An investigation into adaptation in genes associated with exposure to estrogenic pollution in populations of roach (*Rutilus rutilus*) living in English rivers

AUTHORS

Patrick. B. Hamilton¹²*, Anne E. Lockyer³, Tamsyn M. Uren Webster¹⁴, David J. Studholme¹, Josephine R. Paris¹, Alice Baynes³, Elizabeth Nicol³, Deborah A. Dawson⁵, Karen Moore¹, Audrey Farbos¹, Susan Jobling³, Jamie R. Stevens¹, Charles R. Tyler¹

1. Biosciences, College of Life and Environmental Sciences, University of Exeter, Exeter, EX4 4QD, UK

2. College of Medicine and Health, St Luke's Campus, Heavitree Road, Exeter, EX1 2LU
Exposure of male fish to estrogenic substances from wastewater treatment works (WwTWs) results in feminization and reduced reproductive fitness. Nevertheless, self-sustaining populations of roach (Rutilus rutilus) inhabit river stretches polluted with estrogenic WwTW effluents. In this study we examine whether such roach populations have evolved adaptations to tolerate estrogenic pollution by comparing frequency differences in single nucleotide polymorphisms (SNPs) between
populations sampled from rivers receiving either high or low level WwTW discharges. SNPs within 36 ‘candidate’ genes, selected for their involvement in estrogenic responses, and 120 SNPs in reference genes were genotyped in 465 roach. There was no evidence for selection in highly estrogen-dependent candidate genes, including those for the estrogen receptors, aromatases and vitellogenins. The androgen receptor (ar) and cytochrome P450 1A genes were associated with large shifts in allele frequencies between catchments and in individual populations, but there is no clear link to estrogen pollution. Selection at ar in the effluent dominated River Lee may have resulted from historical contamination with endocrine disrupting pesticides. Critically, while our results suggest population-specific selection including at genes related to endocrine disruption, there was no strong evidence the selection resulted from exposure to estrogen pollution.
The occurrence of feminized male fish has been reported in rivers and estuaries on several continents and has been attributed to pollution by natural and synthetic steroid estrogens, including ethinylestradiol (EE2),\(^1\)\(^2\) contained in wastewater treatment work (WwTW) effluents. Feminized male characteristics known to be induced by steroid estrogens include the presence of precursors of egg yolk proteins, such as vitellogenin (VTG), in the blood plasma,\(^3\) feminized reproductive ducts and the presence of developing eggs in otherwise male gonads.\(^4\) This intersex phenomenon associated with exposures to WwTW effluents was first reported to be widespread in roach (Rutilus rutilus) in English rivers in the 1990s and the 2000s,\(^5\)\(^6\) and has since been reported in many species of both riverine and estuarine fish in several countries of the world.

In vitro fertilization studies using wild male roach (Rutilus rutilus)\(^7\) indicate that fish with feminized gonads have reduced fertility, and a competitive breeding study found wild male roach with moderately to severely feminized gonads to have reduced reproductive output.\(^8\) Exposures of roach (Rutilus rutilus) to undiluted effluent\(^9\) or to 4-6 ng/L EE2 over the period of sexual development\(^10\)\(^-\)\(^11\) have been shown to result in full sex reversal and/or breeding failure and long-term laboratory exposures to lower concentrations of 0.47-1 ng/L EE2 (predicted for rivers heavily dominated with WwTW effluents) have resulted in female-skewed sex ratios and
decreased egg fertilization for several fish species. Furthermore, dosing of a lake in Canada with 4-6 ng/L EE2 over a period of three years resulted in the collapse of the fathead minnow (Pimephales promelas) population which subsequently recovered after removal of EE2.

Population genetic studies on wild roach across 28 UK sample sites, however, found no significant negative correlation between effective population sizes and modeled estimates of steroid estrogen exposure, and demonstrated the existence of self-sustaining roach populations over multiple generations. This raises the question of whether such populations have evolved to tolerate the harmful effects of steroid estrogen. Several studies have demonstrated that populations of Atlantic killifish (Fundulus heteroclitus) and Atlantic tomcod (Microgadus tomcod), have developed tolerance to specific pollutant classes including to polycyclic aromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs) and dioxin-like compounds. In these cases, adaptation has involved selection for genes associated with the aryl hydrocarbon receptor (AhR) response that regulates metabolism of hydrocarbon contaminants, including cytochrome P450 1A (cyp1A). No studies have examined whether wild populations of fish have adapted to steroid estrogens found in WwTW effluents, although studies in both mammals and fish show evidence for a genetic influence on responses to estrogen and that polymorphisms in genes for steroid receptors are associated with a variety of impacts on fitness (reproduction and/or likely survival). For the roach, even though prolonged exposure impairs
reproductive fitness, no studies have examined whether genetic differences alter sensitivity to estrogen, or investigated evidence for adaptation to estrogen pollution.

In order to investigate the potential for adaptation, we studied roach populations in two eastern English catchments with well-documented histories of exposure to estrogenic WwTW effluent. An analysis was conducted of frequency differences in single nucleotide polymorphisms (SNPs) in genes involved in estrogen response to test for evidence of directional selection and potential adaptation.25

MATERIALS AND METHODS

Study Species: Roach (Rutilus rutilus). Populations of roach (Rutilus rutilus, a cyprinid fish) occur widely in UK rivers that differ their WwTW effluent content. Numerous obstructions, such as locks and weirs can restrict fish movement, containing populations of roach within defined river stretches.17 See Additional file 1 for a more detailed rationale about the choice of study species.

