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

AUTHORS

*Patrick. B. Hamilton[1,2,*], Anne E. Lockyer[3], Tamsyn M. Uren Webster[1,4], David J. Studholme[1], Josephine R. Paris[1], Alice Baynes[3], Elizabeth Nicol[3], Deborah A. Dawson[5], Karen Moore[1], Audrey Farbos[1], Susan Jobling[3], Jamie R. Stevens[1], Charles R. Tyler[1]*

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
3. Institute of Environment, Health and Societies, Brunel University London, Uxbridge, Middlesex, UB8 3PH, UK
4. Biosciences, College of Science, Swansea University, Swansea, SA2 8PP, UK

5. NERC Biomolecular Analysis Facility, Department of Animal and Plant Sciences, University of Sheffield, S10 2TN, UK

* corresponding author: email p.b.hamilton@exeter.ac.uk

KEYWORDS

adaptation; ecotoxicology; natural selection; contemporary evolution; endocrine disruption; pollution

ABSTRACT

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.

## INTRODUCTION

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 rultilus*) 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.[12-14] 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[15] which subsequently recovered after removal of EE2.[16]

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,[17] and demonstrated the existence of self-sustaining roach populations over multiple generations.[17] 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*).[e.g. 18, 19] 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*).[18, 20-21] 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[22] and that polymorphisms in genes for steroid receptors are associated with a variety of impacts on fitness (reproduction and/or likely survival).[23-24] 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[27] had previously been calculated using the geographical information systems-based model (LF2000-WQX). This model predicts the estradiol equivalents ($E_2$Eq) (see [28]), an estimate of estrogenic potency which correlates with the actual incidence and severity of intersex in fish found downstream of WwTWs.[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.[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).[31] 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.[32] 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 $F_{ST}$ 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 2.1[34] (provided at http://cmpg.unibe.ch/software/BayeScan/) and fdist program[35] implemented in Lositan[36] 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.[37]

**Full Dataset Analysis.** BayeScan and the hierarchical method implemented in Arlequin 3.5.[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. This was done such that each candidate gene was represented by a single value in the statistical analyses and was conducted 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*.[39] We used Stacks version 1.40 [40-41] for building loci and calling SNPs in three populations. BLAST analysis was used to identify the sequence 5 kb [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[43] searches against the zebrafish Ensembl[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),[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*)

212 cytochrome p450 genes, *vtg3*, and the main vitellogenin (*vtg*) locus which includes *vtg 1-2, 4-*
213 *7* genes.

214 **Analysis of Population-Genetic Structure Using DNA Microsatellites and SNPs.** A
215 total of 640 fish were specifically sampled for this study for SNP and/or microsatellite
216 analyses. Microsatellite analyses, based on microsatellite genotypes from 2369 roach from
217 41 sites, revealed groups of populations corresponding to their catchments (Figure 2, Figures
218 S2-S3) previously.[17] With increased sampling of roach populations from the Humber
219 Catchment these are now seen to form a distinct group (Figures 2, S2-S4). Of populations
220 sampled for SNP analysis GraCas, LeeWhe, MolMea grouped with 'samples' previously
221 obtained from these same locations[17] with strong (>86%) bootstrap support (Figure 2),
222 indicating restricted fish migration to and from these locations. Populations from GraBas and
223 CufBro, also used for SNP analyses, also showed genetic isolation from nearby populations
224 (Additional file 1, Table S7, Figures S3-S4). See Additional file 1 for more detailed
225 discussion on population genetic structure.

226 **Identification of SNPs That Correlate with Predicted Estrogen Pollution using Latent**
227 **Factor Mixed Models (LFMM).** The landscape genomics approach implemented in the
228 programme LFMM ("latent factor mixed models")[33] identified seven SNPs that correlated
229 with estrogen pollution status after a stringent Bonferroni corrected p value (< 0.00023) –
230 Table 1. For full list see Additional file 10. The results were influenced by whether the
231 environmental variable used to code for estrogen pollution status was based on predictions of
232 steroid estrogen contamination (E2 equivalents; E2eq), or using a coarser categorical measure
233 of estrogenic pollution (0 for 'clean' and 1 for 'estrogenic'). Three of the 84 successfully
234 genotyped SNPs within 36 estrogen candidates correlated with estrogen exposure, compared
235 to four SNPs in 120 reference genes. These candidate genes were breast cancer anti-estrogen
236 resistance 2 (*brca2*), *vasa* and *ltbp3*, and these correlated using both methods of scoring

237 pollution status. For reference genes erythroid differentiation-related factor (*edrf*) only
238 correlated when E2eq was used as the environmental variable, and *pcdh17*, *rad54b* and
239 *znf518a* correlated only when using the categorical estimate of estrogenic pollution. There
240 were no differences in the proportion of SNPs in candidate and reference genes identified as
241 outliers ($\chi^2$, $p = 0.91$), or in average *p* values between the two groups (Mann-Whitney U tests,
242 E2Eq, $p = 0.66$; categorical, $p = 0.75$).

