Convergent transcriptomic landscapes under polygenic selection accompany inter-continental parallel evolution within a Nearctic *Coregonus* (Salmonidae) sister-species complex.

Clément Rougeux*[1], Pierre-Alexandre Gagnaire[2], Kim Praebel[3], Ole Seehausen[4], Louis Bernatchez[1]

[1]Département de biologie, Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, G1V 0A6, Québec, Canada
[2]Université Montpellier 2, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France
[4]Norwegian College of Fishery Science, UiT The Arctic University of Norway, N-9037 Tromsø, Norway
[5]Aquatic Ecology and Evolution, Institute of Ecology & Evolution, University of Bern-Bern, Switzerland
* Corresponding author: clement.rougeux.1@ulaval.ca

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-species, the European Whitefish (*C. lavaretus*). We analysed six replicated sympatric species-pairs from North America, Norway and Switzerland. Following *de novo* transcriptome assembly based on RNAseq data, we characterized genomic variation and differential gene expression between sympatric limnetic and benthic species but also across regions and continents. We found significantly differentially expressed genes (DEG) between limnetic and benthic whitefish by analyzing orthologous genes. DEG were enriched in shared polymorphism among sister-species, supporting the idea that some ancestral polymorphisms involved in parallel phenotypic divergence between sympatric species-pairs are maintained over long-term by the interaction of selection and gene flow. Using both and genotypes to infer polygenic selection and co-variation of genes expression involved in the same metabolic pathways, we identified parallel outliers and DEG with genes primarily over-expressed in limnetic species relative to the benthic species. Finally, we discovered a *cis*-eQTL associated with variable level of gene expression related to a trait involved in limnetic-benthic species divergence, which is shared in parallel across continents despite 500,000 years of geographic isolation. Overall, comparative transcriptomics across continents allowed identifying biological functions and candidate genes enriched in shared polymorphism associated with parallel phenotypic divergence across species-pairs of the sister-species complex from different continents, consistent with a fundamental role of longstanding natural selection on phenotypic traits involved in the ecological speciation of limnetic/benthic whitefish species-pairs.

Introduction

Decrypting the genomic basis of local adaptation of divergent populations ultimately leading to ecological speciation has been one of foremost interest over the last decade. The concept of local adaptation implies population phenotypic response to constraints associated with selective pressures stemming from their environment. This is particularly well evidenced by the occurrence of independent parallel phenotypic evolution among closely related and locally adapted nascent species (1-3). Parallel phenotypic evolution can emerge from repeated divergence of the same genomic regions (4) or from different genes involved in similar or different biological pathways (3,5), and has been associated with changes in gene expression during adaptive divergence (6-8). Indeed, adaptive genetic changes include variations in regulation of gene expression resulting in phenotypic variation (9) and sequence variation altering proteins structures (10). Genetic variation underlying these traits may originate from parallelism at the molecular level that have arise from *de novo* mutations affecting the same genes (11,12). However, such mutations are generally associated with loci of large effect controlling the expression of a given phenotypic trait (*i.e.*, monogenic architectures) (12). This contrasts with the polygenic architecture of most complex traits, including those generally involved in ecological speciation and in adaptation more generally (13,14). 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 (15). Several recent studies showed that standing variation originated from past admixture events (16-19), 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 (20,21), relatively few 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 (22).

The increased resolution of genomic analyses allowed revealing the importance of complex polygenic genomic architectures of traits involved in local adaptation, sometimes accompanied by genetic redundancy (14,23-28). Despite repeated claims for the importance of investigating the molecular basis of local adaptation and ecological speciation from a polygenic standpoint (23,26,27,29,30), 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). These two closely related taxa became geographically isolated ~500,000 years ago and evolved independently since then (31,32). This system thus offers a valuable model to study the genomic and transcriptomic underpinnings of parallel local adaptation leading to ecological speciation. Indeed, both taxa each comprises several isolated lakes harbouring partially reproductively isolated sympatric benthic (normal) and limnetic (dwarf) species-pairs. The European whitefish appear to be of hybrid origin following contact between glacial lineages (33), resulting in independent events of intra-lacustrine (sympatric) divergence of benthic and limnetic

species across Scandinavian and Alpine lakes (34-36). The North American lake whitefish sympatric species-pairs are also the result of a post-glacial secondary-contact between two glacial 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 (19,31,37,38).

Lake whitefish has been the subject of numerous studies aiming to identify the ecological and genomic basis of adaptive divergence between limnetic and benthic species. The limnetic species differ from the benthic species in its use of habitat and trophic resources, with a higher metabolic rate and more active swimming behaviour for foraging and predator avoidance (39-41), reduced energy allocated to growth relative to benthic whitefish (39,42), differences in morphology and life history (43), which are complex traits underlined by polygenic architectures (44,45).

The main goal of this study was to investigate differential transcriptional signatures between limnetic and benthic sister species to test the general hypothesis that an overlapping polygenic molecular basis underlies the parallel phenotypic divergence observed between sympatric species-pairs from the sister lineages on the two continents. Specifically, i) we first documented the amount and functional role of shared ancestral genetic polymorphism within coding genes among all sympatric species-pairs, 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 higher level of shared polymorphism between sister-species, and finally iv) explored associations between polymorphism and variation in expression at the gene level.

Results

Reference transcriptome assembly

A total of 2.6 x 10^8 100bp single-end reads were generated from 72 individuals. Filtered libraries (1.9 x 10^8 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 (46)). 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 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.

Population genetic structure

The analysis of genetic relationships among the studied populations with TreeMix revealed the presence of shared polymorphisms maintained across the entire hierarchical genetic structure (Fig 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 this Species-pair level reflected the similar degrees of genetic differentiation among species-pairs from different regions consistent with a relatively similar timing of postglacial divergence among species-pairs across the whitefish system. 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 represented by *C. clupeaformis* and *C. lavaretus* species complexes. The sharing of ancestral polymorphism between geographically isolated taxa across continents was also captured by the inferred migration link connecting two limnetic populations from Maine and Norway (Fig 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 (47) or past admixture events.