Sampling and Choice of Rivers. Five of the locations in four rivers (Rivers Aire, Lee, Mole, and Foss) were selected for this study had historically been contaminated with WwTW effluents (for simplicity we refer to as ‘high estrogen/estrogenic’). Studies in all these rivers have demonstrated estrogenic activity of the river water and/or the presence of feminised male roach.6 26 Five locations had low or no WwTW effluent inputs referred to as ‘clean’ (Figure 1), although they may have other
sources of pollution. Modeled estimates of steroid estrogens and estrogenic alkylphenolic chemicals\textsuperscript{27} had previously been calculated using the geographical information systems-based model (LF2000-WQX). This model predicts the estradiol equivalents (E\textsubscript{2}Eq) (see \textsuperscript{28}), an estimate of estrogenic potency which correlates with the actual incidence and severity of intersex in fish found downstream of WwTWs.\textsuperscript{6} See Additional file 1 for further details on study site selection criteria and river history.

The biological material (fin clips) for genetic analysis were obtained from a combination of freshly collected material (from roach captured via an electrofishing in 2012-2014) and samples collected from previous studies between 2010-2012.\textsuperscript{17, 29} A total of 640 individuals were specifically sampled for this study for SNP and/or microsatellite analyses (Additional file 1, Table S1), collected following UK Home Office procedures.
Population Genetic Analysis. To better understand the history of each roach population sampled, population genetic structure was investigated using DNA microsatellite analysis (Additional file 1, Table S2) as described previously.¹⁷ The genotypes obtained were combined with the dataset on 1,769 fish sampled between 1995 and 2011 (a total of 51 population ‘samples’ from 41 sites; see detail in Additional file 1).

The same procedures were used for population-genetic analyses of SNP data. Analyses were based on 217 SNP loci from 465 individuals from nine different sample sites.

Candidate Gene Selection. We adopted a targeted approach to SNP genotyping. Candidate genes were selected from literature searches and published datasets (Additional file 2). These included estrogen receptors, aromatases and other estrogen-regulated genes that play key roles in reproduction, growth and development. These are often found to be differentially regulated following estrogen exposure.³⁰ For some genes, evidence of estrogen regulation is from mammals, and has not yet been investigated in fish eg. brca and bcar genes. In addition, we included genes previously identified as being involved in adaptation in other fish species (see Additional file 1).

Available sequences for these genes in roach, zebrafish and other fish were then used to select orthologous genes in the roach transcriptome using the BLASTn and
tBLASTx algorithms implemented in Seqtools version 8.4.017 (http://www.seqtools.dk/) and the roach transcriptome as a local database.

**Transcriptome Sequencing/Assembly.** The transcriptome of roach was sequenced in order to identify genetic variants for subsequent SNP genotyping. These were submitted to NCBI Short Read Archive (SRA) associated with BioProject PRNJA295813. A de novo transcriptome was generated from the trimmed, filtered and repaired FASTA files using sequences from 8 libraries using Trinity (version:trinityrnaseq_r20140717). The resulting FASTA file was submitted to the Transcriptome Shotgun Assembly sequence database (TSA) associated with BioProject PRJNA295813.

**Roach Genome Sequencing.** The genome of a single male roach was sequenced; reads are available via the Transcriptome Shotgun Assembly sequence database (TSA): PRJEB14887.

**SNP Identification.** Reads from each library were mapped back to the modified transcriptome using the Burrows-Wheeler Aligner (BWA) program version 0.7.5a-r405. Variant sites were identified using a custom Perl script (Additional file 3). The fragmented roach genome sequences were then used to identify intron positions, so that they could be avoided or included in the SNP-genotyping primers. SNPs from the transcriptome were substituted into the corresponding position in contigs assembled from the genome sequencing using a custom script (Additional file 4).
Additional SNPs for priority genes were identified by designing primer sequences from genomic contigs and these were used for Sanger sequencing (Additional file 1, Table S3). The sequences including the SNPs are shown in Additional file 5.

**SNP Genotyping.** Three hundred and fifty SNPs were selected for genotyping using the Kompetitive Allele-Specific PCR (KASP™) assays (LGC genomics), following whole genome amplification (WGA) using the primer extension pre-amplification (PEP-PCR) method (https://www.lgcgroup.com/). Up to 5 SNPs in each candidate gene were chosen whereas a single SNP was chosen from each reference gene by randomly selecting transcripts of named genes from the transcriptome with only one isoform.

**Tests for Selection Using Environmental Correlations LFMM.** The full SNP dataset (Additional file 6) was analyzed using the landscape genomics approach implemented in the programme LFMM (“latent factor mixed models”) 33 (see Additional file 1).

**Tests for Selection Using Pairwise FST Outlier Tests.** Differences in allele frequencies between populations in rivers sites were also used to identify loci under selection. Outliers in multiple comparisons of populations from polluted rivers with those from clean rivers within each catchment would be considered strong candidates of selection resulting from estrogen exposure. BayeScan version 2.134 (provided at http://cmpg.unibe.ch/software/BayeScan/) and fdist program35 implemented in Lositan36 were both used to identify loci exhibiting extreme $F_{ST}$
values. Of the available methods, FDIST2 and BayeScan typically had the lowest type II error, BayeScan had the least type I error.\textsuperscript{37}

**Full Dataset Analysis.** BayeScan and the hierarchical method implemented in Arlequin 3.5,\textsuperscript{38} which is more robust to differences in population history were used to identify loci under selection from analysis of whole dataset.

**Statistical Analysis.** To test for differences between candidate and reference genes, probability/p-values were compared for candidate genes and reference genes using Mann-Whitney U tests (see Additional file 1 for more detail). The test statistics/p-values were averaged for the multiple SNPs for each candidate gene, so that a each candidate gene is represented by a single value in the statistical analyses; this was done to avoid repeated sampling and non-independence.