243 **Table 1. Single nucleotide polymorphisms identified as genetic outliers**

| | Correlation with estrogen content | | Correlation with catchment | Significant within-catchment pairwise analyses | | Tests for selection in whole dataset | | |
|---|---|---|---|---|---|---|---|---|
| **SNPs in targeted genes (estrogen)** | LFMM E2eq P-value | LFMM Pol 0 1 P-value | LFMM Catchment P-value | BayeScan (values > 0.2) | Lositan (values > 0.95) | Hierarchical method $F_{ST}$ P-value | BayeScan prob | BayeScan log10(PO) |
| aqp12_c220_368_R | 0.36 | 0.43 | **1.1E-05** | CufBro vs. others | CufBro vs. others | 0.34 | 0.059 | -1.2 |
| ar_c4_176_M | 0.20 | 0.00077 | **3.7E-13** | LeeWhe vs. others; GraBas vs. FosYor/DerLof | LeeWhe vs. others; GraBas vs. others | **7.1E-15** | **1** | **1000** |
| ar_c6_283_R | 0.17 | 0.021 | **3.0E-13** | LeeWhe/CufBro vs. others | LeeWhe/CufBro vs. others | **2.7E-17** | **1** | **1000** |
| bcar1_c7_408_K | 0.57 | 0.52 | 0.10 | LeeWhe vs. Cuf | CufBro vs. others | 0.062 | 0.19 | -0.65 |
| brca2_c3_251_K | **7.3E-06** | **6.9E-06** | 0.74 | | LeeWhe vs. CufBro/GadCas/KenNor | 0.28 | 0.061 | -1.2 |
| cyp1a_c3_204_S | 0.40 | 0.20 | **1.6E-12** | CufBro vs. others | CufBro vs. others; AirBea vs. FosYor | **5.2E-05** | **1** | **1000** |
| cyp1a_c2_71_R | 0.217 | 0.37 | **1.2E-16** | CufBro vs. others | CufBro vs. others | **6.6E-14** | **1** | **1000** |
| FSHrecptr_c9_294_R | 0.86 | 0.25 | **3.1E-06** | | KenNor vs. CufBro/GadCas | **0.028** | 0.11 | -0.90 |
| FSH_rec9_99_Y | 0.38 | 0.12 | 0.00023 | CufBro vs. LeeWhe | CufBro vs. LeeWhe | **0.016** | **0.41** | **-0.15** |
| ltbp3_c8_110_R | **0.00018** | **1.2E-05** | 0.072 | | LeeWhe vs. CufBro/GadCas | 0.39 | 0.057 | -1.2 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| LHrecptr_c1_17_265_S | 0.50 | 0.56 | **1.1E-05** | | CufBro vs GadCas | **0.038** | 0.12 | -0.88 |
| STAR_c13_128_R | 0.92 | 0.89 | 0.83 | | | **0.046** | **0.66** | **0.28** |
| STAR_c7_307_R | 0.55 | 0.96 | 0.66 | | | **0.036** | **0.70** | **0.37** |
| sox9a_c4_490_R | 0.036 | 0.014 | **5.3E-06** | | | 0.14 | 0.079 | -1.1 |
| tgm2l_c54_509_S | 0.97 | 0.69 | 0.49 | | | **0.034** | **0.66** | **0.29** |
| vasa_c6_145_Y | **7.2E-05** | **6.4E-07** | 0.019 | | AirBea vs. FosYor | 0.14 | 0.10 | -0.95 |
| vtg3_c1593_478_Y | 0.36 | 0.53 | **3.3E-05** | | MolMea vs. LeeWhe/CufBro | **0.021** | 0.14 | -0.78 |
| **SNPs in other targeted genes (unrelated to estrogen)** | | | | | | | | |
| cfB_c8_111_M[a] | 0.012 | 0.21 | **1.2E-05** | | LeeWhe/CufBro vs. KenNor | 0.10 | 0.077 | -1.1 |
| ctnnb1_c39_260_Y | 0.72 | 0.50 | 0.58 | | MolMea vs. GadCas/KenNor | **0.047** | 0.11 | -0.90 |
| **SNPs in reference genes** | | | | | | | | |
| bbs2_c13_244_Y | 0.0014 | 0.0029 | **8.2E-05** | | | 0.30 | 0.054 | -1.3 |
| CIc13_445_M | 0.58 | 0.65 | **3.5E-06** | | | 0.13 | 0.057 | -1.2 |
| EDRF1_c6_129_Y | **0.00018** | 0.075 | **3.8E-13** | | LeeWhe/CufBro vs. KenNor; AirBea vs. FosYor | **0.00039** | **1.0** | **3.7** |
| f9b_c9_102_M | 0.053 | 0.069 | **2.8E-10** | | GadCas vs. MolMea/CufBro | **0.0065** | **0.96** | **1.4** |
| fam171a2_c6_836_S | 0.90 | 0.17 | **3.1E-08** | | | 0.092 | 0.057 | -1.2 |
| INTS4_c2_448_R | 0.34 | 0.11 | **9.8E-05** | | | 0.18 | 0.069 | -1.1 |
| msh2_c10_139_R | 0.60 | 0.36 | 2**.2E-08** | | | 0.14 | 0.057 | -1.2 |
| pcdh17_c3_171_R | 0.0013 | **0.00014** | 0.55 | | LeeWhe vs. KenNor; GraBas vs. DerLof/FosYor | 0.24 | 0.065 | -1.2 |
| pkd2_c39_1061_R | 0.60 | 0.44 | 0.21 | | | **0.050** | **0.29** | **-0.38** |
| rad54b_c16_1215_W | 0.0014 | **0.00013** | 0.026 | | LeeWhe vs. GadCas/KenNor | 0.26 | 0.062 | -1.2 |
| RASGRF1_c157_346_R | 0.017 | 0.83 | **2.4E-07** | | | 0.058 | 0.14 | -0.77 |
| tdp1_c3_284_R | 0.40 | 0.89 | **7.9E-06** | | | 0.30 | 0.055 | -1.2 |
| zc3h4_c3_114_W | 0.35 | 0.11 | **0.00014** | | | 0.24 | 0.060 | -1.2 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| zg109744_c3_524_M | 0.33 | 0.37 | 0.30 | | | 0.064 | **0.38** | **-0.21** |
| ZNF518A_c3_889_M | 0.00065 | **3.0E-05** | 0.47 | | | 0.26 | 0.064 | -1.2 |

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. [a]

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,[34] 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,[35] 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 *ctnnb1*) 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 were outliers 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 between Roach Populations in the Thames and Humber 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 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.[46] 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.[17] 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 $F_{ST}$ values were 0.025, 0.017 and 0.019 respectively with 553, 174, and 266 loci respectively with $F_{ST}$ 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 [45] 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 imply 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 a 650 kb haplotype containing the AIP gene.[21] This raises the possibility that allele shifts observed at the *ar* and *cyp1A* in our study have resulted through 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 fish.[48] 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 differentiation in the *ar* SNPs has occurred at least twice within the Thames Catchment, with a unique allele shift at the LeeWhe population. False positives in $F_{ST}$ outlier tests can also arise from historic demographic events such as recent range expansions, although here average $F_{ST}$ is low, reducing false positives.[49] The key role of *cyp1A* in adaptation to harmful hydrocarbon pollutants in other fish species, and the high likelihood of

contamination with similar pollutants these rivers adds weight to the suggestion that this has resulted from natural selection. Likewise, the independent large allele shifts in *ar* in both catchments adds confidence that at least one of these shifts results from natural selection.

