MOLECULAR ECOLOGY

Molecular Ecology (2012) **21**, 4797–4810 doi: 10.1111/j.1365-294X.2012.05645.x

Transcriptome changes after genome-wide admixture in invasive sculpins (*Cottus*)

TILL CZYPIONKA,* JIE CHENG,* ALEXANDER POZHITKOV*† and ARNE W. NOLTE*

**Max-Planck Institute for Evolutionary Biology, August-Thienemannstrasse 2, 24306 Plön, Germany, †Department of Periodontology, University of Washington, Box 357444, Seattle, WA 98195, USA*

Abstract

Models on hybrid speciation assume that hybridization generates increased phenotypic variance that is utilized to invade new adaptive peaks. We test to what extent this prediction can be traced using gene expression data in the fish species *Cottus perifretum* and *Cottus rhenanus* as well as a natural hybrid lineage referred to as invasive sculpins. In addition, interspecies crosses were used to explore evolutionary trajectories from initial stages to the hybrid lineage. EST (expressed sequence tag) libraries were sequenced to design an oligonucleotide microarray that was calibrated for probe-specific differences in binding behaviour. Levels of gene expression divergence between species correlate with genetic divergence at neutral markers and, accordingly, invasive sculpins were intermediate between the parental species overall. However, the hybrid lineage is distinguished through unique patterns of gene expression that are enriched for biological functions which represent candidates for the fitness properties of invasive sculpins. We compare F_2 crosses with natural invasive sculpins to show that the variance in gene expression decreases in invasives. Moreover, few of the transgressive patterns of gene expression that distinguish invasives can be directly observed in F_2 crosses. This suggests that the invasive transcriptome was subject to secondary changes after admixture. The result is in line with an evolutionary process that reduces maladaptive variance and optimizes the phenotype of an emerging hybrid lineage.

Keywords: adaptation, hybridization, invasive species, microarray calibration, transgressive segregation

Received 31 January 2012; revision received 28 March 2012; accepted 7 April 2012

Introduction

There is a recent interest in the origin of new species through the process of hybridization (Mallet 2007; Abbott *et al.* 2010; Nolte & Tautz 2010) and, from a different perspective, the merging of gene pools induced by ecological change (Seehausen *et al.* 1997; Taylor *et al.* 2006) or by the artificial transplantation of alien taxa (Laikre *et al.* 2010). However, to date, most studies on natural hybridization are relatively uninformed about the functional genetic and ecological consequences of admixture and mostly draw conclusions based on anonymous genetic markers. One alternative to obtain broad insight into functionally relevant traits is given by gene expression profiling. The possibility to test many genes simultaneously reduces the unavoidable bias imposed by selecting phenotypic traits. Therefore, it uncovers 'hidden phenotypes' that were not given attention before (Larsen *et al.* 2011). Given a list of differentially expressed and annotated genes, gene ontology (GO) terms permit inference of functions that have evolved (Conesa *et al.* 2005). Although GO term analysis cannot account for evolutionarily young but important orphan genes (Domazet-Lošo & Tautz 2010), there is a growing number of studies that employ gene expression analysis to study adaptive evolutionary change. For example a series of studies on whitefish (*Coregonus clupeaformis*) have shown that candidate adaptive traits inferred from gene expression data can be confirmed using independent

Correspondence: Arne W. Nolte, Fax: +49 4522 763 281;
E-mail: nolte@evolbio.mpg.de

physiological experiments and convey fitness advantages in a known ecological context (Bernatchez *et al.* 2010). Although it has been documented that gene expression levels differ in a very characteristic fashion between tissues and under different ecological conditions (Xia *et al.* 2007, Cheviron *et al.* 2008), it is reassuring that a considerable fraction of candidate adaptive transcriptomic patterns can be found in different tissues independently of the environment (Nolte *et al.* 2009).

Gene expression data have been recognized as a rich source of information to study fish hybrids (Mavárez *et al.* 2009; Normandeau *et al.* 2009; Renaut *et al.* 2009) and permit testing of predictions related to the process of hybrid speciation. The evolutionary outcome of hybridization is critically determined by the genetic interactions which result from the fusion of different gene pools (Landry *et al.* 2007). Notably, this includes both effects that can reduce or raise fitness. For example Michalak & Noor (2003) have successfully employed a microarray screen to identify genes that show disrupted patterns of expression in *Drosophila* hybrids, and many of these candidate genes indeed contribute to post-zygotic isolation (Michalak & Noor 2004). A fitness-enhancing effect of initial hybridization is given by heterosis (Charlesworth & Willis 2009) or when transgressive segregation generates phenotypes that lie outside the range observed in the parental species (Rieseberg *et al.* 1999, 2003a). This increased phenotypic variance might facilitate the invasion of a new adaptive peak not available to the parents (Barton 2001; Mallet 2007). However, only a limited number of the admixed genotypes will have a high fitness and an admixed population goes through a lag phase in which selection optimizes the composition and overall fitness of the admixed gene pool (Barton 2001). This evolutionary change during the initial phase is what distinguishes this mode of speciation most from other modes of speciation (Nolte & Tautz 2010). We are studying a hybrid lineage of invasive sculpins (Cottus, Cottidae, Teleostei) that lends itself to identify the gene expression changes associated with the ecological transition away from the habitats occupied by the parental species. Invasive sculpins emerged in the wake of man-made environmental changes within the past 200 years that have increased the connectivity of the central European rivers Rhine and Scheldt and massively altered the ecological conditions in the lower parts of these drainages. These perturbations have fostered hybridization of two previously isolated fish species (*Cottus rhenanus* and *Cottus perifretum*) from smaller tributaries to the River Rhine and River Scheldt systems (Nolte *et al.* 2005; Stemshorn *et al.* 2011). Whereas the parental taxa have been estimated to be separated for 2 million years (Englbrecht *et al.* 2000; but see Volckaert *et al.* 2002), the hybrid sculpins are possibly less than 200 generations old and were first detected at a massive scale in the Netherlands after 1980. They are referred to as 'invasive', because they invaded new habitats that are not occupied by the parental species (Nolte *et al.* 2005). Population genetic studies indicate that this is because of a raised fitness of invasive sculpins in large river habitats (Nolte *et al.* 2006; Stemshorn *et al.* 2011). A recent systematic screen for genetic incompatibilities between the ancestral species has revealed no evidence that effective intrinsic barriers to reproduction exist that explain the population structure of *Cottus* in the River Rhine basin (J. Cheng, T. Czypionka, A. W. Nolte unpublished data). Taken together, these findings suggest that ecological selection pressures constitute an essential component to explain the evolution of invasive sculpins and stress the need to assess the adaptive relevance of genetic changes for the adaptation of hybrids (Abbott *et al.* 2010).

Transcriptome analyses are subject to a number of biases that, while not erasing biologically relevant signal, can distort inferred patterns of gene expression massively (Jeukens *et al.* 2010). For oligonucleotide microarrays, this concerns non-responsive or sticky probes (Pozhitkov *et al.* 2006) that do not have a linear relationship between signal intensity and target transcript concentration (A. Pozhitkov, P. A. Noble, D. Tautz unpublished data). We analysed gene expression data using custom oligonucleotide microarrays that were calibrated for the samples analysed in this study. Gene expression divergence was compared among populations representing the parental species *C. perifretum* and *C. rhenanus* as well as their natural hybrids, the invasive sculpins. These comparisons identified gene functions that are correlated with the invasion of new habitats, and were also used to assess functional differentiation at the within and between species level. As the rise of invasive sculpins is very recent, the *Cottus* system is suitable to recreate initial steps that have led to the formation of invasive sculpins and to test hypotheses about the origin of novelty through hybridization. We compare F_2 crosses with natural invasive sculpins to test to what extent the features that distinguish invasives can be explained as a direct consequence of hybridization or, alternatively, have emerged independently in a secondary evolutionary process.