**SNP Genotyping: RAD-Seq.** The population from the polluted River Lee (LeeWhe) was compared with two low effluent river populations (CufBro, KenNor) from the same catchment using RAD-seq in order to examine SNPs throughout the genome. Restriction site associated RAD libraries were as described in Etter et al.\textsuperscript{39} We used Stacks version 1.40\textsuperscript{40-41} for building loci and calling SNPs in three populations. BLAST analysis was used to identify the sequence 5 kb\textsuperscript{42} in either direction in the fathead minnow (P. promelas) genome, a relatively close relative of the roach. For RAD loci which had $F_{ST}$ values of greater than 0.1 BLASTx and BLASTn\textsuperscript{43} searches against the zebrafish Ensembl\textsuperscript{44} peptide and nucleotide databases were used to identify genes within the RAD loci or within the corresponding fathead minnow sequences genes,
using an e value cut off of $< 1 \times 10^{-5}$. To identify the population in which selection is likely to have occurred, $F_{ST}$ values for loci of interest were examined in the other two pairwise comparisons. Less stringent criteria ($F_{ST} > 0.8$, $p < 0.05$) were used for this comparison. Gene ontology (GO) analysis was conducted in Database for Annotation, Visualisation and Integrated Discovery (DAVID),\textsuperscript{45} using Danio rerio as a background.

RESULTS

Single Nucleotide Polymorphism (SNP) Identification and Genotyping.

Transcriptome sequencing yielded 184.5 million reads 150 bp paired-end reads after quality trimming (94.04%) – Table S4. The transcriptome assembly yielded 200,361 transcripts (summary statistics are given in Additional file 1, Table S5). 25,886 genes were identified using the Ensembl peptide database for Danio rerio. Genome sequencing of a single male roach generated 249.7 million reads after removal of low quality sequences.

A total of 217 SNPs were successfully genotyped in 465 fish from 10 locations in 9 rivers with overall genotyping success of 99.24%. Eighty four were in 36 genes related to estrogen response candidate genes, 12 were in four other genes related to selection and 120 were each in a different reference gene (see Table S6 for genotyped candidates - Additional file 1). SNPs within genes of some of the most obvious candidate genes for estrogen adaptation were successfully genotyped
including the three nuclear estrogen receptors, the membrane-bound estrogen receptor (gper), the androgen receptor (ar), brain (cyp19a) and gonadal (cyp19b) cytochrome p450 genes, vtg3, and the main vitellogenin (vtg) locus which includes vtg 1-2, 4-7 genes.

**Analysis of Population-Genetic Structure Using DNA Microsatellites and SNPs.**

A total of 640 fish were specifically sampled for this study for SNP and/or microsatellite analyses. Microsatellite analyses, based on microsatellite genotypes from 2369 roach from 41 sites, revealed groups of populations corresponding to their catchments (Figure 2, Figures S2-S3) previously.\(^\text{17}\) With increased sampling of roach populations from the Humber Catchment these are now seen to form a distinct group (Figures 2, S2-S4). Of populations sampled for SNP analysis GraCas, LeeWhe, MolMea grouped with ‘samples’ previously obtained from these same locations\(^\text{17}\) with strong (>86%) bootstrap support (Figure 2), indicating restricted fish migration to and from these locations. Populations from GraBas and CufBro, also used for SNP analyses, also showed genetic isolation from nearby populations (Additional file 1, Table S7, Figures S3-S4). See Additional file 1 for more detailed discussion on population genetic structure.

**Identification of SNPs That Correlate with Predicted Estrogen Pollution using Latent Factor Mixed Models (LFMM).** The landscape genomics approach implemented in the programme LFMM ("latent factor mixed models")\(^\text{33}\) identified seven SNPs that correlated with estrogen pollution status after a stringent Bonferroni
corrected p value (< 0.00023) – Table 1. For full list see Additional file 10. The results were influenced by whether the environmental variable used to code for estrogen pollution status was based on predictions of steroid estrogen contamination (E2 equivalents; E2eq), or using a coarser categorical measure of estrogenic pollution (0 for ‘clean’ and 1 for ‘estrogenic’). Three of the 84 successfully genotyped SNPs within 36 estrogen candidates correlated with estrogen exposure, compared to four SNPs in 120 reference genes. These candidate genes were breast cancer anti-estrogen resistance 2 (brca2), vasa and ltbp3, and these correlated using both methods of scoring pollution status. For reference genes erythroid differentiation-related factor (edrf) only correlated when E2eq was used as the environmental variable, and pcdh17, rad54b and znf518a correlated only when using the categorical estimate of estrogenic pollution. There were no differences in the proportion of SNPs in candidate and reference genes identified as outliers ($\chi^2$, p = 0.91), or in average p values between the two groups (Mann-Whitney U tests, E2Eq, p = 0.66; categorical, p = 0.75).
Table 1. Single nucleotide polymorphisms identified as genetic outliers