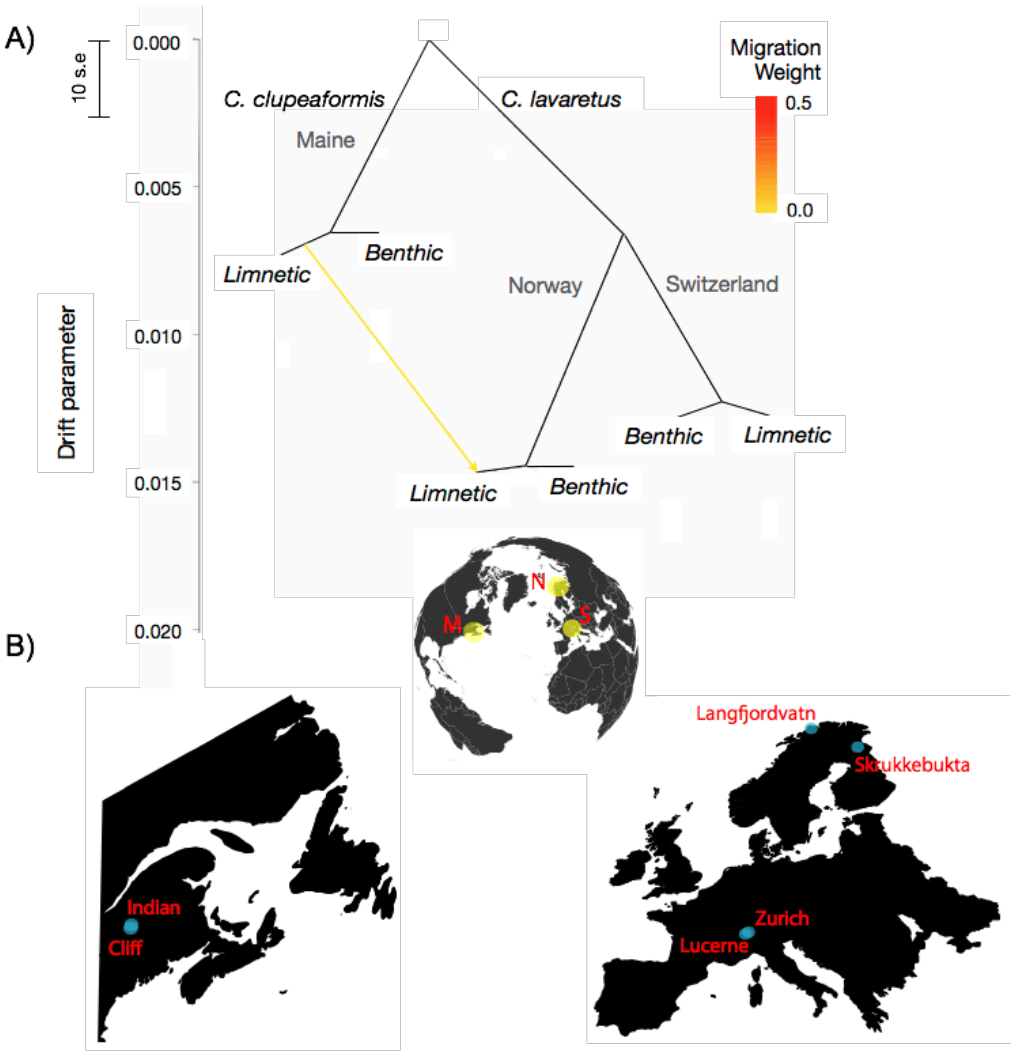

Fig 1. Details about the whitefish study system. A) Treemix analysis illustrating independent differentiation between sympatric Benthic and Limnetic species, from the

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.

Trans-specific polymorphism quantification

Given the evidence for shared polymorphisms maintained among species from different continents, we documented the overall extent of trans-specific polymorphism. Trans-specific polymorphism corresponds to ancestral variation shared among limnetic and/or benthic species from North America (*C. clupeaformis*) and Europe (*C. lavaretus*). Among the 20,911 SNPs initially obtained after genotyping and filtering steps (see Methods), we identified 2241 SNPs (10.7%) distributed among 1,251 genes (3.8%) that met our criteria (*i.e.*, trans-specific polymorphic loci and three possible genotypes). The genes containing trans-specific polymorphisms showed a significantly higher level of nucleotide diversity (π) within species-pairs (Student's *t*-test, $P<0.001$, Fig S6), and higher level of absolute divergence (Dxy) between limnetic/benthic species (Student's *t*-test, $P<0.001$, Fig S7) compared to genes with no trans-specific SNPs. These results indicate that these two categories of genes are influenced by partly different evolutionary processes.

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 (*C. clupeaformis*-*C. lavaretus*). We conditioned the ordination to account for the hierarchical genetic structure of populations (Lake, Region and Continent). The cRDA thus allowed the identification of variants associated with limnetic-benthic divergence, explaining 2.8% of the total genetic variance (ANOVA, $F=1.259$, $P=0.001$), after removing the variance explained by regional and continental population structure (Fig 2A and Fig S5). The distribution of individual locus scores on the first cRDA axis between all limnetic and benthic samples allowed identifying 348 outlier markers ($P<0.001$, 3.0 s.d.) that were congruent with parallel allele frequencies shift across continents. This result illustrates shared genetic bases (15% of shared outliers between replicate species-pairs) of limnetic-benthic divergence across both continents. Gene ontology analysis of transcripts associated with parallel outliers SNPs revealed significant enrichment ($P<0.001$) in metabolic process (*i.e.*, catabolism), immune system process and developmental process, for example (TableS1).

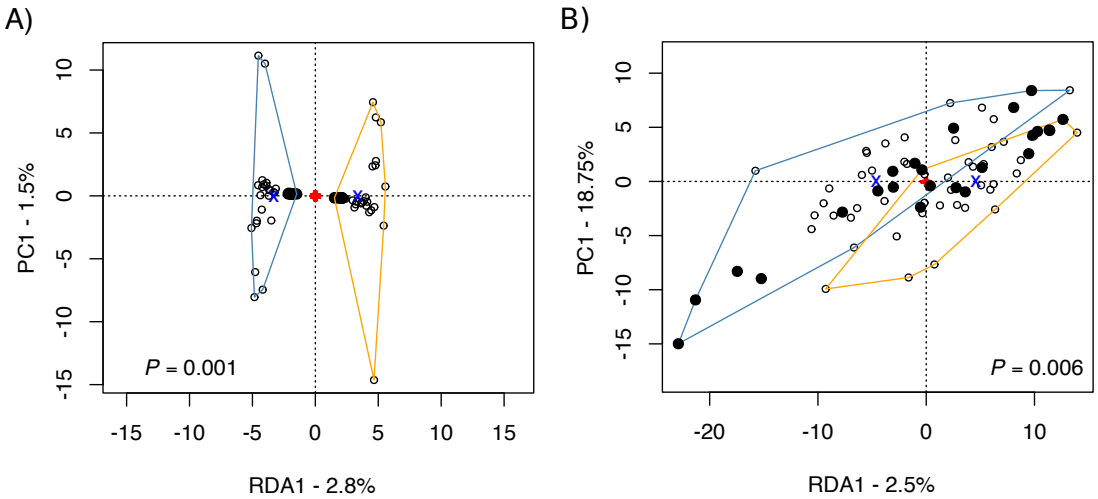

Fig 2. Conditioned redundancy analysis (cRDA) clustering individuals per species (all limnetic vs. all benthic from both continents). The cRDA analysis was used to correct data for three hierarchical levels of population structure (lake, region and continent) on A) genotypes and B) gene expression levels. Orange and blue clusters correspond to benthic and limnetic species, respectively. Individuals are represented by circles and filled circles distinguished *C. clupeaformis* from *C. lavaretus*.