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;[50] 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 adaptation to estrogenic pollution. In males, androgens play key roles in sexual development, puberty, the development of secondary sexual characteristics, and reproductive behaviour.[51] Estrogens are antagonists of AR androgen binding,[52] can reduce androgen levels in male fish[53] and modify *ar* expression[54] 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.[55]

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))[29], endosulfan and lindane[56] 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.[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.[29] Several DDT metabolites are anti-androgenic and some are also estrogenic [57-58] and alter expression of estrogen receptors in fish.[59]

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 in this canal or in the neighboring polluted river Trent[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 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.[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 c*yp1A* 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[63] so the high differentiation at this gene could relate to selection due to differences in oxygen availability; average water temperatures are approximately 2$^{o}$ C higher in the more southerly Thames Catchment[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 resason.[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. Three 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.[66] Expression

of oxysterol genes is modified by estrogen[67] and lindane[68] found at elevated concentrations in tissues from roach from this River Lee location[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 identified by targeted SNP genotyping. Thus a resequencing approach[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,[27] decreased during the 1990s.[69] 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.[29] For further information on history of fish in these rivers see Additional file 1.

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,[70] thus, restocking with locally adapted genotypes may result in greater success.

## ABBREVIATIONS

ar: androgen receptor, ampd1: AMP deaminase 1; AhR: aryl hydrocarbon receptor; aip: aryl hydrocarbon receptor-interacting protein; ampd1: adenosine monophosphate deaminase 1; BLAST: Basic Local Alignment Search Tool; bp: base pairs; brca2: breast cancer 2 (currently BRCA2, DNA repair associated); CfB: complement factor B precursor; ctnnb1: catenin beta 1; Cyp: cytochrome P450; DAVID: database for annotation, visualisation and integrated discovery; DDE: dichlorodiphenyldichloroethylene; DDT: dichlorodiphenyltrichloroethane; EDC: endocrine disrupting chemical; edrf: erythroid differentiation-related factor; ER beta: Estrogen receptor beta; E2: estradiol; EE2: ethinylestradiol; f9b: coagulation factor IXb; $F_{ST}$: fixation index: fh: follicle-stimulating hormone; fshr: follicle-stimulating hormone receptor or FSH receptor; GO: gene ontology; gper: G protein-coupled estrogen receptor-1; hbb1: haemoglobin beta1; LFMM: latent factor mixed model; lhr: luteinizing hormone receptor; ltbp3: latent transforming growth factor beta binding 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

471 The authors declare that they have no competing interests.

472

473 ACKNOWLEDGEMENTS

474

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

491

492 AUTHOR'S CONTRIBUTIONS

493

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

518 Additional file 13: SNPs in the androgen receptor

519 Additional file 14: RAD-seq top hits

520

521 References

522

523 1. Desbrow, C.; Routledge, E. J.; Brighty, G. C.; Sumpter, J. P.; Waldock, M.,
524 Identification of estrogenic chemicals in STW effluent. 1. Chemical fractionation and in vitro
525 biological screening. *Environ. Sci. Technol.* **1998,** *32*, (11), 1549-1558

526 2. Hinck, J. E.; Blazer, V. S.; Schmitt, C. J.; Papoulias, D. M.; Tillitt, D. E., Widespread
527 occurrence of intersex in black basses (*Micropterus* spp.) from US rivers, 1995-2004. *Aquat.*
528 *Toxicol.* **2009,** *95*, (1), 60-70

529 3. Purdom, C. E.; Hardiman, P. A.; Bye, V. V. J.; Eno, N. C.; Tyler, C. R.; Sumpter, J.
530 P., Estrogenic effects of effluents from sewage treatment works. *Chem. Ecol.* **1994,** *8*, (4),
531 275 - 285

532 4. Nolan, M.; Jobling, S.; Brighty, G.; Sumpter, J. P.; Tyler, C. R., A histological
533 description of intersexuality in the roach. *Journal of Fish Biology* **2001,** *58*, (1), 160-176

534 5. Jobling, S.; Nolan, M.; Tyler, C. R.; Brighty, G.; Sumpter, J. P., Widespread sexual
535 disruption in wild fish. *Environ. Sci. Technol.* **1998,** *32*, (17), 2498-2506

536 6. Jobling, S.; Williams, R.; Johnson, A.; Taylor, A.; Gross-Sorokin, M.; Nolan, M.;
537 Tyler, C. R.; van Aerle, R.; Santos, E.; Brighty, G., Predicted exposures to steroid estrogens
538 in UK rivers correlate with widespread sexual disruption in wild fish populations. *Environ.*
539 *Health Persp.* **2006,** *114*, 32-39

540 7. Jobling, S.; Coey, S.; Whitmore, J. G.; Kime, D. E.; Van Look, K. J. W.; McAllister,
541 B. G.; Beresford, N.; Henshaw, A. C.; Brighty, G.; Tyler, C. R.; Sumpter, J. P., Wild intersex
542 roach (*Rutilus rutilus*) have reduced fertility. *Biol. Reprod.* **2002,** *67*, (2), 515-524

8. Harris, C. A.; Hamilton, P. B.; Runnalls, T. J.; Vinciotti, V.; Henshaw, A.; Hodgson, D.; Coe, T. S.; Jobling, S.; Tyler, C. R.; Sumpter, J. P., The consequences of feminization in breeding groups of wild fish. *Environ. Health Persp.* **2011,** *119*, (3), 306-311

9. Lange, A.; Paull, G. C.; Hamilton, P. B.; Iguchi, T.; Tyler, C. R., Implications of Persistent Exposure to Treated Wastewater Effluent for Breeding in Wild Roach (Rutilus rutilus) Populations. *Environ. Sci. Technol.* **2011,** *45*, (4), 1673-1679

10. Nash, J. P.; Kime, D. E.; Van der Ven, L. T. M.; Wester, P. W.; Brion, F.; Maack, G.; Stahlschmidt-Allner, P.; Tyler, C. R., Long-term exposure to environmental concentrations of the pharmaceutical ethynylestradiol causes reproductive failure in fish. *Environ. Health Persp.* **2004,** *112*, (17), 1725-1733

11. Lange, R.; Hutchinson, T. H.; Croudace, C. P.; Siegmund, F.; Schweinfurth, H.; Hampe, P.; Panter, G. H.; Sumpter, J. P., Effects of the synthetic estrogen 17 alpha-ethinylestradiol on the life-cycle of the fathead minnow (*Pimephales promelas*). *Environ. Toxicol. Chem* **2001,** *20*, (6), 1216-1227

12. Parrott, J. L.; Blunt, B. R., Life-cycle exposure of fathead minnows (*Pimephales promelas*) to an ethinylestradiol concentration below 1 ng/L reduces egg fertilization success and demasculinizes males. *Environ. Toxicol.* **2005,** *20*, (2), 131-141