Methods

Study populations and samples

The populations used in this study were also described and used by J. Cheng, T. Czypionka, A. W. Nolte unpublished data. Here, we compared gene expression

among parental species (*Cottus rhenanus* and *Cottus perifretum*), F_2 crosses between these and invasive sculpins to infer patterns of transcriptome evolution. We analysed two independent biological replicate crosses to account for within and between species variance. Previous studies confirmed that the populations used here (Table 1) are representative of the ancestral species that gave rise to invasive sculpins (Knapen *et al.* 2003; Freyhof *et al.* 2005; Nolte *et al.* 2005; Stemshorn *et al.* 2011). *Cottus rhenanus* from the Broel and Naaf populations display significant differentiation with an F_{ST} of 0.3 (Nolte *et al.* 2006) and the same holds for *C. perifretum* populations from Witte Nete (WN) and Laarse Beek (LB) (F_{ST} = 0.368, Knapen *et al.* 2003).

Adult fishes of the study populations were transferred to laboratory aquaria in climate chambers and fed with frozen and live insect larvae and brine shrimp. The room temperature and light regime mimicked the conditions in central Europe. During the winter, the temperature was lowered to 4 °C for at least 1 month. Spawning occurred readily in artificial shelters partially buried in sand when water temperatures were raised to 8–10 °C. We have generated two independent groups of F_2 crosses from interspecies crosses between pure *C. perifretum* and *C. rhenanus*: Broel(♂) × WN(♀) and WN(♂) × Broel(♀) as well as Naaf(♂) × LB(♀) and LB(♂) × Naaf(♀). Larvae were fed initially using live *Artemia* nauplii, and later with frozen insect larvae until a majority had reached a length of approximately 3 cm (September–October 2010). Water temperatures were raised to a maximum of 17 °C during the summer, and fish were kept at this temperature for several weeks before they were killed. Only individuals that weighed 0.4–0.5 g appeared to be normally developed, and healthy individuals were selected for analysis of gene expression. Fishes were anaesthetized with CO_2 and killed by immersing the whole fish in RNA_{later} until RNA extraction. We were concerned that analyses based on single families from aquarium stocks yield biased patterns of gene expression. Thus, we combined unrelated fishes from our aquaria and ones that have grown in natural habitats, where possible (Table 1). For this purpose, wild fish representing invasive *Cottus* as well as the two populations of *C. rhenanus* were collected in the field and adapted to our aquaria for one week before they were killed. Our study is imbalanced with respect to controlling for gene expression changes because of long-term aquarium care effects for ancestral *C. perifretum* as these fish are protected in their native range and could not be resampled. Likewise, synthetic F_2 crosses cannot be obtained from natural habitats.

Development of a microarray

To develop a custom oligonucleotide microarray, we sequenced transcriptome libraries of 22 individuals representing all populations mentioned above (Table 1) using a Roche GS-FLX DNA Sequencer. A total of 27243

Table 1 Origin and affinity of *Cottus* samples used for transcriptome analysis. Two populations were chosen to represent the parental species *Cottus rhenanus* (Broel and Naaf) and *Cottus perifretum* (Laarse Beek and Witte Nete). A hybrid lineage that resulted from natural admixture between the two species is represented by the invasive *Cottus* from the Sieg population. Initial stages of hybridization were re-created in the laboratory as F_2 crosses between the parental species. The samples column shows that each group of F_2 crosses comprises individuals (number given) from two independent families (alternative cross directions) and that all experimental groups contained individuals from unrelated families reared in the laboratory (LR, laboratory reared) or outbred fishes collected in nature (WC, wild caught)

Species	Population	Samples	Origin of populations
C. rhenanus	Broel	2 × WC 6 × LR	Broel (GIS: 50°50′N 7°22′E; River Sieg system, North Rhine-Westphalia, Germany)
	Naaf	2 × LR 6 × WC	Naaf (GIS: 50°52′N 7°16′E River Agger system, southeast of Hausdorp, North Rhine-Westphalia, Germany
C. perifretum	Laarse Beek	4 × LR 4 × LR	Laarse Beek (GIS: 51°17′N 5°04′E; River Scheldt system, Flanders, Belgium)
	Witte Nete	4 × LR 4 × LR	Witte Nete (GIS: 53°14′N 5°04′E; River Scheldt system, Flanders, Belgium)
Invasive *Cottus*	Sieg	2 × LR 6 × WC	Sieg (GIS: 50°48′N 7°9′E; River Rhine system, North Rhine Westphalia, Germany)
F_2 crosses	Broel × Witte Nete	4 × LR 4 × LR	Synthetic laboratory crosses between Broel and WN
F_2 crosses	Naaf × Laarse Beek	4 × LR 4 × LR	Synthetic laboratory crosses between Naaf and LB

LB, Laarse Beek; WN, Witte Nete.

EST contigs were reconstructed from the transcriptome reads using the Roche analysis software Newbler with default criteria and were available together with 87320 unassembled reads to design an Agilent oligonucleotide microarray which requires consideration of the read direction. We performed a BLASTX (Altschul *et al.* 1990) homology search of *Cottus* sequences against the stickleback protein sequences (Version 5.17; http://www.ensembl.org/) and were able to annotate 11430 sequences (*e*-value < 10^{-3}). In a second step, a custom database was built for protein coding nucleotide sequences annotated to the Actinopterygii from the NCBI Taxonomy Browser (http://www.ncbi.nlm.nih.gov/Taxonomy/) (as of 7/6/2010). A total of 713 additional sequences could be annotated by blasting against this database (*e*-value < 10^{-3}). An alternative criterion to identify the directionality of target sequences is given when the ends of the sequences contain a poly-A tail, which agreed in 86% of the cases with the inference from blast results. All reads for which the directionality could be inferred from one of the two methods were submitted to the AGILENT E-ARRAY software (https://earray.chem.agilent.com/earray/) to design 60-mer oligonucleotide probes that are specific to the target (mRNA) sequences. A total of 13 329 different probes and 1879 replicates of these were synthesized to fill the 15 208 features available on an Agilent 15k Expression microarray. We refer to probes and the levels of transcript abundance detected for these as 'gene' and 'gene expression' throughout this study, but acknowledge that some probes may target alternative splice variants, copies or fragments of the same gene.

Raw signal analysis and quality filtering

Total RNA was extracted from whole fishes using a Trizol protocol and labelled using the Agilent Low Input Quick Amp Labeling kit. The microarray was loaded with labelled cRNA as described in the Agilent protocol for One-Color Microarray-Based Gene Expression Analysis (version 6.5, May 2010). Care was taken to randomize the distribution of the eight individuals representing different populations (see Table 1) between different slides. Microarrays were scanned with an Agilent DNA Microarray Scanner type C. Feature extraction was carried out using the Agilent Feature Extraction software (version 10.7.3.1) and the 'gprocessed signal' was used as raw signal for gene expression analysis. Signals flagged for bad quality or saturated intensity were treated as missing data. Probes with missing data for more than two individuals per population were removed from the data set.

