P, Klopp C, Postlethwait JH, et al. 2016. Gene evolution and gene expression after whole genome duplication in fish: the PhyloFish database. BMC Genomics 17:368.

Pavey SA, Collin H, Nosil P, Rogers SM. 2010. The role of gene expression in ecological speciation. Annals of the New York Academy of Sciences 1206:110–129.

Payseur BA. 2010. Using differential introgression in hybrid zones to identify genomic regions involved in speciation. Mol Ecol Resour 10:806–820.

Pearce RJ, Pota H, Evehe M-SB, Bâ E-H, Mombo-Ngoma G, Malisa AL, Ord R, Inojosa W, Matondo A, Diallo DA, et al. 2009. Multiple Origins and Regional Dispersal of Resistant dhps in African

Plasmodium falciparum Malaria.Seidlein von L, editor. PLoS Med 6:e1000055–15.

Peischl S, Dupanloup I, Kirkpatrick M, Excoffier L. 2013. On the accumulation of deleterious mutations during range expansions. Molecular Ecology 22:5972–5982.

Pennings PS. 2012. Standing Genetic Variation and the Evolution of Drug Resistance in HIV.De Boer RJ, editor. PLoS Comput Biol 8:e1002527.

Perry GH, Foll M, Grenier J-C, Patin E, Nédélec Y, Pacis A, Barakatt M, Gravel S, Zhou X, Nsobya SL, et al. 2014. Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. Proc. Natl. Acad. Sci. U.S.A. 111:E3596–E3603.

Pfennig DW, Wund MA, Snell-Rood EC, Cruickshank T, Schlichting CD, Moczek AP. 2010. Phenotypic plasticity's impacts on diversification and speciation. Trends in Ecology & Evolution 25:459–467.

Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet 8:e1002967.

Pigeon D, Chouinard A, Bernatchez L. 1997a. Multiple Modes of Speciation Involved in the Parallel Evolution of Sympatric Morphotypes of Lake Whitefish (Coregonus clupeaformis, Salmonidae). Evolution 51:196.

Pigeon D, Chouinard A, Bernatchez L. 1997b. Multiple modes of speciation involved in the parallel evolution of sympatric morphotypes of lake whitefish (*Coregonus clupeaformis*, Salmonidae). Evolution 51:196.

Pigliucci M, Murren CJ, Schlichting CD. 2006. Phenotypic plasticity and evolution by genetic assimilation. Journal of Experimental Biology 209:2362–2367.

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M, et al. 2012. Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. The Plant Genome Journal 5:103–111.

Praebel K, Knudsen R, Siwertsson A, Karhunen M, Kahilainen KK, Ovaskainen O, stbye K, Peruzzi S, Fevolden S-E, Amundsen P-A. 2013. Ecological speciation in postglacial European whitefish: rapid adaptive radiations into the littoral, pelagic, and profundal lake habitats. Ecology and Evolution 3:4970–4986.

Presgraves DC. 2010. The molecular evolutionary basis of species formation. Nature Publishing Group 11:175–180.

Ramsey J, Schemske DW. 1998. PATHWAYS, MECHANISMS, AND RATES OF POLYPLOID FORMATION IN FLOWERING PLANTS. Annu. Rev. Ecol. Syst. 29:467–501.

Ravinet M, Westram A, Johannesson K, Butlin R, André C, Panova M. 2015. Shared and nonshared genomic divergence in parallel ecotypes of Littorina saxatilisat a local scale. Molecular Ecology 25:287–305.

Rebeiz M, Pool JE, Kassner VA, Aquadro CF, Carroll SB. 2009. Stepwise modification of a modular enhancer underlies adaptation in a Drosophila population. Science 326:1663–1667.

Renaut S, Bernatchez L. 2010. Transcriptome-wide signature of hybrid breakdown associated with intrinsic reproductive isolation in lake whitefish species pairs (Coregonus spp. Salmonidae). Heredity 106:1003–1011.

Renaut S, Bernatchez L. 2011. Transcriptome-wide signature of hybrid breakdown associated with intrinsic reproductive isolation in lake whitefish species pairs (Coregonus spp. Salmonidae). Heredity 106:1003–1011.

Renaut S, Grassa CJ, Yeaman S, Moyers BT, Lai Z, Kane NC, Bowers JE, Burke JM, Rieseberg LH. 2013. Genomic islands of divergence are not affected by geography of speciation in sunflowers. Nature Communications 4:1827.

