An investigation into adaptation in genes associated with exposure to estrogenic pollution in populations of roach (*Rulitus 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

34 populations sampled from rivers receiving either high or low level WwTW discharges.
35 SNPs within 36 'candidate' genes, selected for their involvement in estrogenic
36 responses, and 120 SNPs in reference genes were genotyped in 465 roach. There was
37 no evidence for selection in highly estrogen-dependent candidate genes, including
38 those for the estrogen receptors, aromatases and vitellogenins. The androgen
39 receptor (*ar*) and cytochrome P450 1A genes were associated with large shifts in
40 allele frequencies between catchments and in individual populations, but there is no
41 clear link to estrogen pollution. Selection at *ar* in the effluent dominated River Lee
42 may have resulted from historical contamination with endocrine disrupting
43 pesticides. Critically, while our results suggest population-specific selection including
44 at genes related to endocrine disruption, there was no strong evidence the selection
45 resulted from exposure to estrogen pollution.

46 INTRODUCTION

47

48 The occurrence of feminized male fish has been reported in rivers and estuaries on
49 several continents and has been attributed to pollution by natural and synthetic
50 steroid estrogens, including ethinylestradiol (EE2),[1-2] contained in wastewater
51 treatment work (WwTW) effluents. Feminized male characteristics known to be
52 induced by steroid estrogens include the presence of precursors of egg yolk proteins,
53 such as vitellogenin (VTG), in the blood plasma,[3] feminized reproductive ducts and
54 the presence of developing eggs in otherwise male gonads.[4] This intersex
55 phenomenon associated with exposures to WwTW effluents was first reported to be
56 widespread in roach (*Rutilus rultilus*) in English rivers in the 1990s and the 2000s,[5-6]
57 and has since been reported in many species of both riverine and estuarine fish in
58 several countries of the world.

59 *In vitro* fertilization studies using wild male roach (*Rutilus rutilus*)[7] indicate that fish
60 with feminized gonads have reduced fertility, and a competitive breeding study
61 found wild male roach with moderately to severely feminized gonads to have
62 reduced reproductive output.[8] Exposures of roach (*Rutilus rutilus*) to undiluted
63 effluent[9] or to 4-6 ng/L EE2 over the period of sexual development[10-11] have been
64 shown to result in full sex reversal and/or breeding failure and long-term laboratory
65 exposures to lower concentrations of 0.47-1 ng/L EE2 (predicted for rivers heavily
66 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_2Eq) (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

144 tBLASTx algorithms implemented in Seqtools version 8.4.017
145 (http://www.seqtools.dk/) and the roach transcriptome as a local database.

146 **Transcriptome Sequencing/Assembly.** The transcriptome of roach was
147 sequenced in order to identify genetic variants for subsequent SNP genotyping.
148 These were submitted to NCBI Short Read Archive (SRA) associated with BioProject
149 PRNJA295813. A *de novo* transcriptome was generated from the trimmed, filtered
150 and repaired FASTA files using sequences from 8 libraries using Trinity
151 (version:trinityrnaseq_r20140717).[31] The resulting FASTA file was submitted to the
152 Transcriptome Shotgun Assembly sequence database (TSA) associated with
153 BioProject PRJNA295813.

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

157 **SNP Identification.** Reads from each library were mapped back to the modified
158 transcriptome using the Burrows-Wheeler Aligner (BWA) program version 0.7.5a-
159 r405.[32] Variant sites were identified using a custom Perl script (Additional file 3). The
160 fragmented roach genome sequences were then used to identify intron positions, so
161 that they could be avoided or included in the SNP-genotyping primers. SNPs from
162 the transcriptome were substituted into the corresponding position in contigs
163 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, 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*.[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,

206 using an e value cut off of $< 1 \times 10^{-5}$. To identify the population in which selection is
207 likely to have occurred, F_{ST} values for loci of interest were examined in the other two
208 pairwise comparisons. Less stringent criteria ($F_{ST} > 0.8$, $p < 0.05$) were used for this
209 comparison. Gene ontology (GO) analysis was conducted in Database for Annotation,
210 Visualisation and Integrated Discovery (DAVID),[45] using *Danio rerio* as a background.
211
212 RESULTS
213
214 **Single Nucleotide Polymorphism (SNP) Identification and Genotyping.**
215 Transcriptome sequencing yielded 184.5 million reads 150 bp paired-end reads after
216 quality trimming (94.04%) – Table S4. The transcriptome assembly yielded 200,361
217 transcripts (summary statistics are given in Additional file 1, Table S5). 25,886 genes
218 were identified using the Ensembl peptide database for *Danio rerio*. Genome
219 sequencing of a single male roach generated 249.7 million reads after removal of low
220 quality sequences.
221 A total of 217 SNPs were successfully genotyped in 465 fish from 10 locations in 9
222 rivers with overall genotyping success of 99.24%. Eighty four were in 36 genes
223 related to estrogen response candidate genes, 12 were in four other genes related to
224 selection and 120 were each in a different reference gene (see Table S6 for
225 genotyped candidates - Additional file 1). SNPs within genes of some of the most
226 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.[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[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")[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 *Itbp3*, 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 (χ^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).