Differential gene expression between Limnetic and Benthic Species

Using the benthic whitefish populations as the reference level in each comparisons levels (lake, region), we quantified differentially expressed genes (DEG) between limnetic and benthic whitefish at the Lake and Region levels using a generalized linear model (glm). North American lakes showed 2,038 (6.2%) and 94 (0.3%) of significantly differentially expressed genes (False discovery rate; FDR<0.05) between sympatric limnetic and benthic species in Cliff Lake and Indian Lake, respectively (Fig 3). For Cliff Lake, twice as many genes were up-regulated in the limnetic species compared to the benthic species (1331 *vs*. 707, χ^2 test, P<0.001) (Fig S1), as was also the case in Indian Lake (60 *vs*. 34, χ^2 test, P=0.007). In Norwegian lakes (Langfjordvatn and Skukkebukta), 99 (0.3%) and 47 (0.2%) significant DEGs were identified between limnetic and benthic whitefish, respectively. Limnetic whitefish showed a higher magnitude of expression in over-expressed genes (Fig S1) and as for North America, more genes were up-regulated in limnetic populations (80 *vs*. 19, χ^2 test, P<0.001 in Langfjordvatn; 32 *vs*. 15, χ^2 test, P=0.013, in Skrukkebukta). In Swiss lakes, 2,385 (7.3%) and 849 (2.6%) 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 (1171 *vs*. 1214, χ^2 test, P=0.378 in Lake Lucerne and 415 *vs*. 434, χ^2 test, P=0.514, in Lake Zurich).

The degree of parallelism in DEG was more important in Swiss lakes at the Region level, with 600 (1.8%) parallel DEGs showing equal proportions of down- and up-regulated genes in limnetic species (297 *vs*. 303, χ^2 test, P=0.806). In contrast, we found only 42 (0.1%) parallel DEGs between limnetic and benthic whitefish in Maine,

with a similar number of genes showing lower and higher expression in the limnetic species (17 *vs*. 25, χ^2 test, P=0.2)(Fig S2). A similar situation was found in Norway where only 95 (0.3%) parallel DEGs were found with more up-regulated genes in the limnetic species (65 *vs*. 30, χ^2 test, P<0.001).

GO enrichment analysis at the Lake and Region levels provided evidence for more parallelism at the biological functions level whereby DEGs were significantly enriched (FDR<0.05, see Table S1) in limnetic species for immune system response, detoxification and antioxidant activity in both North America and Europe. 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 (48-50).

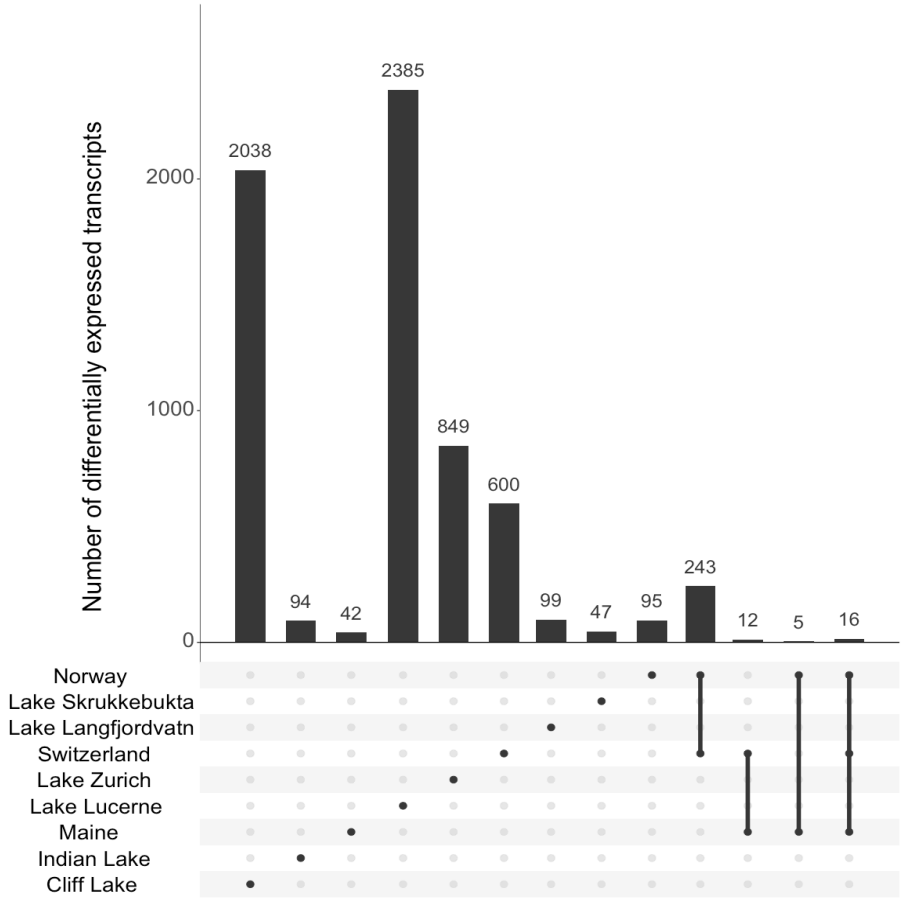

Fig 3. Frequency of shared differentially expressed transcripts between species across hierarchical levels. Limnetic/Benthic comparisons are indicated by the dots (intra-lake and intra-region) or linked dots (inter-regions and inter-continents). The number of significant differentially expressed genes (DEGs) associated to species divergence (FDR<0.05) per comparison is indicated on top of each bar. Results from the univariate gene-by-gene tests.

At the Continent level, we found 16 (0.04%) (hypergeometric test for observation by chance alone, P<0.001) parallel DEGs between limnetic and benthic species between

both continents with significantly more genes being up-regulated in limnetic populations (13 *vs*. 3, χ^2 test, $P<0.012$; Fig S3). Two biological functions were significantly enriched; cell killing ($P=0.028$) and membrane part ($P=0.033$), which are associated with immune response (*e.g.*, Immunoglobulin domain) and cell cycle regulation (*e.g.*, SPRY-associated domain). The two remaining 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-specific polymorphism in DEGs

The excess of ancestral polymorphism shared among 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 population. In order to further test if the retention of ancestral polymorphism could be linked to differential selection on adaptive traits between limnetic/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 (Fig 4).