Renaut S, Maillet N, Normandeau E, Sauvage C, Derome N, Rogers SM, Bernatchez L. 2012. Genome-wide patterns of divergence during speciation: the lake whitefish case study. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 367:354–363.

Renaut S, Nolte AW, Bernatchez L. 2009. Gene Expression Divergence and Hybrid Misexpression between Lake Whitefish Species Pairs (Coregonus spp. Salmonidae). Molecular Biology and Evolution 26:925–936.

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

Rey O, Danchin E, Mirouze M, Loot C, Blanchet S. 2016. Adaptation to Global Change: A Transposable Element–Epigenetics Perspective. Trends in Ecology & Evolution 31:1–13.

Richards E, Poelstra J, Martin C. 2017. Don't throw out the sympatric species with the crater lake water: fine-scale investigation of introgression provides weak support for functional role of secondary gene flow in one of the clearest examples of sympatric speciation. bioRxiv:1–44.

Richards EJ. 2006. Inherited epigenetic variation--revisiting soft inheritance. Nature Reviews Genetics 7:395–401.

Rieseberg LH. 2001. Chromosomal rearrangements and speciation. Trends in Ecology & Evolution 16:351–358.

Rise ML, Douglas SE, Sakhrani D, Williams J, Ewart KV, Rise M, Davidson WS, Koop BF, Devlin RH. 2006. Multiple microarray platforms utilized for hepatic gene expression profiling of GH transgenic coho salmon with and without ration restriction. J. Mol. Endocrinol. 37:259–282.

Roberts A, Pachter L. 2012. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Meth 10:71–73.

Robinson BW. 2013. Evolution of growth by genetic accommodation in Icelandic freshwater stickleback. Proceedings of the Royal Society B: Biological Sciences 280:20132197–20132197.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140.

Robison BD, Wheeler PA, Sundin K, Sikka P, Thorgaard GH. 2001. Composite interval mapping reveals a major locus influencing embryonic development rate in rainbow trout (Oncorhynchus mykiss). Journal of Heredity 92:16–22.

Rockman MV. 2012. The QTN program and the alleles that matter for evolution: all that's gold does not glitter. Evolution 66:1–17.

Roessler K, Takuno S, Gaut BS. 2016. CG Methylation Covaries with Differential Gene Expression between Leaf and Floral Bud Tissues of Brachypodium distachyon.Shiu S-H, editor. PLoS ONE 11:e0150002–e0150020.

Roesti M, Gavrilets S, Hendry AP, Salzburger W, Berner D. 2014. The genomic signature of parallel adaptation from shared genetic variation. Molecular Ecology 23:3944–3956.

Roesti M, Hendry AP, Salzburger W, Berner D. 2012. Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. Molecular Ecology 21:2852–2862.

Rogers S, Bernatchez L. 2007a. 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, Bernatchez L. 2004. FAST-TRACK: Integrating QTL mapping and genome scans towards the characterization of candidate loci under parallel selection in the lake whitefish (*Coregonus clupeaformis*). Molecular Ecology 14:351–361.

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

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

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

Rogers SM, Tamkee P, Summers B, Balabahadra S, Marks M, Kingsley DM, Schluter D. 2012. GENETIC SIGNATURE OF ADAPTIVE PEAK SHIFT IN THREESPINE STICKLEBACK. Evolution 66:2439–2450.

Rose NH, Bay RA, Morikawa MK, Palumbi SR. 2018. Polygenic evolution drives species divergence and climate adaptation in corals. Evolution 72:82–94.

Rosenblum EB, Parent CE, Brandt EE. 2014. The Molecular Basis of Phenotypic Convergence. Annu.

Rev. Ecol. Evol. Syst. 45:203–226.

Rougemont Q, Gagnaire P-A, Perrier C, Genthon C, Besnard A-L, Launey S, Evanno G. 2017. Inferring the demographic history underlying parallel genomic divergence among pairs of parasitic and nonparasitic lamprey ecotypes. Molecular Ecology 26:142–162.

Rougeux C, Bernatchez L, Gagnaire P-A. 2017. Modeling the Multiple Facets of Speciation-with-Gene-Flow toward Inferring the Divergence History of Lake Whitefish Species Pairs (Coregonus clupeaformis). Genome Biology and Evolution 9:2057–2074.