An implicit assumption in normalization procedures and microarray analysis is that there exists a one-to-one relationship between signal intensity and target transcript abundance. However, the binding behaviour of surface-bound oligonucleotides varies greatly (Pozhitkov *et al.* 2006) and a realistic assessment of gene expression changes ideally requires consideration of the binding behaviour of probes (A. Pozhitkov, P. A. Noble, D. Tautz unpublished data). Second, our inference of the antisense probe sequence (required to produce the microarray) may have been wrong. For these two reasons, we have validated and calibrated our custom microarray for a range of concentrations that is appropriate for our experiment. The binding behaviour between a probe and target transcript depends not only on the template concentration, but also on the affinity of the probe and can be described using a dose–response curve that describes the increase in signal intensity as a function of template concentration (A. Pozhitkov, P. A. Noble, D. Tautz unpublished data). A calibration pool containing equimolar amounts of total RNA from representatives of both parental species and the invasive hybrid lineage was prepared, and a calibration series of seven microarrays were loaded with 0.125, 0.25, 0.5, 1, 2, 4 and 8 times the amount of recommended cRNA (600 ng) to detect the hybridization signal. The binding behaviour of the individual probes was assessed using custom R code (R Development Core Team 2010). For the purpose of this study, we were interested in identifying probes with a positive and approximately linear dose response and to exclude probes that did not detect a change in concentration or for which the signal approaches saturation. A linear function

$$SI_{calibration} = m_{probe} * c_{calibration} + b_{probe} \quad (1)$$

was fitted for each probe, where $SI_{calibration}$ is the raw signal and c is the amount of calibration pool (ng) for which the $SI_{calibration}$ was obtained. The slope m_{probe} and the intercept b_{probe} derived from this calibration step were used in the normalization of raw-signal measurements (see below). Probes with a non-linear binding behaviour (R^2 of the linear fit <0.95) and the most responsive probes with the 2.5% highest dose–response slopes were excluded. Furthermore, probes with slopes smaller than 0.3 were not used as a result of their unfavourable noise-to-signal ratios (see Fig. 1). The noise-to-signal ratio was defined as twice the average of the residual absolute values divided by the overall change in signal.

Data normalization

To make changes in signal intensity comparable among target transcripts, raw-signal intensities were normalized

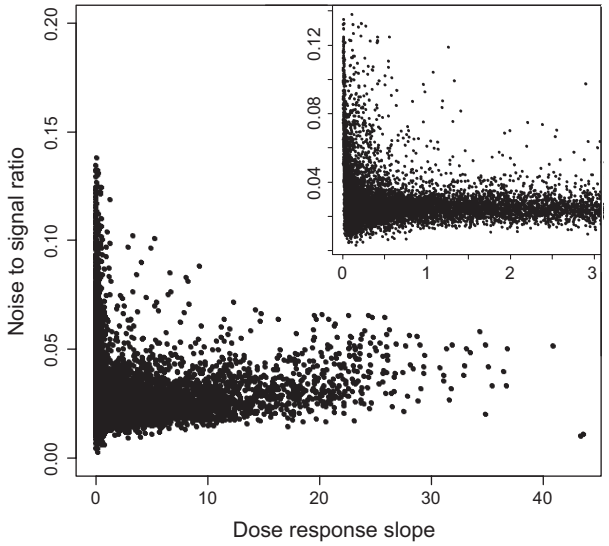

Fig. 1 Relation between slopes of the linear dose–response functions (*x*-axis) to the noise-to-signal ratio (*y*-axis) for all probes with a linear fit ($R^2 > 0.95$). Dots represent individual probes. Dose–response slopes vary greatly and slopes that exceed a value of 1 are common. The highest noise-to-signal ratios are found mostly at probes with a slope of less than ~0.3 (enlarged in small graph).

between probes to correct for different binding behaviours. For this purpose, raw-signal intensities were transformed such that they have a common signal to concentration change relationship in two steps.

First, $c_{detected}$ (in ng cRNA) corresponding to the raw signal intensities obtained in an experiment ($SI_{detected}$) were determined based on the coefficients from the dose–response functions derived in the calibration step (eqn 1).

$$c_{detected} = (SI_{detected} - b_{probe})/m_{probe} \quad (2)$$

This step already enables unbiased calculation of the gene expression fold changes from the same array. Second, based on $c_{detected}$, we calculated signal intensities expected ($SI_{corrected}$) under a uniform linear dose–response behaviour with slope one and intercept 0 for all probes.

$$SI_{corrected} = c_{detected} * 1/600\,ng * SI_{600ng}, \quad (3)$$

where SI_{600ng} is the signal intensity expected at 600 ng calibration pool ($SI_{600ng} = m_{probe}$*600 ng + b_{probe}). Thus, for example for $c_{detected}$ = 300 ng, $SI_{corrected}$ becomes exactly half SI_{600ng}. This last step reintroduces signal intensity as a proxy for transcript abundance to the normalized expression values, which allowed us to perform the between array normalization by dividing all normalized expression signal intensities of an array by their 75 percentile as recommended in the 'One-Color Microarray-Based Gene Expression Analysis—Low Input Quick Amp Labelling Protocol' (Version 6.0. December 2009 Agilent Technologies). Negative normalized signal intensity values were set to zero but not removed from the data set. The normalization procedure was performed with custom R code.

Patterns of gene expression

Our experimental design introduces genotype and environmentally induced effects which are controlled for by (i) using a covariate in the analysis and (ii) by visually checking resulting candidate genes of special interest for remnant effects associated with the origin of the samples. Variance in gene expression was explored using a principal component analysis (PCA) in R to identify major axis of variation in gene expression and to remove sources of variation that are not of interest in this study (Pickrell *et al.* 2010). The R-package R/MAANOVA (Wu 2010) was used to perform a two-stage ANOVA to test for differentiation in gene expression between 'species' (*C. rhenanus*, *C. perifretum* and the invasive hybrid lineage) and 'populations' (Broel, Naaf, LB, WN, Sieg). Differential gene expression was assessed using fixed effect models with the factor species or population respectively. The scores of PC 1 were used as a cofactor in the models to remove effects obscuring the signal suggestive of fixed between species differences. Significance of the pairwise comparisons was determined using a contrasting function implemented in R/MAANOVA ('*t*-test' option with 2000 permutations, Fs test option) with False Discovery Rate (FDR) estimation as described by Storey (2002).

The invasive sculpins in our analysis comprised six individuals sampled from nature and two individuals raised in our experimental aquaria. Genes that were transgressively over-expressed in the invasive turned out to be of major interest in the functional analysis (see below). Therefore, we examined for these genes if transgressive expression patterns were a result of a plastic response to environmental conditions or are a ubiquitous feature of the invasive lineage. According to scenario 1 (plastic response), individuals derived from nature should show transgressive levels, whereas laboratory-raised fish should not. Scenario 2 (static expression level) posits that no general difference should be observed in expression between individuals derived from nature or individuals raised in the laboratory. To remove the effects of PC 1 from gene expression values, a linear model was fitted with PC 1 being the only explanatory variate, and the residuals from each sample were used as a surrogate of its PC 1 independent expression value. These surrogates were plotted and the

genes transgressively over-expressed in the invasive were examined by eye into the categories 'genes showing a plastic response' (scenario 1) and 'genes with static expression level' (scenario 2), and re-analysed separately.