13. Armstrong, B. M.; Lazorchak, J. M.; Murphy, C. A.; Haring, H. J.; Jensen, K. M.; Smith, M. E., Determining the effects of a mixture of an endocrine disrupting compound, 17 alpha-ethinylestradiol, and ammonia on fathead minnow (*Pimephales promelas*) reproduction. *Chemosphere* **2015,** *120*, 108-114.10.1016/j.chemosphere.2014.06.049

14. Zha, J.; Sun, L.; Zhou, Y.; Spear, P. A.; Ma, M.; Wang, Z., Assessment of 17 alpha-ethinylestradiol effects and underlying mechanisms in a continuous, multigeneration

566 exposure of the Chinese rare minnow (*Gobiocypris rarus*). *Toxicol. Appl. Pharm.* **2008,** *226*,
567 (3), 298-308.10.1016/j.taap.2007.10.006

568 15. Kidd, K. A.; Blanchfield, P. J.; Mills, K. H.; Palace, V. P.; Evans, R. E.; Lazorchak, J.
569 M.; Flick, R. W., Collapse of a fish population after exposure to a synthetic estrogen. *PNAS*
570 **2007,** *104*, (21), 8897-8901.10.1073/pnas.0609568104

571 16. Blanchfield, P. J.; Kidd, K. A.; Docker, M. F.; Palace, V. P.; Park, B. J.; Postma, L.
572 D., Recovery of a wild fish population from whole-lake additions of a synthetic estrogen.
573 *Environ. Sci. Technol.* **2015,** *49*, (5), 3136-3144.10.1021/es5060513

574 17. Hamilton, P.; Nicol, E.; De-Bastos, E.; Williams, R.; Sumpter, J.; Jobling, S.; Stevens,
575 J.; Tyler, C., Populations of a cyprinid fish are self-sustaining despite widespread
576 feminization of males. *BMC Biol.* **2014,** *12*, (1), 1

577 18. Wirgin, I.; Roy, N. K.; Loftus, M.; Chambers, R. C.; Franks, D. G.; Hahn, M. E.,
578 Mechanistic Basis of Resistance to PCBs in Atlantic Tomcod from the Hudson River. *Science*
579 **2011,** *331*, (6022), 1322-1325

580 19. Oziolor, E. M.; Bigorgne, E.; Aguilar, L.; Usenko, S.; Matson, C. W., Evolved
581 resistance to PCB- and PAH-induced cardiac teratogenesis, and reduced CYP1A activity in
582 Gulf killifish (*Fundulus grandis*) populations from the Houston Ship Channel, Texas. *Aquat.*
583 *Toxicol.* **2014,** *150*, (0), 210-219.http://dx.doi.org/10.1016/j.aquatox.2014.03.012

584 20. Williams, L. M.; Oleksiak, M. F., Ecologically and evolutionarily important SNPs
585 identified in natural populations. *Mol. Biol. Evol.* **2011,** *28*, (6), 1817-
586 1826.10.1093/molbev/msr004

587 21. Reid, N. M.; Proestou, D. A.; Clark, B. W.; Warren, W. C.; Colbourne, J. K.; Shaw, J.
588 R.; Karchner, S. I.; Hahn, M. E.; Nacci, D.; Oleksiak, M. F.; Crawford, D. L.; Whitehead, A.,
589 The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild
590 fish. *Science* **2016,** *354*, (6317), 1305-1308.10.1126/science.aah4993

591 22. Brazzola, G.; Chèvre, N.; Wedekind, C., Additive genetic variation for tolerance to
592 estrogen pollution in natural populations of Alpine whitefish (*Coregonus* sp., Salmonidae).
593 *Evol Appl* **2014,** *7*, (9), 1084-1093.10.1111/eva.12216

594 23. Shi, B.; Wen, H. S.; He, F.; Dong, S. L.; Ma, S.; Chen, C. F.; Chen, X. Y.; Zhang, J.
595 R.; Jin, G. X., Single nucleotide polymorphisms within the estrogen receptor beta gene are
596 linked with reproductive indices in Japanese flounder, *Paralichthys olivaceus*. *Comp*.
597 *Biochem*. *Phys*. *B*. **2009,** *154*, (1), 62-67

598 24. Yong , E. L.; Loy, C. J.; Sim, K. S., Androgen receptor gene and male infertility.
599 *Human Reproduction Update* **2003,** *9*, (1), 1-7.10.1093/humupd/dmg003

600 25. Foll, M.; Gaggiotti, Oscar E.; Daub, Josephine T.; Vatsiou, A.; Excoffier, L.,
601 Widespread signals of convergent adaptation to high altitude in Asia and America. *Am J Hum*
602 *Genet* **2014,** *95*, (4), 394-407.10.1016/j.ajhg.2014.09.002

603 26. Harries, J. E.; Sheahan, D. A.; Jobling, S.; Matthiessen, P.; Neall, P.; Routledge, E. J.;
604 Rycroft, R.; Sumpter, J. P.; Tylor, T., A survey of estrogenic activity in United Kingdom
605 inland waters. *Environ*. *Toxicol*. *Chem* **1996,** *15*, (11), 1993-2002

606 27. Harries, J. E.; Sheahan, D. A.; Jobling, S.; Matthiessen, P.; Neall, M.; Sumpter, J. P.;
607 Taylor, T.; Zaman, N., Estrogenic activity in five United Kingdom rivers detected by
608 measurement of vitellogenesis in caged male trout. *Environ*. *Toxicol*. *Chem* **1997,** *16*, (3),
609 534-542

610 28. Williams, R. J.; Keller, V. D. J.; Johnson, A. C.; Young, A. R.; Holmes, M. G. R.;
611 Wells, C.; Gross-Sorokin, M.; Benstead, R., A National Risk Assessment For Intersex In Fish
612 Arising From Steroid Estrogens. *Environ. Toxicol. Chem* **2009,** *28*, (1), 220-230

613 29. Jürgens, M. D.; Crosse, J.; Hamilton, P. B.; Johnson, A. C.; Jones, K. C., The long
614 shadow of our chemical past – High DDT concentrations in fish near a former agrochemicals
615 factory in England. *Chemosphere* **2016,** *162*, 333-
616 344.https://doi.org/10.1016/j.chemosphere.2016.07.078