Rougeux C, Gagnaire P-A, Praebel K, Seehausen O, Bernatchez L. 2018. Convergent transcriptomic landscapes under polygenic selection accompany inter-continental parallel evolution within a Nearctic Coregonus (Salmonidae) sister-species complex. bioRxiv:1–27.

Roux C, Fraïsse C, Romiguier J, Anciaux Y, Galtier N, Bierne N. 2016a. Shedding light on the grey zone of speciation along a continuum of genomic divergence.

Roux C, Fraïsse C, Romiguier J, Anciaux Y, Galtier N, Bierne N. 2016b. Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence.Moritz C, editor. PLoS Biol 14:e2000234–22.

Roux C, Tsagkogeorga G, Bierne N, Galtier N. 2013. Crossing the Species Barrier: Genomic Hotspots of Introgression between Two Highly Divergent Ciona intestinalis Species. Molecular Biology and Evolution 30:1574–1587.

Rundle HD, Nagel L, Wenrick Boughman J, Schluter D. 2000. Natural selection and parallel speciation in sympatric sticklebacks. Science 287:306–308.

Savolainen O, Lascoux M, Merilä J. 2013a. Ecological genomics of local adaptation. Nature Publishing Group 14:807–820.

Savolainen O, Lascoux M, Merilä J. 2013b. Ecological genomics of local adaptation. Nature Reviews Genetics 14:807–820.

Schluter D, Clifford EA, Nemethy M, McKinnon JS. 2004. Parallel evolution and inheritance of quantitative traits. Am Nat 163:809–822.

Schluter D, Conte GL. 2009. Genetics and ecological speciation. Proc. Natl. Acad. Sci. U.S.A. 106 Suppl 1:9955–9962.

Schluter D. 1996. Ecological Speciation in Postglacial Fishes. Philosophical Transactions of the Royal Society of London B: Biological Sciences 351:807–814.

Schluter D. 2000. Ecological character displacement in adaptive radiation. Am Nat.

Schluter D. 2001. Ecology and the origin of species. Trends in Ecology & Evolution 16:372–380.

Schrider DR, Kern AD. 2017. Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. Molecular Biology and Evolution 34:1863–1877.

Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. 2011. Global quantification of mammalian gene expression control. Nature 473:337–342.

Servedio MR, van Doorn GS, Kopp M, Frame AM, Nosil P. 2011. Magic traits in speciation: "magic" but not rare? Trends in Ecology & Evolution 26:389–397.

Shabalin AA. 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics 28:1353–1358.

Shaw KL, Mullen SP. 2014. Speciation Continuum. Journal of Heredity 105:741–742.

Signor SA, Nuzhdin SV. 2018. The Evolution of Gene Expression in cis and trans. Trends in Genetics:1–13.

Simons YB, Bullaughey K, Hudson RR, Sella G. 2018. A population genetic interpretation of GWAS findings for human quantitative traits. PLoS Biol 16:e2002985–20.

Slatkin M. 1978. Spatial patterns in the distributions of polygenic characters. Journal of Theoretical Biology 70:213–228.

Smadja C, Butlin RK. 2009. On the scent of speciation: the chemosensory system and its role in premating isolation. Heredity 102:77–97.

Smadja CM, Butlin RK. 2011. A framework for comparing processes of speciation in the presence of gene flow. Molecular Ecology 20:5123–5140.

Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet. Res. 23:23–35.

Smith SD. 2015. Pleiotropy and the evolution of floral integration. New Phytologist 209:80–85.

Song S, Dey DK, Holsinger KE. 2006. Differentiation among populations with migration, mutation, and drift: implications for genetic inference. Evolution 60:1–12.

Soria-Carrasco V, Gompert Z, Comeault AA, Farkas TE, Parchman TL, Johnston JS, Buerkle CA, Feder JL, Bast J, Schwander T, et al. 2014. Stick insect genomes reveal natural selection's role in parallel speciation. Science 344:738–742.

Sousa A, Magalhães S, Gordo I. 2011. Cost of Antibiotic Resistance and the Geometry of Adaptation. Molecular Biology and Evolution 29:1417–1428.

Sousa V, Hey J. 2013. Understanding the origin of species with genome-scale data: modelling gene flow. Nature Publishing Group 14:404–414.