<table>
<thead>
<tr>
<th>SNPs in targeted genes (estrogen)</th>
<th>Correlation with estrogen content</th>
<th>Correlation with catchment</th>
<th>Significant within-catchment pairwise analyses</th>
<th>Tests for selection in whole dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LFMM E2eq P-value</td>
<td>LFMM Pol 0 1 P-value</td>
<td>LFMM Catchment P-value</td>
<td>BayeScan (values &gt; 0.2)</td>
</tr>
<tr>
<td>aqp12_c220_368_R</td>
<td>0.36</td>
<td>0.43</td>
<td>1.1E-05</td>
<td>CufBro vs. others</td>
</tr>
<tr>
<td>ar_c4_176_M</td>
<td>0.20</td>
<td>0.00077</td>
<td>3.7E-13</td>
<td>LeeWhe vs. others; GraBas vs. FosYor/DerLof</td>
</tr>
<tr>
<td>ar_c6_283_R</td>
<td>0.17</td>
<td>0.021</td>
<td>3.0E-13</td>
<td>LeeWhe/CufBro vs. others</td>
</tr>
<tr>
<td>bcar1_c7_408_K</td>
<td>0.57</td>
<td>0.52</td>
<td>0.10</td>
<td>LeeWhe vs. Cuf</td>
</tr>
<tr>
<td>brca2_c3_251_K</td>
<td>7.3E-06</td>
<td>6.9E-06</td>
<td>0.74</td>
<td>LeeWhe vs. CufBro/GadCas/KenNor</td>
</tr>
<tr>
<td>cyp1a_c3_204_S</td>
<td>0.40</td>
<td>0.20</td>
<td>1.6E-12</td>
<td>CufBro vs. others</td>
</tr>
<tr>
<td>cyp1a_c2_71_R</td>
<td>0.217</td>
<td>0.37</td>
<td>1.2E-16</td>
<td>CufBro vs. others</td>
</tr>
<tr>
<td>FSHrecptr_c9_294_R</td>
<td>0.86</td>
<td>0.25</td>
<td>3.1E-06</td>
<td></td>
</tr>
<tr>
<td>FSH_rec9_99_Y</td>
<td>0.38</td>
<td>0.12</td>
<td>0.00023</td>
<td>CufBro vs. LeeWhe</td>
</tr>
<tr>
<td>ltbp3_c8_110_R</td>
<td>0.00018</td>
<td>1.2E-05</td>
<td>0.072</td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>LeeWhe/CufBro</td>
<td>GadCas/KenNor</td>
<td>MolMea vs. LeeWhe/CufBro</td>
<td>LeeWhe/CufBro vs. KenNor</td>
</tr>
<tr>
<td>--------------</td>
<td>---------------</td>
<td>---------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>LHrecptr_c1_17_265_S</td>
<td>0.92</td>
<td>0.038</td>
<td>0.046</td>
<td>0.036</td>
</tr>
<tr>
<td>STAR_c13_128_R</td>
<td>0.96</td>
<td>0.66</td>
<td>0.70</td>
<td>0.56</td>
</tr>
<tr>
<td>STAR_c7_307_R</td>
<td>0.14</td>
<td>0.12</td>
<td>0.10</td>
<td>0.077</td>
</tr>
<tr>
<td>sox9a_c4_490_R</td>
<td>0.014</td>
<td>5.3E-06</td>
<td>0.00018</td>
<td>0.001</td>
</tr>
<tr>
<td>tgm2l_c54_509_S</td>
<td>0.97</td>
<td>0.49</td>
<td>0.66</td>
<td>0.37</td>
</tr>
<tr>
<td>vasa_c6_145_Y</td>
<td>0.36</td>
<td>3.3E-05</td>
<td>MolMea vs. LeeWhe/CufBro</td>
<td>0.014</td>
</tr>
<tr>
<td>vtg3_c1593_478_Y</td>
<td>0.53</td>
<td>0.30</td>
<td>0.054</td>
<td>1.3</td>
</tr>
<tr>
<td>Cln13_445_M</td>
<td>0.059</td>
<td>0.057</td>
<td>0.065</td>
<td>0.38</td>
</tr>
<tr>
<td>EDRF1_c6_129_Y</td>
<td>0.069</td>
<td>2.8E-10</td>
<td>0.0065</td>
<td>0.065</td>
</tr>
<tr>
<td>f9b_c9_102_M</td>
<td>0.053</td>
<td>0.069</td>
<td>0.092</td>
<td>0.057</td>
</tr>
<tr>
<td>fam17a2_c6_836_S</td>
<td>0.89</td>
<td>3.1E-08</td>
<td>0.065</td>
<td>0.086</td>
</tr>
<tr>
<td>INTS4_c2_448_R</td>
<td>0.90</td>
<td>9.8E-05</td>
<td>0.092</td>
<td>0.057</td>
</tr>
<tr>
<td>msh2_c10_139_R</td>
<td>0.40</td>
<td>0.21</td>
<td>0.092</td>
<td>0.057</td>
</tr>
<tr>
<td>pcdh17_c3_171_R</td>
<td>0.0013</td>
<td>0.00014</td>
<td>0.092</td>
<td>0.057</td>
</tr>
<tr>
<td>pkd2_c39_1061_R</td>
<td>0.0014</td>
<td>0.00013</td>
<td>0.092</td>
<td>0.057</td>
</tr>
<tr>
<td>rad54b_c16_1215_W</td>
<td>0.0014</td>
<td>0.00013</td>
<td>0.092</td>
<td>0.057</td>
</tr>
<tr>
<td>RASGRF1_c157_346_R</td>
<td>0.017</td>
<td>2.4E-07</td>
<td>0.092</td>
<td>0.057</td>
</tr>
<tr>
<td>tdp1_c3_284_R</td>
<td>0.40</td>
<td>7.9E-06</td>
<td>0.092</td>
<td>0.057</td>
</tr>
</tbody>
</table>
Differentiated loci were identified (1) using LFMM correlating with predicted estrogen exposure (E2eq) and also by categorical coding of estrogen pollution (1 for rivers with E2eq > 1 and 0 for all others), and catchment (Thames vs. Humber); (2) in pairwise comparisons; and (3) analysis of complete dataset for loci under selection using the hierarchical method and BayeScan. For LFMM analysis, which is susceptible to false positives, those that are significant after Bonferroni correction (corrected p value = 0.00023) are in bold. For within-catchment pairwise comparisons, “CufBro vs. others” indicates significant values for all comparisons of the CufBro population with all other populations from the same catchment. BayeScan probability values above 0.2 are in bold. 