To test whether hybridization affects the variance of gene expression, we calculated the variance in gene expression on a per gene basis for all experimental groups. The variance describes how far data points are spread around their mean on an absolute basis and thus increases with expression value. To disentangle the differences in variance between the tested groups from the effect of the mean expression value on variance, we included the average expression value for each gene and group tested as a cofactor in the analysis. First, we fitted the same model which already was used for species level analysis with the R/MAANOVA package in R. The variance in gene expression was calculated individually per gene for each group from the residuals of the fitted model. The mean expression values per group were calculated for each gene excluding the influence of the first principal component from the same model. To prevent that outliers blur the general patterns in the downstream analysis, we excluded the genes responsible for the 5% highest variances and/or mean expression values for every group. A linear model was fitted in R for the remaining genes with the explanatory variable being 'group' as a categorical variable with the levels (*C. rhenanus* and *C. perifretum*, invasives and F_2 crosses) and 'mean expression value excluding PC 1' of the respective group as a covariate. The response variable was the variance in gene expression of the genes for the respective groups.

In this study, genes were considered to show a transgressive pattern of gene expression when gene expression levels were significantly different from both parental species (*C. rhenanus* and *C. perifretum*) and when the levels of gene expression were not intermediate between the parental species. We distinguished between genes that were transgressively up- and down-regulated relative to the parental populations. Transgressivity was tested for F_2 crosses as well as for invasive sculpins to analyse to what extent the patterns in the artificial hybrids and natural hybrids agree. To test whether the sets of genes transgressively expressed in the invasives and in the F_2 crosses overlapped more than expected by chance, we performed a one-tailed Fisher's exact test using the phyper function in R.

Annotation of genes and GO term enrichment analysis

For the purpose of the GO term enrichment analysis, we lowered the stringency of the statistical significance thresholds ($P < 0.05$ and $q < 0.05$) to use larger sets of genes to test for the over-representation of functional groups based on biological effects. EST contigs that were targeted by probes on our microarray were annotated using the software BLAST2GO (Conesa *et al.* 2005) using the default parameters. A homology search was performed with BLASTX against the NCBI nr database (as of 2/17/2010), and hits with an *e*-value $< 10^{-6}$ were used to extract GO annotations from several repositories (Conesa *et al.* 2005). We used the one-tailed Fisher's exact test with FDR correction implemented in Blast2GO to test if any GO terms were significantly over-represented in a candidate set of genes compared to all genes for which we had gene expression data after removing bad quality probes.

Results

Microarray calibration and normalization

Fitting of a linear model to the raw signal intensities obtained in the calibration experiment revealed that 11 630 of the 13 329 probes on the custom microarray showed an approximately linear dose–response behaviour. The slopes of the linear dose–response relationship with an $R^2 > 0.95$ ranged from 0.001 to 48.5. We excluded those with a slope smaller than 0.3 below which unfavourable noise-to-signal ratios increased (Fig. 1). Probes with a dose–response slope of above 20 were also removed, thus excluding the 2.5% most responsive probes from the analysis. As a last step, probes with more than two (out of eight) individuals per population having missing data were removed. Out of the 13 329 probes, 7479 could be analysed for differentiation in gene expression (Table S1, Supporting information).

Differentiation in gene expression

A PCA using the individuals as variables and expression data (4622 probes without missing data) as observations was used to explore major axes of variation in gene expression. PC1 explained ~28% of the total variance, whereas PC2 explained ~10% of the total variance. Individuals belonging to all experimental groups overlap greatly along axis one (Fig. 2). The ancestral species *Cottus perifretum* and *Cottus rhenanus* separate along axis two and F_2 crosses as well as invasive *Cottus* assume an intermediate position. Figure 2 suggests that the environment in which individuals have been raised contributes to PC1, which is supported when comparing the PC1 scores of the laboratory-reared individuals with those from the wild (Welch's test *P*-value < 0.001). However, the natural environment did not affect artificial F_2 crosses and *C. perifretum*, but these groups still

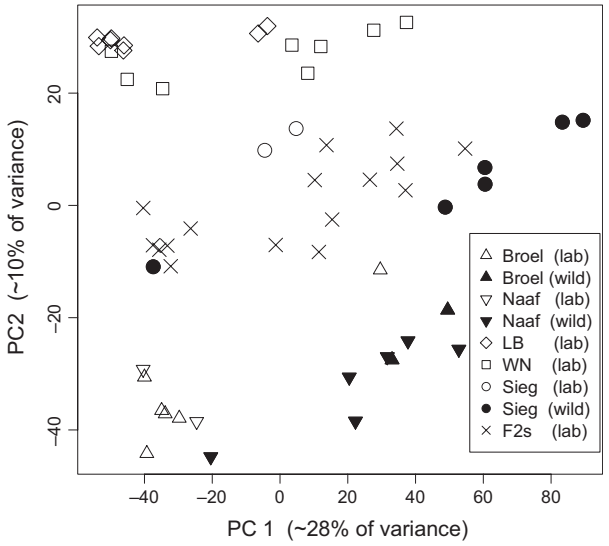

Fig. 2 Plot of individual scores for principal components of gene expression. *Cottus perifretum* is represented by the populations Witte Nete and Laarse Beek, and *Cottus rhenanus* is represented by the populations Naaf and Broel. Symbols are filled or empty depending on whether individuals grew up in the laboratory or in natural environments (legend). The environment contributes to the variance within groups along PC axis 1 (*x*-axis), whereas PC axis 2 (*y*-axis) separates species. Artificial F_2 crosses as well as invasive sculpins take an intermediate position between the parental species along PC axis 2.

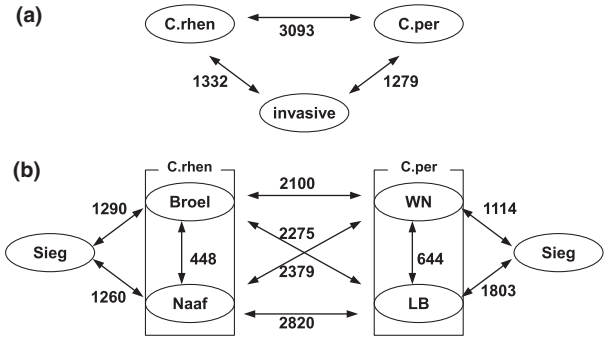

Fig. 3 Number of differentially expressed genes ($P < 0.01$; $q < 0.01$) between species (a) and populations (b). Broel and Naaf = *Cottus rhenanus*; Witte Nete and Laarse Beek = *Cottus perifretum*; A single population from the River Sieg represents the invasive hybrid lineage and is used in all comparisons. Numbers near arrows indicate differentially expressed genes.

vary along PC axis 1. Thus, it is likely that PC1 is affected by unknown factors. Taken together, these factors obscure the differences between species. We thus use PC1 scores as covariates to test for differences between species.

Differential expression between species and populations

The number of genes differentially expressed between *C. rhenanus* and *C. perifretum* was assessed for the species level and compared with invasive sculpins (Fig. 3a, Table S1, Supporting information), and also at the population level (Fig. 3b). Both analyses revealed that the number of genes differentially expressed between the parental species were in the order of twofold higher than the differentiation with the invasive hybrid lineage. Moreover, the comparison at the population level demonstrated that the within species differentiation (448 and 664 differentially expressed genes) is lower than the between species comparison (2100 and 2820 differentially expressed genes) and also lower than any comparisons with the invasive *Cottus* (1114–1803 differentially expressed genes) (Fig. 3b). The relative proportions of differentially expressed genes reflect the current classification of species and populations and suggest that the hybrid lineage shows intermediate levels of differentiation to both of its parental taxa. This pattern corresponds with the differentiation observed along PC axis two.