Sousa VMC, Carneiro M, Ferrand N, Hey J. 2013. Identifying Loci Under Selection Against Gene Flow in Isolation-with-Migration Models. Genetics 194:211–233.

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

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A,

Lee C, et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315:848–853.

Sun X-J, Wang M-C, Zhang F-H, Kong X. 2018. An integrated analysis of genome-wide DNA methylation and gene expression data in hepatocellular carcinoma. FEBS Open Bio 8:1093–1103.

Swenson NG, Howard DJ. 2005. Clustering of Contact Zones, Hybrid Zones, and Phylogeographic Breaks in North America. Am Nat 166:581–591.

Tanay A, Regev A, Shamir R. 2005. Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. Proceedings of the National Academy of Sciences 102:7203–7208.

Tang H, Klopfenstein D, Pedersen B, Flick P, Sato K, Ramirez F, Yunes J, Mungall C. 2015. GOATOOLS: Tools for Gene Ontology.

Tang H, Lyons E, Town CD. 2015. Optical mapping in plant comparative genomics. Gigascience 4:3.

Taudt A, Colomé-Tatché M, Johannes F. 2016. Genetic sources of population epigenomic variation. Nature Publishing Group 17:319–332.

Taylor EB, Bentzen P. 1993. Evidence for Multiple Origins and Sympatric Divergence of Trophic Ecotypes of Smelt (Osmerus) in Northeastern North America. Evolution 47:813.

Taylor EB, Donald McPhail J. 2000. Historical contingency and ecological determinism interact to prime speciation in sticklebacks, Gasterosteus. Proceedings of the Royal Society B: Biological Sciences 267:2375–2384.

Taylor EB. 1999. Species pairs of north temperate freshwater fishes: evolution, taxonomy, and conservation. Reviews in Fish Biology and Fisheries 9:299–324.

Tenaillon O. 2014. The Utility of Fisher's Geometric Model in Evolutionary Genetics. Annu. Rev. Ecol. Evol. Syst. 45:179–201.

Tine M, Kuhl H, Gagnaire P-A, Louro B, Desmarais E, Martins RST, Hecht J, Knaust F, Belkhir K, Klages S, et al. 2014. European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nature Communications 5:1–10.

TransDecoder. 2016. https://transdecoder.github.io/. Accessed Dec 2016.

Trudel M, Tremblay A, Schetagne R, Rasmussen JB. 2001. Why are dwarf fish so small? An energetic analysis of polymorphism in lake whitefish (*Coregonus clupeaformis*). Can. J. Fish. Aquat. Sci. 58:394–405.

Tung J, Zhou X, Alberts SC, Stephens M, Gilad Y. 2014. The Genetic Architecture of Gene Expression Levels in Wild Baboons.

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

Turgeon J, Bernatchez L. 2003. Reticulate evolution and phenotypic diversity in North American ciscoes, <Emphasis Type="Italic">Coregonus</Emphasis> ssp. (Teleostei: Salmonidae): implications for the conservation of an evolutionary legacy. Conserv Genet 4:67–81.

Turner TL, Hahn MW, Nuzhdin SV. 2005. Genomic Islands of Speciation in Anopheles gambiae. PLoS Biol 3:e285.

Uebbing S, Künstner A, Mäkinen H, Backström N, Bolivar P, Burri R, Dutoit L, Mugal CF, Nater A, Aken B, et al. 2016. Divergence in gene expression within and between two closely related flycatcher species. Molecular Ecology 25:2015–2028.

Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, et al. 2015. Tissue-based map of the human proteome. Science 347:1260419–1260419.

Van Belleghem SM, Vangestel C, De Wolf K, De Corte Z, Möst M, Rastas P, De Meester L, Hendrickx F. 2018. Evolution at two time frames: Polymorphisms from an ancient singular divergence event fuel contemporary parallel evolution.Schierup MH, editor. PLoS Genet 14:e1007796–26.

Verta J-P, Landry CR, MacKay J. 2016. Dissection of expression-quantitative trait locus and allele specificity using a haploid/diploid plant system - insights into compensatory evolution of transcriptional regulation within populations. New Phytologist 211:159–171.

Via S. 2009. Natural selection in action during speciation. Proc. Natl. Acad. Sci. U.S.A. 106 Suppl 1:9939–9946.

Via S. 2012. Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 367:451–460.