cfb_c8_111_M indicates cfb (gene code), c8 = (clone 8), 111 (position 111) M (IUPC degenerate code for base M = A or C).
Within-Catchment Pairwise Comparisons. Seven SNPs were identified as outliers in at least one pairwise comparison within each catchment (Table 1) using BayeScan, and all were within five estrogen candidate genes: aquaporin 12 (aqp12), ar, bcar1, cyp1a and fsh receptor (for full list of values see Additional file 11). 18 SNPs were identified as outliers using the less stringent fdist program, 12 in estrogen candidates (those identified using BayeScan and brca2, fsh receptor, ltbp3, lh receptor, and vtg3); two in genes previously associated with adaptation in other fish species unrelated to pollution (cfB and cttnb1) and four in ‘reference’ genes: edrf, f9b, pcdh17 and rad54b (Table 1, for full list of Lositan values, see Additional file 12). For both BayeScan and Lositan analyses significantly higher proportions of SNPs in candidate genes relative to reference genes had signatures of selection in at least 1 pairwise comparison (e.g. for Lositan ($\chi^2 (1) = 5.39, n = 205, p = 0.021$).

The only evidence for directional selection at a high estrogen site (outlier compared to at least 2 clean sites within the catchment) was within the LeeWhe population with large shifts in the allele frequencies of two ar SNPs (Figure 3, Additional file 1, Figure S5) and smaller shifts in ltbp3, brca2, rad54b (Table 1). Pairwise comparisons indicated that large shifts in allele frequency within other genes related to estrogen response had also occurred in populations at ‘clean’ sites; notably one SNP within the ar and two in cyp1a had large allele shifts in the CufBro population and there were smaller shifts for aqp12, bcar1 in this population. Within the Humber Catchment, a single ar SNP had a large allele shift within the ‘clean’
Grantham Canal (GraBas). The large differences in allele frequencies for ar and cyp1A can be seen in Figure 1 and Additional file 1, Figure S5.

The SNPs found to correlate with estrogen pollution using LFMM (e.g. brca2, vasa, ltbp3) were only identified as outliers using the less stringent method (Lositan) in a maximum of three pairwise comparisons, suggesting small but consistent shifts in allele frequency in populations in estrogenic rivers. Likewise ar and cyp1a were not identified using LFMM, indicating that these genes are not consistently under selection across the populations from these estrogenic river stretches.

**Differentiated Loci Identified in Comparisons between the Catchments.**

Twenty SNPs in 18 genes correlated with catchment (Thames vs Humber) using LFMM (Table 1). There were no differences in the proportion of candidate genes and reference genes reaching the threshold of significance ($\chi^2$, $p = 0.92$) or in average p-values ($p=0.097$). Notably SNPs in the androgen receptor (ar), cyp1A, edrf and coagulation factor IXb (f9b) had very low p values ($p < 2 \times 10^{-10}$) – Table 1. This is consistent with analyses of the combined SNP data from all 10 populations using BayeScan and the Hierarchical method\(^{38}\) that revealed that six SNPs in four genes - ar, cyp1A, coagulation factor IXb (f9b) and edrf - were clear outliers(Figure 3, Table 1, see Additional files 10-11 for full lists). However, for both these analyses there were significant differences in the probabilities/p-values between the candidate and the reference genes (e.g. Mann-Whitney U tests: BayeScan, $p = 0.0018$, Hierarchical, $p = 0.011$).
Analysis of androgen receptor SNPs. The two SNPs in the ar identified as genetic outliers did not alter the amino acid sequence. Sequence analysis of exons 5 and 8 that encode the ligand-binding domain from 15 and 9 fish, respectively, revealed only one variant in exon 5 to alter the amino acid sequence from gly -> ser (position 1081 in sequence accession = GQ161219) of the gene, but not in a position known to affect androgen binding. See Additional file 13 for SNPs identified in the androgen receptor.

Analysis of a River Lee Population using RAD-Seq. The LeeWhe sample site in the River Lee has a predicted exposure of 6.6 ng/L E2Eq (28% effluent), exceeding an E2Eq of 11 ng/L 10% of the time. This population was compared to those from two ‘clean’ rivers in the Thames Catchment using RAD-seq analysis. The final sample sizes were as follows: LeeWhe (18 fish), KenNor (20 fish) and CufBro (24 fish). A total of 543,887 catalogue RAD loci were assembled of which 45,607 were polymorphic (summary statistics of raw sequencing reads are given in Additional file 1, Table S8). There were 11,860 loci for the LeeWhe-CufBro comparison, 11,387 loci for the LeeWhe-KenNort comparison and 11,947 loci for the KenNor-CufBro comparison. Average FST values were 0.025, 0.017 and 0.019 respectively with 553, 174, and 266 loci respectively with FST values of over 0.1 with p-values < 0.01. BLAST analysis revealed 208, 54 and 65 loci respectively had hits on genes either directly, or by searching by 5000 bp either side of the RAD locus in the fathead minnow genome (Additional file 14– list of top hits for RAD data). The androgen receptor was among
those identified in the LeeWhe-CufBro comparison. No enriched GO terms in DAVID were identified.