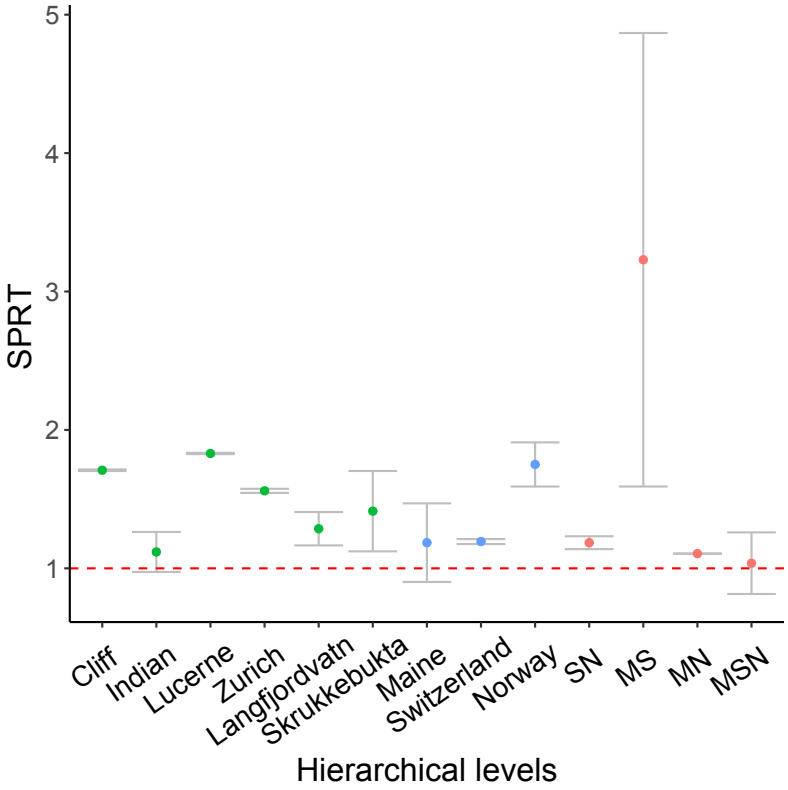

Fig 4. Shared polymorphism enrichment among DEGs compared to Non-DEGs. Test ratio (SPRT) of the proportion of DEGs against non-DEGs, for genes with trans-specific polymorphism, for intra-lake (green dots), inter-lake within regions (blue dots) and inter-regions (SN: Switzerland/Norway, MS: Maine/Switzerland, 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. Grey bars indicate the 95% confidence interval around observed value obtained using 1000 bootstrap resampling of the dataset.

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 across continents (ANOVA, F=2.516, P=0.006, Fig 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]. Applying two different significance thresholds (P<0.01 and P<0.001) allowed identifying 272 (P<0.01, Fig S4) and 66 (P<0.001, Fig S4) 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 S1). Nine out of the 16 DEGs (56%; 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 extant of parallel expression between whitefish limnetic-benthic species 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 S2). Five of these pathways shared genes and were therefore merged together to identify genes showing convergent patterns among the significant sub-networks (Fig 5).

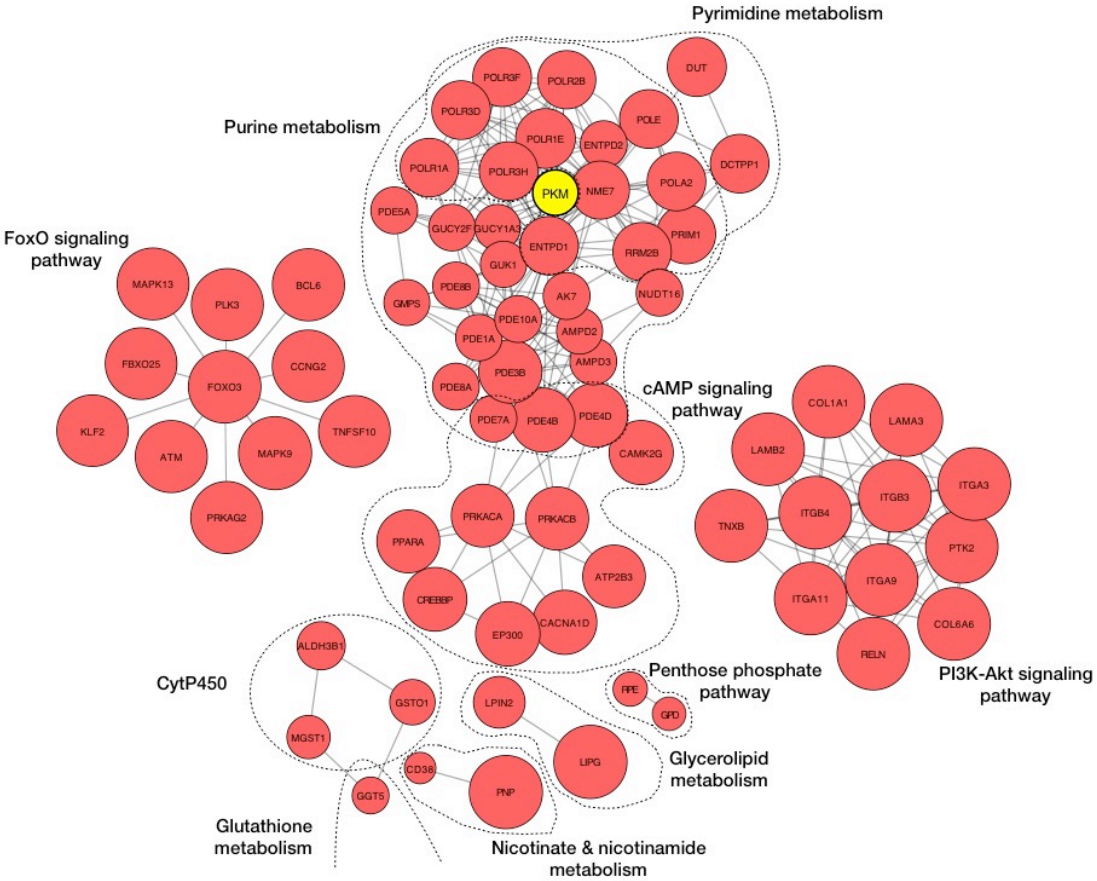

Fig 5. Merged significant subnetworks for Limnetic/Benthic comparisons. Detailed subset of significant differentially expressed genes (DEGs), between species, among corresponding metabolic pathways. Pyruvate kinase (PKM) gene expression (in yellow) is associated with a *cis*-eQTL in its 3'UTR. Gene identification is based on the Ensembl nomenclature.

From the ten identified metabolic pathways composed of 73 parallel DEGs, three categories of metabolic functions were represented: 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), detoxification metabolism and immune system (*e.g.*, CytP450 and glutathione metabolism) associated to detoxification and oxidative stress, maintaining cell integrity by preventing damage caused by reactive oxygen species (ROS), and 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 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). That is, variable level of expression was

associated with the three genotypes at a given SNP. Controlling for multiple tests, we retained 134 significant (FDR<0.05) *cis*-eQTL across continents.

We identified SNPs and genes showing redundant patterns across the different analyses. Indeed, two significant *cis*-eQTL overlapped the 20 outliers SNPs from the redundancy analysis (hypergeometric test, $P<0.001$), were physically linked to the complement factor H (*CFH*, Entrez 3075; Fig 6B) and Protein Kinase AMP-Activated Non-Catalytic Subunit Beta 1 (*Prkab1*, Entrez 19079; Fig 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*, Entrez 5315) remained significant in both (sub-networks and eQTL) analysis (Fig 4). This gene encodes a protein, involved in glycolysis, which generates ATP and pyruvate. Moreover, the level of expression of this gene was higher in heterozygous and homozygous individuals for the minor allele (Fig 6A and Fig S8, 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.