617 30. Filby, A. L.; Santos, E. M.; Thorpe, K. L.; Maack, G.; Tyler, C. R., Gene expression
618 profiling for understanding chemical causation of biological effects for complex mixtures: A
619 case study on Estrogens. *Environ. Sci. Technol.* **2007,** *41*, (23), 8187-8194

620 31. Grabherr, M. G.; Haas, B. J.; Yassour, M.; Levin, J. Z.; Thompson, D. A.; Amit, I.;
621 Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; Chen, Z.; Mauceli, E.; Hacohen, N.;
622 Gnirke, A.; Rhind, N.; di Palma, F.; Birren, B. W.; Nusbaum, C.; Lindblad-Toh, K.;
623 Friedman, N.; Regev, A., Full-length transcriptome assembly from RNA-Seq data without a
624 reference genome. *Nat Biotech* **2011,** *29*, (7), 644-
625 652.http://www.nature.com/nbt/journal/v29/n7/abs/nbt.1883.html#supplementary-information

626 32. Li, H.; Durbin, R., Fast and accurate short read alignment with Burrows–Wheeler
627 transform. *Bioinformatics* **2009,** *25*, (14), 1754-1760.10.1093/bioinformatics/btp324

628 33. Frichot, E.; François, O., LEA: An R package for landscape and ecological
629 association studies. *Methods in Ecology and Evolution* **2015,** *6*, (8), 925-929.10.1111/2041-
630 210x.12382

631 34. Foll, M.; Gaggiotti, O., A genome-scan method to identify selected loci appropriate
632 for both dominant and codominant markers: a Bayesian perspective. *Genetics* **2008,** *180*, (2),
633 977-993

634 35. Beaumont, M. A.; Nichols, R. A., Evaluating loci for use in the genetic analysis of
635 population structure. *P. Roy. Soc. B-Biol. Sci.* **1996,** *263*, (1377), 1619-1626

636 36. Antao, T.; Lopes, A.; Lopes, R. J.; Beja-Pereira, A.; Luikart, G., LOSITAN: A
637 workbench to detect molecular adaptation based on a F(st)-outlier method. *BMC*
638 *Bioinformatics* **2008,** *9*, 323.doi.org/10.1186/1471-2105-9-323

639 37. Narum, S. R.; Hess, J. E., Comparison of F(ST) outlier tests for SNP loci under
640 selection. *Mol Ecol Resour* **2011,** *11*, 184-194

641 38. Excoffier, L.; Lischer, H. E. L., Arlequin suite ver 3.5: a new series of programs to
642 perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* **2010,** *10*,
643 (3), 564-567.10.1111/j.1755-0998.2010.02847.x

644 39. Etter, P.; Bassham, S.; Hohenlohe, P.; Johnson, E.; Cresko, W., SNP Discovery and
645 Genotyping for Evolutionary Genetics Using RAD Sequencing. In *Molecular Methods for*
646 *Evolutionary Genetics*, Orgogozo, V.; Rockman, M. V., Eds. Humana Press: 2011; Vol. 772,
647 pp 157-178.

648 40. Catchen, J.; Hohenlohe, P. A.; Bassham, S.; Amores, A.; Cresko, W. A., Stacks: an
649 analysis tool set for population genomics. *Mol. Ecol.* **2013,** *22*, (11), 3124-
650 3140.10.1111/mec.12354

651 41. Catchen, J. M.; Amores, A.; Hohenlohe, P.; Cresko, W.; Postlethwait, J. H., Stacks:
652 building and genotyping loci de novo from short-read sequences. *G3-Genes Genom. Genet.*
653 **2011,** *1*, (3), 171-182.10.1534/g3.111.000240

654 42. Fraser, B. A.; Künstner, A.; Reznick, D. N.; Dreyer, C.; Weigel, D., Population
655 genomics of natural and experimental populations of guppies (*Poecilia reticulata*). *Mol. Ecol.*
656 **2015,** *24*, (2), 389-408.10.1111/mec.13022

657 43. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., Basic local
658 alignment search tool. *J. Mol. Biol.* **1990,** *215*, (3), 403-410.10.1006/jmbi.1990.9999

659 44. Flicek, P.; Amode, M. R.; Barrell, D.; Beal, K.; Billis, K.; Brent, S.; Carvalho-Silva,
660 D.; Clapham, P.; Coates, G.; Fitzgerald, S.; Gil, L.; Girón, C. G.; Gordon, L.; Hourlier, T.;
661 Hunt, S.; Johnson, N.; Juettemann, T.; Kähäri, A. K.; Keenan, S.; Kulesha, E.; Martin, F. J.;
662 Maurel, T.; McLaren, W. M.; Murphy, D. N.; Nag, R.; Overduin, B.; Pignatelli, M.;
663 Pritchard, B.; Pritchard, E.; Riat, H. S.; Ruffier, M.; Sheppard, D.; Taylor, K.; Thormann, A.;
664 Trevanion, S. J.; Vullo, A.; Wilder, S. P.; Wilson, M.; Zadissa, A.; Aken, B. L.; Birney, E.;
665 Cunningham, F.; Harrow, J.; Herrero, J.; Hubbard, T. J. P.; Kinsella, R.; Muffato, M.; Parker,
666 A.; Spudich, G.; Yates, A.; Zerbino, D. R.; Searle, S. M. J., Ensembl 2014. *Nucleic Acids*
667 *Res.* **2014,** *42*, (D1), D749-D755.10.1093/nar/gkt1196

668 45. Huang, D. W.; Sherman, B. T.; Lempicki, R. A., Systematic and integrative analysis
669 of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **2009,** *4*, (1), 44-
670 57.10.1038/nprot.2008.211

671 46. Gottlieb, B.; Beitel, L. K.; Nadarajah, A.; Paliouras, M.; Trifiro, M., The androgen
672 receptor gene mutations database: 2012 update. *Human Mutation* **2012,** *33*, (5), 887-
673 894.10.1002/humu.22046

674 47. Grummer, J. A.; Beheregaray, L. B.; Bernatchez, L.; Hand, B. K.; Luikart, G.; Narum,
675 S. R.; Taylor, E. B., Aquatic landscape genomics and environmental effects on genetic
676 variation. *Trends Ecol. Evol.* **2019,** *34*, (7), 641-654.10.1016/j.tree.2019.02.013