The only gene potentially related to endocrine disruption showing directional selection within the LeeWhe population was oxysterol binding protein 7 (osbp7). Two SNPs showed evidence for directional selection in the CufBro population: bard1 and sox9b. Other genes potentially related to endocrine disruption were identified in the LeeWhe-CufBro comparison (ar, osbp5 osbp8 and srd5a1), but there was no clear evidence of directionality (Additional file 14).

DISCUSSION

Understanding the impacts of chemical pollution on fish populations requires knowledge of the ability of fish to tolerate and/or adapt to the harmful effects of exposure. Our results identified several genes involved in responses to endocrine disrupting pollutants which were highly differentiated between populations, a potential result of selection. However, there was no evidence that these allele shifts resulted from adaptation to estrogen pollution, as there were no consistent allele shifts in the most obvious candidate genes between populations in clean and effluent dominated rivers stretches within catchments. This is despite the inclusion of some populations restricted to river stretches with some of the highest known proportions of WwTW effluent in UK rivers. The androgen receptor (ar) and cyp1A
exhibited large shifts in allele frequency both between individual populations of roach within catchments and between catchments. Though our study provided no clear link with estrogen pollution, to our knowledge the androgen receptor has not previously been implicated in local adaptation in fish. Cyp1A has previously been associated in adaptation to hydrocarbon pollutants in other fish species, e.g. 20, 21 although the pattern here does not implicate selection resulting from WwTW pollution.

In fish, linkage blocks can range from 1 kb in zebrafish (Danio rerio) to 1 Mb in lake whitefish (Coregonus clupeaformis)47 Under strong, recent selection, linkage blocks can be large; in killifish the median lengths outlier windows at polluted sites were 50-62 kb but some haplotypes were larger including 650 kb haplotype containing the AIP gene {Reid, 2016 #2377}. This raises the possibility that allele shifts observed at the ar and cyp1A had occurred by selection in linked genes. Our data, however, suggest this is not the case for ar, as the two SNPs have different patterns of selection in both catchments. These SNPs are separated by 7 kb in the zebrafish genome, which has synteny with other cyprinid fish48 . In contrast, the two cyp1A SNPs are separated by only 145 bp and have the same patterns of selection. The closest genes to these SNPs are 67 kb for the ar and 29 kb for the cyp1A. Indeed, our results suggest that large allele shifts have occurred in the ar at least twice within the Thames Catchment, with a unique allele shift at the LeeWhe population.
The results of the correlation analysis (using LFMM) did not provide strong evidence for adaptation to steroid estrogen pollution. There was no difference in the proportion of candidates and reference genes identified under selection using this method. Furthermore, none of the obvious candidate genes known for estrogen response (e.g. estrogen receptors, aromatases and vitellogenins) showed correlations with estrogenic pollution. Additionally, the estrogen-adaptation candidate genes (vasa, bcra2 and ltbp3) identified were not subject to large shifts in allele frequency in any population. Of the four reference genes that correlated with estrogen pollution, three had no obvious link with estrogen pollution (edrf, pcdh17, and znf518B). The fourth, rad54b, is involved in DNA repair, but humans variants have been associated with excessive levels of androgens in females;\textsuperscript{49} so variants could potentially modify responses to EDCs in fish. Thus, overall these results do not provide strong evidence for parallel selection related to estrogen pollution, but do not exclude an influence.

It is possible that some, but not all, populations of roach have adapted to estrogenic pollution, or that different populations have adapted, but through different mechanisms. Such patterns would not have been identified in the correlation analysis. For instance, the large allele shift at the ar in the population from the River Lee (LeeWhe) could be a consequence of estrogenic pollution. In males, androgens play key roles in sexual development, puberty, the development of secondary sexual characteristics, and reproductive behaviour.\textsuperscript{50} Estrogens are antagonists of AR androgen binding,\textsuperscript{51} can reduce androgen levels in male fish\textsuperscript{52} and
modify ar expression\textsuperscript{53} at an estrogenic potency (5 ng/L E2Eq) similar to the average (6.6 E2Eq ng/L) predicted for this river site. The effect of ar polymorphisms in fish is not known, but in humans they modify susceptibility to the effects of estrogen exposure.\textsuperscript{54}

Adaptation to pollution from other endocrine disrupting chemicals could also explain differentiation of the ar in the LeeWhe population. Elevated concentrations of pesticides including dichlorodiphenyltrichloroethane (DDT) metabolites (e.g. p,p’ dichlorodiphenyldichloroethylene (DDE))\textsuperscript{29}, endosulfan and lindane\textsuperscript{55} were detected in the tissues of roach sampled at this location 20 years after a pesticide formulation factory next to this site closed in 1982.\textsuperscript{29} The p,p’DDE concentrations equated to those known to affect the early life stages of fish (gene expression and gonadal intersex) and approaching reported effect concentrations for adult fish.\textsuperscript{29} Several DDT metabolites are anti-androgenic and some are also estrogenic\textsuperscript{56-57} and alter expression of estrogen receptors in fish.\textsuperscript{58}

It is also possible that the shifts in allele frequency at the androgen receptor do not relate to adaptation to pollution as one ar SNP is also highly differentiated in the population from an isolated stretch of the Grantham Canal (GraBas) with no known WwTW inputs (see Additional file 1 Table S1). We cannot exclude selection from other EDC pollutant from an unidentified source or in the neighboring polluted river Trent\textsuperscript{60} before the separation of these waterways approximately 50 years ago. Thus while our study suggests the ar is important for local adaptation, the cause of the
selection is unclear and it may be independent of the effects of endocrine disruption, or pollution. It could, for example, relate to differences in sexual selection between populations. Experiments are required to assess whether these genotypes associate with susceptibility to EDC pollution. Further sequencing of the wider genomic region is required to identify the linked genetic variants that are responsible for the suspected adaptation.