Villa SM, Altuna JC, Ruff JS, Ruff JS, Beach AB, Beach AB, Mulvey LI, Mulvey LI, Poole EJ, poole EJ, et al. 2018. Experimental evolution of reproductive isolation from a single natural population. :1–30.

Vonlanthen P, Roy D, Hudson AG, Largiader CR, Bittner D, Seehausen O. 2009. Divergence along a steep ecological gradient in lake whitefish (Coregonussp.). J. Evol. Biol. 22:498–514.

Waddington CH. 1961. Genetic Assimilation. In: Vol. 10. Advances in Genetics. Elsevier. pp. 257–293.

Wagner GP, Kenney-Hunt JP, Pavlicev M, Peck JR, Waxman D, Cheverud JM. 2008. Pleiotropic scaling of gene effects and the "cost of complexity." Nature 452:470–472.

Wagner GP, Pavlicev M, Cheverud JM. 2007. The road to modularity. Nature Reviews Genetics 8:921–931.

Wagner GP, Zhang J. 2011. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. Nature Publishing Group 12:204–213.

Wagner GP. 1996. Homologues, natural kinds and the evolution of modularity. American Zoologist

36:36–43.

Wang M, Uebbing S, Ellegren H. 2017. Bayesian Inference of Allele-Specific Gene Expression Indicates Abundant Cis-Regulatory Variation in Natural Flycatcher Populations. Genome Biology and Evolution 9:1266–1279.

Wang Z, Liao BY, Zhang J. 2010. Genomic patterns of pleiotropy and the evolution of complexity. Proceedings of the National Academy of Sciences 107:18034–18039.

Weaver ICG, Cervoni N, Champagne FA, D'Alessio AC, Sharma S, Seckl JR, Dymov S, Szyf M, Meaney MJ. 2004. Epigenetic programming by maternal behavior. Nat Neurosci 7:847–854.

Welch JJ, Jiggins CD. 2014. Standing and flowing: the complex origins of adaptive variation. Molecular Ecology 23:3935–3937.

Welch JJ, Waxman D. 2003. MODULARITY AND THE COST OF COMPLEXITY. Evolution 57:1723–1734.

West-Eberhard M. 1989. Phenotypic Plasticity And The Origins Of Diversity. Annu. Rev. Ecol. Syst. 20:249–278.

West-Eberhard MJ. 2005. Developmental plasticity and the origin of species differences. Proceedings of the National Academy of Sciences 102 Suppl 1:6543–6549.

Westram AM, Galindo J, Alm Rosenblad M, Grahame JW, Panova M, Butlin RK. 2014. Do the same genes underlie parallel phenotypic divergence in different Littorina saxatilispopulations? Molecular Ecology 23:4603–4616.

Whitlock MC. 2003. Fixation probability and time in subdivided populations. Genetics 164:767–779.

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

Wittkopp PJ, Kalay G. 2011. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nature Reviews Genetics 13:59–69.

Wlliams GC. 1966. Adaptation and Natural Selection. Princeton University Press. Princeton

Wolf JBW, Ellegren H. 2017. Making sense of genomic islands of differentiation in light of speciation. Nature Reviews Genetics 18:87–100.

Wood CC, Foote CJ. 1996. Evidence for Sympatric Genetic Divergence of Anadromous and Nonanadromous Morphs of Sockeye Salmon (Oncorhynchus nerka). Evolution 50:1265.

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. Proc. Natl. Acad. Sci. U.S.A. 106:13875–13879.

Woods PJ, Müller R, Seehausen O. 2009. Intergenomic epistasis causes asynchronous hatch times in whitefish hybrids, but only when parental ecotypes differ. J. Evol. Biol. 22:2305–2319.

Wray GA. 2003. The Evolution of Transcriptional Regulation in Eukaryotes. Molecular Biology and

Evolution 20:1377–1419.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nature Publishing Group 8:206–216.

Wu C-I. 2001. The genic view of the process of speciation. J. Evol. Biol. 14:851–865.

Wu H, Wang C, Wu Z. 2013. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. Biostatistics 14:232–243.

Xie KT, Wang G, Thompson AC, Wucherpfennig JI, Reimchen TE, MacColl ADC, Schluter D, Bell MA, Vasquez KM, Kingsley DM. 2019. DNA fragility in the parallel evolution of pelvic reduction in stickleback fish. Science 363:81–84.