677 48. Kuang, Y.-Y.; Zheng, X.-H.; Li, C.-Y.; Li, X.-M.; Cao, D.-C.; Tong, G.-X.; Lv, W.-
678 H.; Xu, W.; Zhou, Y.; Zhang, X.-F.; Sun, Z.-P.; Mahboob, S.; Al-Ghanim, K. A.; Li, J.-T.;
679 Sun, X.-W., The genetic map of goldfish (*Carassius auratus*) provided insights to the
680 divergent genome evolutions in the Cyprinidae family. *Scientific Reports* **2016,** *6*, (1),
681 34849.10.1038/srep34849

682 49. Hoban, S.; Kelley, J. L.; Lotterhos, K. E.; Antolin, M. F.; Bradburd, G.; Lowry, D. B.;
683 Poss, M. L.; Reed, L. K.; Storfer, A.; Whitlock, M. C., Finding the genomic basis of local
684 adaptation: pitfalls, practical solutions, and future directions. *The American Naturalist* **2016,**
685 *188*, (4), 379-397.10.1086/688018

686 50. Wang, Z.; Li, T.; Xing, X.; Gao, X.; Zhang, X.; You, L.; Zhao, H.; Ma, J.; Chen, Z.-
687 J., Replication study of RAD54B and GREB1 polymorphisms and risk of PCOS in Han
688 Chinese. *Reprod. Biomed. Online* **2013,** *27*, (3), 316-321.10.1016/j.rbmo.2013.05.007

689 51. Borg, B., Androgens in teleost fishes. *Comp. Biochem. Physiol. C-Toxicol.*
690 *Pharmacol.* **1994,** *109*, (3), 219-245.https://doi.org/10.1016/0742-8413(94)00063-G

691 52. Wilson, E. M.; French, F. S., Binding properties of androgen receptors. Evidence for
692 identical receptors in rat testis, epididymis, and prostate. *J Biol Chem* **1976,** *251*, (18), 5620-
693 5629

694 53. Coe, T. S.; Hamilton, P. B.; Hodgson, D.; Paull, G. C.; Stevens, J. R.; Sumner, K.;
695 Tyler, C. R., An environmental estrogen alters reproductive hierarchies, disrupting sexual
696 selection in group-spawning fish. *Environ. Sci. Technol.* **2008,** *42*, (13), 5020-5025

54. Nikoleris, L.; Hultin, C. L.; Hallgren, P.; Hansson, M. C., 17α-Ethinylestradiol (EE2) treatment of wild roach (*Rutilus rutilus*) during early life development disrupts expression of genes directly involved in the feedback cycle of estrogen. *Comp. Biochem. Phys. C Toxicol. Pharmacol. 2* **2016,** *180*, 56-64.http://dx.doi.org/10.1016/j.cbpc.2015.12.002

55. The Marie-Genica Consortium on Genetic Susceptibility for Menopausal Hormone Therapy Related Breast Cancer Risk, Polymorphisms in genes of the steroid receptor superfamily modify postmenopausal breast cancer risk associated with menopausal hormone therapy. *Int J Cancer* **2010,** *126*, (12), 2935-2946.10.1002/ijc.24892

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

57. Kojima, H.; Katsura, E.; Takeuchi, S.; Niiyama, K.; Kobayashi, K., Screening for estrogen and androgen receptor activities in 200 pesticides by in vitro reporter gene assays using Chinese hamster ovary cells. *Environ. Health Persp.* **2004,** *112*, (5), 524-531.10.1289/ehp.6649

58. Kelce, W. R.; Stone, C. R.; Laws, S. C.; Gray, L. E.; Kemppainen, J. A.; Wilson, E. M., Persistent DDT metabolite p,p'–DDE is a potent androgen receptor antagonist. *Nature* **1995,** *375*, (6532), 581-585.10.1038/375581a0

59. Garcia-Reyero, N.; Barber, D. S.; Gross, T. S.; Johnson, K. G.; Sepúlveda, M. S.; Szabo, N. J.; Denslow, N. D., Dietary exposure of largemouth bass to OCPs changes expression of genes important for reproduction. *Aquat. Toxicol.* **2006,** *78*, (4), 358-369.http://dx.doi.org/10.1016/j.aquatox.2006.05.003

718 60. Cowx, I. G.; Broughton, N. M., Changes In the species composition of anglers'
719 catches in the River Trent (England) Between 1969 and 1984. *J. Fish Biol.* **1986,** *28*, (5),
720 625-636

721 61. Williams, L. M.; Oleksiak, M. F., Evolutionary and functional analyses of cytochrome
722 P4501A promoter polymorphisms in natural populations. *Mol. Ecol.* **2011,** *20*, (24), 5236-
723 5247.10.1111/j.1365-294X.2011.05360.x

724 62. Wirgin, I.; Waldman, J. R., Resistance to contaminants in North American fish
725 populations. *Mutation Research-Fundamental And Molecular Mechanisms Of Mutagenesis*
726 **2004,** *552*, (1-2), 73-100

727 63. Wang, D.; Yang, X.; Shen, B., A novel erythroid differentiation related gene EDRF2
728 inhibited α-globin gene expression in K562 cells. *Chin. Sci. Bull.* **2002,** *47*, (5), 398-
729 402.10.1360/02tb9093

730 64. Hammond, D.; Pryce, A. R. *Climate change impacts and water temperature*, UK,
731 Environment Agency: Bristol, 2007.

732 65. Rallapalli, P. M.; Orengo, C. A.; Studer, R. A.; Perkins, S. J., Positive selection
733 during the evolution of the blood coagulation factors in the context of their disease-causing
734 mutations. *Mol. Biol. Evol.* **2014,** *31*, (11), 3040-3056.10.1093/molbev/msu248

735 66. Nelson, E. R.; DuSell, C. D.; Wang, X.; Howe, M. K.; Evans, G.; Michalek, R. D.;
736 Umetani, M.; Rathmell, J. C.; Khosla, S.; Gesty-Palmer, D.; McDonnell, D. P., The
737 oxysterol, 27-hydroxycholesterol, links cholesterol metabolism to bone homeostasis through
738 its actions on the estrogen and liver x receptors. *Endocrinology* **2011,** *152*, (12), 4691-
739 4705.10.1210/en.2011-1298

67. Martyniuk, C. J.; Gerrie, E. R.; Popesku, J. T.; Ekker, M.; Trudeau, V. L., Microarray analysis in the zebrafish (*Danio rerio*) liver and telencephalon after exposure to low concentration of 17alpha-ethinylestradiol. *Aquat. Toxicol.* **2007,** *84*, (1), 38-49.https://doi.org/10.1016/j.aquatox.2007.05.012