Variance in gene expression

A linear model was used to assess differences in the variance in gene expression. This model tested for the relationship between the categorical variable 'group' (levels: *C. perifretum*, *C. rhenanus*, invasives, F_2 crosses) and the covariate 'expression value' (mean level of gene expression per gene and group after removal of PC1) vs. variance in gene expression. *Cottus rhenanus* was used as the base category from which the other terms deviated. The model is highly significant ($P < 2*10^{-16}$) with an adjusted R^2-value of 0.4055. The estimated parameters indicate a significant effect of the 'expression value' term ($P < 2*10^{-16}$; estimate = 0.1150252) and indicate significant differences between the different levels of the 'group' term. The estimates for the group term reveal that the variance in gene expression is lowest in the parental species, being slightly lower for *C. perifretum* (intercept + estimate = −0.0303738) than for *C. rhenanus* (estimate = −0.0226103). The highest variance in gene expression was found for the F_2 crosses (intercept + estimate = 0.0158769), whereas for the invasive lineage variance in gene expression (intercept + estimate = 0.0024742) was intermediate between that of parental species and F_2 crosses. The *P*-values indicate that all estimates are significantly different from *C. rhenanus*. Using the invasive lineage as the base of the model, significant differences in gene expression variance are detected from both parental species ($P < 2*10^{-16}$) and from the F_2 crosses ($P = 1.12*10^{-10}$).

Transgressivity

The number of genes that were transgressively expressed in the invasive lineage and the F_2 crosses relative to the parental species was assessed separately for genes that have an up-regulated or down-regulated pattern (Table S1, Supporting information). A balanced proportion of 237 and 264 genes were up- and down-regulated in F_2 crosses, whereas an excess of 580 genes was transgressively up-regulated in invasive sculpins as opposed to 120 genes that were transgressively down-regulated in that group. When the intersecting sets of transgressively expressed genes among F_2 crosses and invasive sculpins were identified, 18 and 15 genes were found to be shared between up- and down-regulated genes respectively. Interestingly, transgressively down-regulated genes are shared more often between F_2 crosses and invasive sculpins than expected by chance ($P = 4.23*10^{-6}$), whereas sets of transgressively up-regulated genes do not have more genes in common than expected by chance ($P = 0.48$).

Functional annotation and GO term enrichment of candidate gene sets

We were able to annotate 4942 of the 7479 features entering analysis (Table S1, Supporting information) based on database searches (distant homologies) with model organisms, and this set of features served as a reference data when testing for the over-representation of GO terms. The largest fraction of 3952 features was found to be differentially expressed between populations representing the ancestral species *C. rhenanus* and *C. perifretum*. However, no enrichment of GO terms was detected for these genes. There are considerable numbers of features in which the invasive sculpins retain expression profiles from one parental species but not the other one (1203 and 1229 genes respectively), but none of these sets revealed over-represented GO terms. The analysis revealed 950 genes in which the invasive sculpins differed significantly from both ancestral species, and for this set, we detected an over-representation of the GO terms 'proteinaceous extracellular matrix', 'embryonic development' and 'nucleic acid binding' (Table 2). For the genes, transgressively up-regulated in the invasives, the same GO terms were found to be enriched together with the additional terms 'anatomical structure morphogenesis', 'transcription factor activity', 'nucleus', 'cellular component organization', 'DNA metabolic process', 'actin binding', 'motor activity' and 'cell cycle' (Table 2). In contrast, the genes transgressively down-regulated in the invasive were not enriched in any GO term. Analogous analyses were carried out for each population and cross separately which corroborated the GO terms and trends found with the inference based on pooled populations (not shown).

The samples from the invasive lineage consisted in their majority (six out of eight individuals) from captures from the wild but contained two laboratory-raised fish. We visually inspected levels of gene expression (after correcting for the variance explained by PC1) for all transgressively over-expressed genes and separated them into two groups. The first group contained 388 genes for which expression values of individuals from the laboratory-reared fish fell into the distribution of values of the wild-caught fish. The second group consisted of 192 genes where the significant transgressive pattern was caused by the fishes from the wild, and where our laboratory-reared fish showed a trend to be less differentiated from the parental populations. GO term enrichment analysis was performed for these two sets (see Table 2). For the 'genes with a static expression level' the GO terms 'embryonic development', 'nucleus', 'transcription factor activity', 'cell cycle', 'DNA metabolic process' and 'anatomical structure morphogenesis' were found to be enriched. This recovers GO terms already found during the GO term analysis of all genes which are transgressively over-expressed in invasive sculpins, although the FDR-corrected *P*-values are slightly higher. This situation was mirrored for the 'genes with a plastic response' for which the GO terms 'proteinaceous extracellular matrix', 'structural molecule activity', 'calcium ion binding', 'motor activity' and 'actin binding' were enriched. With the exception of 'calcium ion binding, these terms were already found to be enriched in the set of all genes that are transgressively over-expressed in the invasive sculpins.

Discussion

The theoretical appeal of hybridization lies in the fact that it may facilitate more complex evolutionary transitions than could be attained through the sequential accumulation of new mutations. A possible driver of such changes is given by the common process of transgressive segregation (Rieseberg *et al.* 2003a; Stelkens & Seehausen 2009; Stelkens *et al.* 2009), whereby recombined alleles in admixed individuals interact to create phenotypic values that lie outside the range observed in the ancestral species. Notably, transgressive patterns of gene expression are common in fish hybrids (Mavárez *et al.* 2009; Normandeau *et al.* 2009; Renaut *et al.* 2009; Bougas *et al.* 2010). This not only makes transcriptome analysis particularly useful to document the effects of hybridization, but also poses questions related to the evolutionary significance of such patterns. In this study,

Table 2 GO term enrichment of sets of differentially expressed genes. The column 'gene set' specifies pairwise comparisons and the nature of transgressive patterns, whereas the numbers of genes with GO annotation are given in column two

Gene set	Features/genes (with GO annotation)	Enriched GO terms	FDR corrected *P*-value
Differentially expressed between parental species *Cottus rhenanus* and *Cottus perifretum*	3952 (2587)	None	
Differentially expressed between invasive sculpins and both parental species	950 (611)	Proteinaceous extracellular matrix (C)	1.5 E-4
		Embryonic development (P)	7.9 E-3
		Nucleic acid binding (F)	7.9 E-3
Transgressively over-expressed in invasive sculpins: all genes	580 (387)	Proteinaceous extracellular matrix (C)	4.0 E-7
		Embryonic development (P)	3.0 E-6
		Anatomical structure morphogenesis (P)	7.2 E-5
		Transcription factor activity (F)	3.9 E-3
		Nucleus (C)	4.1 E-3
		Cellular component organization (P)	4.1 E-3
		DNA metabolic process (P)	4.1 E-3
		Actin binding (F)	4.1 E-3
		Motor activity (F)	4.5 E-3
		Cell cycle (P)	4.6 E-3
Transgressively over-expressed in invasive sculpins: genes with a static expression level	388 (265)	Embryonic development (P)	3.5 E-5
		Nucleus(C)	2.5 E-3
		Transcription factor activity (F)	2.7 E-3
		Cell cycle (P)	3.8 E-3
		DNA metabolic process (P)	6.6 E-3
		Anatomical structure morphogenesis (P)	8.5 E-3
Transgressively over-expressed in invasive sculpins: genes showing a plastic response	192 (122)	Proteinaceous extracellular matrix (C)	6.8 E-9
		Structural molecule activity (F)	1.6 E-3
		Calcium ion binding (F)	2.3 E-3
		Motor activity (F)	3.9 E-3
		Actin binding (F)	7.9 E-3
Transgressively under expressed in invasive sculpins	120 (73)	None	
Expression in invasive sculpins matches *C. rhenanus*, but not *C. perifretum*	1203 (794)	None	
Expression in invasive sculpins matches *C. perifretum*, but not *C. rhenanus*	1229 (782)	None	