Yeaman S, Aeschbacher S, Bürger R. 2016. The evolution of genomic islands by increased establishment probability of linked alleles.Abbott RJ, Barton NH, Good JM, editors. Molecular Ecology 25:2542–2558.

Yeaman S, Gerstein AC, Hodgins KA, Whitlock MC. 2018. Quantifying how constraints limit the diversity of viable routes to adaptation. PLoS Genet 14:e1007717.

Yeaman S, Hodgins KA, Lotterhos KE, Suren H, Nadeau S, Degner JC, Nurkowski KA, Smets P, Wang T, Gray LK, et al. 2016. Convergent local adaptation to climate in distantly related conifers. Science 353:1431–1433.

Yeaman S, Whitlock MC. 2011. The genetic architecture of adaptation under migration-selection balance. Evolution 65:1897–1911.

Yeaman S. 2015. Local Adaptation by Alleles of Small Effect. Am Nat 186.

Youngson NA, Whitelaw E. 2008. Transgenerational epigenetic effects. Annu. Rev. Genom. Human Genet. 9:233–257.

Zaghlool SB, Al-Shafai M, Muftah Al WA, Kumar P, Gieger C, Waldenberger M, Falchi M, Suhre K. 2016. Mendelian inheritance of trimodal CpG methylation sites suggests distal cis-acting genetic effects. Clinical Epigenetics 8:1–10.

Zhao L, Begun DJ. 2017. Genomics of parallel adaptation at two timescales in Drosophila.Matzkin L, editor. PLoS Genet 13:e1007016–e1007027.

Zhao L, Wit J, Svetec N, Begun DJ. 2015. Parallel Gene Expression Differences between Low and High Latitude Populations of Drosophila melanogaster and D. simulans.Nuzhdin SV, editor. PLoS Genet 11:e1005184–25.

Zheng W, Gianoulis TA, Karczewski KJ, Zhao H, Snyder M. 2011. Regulatory Variation Within and Between Species. Annu. Rev. Genom. Human Genet. 12:327–346.

Østbye K, Amundsen P-A, Bernatchez L, Klemetsen A, Knudsen R, Kristoffersen R, Naesje TF, Hindar K. 2006. Parallel evolution of ecomorphological traits in the European whitefish *Coregonus lavaretus* (L.) species complex during postglacial times. Molecular Ecology 15:3983–4001.

Østbye K, Bernatchez L, Naesje TF, Himberg KJM, Hindar K. 2005. Evolutionary history of the European whitefish *Coregonus lavaretus* (L.) species complex as inferred from mtDNA phylogeography and gill-raker numbers. Molecular Ecology 14:4371–4387.

Contributions scientifiques durant le doctorat

Laporte, M., Le Luyer, J., Rougeux, C., Dion-Côté, A.-M., Krick, M., Bernatchez, L. (en révision). DNA methylation reprogramming, TEs derepression and postzygotic isolation of nascent species. *Science Advances*.

Hashemzadeh Segherloo, I., Normandeau, E., Benestan, L., Rougeux, C., Cote, G., Moore, J.-S., et al. (2018). Genetic and morphological support for possible sympatric origin of fish from subterranean habitats. *Scientific Reports*, *8*(1), 2909. http://doi.org/10.1038/s41598-018-20666-w

Benestan, L., Moore, J.-S., Sutherland, B. J. G., Le Luyer, J., Maaroufi, H., Rougeux, C., et al. (2017). Sex matters in massive parallel sequencing: Evidence for biases in genetic parameter estimation and investigation of sex determination systems. *Molecular Ecology*, *26*, 6767–6783. http://doi.org/10.1111/mec.14217

Dalziel, A. C., Laporte, M., Rougeux, C., Guderley, H., & Bernatchez, L. (2016). Convergence in organ size but not energy metabolism enzyme activities among wild Lake Whitefish (*Coregonus clupeaformis*) species-pairs. *Molecular Ecology*, *26*(1), 225–244. http://doi.org/10.1111/mec.13847

Laporte, M., Pavey, S. A., Rougeux, C., Pierron, F., Lauzent, M., Budzinski, H., et al. (2016). RAD sequencing reveals within-generation polygenic selection in response to anthropogenic organic and metal contamination in North Atlantic Eels. *Molecular Ecology*, *25*(1), 219–237. http://doi.org/10.1111/mec.13466