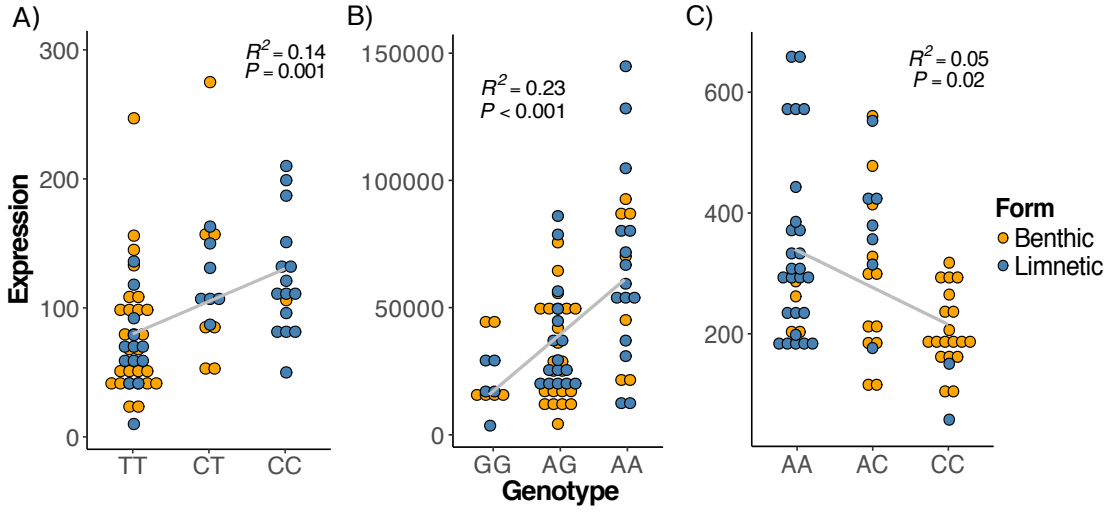

Fig 6. Association between significant *cis*-eQTL genotypes and level of expression of three genes between limnetic and benthic whitefish, independently of geographic origin and species. Three examples of transcripts abundance per individual (circles) varies with genotypes for a *cis*-eQTL in A) pyruvate 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.

Discussion

C. clupeaformis and *C. lavaretus* respectively found in North America and Eurasia diverged since at least 500,000 years ago (31,32), yet maintained similar habitat preferences in cold freshwater lakes (31,34-36), with benthic and limnetic species being respectively associated to benthic and limnetic ecological and trophic niches (51-55). 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 (19), 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-pairs 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 radiations 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 or cell cycle is congruent with previous transcriptomic analysis conducted in a single species (*i.e.*, *C. clupeaformis*) on the same organ tissue, albeit focusing on reduced transcriptome representation (low resolution microarrays developed for another species) and candidate genes (56-59). 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 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 (14,28), 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 genic 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. 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 multigenic regulatory systems (14). The accumulated results coupled with previous analyses on this system thus highlight the repeated action of selective pressures on similar polygenic bases of phenotypic traits (44,56,57,60,61).

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 have been found in other systems (ranging from 6% to 28%; (62-66)). We also identified a substantial amount of genes exhibiting shared (*i.e.*, *trans*-specific) polymorphism across the complex of *Coregonus* lineages and radiations. 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 (67) 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*-specific polymorphism was elevated compared to the mean level of absolute divergence in genes without *trans*-specific polymorphism may reflect the balanced maintenance of these divergent alleles between sympatric species-pairs over the long term by the interaction of selection and admixture (68,69). One possible explanation for this apparent paradox could be that in all sympatric pairs studied here, gene flow is being maintained between limnetic and benthic whitefish within each lake, therefore contributing to maintain alleles favoured in each species in a balanced polymorphism within each given lake. This could be eased by the apparently highly polygenic nature of traits under divergent selection meaning that the intensity of selection acting on each underlying locus will be weak (70) such that even a modest amount of gene flow could possibly maintained a balanced polymorphism within each species-pairs. This however, remains to be investigated more formally. An alternative and non-exclusive explanation could be that balancing selection among sister species acting on loci underlying polygenic traits overwhelmed stabilizing selection allowing variation and maintained polymorphism by overwhelming weak divergent selection acting on each underlying locus (47).

The identification of orthologous genes with *trans*-specific polymorphism associated with their differential expression between benthic and limnetic species supports the existence of *cis*-acting SNPs on transcripts abundance. Moreover, characterisation of DEGs enriched in shared polymorphism (*trans*-specific polymorphism) 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 (71), 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 on both continents (71). 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. Given the importance of 3'UTRs in regulating the transcription process and transcripts abundance (72,73), such 3'UTR SNP could be under divergent selection between limnetic and benthic species and therefore protected from being lost by drift within population over the long term and maintained at balanced frequency 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.

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 diverging between sympatric benthic and limnetic species within both the *C. clupeaformis* and *C. lavaretus* species radiations (44,45). Our results confirm previous studies of differential gene expression between sympatric limnetic and benthic species (56,57,60). They moreover extend previous findings by revealing patterns of parallelism between species radiations on two continents, derived from taxa that diverged at least half a million years ago. Furthermore, they show the effects of polygenic selection on genes associated with fundamental and constrained metabolic pathways, such as functions associated with energetic metabolism (59). 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 increases 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 (74), as 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 (11,75,76). 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 the phenotypic differences between sister species, and the respective role of long standing variants in ecological speciation. Moreover, given the apparently important role of divergent gene expression in the whitefish adaptive radiation, future studies should also investigate variation in regulatory mechanisms (*e.g.*, epigenetic variation) involved in controlling levels of gene expression.

Materials and methods

Sample collection, library preparation and sequencing

C. clupeaformis samples were collected from Indian Lake and Cliff Lake, Maine (USA) (Fig 1) in 2010, and correspond to samples used in previous RAD-Seq studies (19,44). These lakes are part of a well-studied lake whitefish system (77) and comprise the most divergent species-pairs along the divergence continuum described in previous studies (19,44,78). In parallel, *C. lavaretus* individuals were sampled in two Scandinavian lakes in Norway: Skrukkebukta, Langfjordvatn and two alpine lakes in Switzerland: Zurich and Lucerne (Fig 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 an independent evolutionary origin from those from the central Alpine lakes region (36). 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 (79). FLASH-merged reads (80) from all individuals were used to assemble a *de novo* reference transcriptome using Trinity version 2.2.0 (80). We aimed to build an orthologous gene composite reference for the two sister-species. Contigs lacking an ORF longer than 200bp were discarded and, in absence of reference genome, only the longest isoform per transcript were kept. 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. From normalized transcripts, we considered reciprocal best hit within each transcriptome, for both *C. clupeaformis and C. lavaretus* to identify and discard paralogous genes. We then blasted each reference transcriptome to the common reference and discarded unmapped contigs from each species to keep only common transcripts (*i.e.*, found in both species) 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. Transdecoder (81) was used to identify protein coding regions within the orthologous gene composite reference. 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 (82) and resulting Bam files were parsed to estimate individual reads counts with eXpress v1.5.1 (83). Differential expression analysis were conducted with the R packages edgeR v3.18.1 (84) and DESeq2 v1.14.1 (85), but only results from DESeq2 analysis are presented, given to the similarities between outputs of the two tools (results not shown). 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 lakes, regions and continents (hereafter called 'phylogeographic') effects on gene expression. Differentially

expressed genes (DEGs) were determined based on a FDR<0.05. To focus on differential expression between limnetic and benthic species, transcripts associated with phylogeographic effects, in addition or in interaction with differential expression between limnetic and benthic whitefish were discarded. Then, GO enrichment analysis was performed with GOATOOLS (81). For all tested lists of genes, GO enrichment was associated with FDR<0.05 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 (86). Resulting SAM files were converted to BAM files and sorted using Samtools v1.3 (87) 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 (88). Variable sites were filtered for a minimum coverage of three reads per individual. Further steps of genotype quality filtering were then applied to the data. We used the *vcffilter* program from *vcflib* (89) 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* (90), 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.