The large allele shifts in two SNPs in cyp1A in the genetically isolated CufBro population may have resulted from adaptation to pollution at this site. This could not have been driven by the effects of WwTW pollution as there are no known upstream inputs. This gene has an important role in detoxification of a wide range of contaminants and is involved in adaptation of F. heteroclitus and M. tomcod to hydrocarbon pollutants such as PAHs and PCBs/dioxin-like compounds.\textsuperscript{18, 61-62}, and this may be the case here.

Our analysis identified large shifts in SNP frequencies related to catchment, particularly at ar, cyp1A, edrf and f9b. As these populations have potentially been separated from each other since the end of the last ice age, these allele shifts could have occurred in either catchment over a long time scale. The inclusion of cyp1A among these suggests that allele shifts may have, in part, been driven by pollution-related selection although there was no evidence estrogen-pollution had driven this, as we had originally hypothesised. In humans, edrf is involved in the regulation of alpha-globin expression\textsuperscript{63} so the high differentiation at this gene could relate to
selection due to differences in oxygen availability; average water temperatures are approximately 2°C higher in the more southerly Thames Catchment\textsuperscript{64} and rivers in both catchments would have suffered from nutrient-rich pollution e.g. from fertilizers and poorly treated sewage. High differentiation at coagulation factor IXb (f9b) may relate to adaptation against blood pathogens; the coagulation system has been under strong selective pressure in primates, possibly for this reason.\textsuperscript{65}

Analysis of the population from the estrogenic River Lee (LeeWhe) using RAD-seq provided no evidence for adaptation to estrogen pollution, as genes involved in estrogen response were not overrepresented among loci with elevated $F_{ST}$ values in comparisons with populations from clean sites. Indeed the only gene that was found to be related to endocrine disruption under directional selection in the LeeWhe population was oxysterol binding protein 7. Interestingly 3 other oxysterol binding proteins were also identified in the LeeWhe-CufBro comparison but the direction of selection was not determined. Oxysterols modify estrogen receptor function and can bind to, and modulate, the activity of ERα and ERβ.\textsuperscript{66} Expression of genes for these proteins is modified by estrogen\textsuperscript{67} and lindane\textsuperscript{68} found at elevated concentrations in tissues from roach from this River Lee location\textsuperscript{29} The LeeWhe-CufBro comparison identified ar, confirming the result from the targeted gene analysis. Nevertheless, cyp1A was not identified using this method, despite the large allele shift in the CufBro population. Thus a resequencing approach\textsuperscript{21} would enable a more complete and detailed analysis of genes under selection.
Limitations of this study include that the full history of roach within these rivers is not known. Each population will have had different levels of immigration, most restocking events are undocumented and the success of this restocking is unknown. Levels of estrogen contamination will have varied over time with changes in waste-water treatments processes and changes in industry chemical use. For instance the concentration of nonylphenol, responsible for a major part of the estrogenicity in the River Aire, decreased during the 1990s. Levels of other EDC pollutants have not been recorded; the high levels of DDT metabolites for fish in the River Lee were only discovered accidently. For further information on history of fish in these rivers see Additional file 1. Another limitation is that false positives in FST outlier tests can arise from historic demographic events such as recent range expansions. Here, average FST is low, and all populations have high genetic diversity, reducing false positives. The key role of cyp1A in adaptation in other fish species, and the near certainty of the occurrence similar pollutants in some of these rivers adds confidence that this has resulted from natural selection. Likewise, the occurrence of independent large allele shifts in ar in both catchments adds confidence that at least one of these shifts results from natural selection. Irrespective of the cause of the highly differentiated loci observed in this study, our results caution against extrapolating effects from fish derived from only one population for assessing the impacts of endocrine disrupting chemicals on the health of fish. Selection of EDC responsive genes may indicate different fish populations
could respond differently to EDC exposure. This also has implications for the
management of fish stocks. For instance, failure of restocking programs for
salmonids has been attributed to local adaptation,\textsuperscript{71} thus, restocking with locally
adapted genotypes may result in greater success.

ABBREVIATIONS

\begin{itemize}
\item ar: androgen receptor, ampd1: AMP deaminase 1; AhR: aryl hydrocarbon receptor;
\item aip: aryl hydrocarbon receptor-interacting protein; ampd1: adenosine
\item monophosphate deaminase 1; BLAST: Basic Local Alignment Search Tool; bp: base
\item pairs; brca2: breast cancer 2 (currently BRCA2, DNA repair associated); CfB:
\item complement factor B precursor; ctnnb1: catenin beta 1; Cyp: cytochrome P450;
\item DAVID: database for annotation, visualisation and integrated discovery; DDE:
\item dichlorodiphenyldichloroethylene; DDT: dichlorodiphenyltrichloroethane; EDC:
\item endocrine disrupting chemical; edrf: erythroid differentiation-related factor; ER beta:
\item Estrogen receptor beta; E2: estradiol; EE2: ethinylestradiol; f9b: coagulation factor IXb
\item ; F\textsubscript{ST}: fixation index; fh: follicle-stimulating hormone; fshr: follicle-stimulating hormone
\item receptor or FSH receptor; GO: gene ontology; gper: G protein-coupled estrogen
\item receptor-1; hbb1: haemoglobin beta1; LFMM: latent factor mixed model; lhr:
\item luteinizing hormone receptor; ltbp3: latent transforming growth factor beta binding
\end{itemize}
protein; osbp8: oxysterol binding protein like 8; MEGA: Molecular Evolutionary
Genetic Analysis; PAHs: polycyclic aromatic hydrocarbons; PCA: principal component
analysis; PCBs: poly chlorinated biphenyls; pcdh17: protocadherin-17, RAD-seq:
Restriction site Associated DNA Sequencing; rad54b: DNA repair and recombination
protein 54b; SNPs: single nucleotide polymorphisms; star: steroidogenic acute
regulatory protein; TELO2: telomere length regulation protein; TGFβ: Transforming
growth factor beta; VTG: vitellogenin; WwTW: waste-water treatment works; znf518a:
zinc finger protein 518A.