FDR, false discovery rate; GO, gene ontology.
Significant enrichment (cut-off: FDR-corrected $P < 0.01$) of specific GO terms of these sets compared to all genes entering analysis is reported. GO terms are indexed as pertaining to the categories cellular component (C), molecular function (F) or biological process (P). There is no enrichment for genes differentially expressed between parental species. However, gene sets in which the invasive lineage has a higher expression than both parental species are enriched for several GO terms. This holds for genes with a static expression level and for genes showing a plastic response. In stark contrast, genes transgressively under-expressed in invasives or genes in which the invasives differ from only one parental species do not show signs for enrichment. Note that differential expression was called at $P < 0.05$ and $q < 0.05$ for the purpose of this analysis.

we document the emergence of novel patterns of gene expression in invasive *Cottus* from the river Rhine basin that is consistent with an evolutionary optimization process. Our results indicate that the majority of the distinguishing patterns of gene expression of invasive sculpins are evolutionarily modified after the initial admixture event rather than being an immediate feature of early generation hybrids. As for the technical aspects, we found non-linear binding behaviour and a wide range of dose–response slopes for the microarray probes that, unless taken into account, bias the inference of gene expression.

Microarray calibration

The microarray analysis described here followed established standard protocols with respect to the preparation of template, the hybridization and signal detection procedures. The analysis was complemented by performing a calibration of our custom array that assesses

the relationship of signal intensity and transcript abundance to identify probes that produced a reliable signal. As our calibration pool originated from the species analysed here, the calibration is specifically tuned to the range of transcript concentrations encountered in this experiment. Even after excluding probes that were not responsive or did not have a linear dose response, we observed a broad range of dose–response behaviours that vary more than 60-fold (Fig. 1), although the dilution series was identical for all target sequences. It is likely that such distortions of the signal-to-gene expression ratio complicate comparisons of different approaches to study transcriptomes (Jeukens *et al.* 2010). This finding is in line with our previous results (Pozhitkov *et al.* 2006) and has prompted our claim that the diversity of binding behaviours of probes needs to be considered in microarray experiments (A. Pozhitkov, P. A. Noble, D. Tautz unpublished data). We have implemented a normalization procedure that establishes a common signal-to-fold change ratio for all probes and thus disentangles probe effects from changes in expression.

Hybrid origin and adaptive differentiation of invasive sculpins

The PCA of the normalized expression values reveals a first major axis of variation that is correlated with the environment in which the fish were raised and possibly other factors (Fig. 2). Ancestral species do not differentiate along this axis and it does not capture features that make the hybrid transcriptome different from what is observed in the parental species. In contrast, PC axis 2 clearly separated the ancestral species and indicates that the overall gene expression profile of the F_2 crosses and the invasive hybrid lineage is intermediate between the two parental species. Hybrid intermediacy is also reflected in the number of genes differentially expressed between the species (Fig. 3). The number of significant genes between the *Cottus rhenanus* and *Cottus perifretum* is more than twice as high as the number of genes differentially expressed between the hybrid species and either parental species which is corroborated in the analysis at the population level. The relative proportions of genes differentially expressed in pairwise comparisons between species and populations agree with the current taxonomic distinction of the ancestral species (Freyhof *et al.* 2005) and levels of genetic differentiation among different populations observed at SNP (single nucleotide polymorphism) or microsatellite loci (Nolte *et al.* 2005; Stemshorn *et al.* 2011). The number of differentially expressed genes and the divergence time (~2 Myr, Englbrecht *et al.* 2000) between *C. rhenanus* and *C. perifretum* are the largest in this analysis, but we have not found over-representation of GO terms that would indicate biological functions that play a special role in this divergence.

Although the invasive lineage of sculpins phenotypically resembles *C. perifretum*, it possesses a unique potential to invade summer warm river habitats (Nolte *et al.* 2005, 2006). Although this inference is supported by general patterns of distribution and gene flow, specifically across hybrid zones, we have not identified the traits or precise ecological factors that cause the particular fitness properties of invasive sculpins. This study provides a first inventory of functional genetic variance that may contribute to novel adaptive phenotypes. The identification of the respective traits also represents a step towards understanding whether ecological changes in *Cottus* can be induced by the process of hybridization (Abbott *et al.* 2010, Nolte & Tautz 2010). In contrast to the within and between species comparisons, the invasive transcriptome is distinguished through differential expression of a set of genes that is significantly biased towards several functions (Table 2). This suggests that a significant proportion of the changes in gene expression is functionally related and therefore affects common, more complex, phenotypes. Whereas ideas about hybrid speciation centre on an instrumental role of hybrid ancestry and the recombinant nature of novel adaptive traits (Mallet 2007; Abbott *et al.* 2010), an alternative explanation for the presence of an adaptive potential relates to the evolution of traits in allopatry within the parental species. Such a preadaptation hypothesis posits that traits in invasive sculpins convey a fitness advantage without the genetic effects of admixture. Invasive sculpins shared expression patterns at 1203 and 1229 genes with *C. rhenanus* and C. *perifretum* respectively. Although these genes could in principle represent characters that are retained from the ancestral species through a selective process, GO term enrichment analysis could not detect any enrichment in these gene sets (see Table 2). Hence, the biological functions that are distinct and enriched in invasive sculpins are most likely not characters that were evolutionarily retained from one of the ancestors but have emerged in invasives. The fact that some transgressive patterns of gene expression of invasives could be recreated in F_2 crosses suggests that the process of hybridization has contributed to functional changes.

The most significantly enriched biological process was 'embryonic development'. This is surprising, as all fish in the analysis were juveniles transitioning to the adult stage. On the contrary, this can be explained because GO term analysis, although a valuable tool for meta-analysis, suffers from annotation biases as a result of idiosyncrasies of different research communities (Leong & Kipling 2009). However, screening sets of

enriched genes in the light of evolutionary ecological hypothesis yields promising candidate genes. For example invasive sculpins are found in parts of rivers that have a different temperature regime (Nolte *et al.* 2005, 2006). The influence of temperature on the distribution of species is described by the concept of oxygen and capacity limited thermal tolerance (OCLTT) (Pörtner & Knust 2007). It states for aquatic organisms that the capability to perform at high or low temperatures is determined by a mismatch between the demand for oxygen and the capacity to supply oxygen to the tissues. Accordingly, Eliason *et al.* (2011) showed that selection pressures associated with the temperature regime led to fast evolutionary modifications of the blood transportation system in salmonids. A re-evaluation of the genes causing enrichment of the GO term 'proteinaceous extracellular matrix' in invasive sculpins with respect to functions expected under the OCLTT concept yielded several genes (e.g. matrix metallopeptidase 2, transforming growth factor beta) that are known to play a role in the formation of blood vessels (Fisher & Berger 2003) but were not annotated for the GO term 'angiogenesis'. Hence, it would be possible that an adaptive modification of the oxygen transportation system escapes GO term analyses.