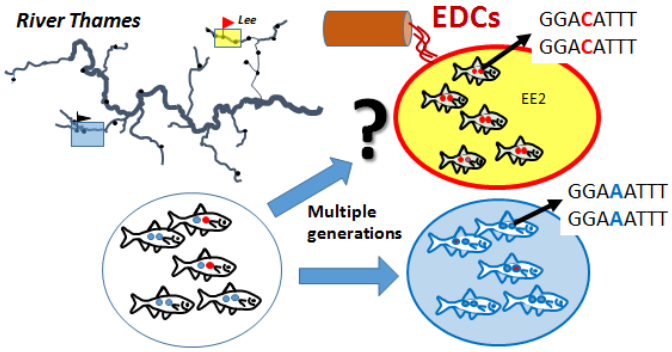
68. Baillon, L.; Pierron, F.; Coudret, R.; Normendeau, E.; Caron, A.; Peluhet, L.; Labadie, P.; Budzinski, H.; Durrieu, G.; Sarraco, J.; Elie, P.; Couture, P.; Baudrimont, M.; Bernatchez, L., Transcriptome profile analysis reveals specific signatures of pollutants in Atlantic eels. *Ecotoxicology* **2015,** *24*, (1), 71-84.10.1007/s10646-014-1356-x

69. Sheahan, D. A.; Brighty, G. C.; Daniel, M.; Jobling, S.; Harries, J. E.; Hurst, M. R.; Kennedy, J.; Kirby, S. J.; Morris, S.; Routledge, E. J.; Sumpter, J. P.; Waldock, M. J., Reduction in the estrogenic activity of a treated sewage effluent discharge to an English river as a result of a decrease in the concentration of industrially derived surfactants. *Environ. Toxicol. Chem* **2002,** *21*, (3), 515-519

70. Griffiths, A. M.; Ellis, J. S.; Clifton-Dey, D.; Machado-Schiaffino, G.; Bright, D.; Garcia-Vazquez, E.; Stevens, J. R., Restoration versus recolonisation: The origin of Atlantic salmon (*Salmo solar* L.) currently in the River Thames. *Biol. Conserv.* **2011,** *144*, (11), 2733-2738

71. Cavalli-Sforza , L.; Edwards, A., Phylogenetic analysis: models and estimation procedures. *Am J Hum Genet* **1967,** *19*, 233-257