Shared polymorphism and historical relationship among species

The inference of historical relationship among all populations was performed using *Treemix* v1.12 (90) 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 events was determined empirically to improve the fit to the inferred tree. Links between populations may either reflect gene flow, or the retention of shared ancestral polymorphism among geographically isolated populations (91,92).

We then quantified and compared the amount of shared polymorphism retained at the lake, region and continental hierarchical levels. 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 (91,92). Redundancy Analysis (RDA) (93) 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 (94). 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 (95) R package, to identify genes that diverge the most between limnetic and benthic whitefish, independent of population structure. 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 and 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 (96).

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 populations after correction for hierarchical population structure and a significance threshold on expression scores (*p*-value of 0.01 and of 0.001) was applied. 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 to detect polygenic

selection (97). 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* (97) 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-specific 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 controlling for the covariates lake, region and continent as environmental effects. This analysis was run with the R package *MatrixEQTL v2.1.1* (98). 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.

Data accessibility

Raw sequence data are available through the NCBI sequence read archive (SRA) database under submission n°SUB3846793, upon acceptance of publication.

Author contributions

C.R. and L.B. designed the study. CR performed bench work for RNAseq, analysed the data and lead writing of the manuscript in collaboration of P-A.G. and L.B. K.P. and O.S. shared samples from Norway and Switzerland, respectively. All authors contributed to writing and critically commented the final version of the manuscript.

Acknowledgements

Jérémy Le Luyer, Ben J. G. Sutherland provided valuable advice 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* 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. This research was supported by a Discovery Research grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to L.B and strategic funding from UiT to KP. L.B also holds the Canadian Research Chair in genomics and conservation of aquatic resources, which funded the research infrastructure for this project.

References

1. Endler JA. Natural selection in the wild. Princeton Univ. Press: Princeton. 1986. 336 p.

2. Orr HA. The probability of parallel evolution. Evolution. 2005 Jan;59(1):216–20.

3. Losos JB. Convergence, adaptation, and constraint. Evolution. 2011 Jul;65(7):1827–40.

4. Conte GL, Arnegard ME, Peichel CL, Schluter D. The probability of genetic parallelism and convergence in natural populations. Proc Biol Sci. 2012 Nov 12;279(1749):5039–47.

5. Cohan FM, Hoffmann AA. Uniform Selection as a Diversifying Force in Evolution: Evidence from Drosophila. American Naturalist. 1989;134(4):613–37.

6. Manceau M, Domingues VS, Mallarino R, Hoekstra HE. The developmental role of Agouti in color pattern evolution. Science. 2011 Feb 25;331(6020):1062–5.

7. Harrison PW, Wright AE, Mank JE. The evolution of gene expression and the transcriptome–phenotype relationship. Seminars in Cell and Developmental Biology. Elsevier Ltd; 2012 Apr 1;23(2):222–9.

8. Pavey SA, Collin H, Nosil P, Rogers SM. The role of gene expression in ecological speciation. Annals of the New York Academy of Sciences. 2010 Sep 22;1206(1):110–29.

9. Rebeiz M, Pool JE, Kassner VA, Aquadro CF, Carroll SB. Stepwise modification of a modular enhancer underlies adaptation in a Drosophila population. Science. 2009 Dec 18;326(5960):1663–7.

10. Hoekstra HE, Hirschmann RJ, Bundey RA, Insel PA, Crossland JP. A single amino acid mutation contributes to adaptive beach mouse color pattern. Science. 2006 Jul 7;313(5783):101–4.

11. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature.

Nature Publishing Group; 2012 Apr 4;484(7392):55–61.

12. Manceau M, Domingues VS, Linnen CR, Rosenblum EB, Hoekstra HE. Convergence in pigmentation at multiple levels: mutations, genes and function. Philosophical Transactions of the Royal Society of London B: Biological Sciences. 2010 Jul 19;365(1552):2439–50.

13. Gagnaire P-A, Gaggiotti OE. Detecting polygenic selection in marine populations by combining population genomics and quantitative genetics approaches. Current Zoology. 2016 Dec 28;62(6):603–16.

14. Yeaman S. Local Adaptation by Alleles of Small Effect. Am Nat. The American Society of Naturalists; 2015 Oct 1;186(S1).

15. Welch JJ, Jiggins CD. Standing and flowing: the complex origins of adaptive variation. Molecular Ecology. 2014 Aug;23(16):3935–7.

16. Roesti M, Gavrilets S, Hendry AP, Salzburger W, Berner D. The genomic signature of parallel adaptation from shared genetic variation. Molecular Ecology. 2014 Aug;23(16):3944–56.

17. Martin CH, Cutler JS, Friel JP, Touokong CD, Coop G, Wainwright PC. Complex histories of repeated gene flow in Cameroon crater lake cichlids cast doubt on one of the clearest examples of sympatric speciation. Evolution. 2015 Jun;69(6):1406–22.

18. Meier JI, Sousa VC, Marques DA, Selz OM, Wagner CE, Excoffier L, et al. Demographic modelling with whole-genome data reveals parallel origin of similar *Pundamilia* cichlid species after hybridization. Molecular Ecology. 2017 Jan;26(1):123–41.

19. Rougeux C, Bernatchez L, Gagnaire P-A. 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. 2017 Aug 2;9(8):2057–74.

20. Barrett RDH, Schluter D. Adaptation from standing genetic variation. Trends in Ecology & Evolution. 2008 Jan;23(1):38–44.

21. Schrider DR, Kern AD. Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. Molecular Biology and Evolution. 2017 May 8;34(8):1863–77.

22. Nelson TC, Cresko WA. Ancient genomic variation underlies repeated ecological adaptation in young stickleback populations. Evolution Letters. 2018 Jan 26;114(Suppl 1):7061–13.

23. Yeaman S, Hodgins KA, Lotterhos KE, Suren H, Nadeau S, Degner JC, et al.

Convergent local adaptation to climate in distantly related conifers. Science. 2016 Sep 23;353(6306):1431–3.

24. Berg JJ, Coop G. A Population Genetic Signal of Polygenic Adaptation. W Feldman M, editor. PLoS Genet. 2014 Aug 7;10(8):e1004412–25.