From initial hybridization to the invasive phenotype

From an evolutionary ecological perspective, an emerging hybrid lineage must separate from the parental species to avoid recurrent backcrossing and direct competition (Buerkle *et al.* 2000; Barton 2001) Seehausen (2004) and Stemshorn *et al.* (2011) discuss examples for how new adaptive peaks are exploited through the increased genetic variance present in a 'hybrid swarm'. Barton (2001) suggested that the rise and fixation of newly adapted hybrid genotypes with a superior fitness will take some time and that a population of admixed, outbreeding hybrids will go through a lag phase before selection has optimized novel genotypic combinations. Although evidence that hybridization has played a key role in the evolution of hybrid species was obtained through experimental re-creation of hybrid phenotypes in the laboratory for sunflowers (Rieseberg *et al.* 2003b) and *Heliconius* butterflies (Mavárez *et al.* 2006), empirical studies that document the evolutionary transition from initial hybrids to an admixed lineage are still scarce. The sequence of events that occur in consecutive generations following initial hybridization was studied in *Senecio squalidus*, a hybrid species of plant that became invasive in the UK after being transplanted from a hybrid zone of its parental species, *Senecio aethnensis* and *Senecio chrysanthemifolius*. Hegarty *et al.* (2009) experimentally re-created crosses of the parental species and traced changes in gene expression from the F_1 to the F_5 generation. They detected a relatively high variance in gene expression and patterns of non-additive and transgressive gene expression in initial generations and found that these patterns were considerably reduced in the F_4 and F_5 generations post-hybridization. As seed germination rates sank from the F_1 to the F_3 generation, but increased later on, Hegarty *et al.* (2009) suspected that unintentional selection may have removed maladaptive patterns of gene expression. This indicates that phenotypic variability of admixed populations can be significantly altered after only a few generations in a particularly dynamic process (Nolte & Tautz 2010).

Although we have not analysed consecutive generations of the artificial *Cottus* hybrids, an analysis can reveal to what extent patterns of gene expression in the natural hybrids are adumbrated in the laboratory-reared hybrids. Invasive sculpin gene expression was less variable than that of the F_2 generation, but was clearly more variable than in its parental species. This corresponds with a loss of heterozygosity across the invasive sculpin genome over time (Stemshorn *et al.* 2011) and agrees with the prediction that an initially maximal genetic diversity is reduced as maladaptive genetic variance is lost. Moreover, we have compared the numbers and types of transgressively over- and under-expressed genes between F_2 crosses and invasive sculpins to test how patterns that reflect initial admixture may have been modified as invasive sculpins evolved. The numbers of transgressively up- and down-regulated genes in the F_2 crosses are approximately balanced, but this pattern did not prevail in invasive *Cottus*. The number of transgressively down-regulated genes dropped, whereas the number of transgressively up-regulated genes more than doubled from the F_2 generation to invasive *Cottus*. At the same time, many of the initial transgressive patterns were lost and new ones emerged. Interestingly, the set of transgressively up-regulated genes in invasive sculpins turned out to be the set of genes that showed the strongest enrichment in our study (Table 2), whereas down-regulated genes in invasives were not significantly enriched for any GO term. Enrichment analysis of the transgressively over-expressed genes reproduced GO terms that were previously found to be characteristic for the whole transcriptome of invasive *Cottus*, albeit with higher significance levels (Table 2). This suggests that the functional enrichment detected in the differentially expressed genes is mainly caused by the genes which are transgressively over-expressed in the invasive hybrid lineage and that this set represents a non-random set that is functionally linked. Intriguingly,

transgressive over-expression seemed to be plastic for a large subset of these genes, which may suggest that phenotypic plasticity interacts with fixed traits to create evolutionary novelty (Moczek *et al.* 2011). This aspect remains unexplored for the origin of the evolutionary novelty and adaptation of the invasive *Cottus* hybrid lineage. To date, the set of transgressively over-expressed genes represents the best candidates to further study the adaptive advantage that invasive sculpins have over their ancestors in the lower River Rhine basin (Nolte *et al.* 2005; Stemshorn *et al.* 2011). Conversely, a significant number of transgressively down-regulated genes in invasive sculpins were retained from the pool of genes that showed this pattern initially. We did not detect any enrichment of GO terms for this set of genes, but it is possible that selection has contributed to maintain these patterns and that they represent transgressive traits that have immediately contributed positively to the fitness of invasive sculpins.

One of the most conspicuous results of this study is that the gene expression features that distinguish invasive sculpins are mostly up-regulated. Such a concerted change would not be expected if regulatory changes evolved independently to modify adaptive phenotypes as it was found by Lai *et al.* (2006). We consider it more likely that this result involves the exaggerated growth of some organs or tissues, which would result in a concerted rise of gene expression for genes specifically expressed in such tissues. On the one hand, this impairs the detection of the primary genetic factors that cause changes in gene expression. On the other hand, it provides us with a more complex phenotype that can be studied experimentally. Besides mapping of the underlying genes (J. Cheng, T. Czypionka, A. W. Nolte unpublished data), population genetic signatures of fixation in the invasive gene pool can be exploited to identify genomic regions (Stemshorn *et al.* 2011) that contribute to emerging phenotypes in the course of genomic admixture events.

Acknowledgements

We thank F. Volckaert, K. de Gelas, A. Kobler and L. Bervoets for their benevolent help in obtaining breeding stocks of *Cottus perifretum*. K. Lessenich and A. Mellin from Bezirksregierung Köln, Nordrhein—Westfalen, and T. Heilbronner from the Siegfischereigenossenschaft have given permits to conduct the research. *Cottus* were kept in the laboratory of the MPI in Plön with permission from S. Hauschildt (Veterinaeramt in Plön, Schleswig—Holstein). E. Blohm-Sievers, E. Bustorf, H. Buhtz, S. Dembeck, H. Harre, G. Augustin, D. Mertens and D. Lemke have contributed to the laboratory work and fish care. We thank K. Konrad, T. Domazet-Lošo and D. Tautz for comments and discussions, and D.T. and the Max-Planck Society for generous support. The project was funded by a DFG grant to A.N.

References

Abbott RJ, Hegarty MJ, Hiscock SJ, Brennan AC (2010) Homoploid hybrid speciation in action. *Taxon*, **59**, 1375–1386.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

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

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

Bougas B, Granier S, Audet C, Bernatchez L (2010) The transcriptional landscape of cross-specific hybrids and its possible link with growth in Brook Charr (*Salvelinus fontinalis* Mitchill). *Genetics*, **186**, 97–107.

Buerkle CA, Morris RJ, Asmussen MA, Rieseberg LH (2000) The likelihood of homoploid hybrid speciation. *Heredity*, **84**, 441–451.

Charlesworth D, Willis JH (2009) The genetics of inbreeding depression. *Nature Reviews Genetics*, **10**, 783–796.

Cheviron ZA, Whitehead A, Brumfield RT (2008) Transcriptomic variation and plasticity in rufous-collared sparrows (*Zonotrichia capensis*) along an altitudinal gradient. *Molecular Ecology*, **17**, 4556–4569.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Domazet-Lošo T, Tautz D (2010) A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*, **468**, 815–818.

Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, **30**, 207–210.