763 GRAPHICAL ABSTRACT




764

765

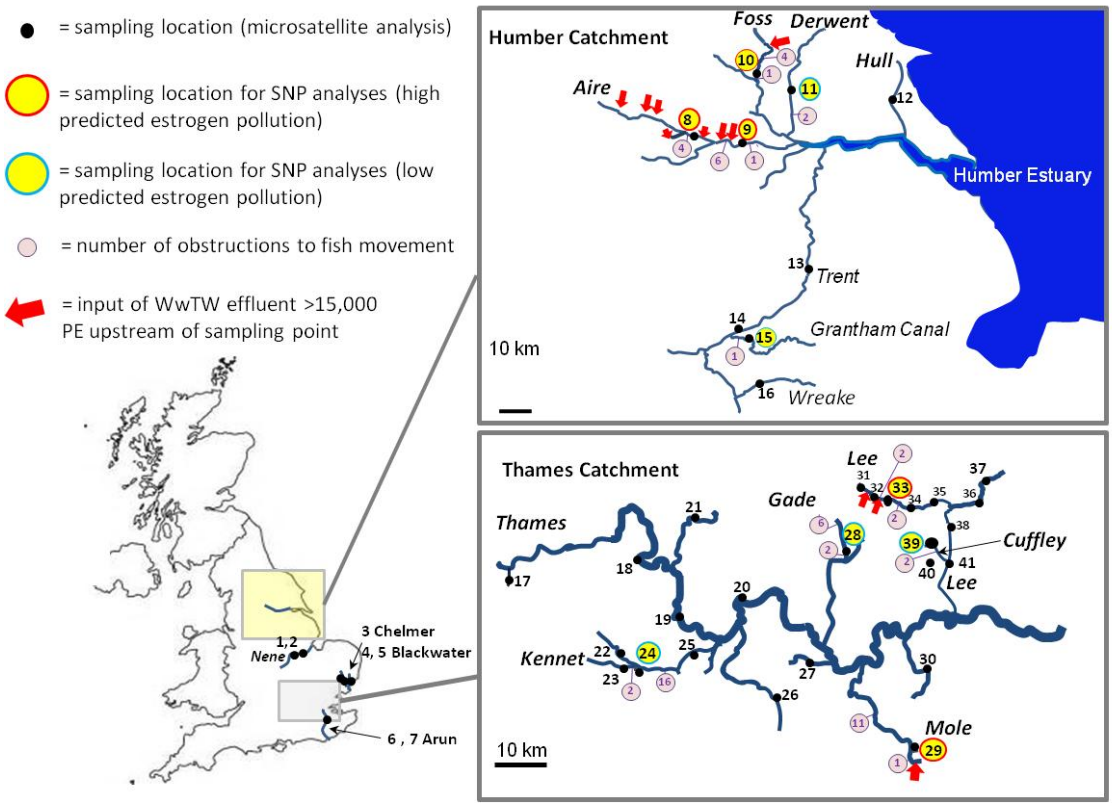
766

767

768

769

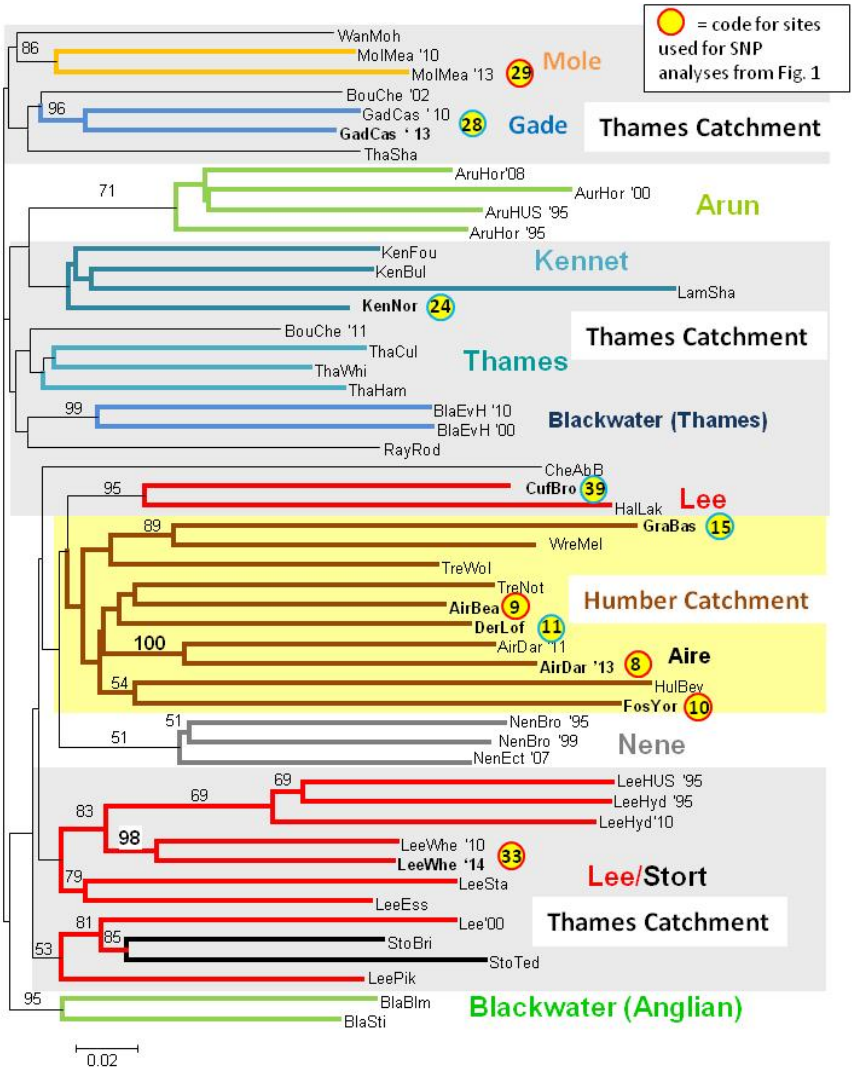
770 FIGURES




771

772 Figure 1. Locations of river sample sites in England genotyped in this study. Sample codes: 1.
773 NenBro; 2. NenEct; 3. CheAbB; 4. BlaBlM; 5. BlaSti, 6. AruHor; 7. AruHUS; 8. AirDar; 9.
774 AirBea; 10. FosYor; 11. DerLof; 12. HulBev; 13. TreWol; 14. TreNot; 15 GraBas; 16
775 WreMel; 17. RayRod; 18. ThaCul; 19. ThaWhi; 20. ThaHam; 21. ThaSha 22. LamSha 23.
776 KenBul 24. KenNor; 25. KenFou; 26. BlaEvH; 27. BouChe; 28. GadCas; 29. MolMea; 30.
777 WanMoh; 31. LeeHUS; 32. LeeHyd; 33. LeeWhe; 34. LeeSta; 35. LeeEss; 36. StoBri; 37.
778 StoTed; 38. Lee'00 (exact location uncertain); 39. CufBro; 40. HalLak; 41. LeePik. Details of
779 newly sampled locations are given in Additional file 1, Table S1. For the River Aire locations
780 there are 9 and 15 WwTWs with a population served greater than 15,000 upstream of AirDar
781 and AirBea respectively. Further details on sample sites and obstructions to fish movement
782 (locks and weirs) in the Thames Catchment are given in the map figure in Hamilton *et al*.[17]

783

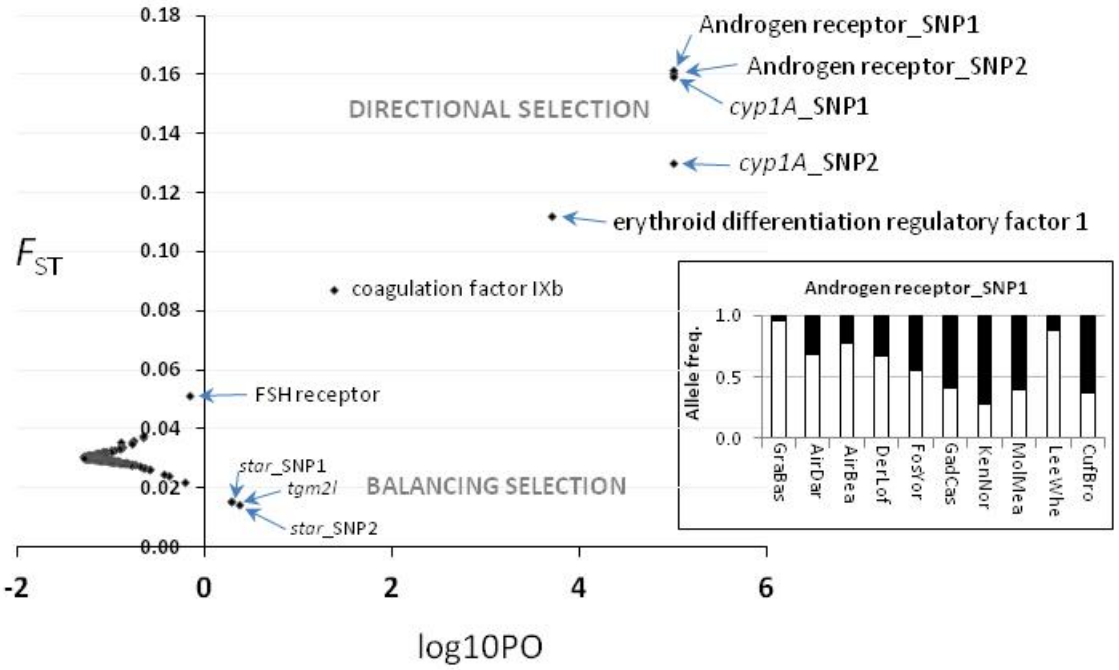



784

785

786 **Figure 2.** Neighbor-joining tree for roach population samples produced from data from 2369
787 roach from 41 sample sites. Several locations were sampled in different years, producing a
788 total of 51 'samples'. The tree is based on the data from 14 microsatellite loci using Cavalli-
789 Sforza and Edwards' chord distance measure, $D_C$[71]. Only bootstrap values above 50% are
790 shown. Numbers at the end of sample codes indicate years in which populations were
791 sampled (where the same location was sampled in different years). Locations of rivers used
792 are shown in the map (Figure 1).

793




794

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