25. Laporte M, Pavey SA, Rougeux C, Pierron F, Lauzent M, Budzinski H, et al. RAD sequencing reveals within-generation polygenic selection in response to anthropogenic organic and metal contamination in North Atlantic Eels. Molecular Ecology. 2nd ed. 2016 Jan;25(1):219–37.

26. Babin C, Gagnaire P-A, Pavey SA, Bernatchez L. RAD-Seq Reveals Patterns of Additive Polygenic Variation Caused by Spatially-Varying Selection in the American Eel (*Anguilla rostrata*). Genome Biology and Evolution. 2017 Nov 10;9(11):2974–86.

27. Bay RA, Arnegard ME, Conte GL, Best J, Bedford NL, McCann SR, et al. Genetic Coupling of Female Mate Choice with Polygenic Ecological Divergence Facilitates Stickleback Speciation. Current Biology. Elsevier Ltd; 2017 Nov 6;27(21):3344–4.

28. Slatkin M. Spatial patterns in the distributions of polygenic characters. Journal of Theoretical Biology. 1978 Jan 20;70(2):213–28.

29. Harrisson KA, Amish SJ, Pavlova A, Narum S, Telonis-Scott M, Rourke ML, et al. Signatures of polygenic adaptation associated with climate across the range of a threatened fish species with high genetic connectivity. Molecular Ecology. 2017 Oct 4.

30. Jain K, Stephan W. Modes of Rapid Polygenic Adaptation. Molecular Biology and Evolution. 2017 Sep 11;:1–7.

31. Bernatchez L, Dodson JJ. Phylogeographic structure in mitochondrial DNA of the lake whitefish (*Coregonus clupeaformis*) and its relation to Pleistocene glaciations. Evolution. 1991;45(4):1016–35.

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

33. Hudson AG, Vonlanthen P, Seehausen O. Rapid parallel adaptive radiations from a single hybridogenic ancestral population. Proc Biol Sci. 2010 Nov 23;278(1702):58–66.

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

35. Østbye K, Amundsen P-A, Bernatchez L, Klemetsen A, Knudsen R, Kristoffersen R, et al. Parallel evolution of ecomorphological traits in the European whitefish *Coregonus lavaretus* (L.) species complex during postglacial times. Molecular Ecology. 2006 Nov;15(13):3983–4001.

36. Østbye K, Bernatchez L, Naesje TF, Himberg KJM, Hindar K. Evolutionary history of the European whitefish *Coregonus lavaretus* (L.) species complex as inferred from mtDNA phylogeography and gill-raker numbers. Molecular Ecology. 2005 Oct 3;14(14):4371–87.

37. Bernatchez L, Dodson JJ. Allopatric origin of sympatric populations of lake whitefish (*Coregonus clupeaformis*) as revealed by mitochondrial-DNA restriction analysis. Evolution. 1990;24(4):890–908.

38. Pigeon D, Chouinard A, Bernatchez L. Multiple modes of speciation involved in the parallel evolution of sympatric morphotypes of lake whitefish (*Coregonus clupeaformis*, Salmonidae). Evolution. 1997;51(1):196.

39. Trudel M, Tremblay A, Schetagne R, Rasmussen JB. Why are dwarf fish so small? An energetic analysis of polymorphism in lake whitefish (*Coregonus clupeaformis*). Can J Fish Aquat Sci. 2001;58(2):394–405.

40. Bernatchez L, Chouinard A, Lu G. Integrating molecular genetics and ecology in studies of adaptive radiation: whitefish, *Coregonus* sp., as a case study. Biological Journal of the Linnean Society. 1999;68:173–94.

41. Rogers SM, Gagnon V, Bernatchez L. Genetically based phenotype-environment association for swimming behavior in lake whitefish ecotypes (*Coregonus clupeaformis* Mitchill). Evolution. 2002 Nov;56(11):2322–9.

42. Rogers SM, Bernatchez L. 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. 2004 Dec 7;14(2):351–61.

43. Rogers S, Bernatchez L. 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. 2007 Mar 10;24(6):1423–38.

44. Gagnaire P-A, Pavey SA, Normandeau E, Bernatchez L. The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by RAD sequencing. Evolution. 2013 Sep;67(9):2483–97.

45. Laporte M, Rogers SM, Dion-Côté A-M, Normandeau E, Gagnaire P-A, Dalziel AC, et al. 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). 2015 Jul;5(7):1481–91.

46. Pasquier J, Cabau C, Nguyen T, Jouanno E, Severac D, Braasch I, et al. Gene evolution and gene expression after whole genome duplication in fish: the PhyloFish database. BMC Genomics. BioMed Central; 2016 May 18;17(1):368.

47. Turelli M, Barton NH. Polygenic variation maintained by balancing selection: pleiotropy, sex-dependent allelic effects and G x E interactions. Genetics. 2004 Feb;166(2):1053–79.

48. Dalziel AC, Laporte M, Guderley H, Bernatchez L. Do differences in the activities of carbohydrate metabolism enzymes between Lake Whitefish ecotypes match predictions from transcriptomic studies? Comparative Biochemistry and Physiology, Part B. Elsevier; 2017 Aug 17;:1–12.

49. Laporte M, Dalziel AC, Martin N, Bernatchez L. Adaptation and acclimation of traits associated with swimming capacity in Lake Whitefish (*Coregonus clupeaformis*) ecotypes. BMC Evol Biol. BMC Evolutionary Biology; 2016 Aug 10;16(1):1–13.

50. Dalziel AC, Laporte M, Rougeux C, Guderley H, Bernatchez L. Convergence in organ size but not energy metabolism enzyme activities among wild Lake Whitefish (*Coregonus clupeaformis*) species-pairs. Molecular Ecology. 2016 Sep 23;26(1):225–44.

51. Häkli K, Ostbye K, Kahilainen KK, Amundsen P-A, Praebel K. Diversifying selection drives parallel evolution of gill raker number and body size along the speciation continuum of European whitefish. Ecology and Evolution. 2018 Feb 5;8(5):2617–31.

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

53. Lu G, Bernatchez L. Correlated trophic specialization and genetic divergence in sympatric lake whitefish ecotypes (*Coregonus clupeaformis*): support for the ecological speciation hypothesis. Evolution. 1999;53:1491–505.

54. Landry L, Vincent WF, Bernatchez L. Parallel evolution of lake whitefish dwarf ecotypes in association with limnological features of their adaptive landscape. J Evol Biol. 2007 May;20(3):971–84.

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

56. St-Cyr J, Derome N, Bernatchez L. The transcriptomics of life-history trade-offs in whitefish species pairs (*Coregonus* sp.). Molecular Ecology. 2008 Apr;17(7):1850–70.

57. Jeukens J, Bittner D, Knudsen R, Bernatchez L. Candidate genes and adaptive radiation: insights from transcriptional adaptation to the limnetic niche among coregonine fishes (*Coregonus* spp., Salmonidae). Molecular Biology and Evolution. 2009 Jan;26(1):155–66.