Eliason EJ, Clark TD, Hague MJ *et al.* (2011) Differences in thermal tolerance among Sockeye Salmon populations. *Science*, **332**, 109–112.

Englbrecht CC, Freyhof J, Nolte AW, Rassmann K, Schliewen U, Tautz D (2000) Phylogeography of the bullhead *Cottus gobio* (Pisces: Teleostei: Cottidae) suggests a pre-pleistocene origin of the major central European populations. *Molecular Ecology*, **9**, 709–722.

Fisher WE, Berger DH (2003) Angiogenesis and antiangiogenic strategies in pancreatic cancer. *International Journal of Gastrointestinal Cancer*, **33**, 79–88.

Freyhof J, Kottelat M, Nolte AW (2005) Taxonomic diversity of European *Cottus* with description of eight new species (Teleostei: Cottidae). *Ichthyological Exploration of Freshwaters*, **16**, 107–172.

Hegarty MJ, Barker GL, Brennan AC, Edwards KJ, Abbott RJ, Hiscock SJ (2009) Extreme changes to gene expression associated with homoploid hybrid speciation. *Molecular Ecology*, **18**, 877–889.

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.

Knapen D, Knaepens G, Bervoets L, Taylor MI, Eens M, Verheyen E (2003) Conservation units based on mitochondrial and nuclear DNA variation among European bullhead populations (*Cottus gobio* L., 1758) from Flanders, Belgium. *Conservation Genetics*, **4**, 129–140.

Lai Z, Gross BL, Zou Y, Andrews J, Rieseberg LH (2006) Microarray analysis reveals differential gene expression in hybrid sunflower species. *Molecular Ecology*, **15**, 1213–1227.

Laikre L, Schwartz MK, Waples RS, Ryman N (2010) Compromising genetic diversity in the wild: unmonitored large-scale release of plants and animals. *Trends in Ecology & Evolution*, **25**, 520–529.

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

Larsen PF, Schulte PM, Nielsen EE (2011) Gene expression analysis for the identification of selection and local adaptation in fishes. *Journal of Fish Biology*, **78**, 1–22.

Leong HS, Kipling D (2009) Text-based over-representation analysis of microarray gene lists with annotation bias. *Nucleic Acids Research*, **37**, e79.

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

Mavárez J, Salazar CA, Bermingham E, Salcedo C, Jiggins CD, Linares M (2006) Speciation by hybridization in *Heliconius* butterflies. *Nature*, **441**, 868–871.

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

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

Michalak P, Noor MAF (2004) Association of misexpression with sterility in hybrids of *Drosophila simulans* and *D. mauritiana*. *Journal of Molecular Evolution*, **59**, 277–282.

Moczek AP, Sultan S, Foster S *et al.* (2011) The role of developmental plasticity in evolutionary innovation. *Proceedings of the Royal Society B: Biological Sciences*, **278**, 2705–2713.

Nolte AW, Tautz D (2010) Understanding the onset of hybrid speciation. *Trends in Genetics*, **26**, 54–58.

Nolte AW, Freyhof J, Stemshorn KC, Tautz D (2005) An invasive lineage of sculpins, *Cottus* sp. (Pisces, Teleostei) in the Rhine with new habitat adaptations has originated from hybridization between old phylogeographic groups. *Proceedings of the Royal Society B: Biological Sciences*, **272**, 2379–2387.

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

Nolte AW, Renaut S, Bernatchez L (2009) Divergence in gene regulation at young life history stages of whitefish (*Coregonus sp.*) and the emergence of genomic isolation. *BMC Evolutionary Biology*, **9**, 59.

Normandeau E, Hutchings JA, Fraser DJ, Bernatchez L (2009) Population-specific gene expression responses to hybridization between farm and wild Atlantic salmon. *Evolutionary Applications*, **2**, 489–503.

Pickrell JK, Marioni JC, Pai AA *et al.* (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**, 768–772.

Pörtner HO, Knust R (2007) Climate change affects marine fishes through the oxygen limitation of thermal tolerance. *Science*, **315**, 95–97.

Pozhitkov A, Noble PA, Domazet-Lošo T *et al.* (2006) Tests of rRNA hybridization to microarrays suggest that hybridization characteristics of oligonucleotide probes for species discrimination cannot be predicted. *Nucleic Acids Research*, **34**, e66.

R Development Core Team (2010) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

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.

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

Rieseberg LH, Widmer A, Arntz AM, Burke B (2003a) The genetic architecture necessary for transgressive segregation is common in both natural and domesticated populations. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, **358**, 1141–1147.

Rieseberg LH, Raymond O, Rosenthal DM *et al.* (2003b) Major ecological transitions in wild sunflowers facilitated by hybridization. *Science*, **301**, 1211–1216.

Seehausen O (2004) Hybridization and adaptive radiation. *Trends in Ecology & Evolution*, **19**, 198–207.

Seehausen O, van Alphen JJM, Witte F (1997) Cichlid fish diversity threatened by eutrophication that curbs sexual selection. *Science*, **277**, 1808–1811.

Stelkens R, Seehausen O (2009) Genetic distance between species predicts novel trait expression in their hybrids. *Evolution*, **63**, 884–897.

Stelkens RB, Schmid C, Selz O, Seehausen O (2009) Phenotypic novelty in experimental hybrids is predicted by the genetic distance between species of cichlid fish. *BMC Evolutionary Biology*, **9**, 283.

Stemshorn KC, Reed FA, Nolte AW, Tautz D (2011) Rapid formation of distinct hybrid lineages after secondary contact of two fish species (*Cottus sp.*). *Molecular Ecology*, **20**, 1475–1491.

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

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

Volckaert F, Hänfling B, Hellemanns B, Carvalho G (2002) Timing of the population dynamics of bullhead *Cottus gobio* (Teleostei: Cottidae) during the Pleistocene. *Journal of Evolutionary Biology*, **15**, 930–944.

Wu H (2010). maanova: Tools for analyzing microarray experiments. R package version 1.20.0. Modified by Yang H,

Sheppard K, with ideas from Churchill G, Kerr K and Cui X. http://CRAN.R-project.org/package=maanova

Xia Q, Cheng D, Duan J *et al.* (2007) Microarray-based gene expression profiles in multiple tissues of the domesticated silkworm, *Bombyx mori*. *Genome Biology*, **8**, R162.

This paper is part of the PhD project of T.C. who studies the evolution of gene expression in hybrid fish in the lab of A.W.N. T.C. and J.C. have jointly performed the laboratory crosses, EST sequencing and raw sequence analysis as a resource for evolutionary genetic analyses in Cottus. A.P. conducts interdisciplinary research on physical chemistry of nucleic acids and bioinformatics; for this paper he developed the microarray calibration approach. T.C. and A.W.N. have conceived and implemented the research and written this paper.

Data accessibility

Gene expression data have been deposited in NCBI's Gene Expression Omnibus (Edgar *et al.* 2002). Expression data used for the calibration of the microarray are accessible through GEO Series accession number GSE36800. Expression data used for the comparative analysis of *Cottus rhenanus*, *Cottus perifretum*, the F_2 crosses and the invasive hybrid lineage are accessible through GEO Series accession number GSE36755. Raw data only contain sequence information for the probes on the microarray; the full target sequence information for transcripts analysed in this study is provided in Table S1 (Supporting information).

Supporting information

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

Table S1 Target transcripts, microarray probes and differences in gene expression between *C. rhenanus*, *C. perifretum*, invasive sculpins and laboratory-bred F_2 crosses.

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