COMPETING INTERESTS

The authors declare that they have no competing interests.

ACKNOWLEDGEMENTS

We thank Richard Williams and Dr Virginie Keller (CEH) for modelling estrogen
exposure in rivers. Alan Henshaw (Calverton fish) farm for advice and assistance with
sample collection, Environment Agency Fisheries Teams (Pete Turner, Michael Lee).
Katie Sumner (Environment Agency) for helping to co-ordinate sample collection.
Monica Juergens at CEH for providing roach tissue. The Grantham Canals Trust, for
providing information on the history of rivers, Graeme Peirson for providing information on historic fish stocks in the Humber Catchment and Darren Rowe for helping to sample from rivers in the Humber Catchment. Finlay Maguire for help with coding. Jane McDougall, Aron Simpson and Pawel Gardzielewski at LGC assisted with KASP SNP assay design. Gavin Horsburgh prepared the samples for SNP typing. We thank the UK Natural Environmental Research Council (NERC; NE/K004263/1) and the NERC Biomolecular Analysis Facility for funding (NBAF866). Exeter Sequencing Service and Computational core facilities at the University of Exeter were funded by the Medical Research Council Clinical Infrastructure award (MR/M008924/1), Wellcome Trust Institutional Strategic Support Fund (WT097835MF), Wellcome Trust Multi User Equipment Award (WT101650MA) and BBSRC LOLA award (BB/K003240/1). We thank three anonymous reviewers for insightful suggestions on our manuscript.

AUTHOR’S CONTRIBUTIONS

PH, JRS, SJ and CT obtained the funding for this research.

PH, JRS, AEL, SJ and CT participated in the design of the research.

PH, AEL, TUW, EN, DAD, AB, DS, KM and JP participated in data collection, generation and analysis of SNP data and statistical analysis.

PH, AEL, JRS, TUW, SJ and CT wrote the paper.
All authors read and approved the final manuscript

AUTHOR’S INFORMATION

PH, the corresponding author, is at the School of Medicine and Health at the University of Exeter with interests including Ecotoxicology and Molecular Ecology.

ELECTRONIC SUPPLEMENTARY MATERIAL

Additional file 1: Details of microsatellite genotyping methods, population-genetic analyses, river histories, Tables S1-S8, Figures S1-S5.
Additional file 2: List of estrogen candidate genes
Additional file 3: Perl script to identify SNPs
Additional file 4: Script to reconstruct intron positions and genes from genomic reads
Additional file 5: Sequences flanking the SNPs genotyped in this study
Additional file 6: SNP genotypes 217 SNPs genotyped in 465 fish from 10 locations
Additional file 7: Microsatellite genotypes from 2369 roach genotyped at 17 loci
Additional file 8: Summary statistics based on microsatellite data
Additional file 9: $F_{ST}$ values for the full microsatellite dataset
Additional file 10: LFMM and Arlequin results
Additional file 11: BayeScan and Hierarchical method results
Additional file 12: Lositan results
Additional file 13: SNPs in the androgen receptor

Additional file 14: RAD-seq top hits
References


34. Foll, M.; Gaggiotti, O., A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **2008**, *180*, (2), 977-993


55. Jürgens, M. Biomonitoring of wild fish to assess chemical pollution in English rivers: an application of a fish tissue archive. PhD, Lancaster University, 2015.


in the concentration of industrially derived surfactants. Environ. Toxicol. Chem. 2002, 21, (3), 515-
519


GRAPHICAL ABSTRACT
sites and obstructions to fish movement (locks and weirs) in the Thames Catchment are given in the map figure in Hamilton et al.\textsuperscript{17}

\textbf{Figure 2.} Neighbor-joining tree for roach population samples produced from data from 2369 roach from 41 sample sites. Several locations were sampled in different years, producing a total of 51 ‘samples’. The tree is based on the data from 14 microsatellite loci using Cavalli-Sforza and Edwards’ chord distance measure, $D_C^{72}$. Only bootstrap values above 50% are shown. Numbers at the end of sample codes
indicate years in which populations were sampled (where the same location was sampled in different years). Locations of rivers used are shown in the map (Figure 1).

Figure 3. Identification of $F_{ST}$ outlier loci potentially subject to differential selection constructed using data from 217 SNPs loci and 10 sample sites using BayeScan. The x axis represents Log transformed Bayes factors and the y axis represents locus specific $F_{ST}$ from BayeScan. Loci with a posterior probability of 1 (corresponding to a PO of infinity), were ascribed a Log10(BF) arbitrary values of 5. Codes for SNPs: androgen receptor_SNP1, ar_c4_176_M; androgen receptor 2, ar_c6_283_R; Cyp1A_SNP1 - cyp1a_c2_71_R; Cyp1a_SNP2 - cyp1a_c3_204_S; erythroid differentiation regulatory factor - EDRF1_c6_129_Y; f9b, f9b_c9_102_M; STAR_SNP1 - STAR_c7_307_R; STAR_SNP2 - STAR_c13_128_R; tgm2l, tgm2l_c54_509_S; FSHreceptor -
FSH_rec_c9_99_Y. Allele frequencies of the androgen receptor SNP 1 in each population are shown in the inset box.