58. Jeukens J, Renaut S, St-Cyr J, Nolte AW, Bernatchez L. The transcriptomics of sympatric dwarf and normal lake whitefish (*Coregonus clupeaformis* spp., Salmonidae) divergence as revealed by next-generation sequencing. Molecular Ecology. 2010 Dec;19(24):5389–403.

59. Dalziel AC, Laporte M, Guderley H, Bernatchez L. Do differences in the activities of carbohydrate metabolism enzymes between Lake Whitefish ecotypes match predictions from transcriptomic studies? Comparative Biochemistry and Physiology, Part B. Elsevier Inc; 2017 Aug 7;:1–33.

60. Jeukens J, Bernatchez L. Regulatory versus coding signatures of natural selection in a candidate gene involved in the adaptive divergence of whitefish species pairs (*Coregonus* spp.). Ecology and Evolution. 2011 Dec 13;2(1):258–71.

61. Gagnaire P-A, Normandeau E, Pavey SA, Bernatchez L. Mapping phenotypic, expression and transmission ratio distortion QTL using RAD markers in the Lake Whitefish (*Coregonus clupeaformis*). Molecular Ecology. 2013 Jun;22(11):3036–48.

62. Deagle BE, Jones FC, Chan YF, Absher DM, Kingsley DM, Reimchen TE. Population genomics of parallel phenotypic evolution in stickleback across stream-lake ecological transitions. Proc Biol Sci. 2012 Feb 24;279(1732):1277–86.

63. Ravinet M, Westram A, Johannesson K, Butlin R, André C, Panova M. Shared and nonshared genomic divergence in parallel ecotypes of *Littorina saxatilis* at a local scale. Molecular Ecology. 2015 Sep 3;25(1):287–305.

64. Westram AM, Galindo J, Alm Rosenblad M, Grahame JW, Panova M, Butlin RK. Do the same genes underlie parallel phenotypic divergence in different *Littorina saxatilis* populations? Molecular Ecology. 2014 Sep 8;23(18):4603–16.

65. Le Moan A, Gagnaire PA, Bonhomme F. Parallel genetic divergence among coastal-marine ecotype pairs of European anchovy explained by differential introgression after secondary contact. Molecular Ecology. 2016 Jul;25(13):3187–202.

66. Meier JI, Marques DA, Wagner CE, Excoffier L, Seehausen O. Genomics of parallel ecological speciation in Lake Victoria cichlids. Molecular Biology and Evolution. 2018 Mar 16;:1–37.

67. Charlesworth B, Nordborg M, Charlesworth D. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet Res. 1997 Oct;70(2):155–74.

68. Ma T, Wang K, Hu Q, Xi Z, Wan D, Wang Q, et al. Ancient polymorphisms and divergence hitchhiking contribute to genomic islands of divergence within a poplar species complex. Proceedings of the National Academy of Sciences. 2017 Dec 26;5:201713288–8.

69. Han F, Lamichhaney S, Grant BR, Grant PR, Andersson L, Webster MT. Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. Genome Research. 2017 Jun 1;27(6):1004–15.

70. Le Corre V, Kremer A. The genetic differentiation at quantitative trait loci under local adaptation. Molecular Ecology. 2012 Feb 14;21(7):1548–66.

71. Zheng W, Gianoulis TA, Karczewski KJ, Zhao H, Snyder M. Regulatory Variation Within and Between Species. Annu Rev Genom Human Genet. Annual Reviews; 2011 Sep 22;12(1):327–46.

72. Merritt C, Rasoloson D, Ko D, Seydoux G. 3' UTRs are the primary regulators of gene expression in the *C. elegans* germline. Current Biology. 2008 Oct 14;18(19):1476–82.

73. Wittkopp PJ, Kalay G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nature Reviews Genetics. Nature Publishing Group; 2011 Dec 6;13(1):59–69.

74. West-Eberhard MJ. Developmental plasticity and the origin of species differences. Proceedings of the National Academy of Sciences. 2005 May 3;102 Suppl 1(Supplement 1):6543–9.

75. Schluter D, Clifford EA, Nemethy M, McKinnon JS. Parallel evolution and inheritance of quantitative traits. Am Nat. 2004 Jun;163(6):809–22.

76. Wittkopp PJ, Haerum BK, Clark AG. Evolutionary changes in cis and trans gene regulation. Nature. Nature Publishing Group; 2004 Jul 1;430(6995):85–8.

77. Bernatchez L, Renaut S, Whiteley AR, Derome N, Jeukens J, Landry L, et al. On the origin of species: insights from the ecological genomics of lake whitefish. Philos Trans R Soc Lond, B, Biol Sci. 2010 Jun

12;365(1547):1783–800.

78. Renaut S, Maillet N, Normandeau E, Sauvage C, Derome N, Rogers SM, et al. Genome-wide patterns of divergence during speciation: the lake whitefish case study. Philos Trans R Soc Lond, B, Biol Sci. 2012 Feb 5;367(1587):354–63.

79. Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011 Oct 21;27(21):2957–63.

80. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013 Jul 11;8(8):1494–512.

81. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Meth. 2012 Apr;9(4):357–9.

82. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Meth. 2012 Nov 18;10(1):71–3.

83. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139–40.

84. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014 Dec 5;15(12):31–21.

85. Tang H, Klopfenstein D, Pedersen B, Flick P, Sato K, Ramirez F, et al. GOATOOLS: Tools for Gene Ontology. 2015 Sep.

86. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 7;25(16):2078–9.

87. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv. 2012;arXiv:1207.3907.

88. Garrison E. Vcflib, a simple C++ library for parsing and manipulating VCF files [Internet]. Available from: https://github.com/vcflib/vcflib

89. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156–8.

90. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 2012;8(11):e1002967.

91. Duforet-Frebourg N, Luu K, Laval G, Bazin É, Blum MGB. Detecting Genomic Signatures of Natural Selection with Principal Component Analysis: Application to the 1000 Genomes Data. Molecular Biology and Evolution. 2016 Apr;33(4):1082–93.

92. Luu K, Bazin É, Blum MGB. pcadapt: an R package to perform genome scans for selection based on principal component analysis. Mol Ecol Resour. 2017 Jan;17(1):67–77.

93. Legendre P, Legendre L. Numerical ecology. Elsevier. Amsterdam.

94. Capblancq T, Luu K, Blum M, Bazin E. How to make use of ordination methods to identify local adaptation: a comparison of genome scans based on PCA and RDA. bioRxiv. 2018 Feb3;:1-26.

95. Jari Oksanen F, Blanchet G, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community Ecology Package. R package version 2.4-4. https://CRAN.R-project.org/package=vegan.

96. Forester BR, Lasky JR, Wagner HH, Urban DL. Using genotype-environment associations to identify multilocus local adaptation. bioRxiv. 2017 Apr 21;:1–24.

97. Gouy A, Daub JT, Excoffier L. Detecting gene subnetworks under selection in biological pathways. Nucl Acids Res. 2017 Jul 18;:1–11.

98. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012 May 9;28(10):1353–8.