263 **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**
Itbp3_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
Clc13_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 had signatures of selection in at least 1 pairwise comparison (e.g. for Lositan (χ^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*, *Itbp3*) 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 (χ^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.[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

333 those identified in the LeeWhe-CufBro comparison. No enriched GO terms in DAVID
334 [45] were identified.

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

341

342 DISCUSSION

343

344 Understanding the impacts of chemical pollution on fish populations requires
345 knowledge of the ability of fish to tolerate and/or adapt to the harmful effects of
346 exposure. Our results identified several genes involved in responses to endocrine
347 disrupting pollutants which were highly differentiated between populations, a
348 potential result of selection. However, there was no evidence that these allele shifts
349 resulted from adaptation to estrogen pollution, as there were no consistent allele
350 shifts in the most obvious candidate genes between populations in clean and
351 effluent dominated rivers stretches within catchments. This is despite the inclusion of
352 some populations restricted to river stretches with some of the highest known
353 proportions of WwTW effluent in UK rivers. The androgen receptor (*ar*) and *cyp1A*

354 exhibited large shifts in allele frequency both between individual populations of
355 roach within catchments and between catchments. Though our study provided no
356 clear link with estrogen pollution. to our knowledge the androgen receptor has not
357 previously been implicated in local adaptation in fish. *Cyp1A* has previously been
358 associated in adaptation to hydrocarbon pollutants in other fish species,[e.g. 20, 21]
359 although the pattern here does not implicate selection resulting from WwTW
360 pollution.

361 In fish, linkage blocks can range from 1 kb in zebrafish (*Danio rerio*) to 1 Mb in lake
362 whitefish (*Coregonus clupeaformis*)[47] Under strong, recent selection, linkage blocks
363 can be large; in killifish the median lengths outlier windows at polluted sites were 50-
364 62 kb but some haplotypes were larger including 650 kb haplotype containing the
365 AIP gene {Reid, 2016 #2377}. This raises the possibility that allele shifts observed at
366 the *ar* and *cyp1A* had occurred by selection in linked genes. Our data, however,
367 suggest this is not the case for *ar*, as the two SNPs have different patterns of
368 selection in both catchments. These SNPs are separated by 7 kb in the zebrafish
369 genome, which has synteny with other cyprinid fish[48] . In contrast, the two *cyp1A*
370 SNPs are separated by only 145 bp and have the same patterns of selection. The
371 closest genes to these SNPs are 67 kb for the *ar* and 29 kb for the *cyp1A*. Indeed, our
372 results suggest that large allele shifts have occurred in the *ar* at least twice within the
373 Thames Catchment, with a unique allele shift at the LeeWhe population.

374 The results of the correlation analysis (using LFMM) did not provide strong
375 evidence for adaptation to steroid estrogen pollution. There was no difference in the
376 proportion of candidates and reference genes identified under selection using this
377 method. Furthermore, none of the obvious candidate genes known for estrogen
378 response (e.g. estrogen receptors, aromatases and vitellogenins) showed correlations
379 with estrogenic pollution. Additionally, the estrogen-adaptation candidate genes
380 (*vasa*, *bcra2* and *Itbp3*) identified were not subject to large shifts in allele frequency in
381 any population. Of the four reference genes that correlated with estrogen pollution,
382 three had no obvious link with estrogen pollution (*edrf*, *pcdh17,* and *znf518B*). The
383 fourth, *rad54b*, is involved in DNA repair, but humans variants have been associated
384 with excessive levels of androgens in females;[49] so variants could potentially modify
385 responses to EDCs in fish. Thus, overall these results do not provide strong evidence
386 for parallel selection related to estrogen pollution, but do not exclude an influence.
387 It is possible that some, but not all, populations of roach have adapted to
388 estrogenic pollution, or that different populations have adapted, but through
389 different mechanisms. Such patterns would not have been identified in the
390 correlation analysis. For instance, the large allele shift at the *ar* in the population from
391 the River Lee (LeeWhe) could be a consequence of estrogenic pollution. In males,
392 androgens play key roles in sexual development, puberty, the development of
393 secondary sexual characteristics, and reproductive behaviour.[50] Estrogens are
394 antagonists of AR androgen binding,[51] can reduce androgen levels in male fish[52] and

395 modify *ar* expression[53] at an estrogenic potency (5 ng/L E2Eq) similar to the average
396 (6.6 E2Eq ng/L) predicted for this river site. The effect of *ar* polymorphisms in fish is
397 not known, but in humans they modify susceptibility to the effects of estrogen
398 exposure.[54]
399 Adaptation to pollution from other endocrine disrupting chemicals could also explain
400 differentiation of the *ar* in the LeeWhe population. Elevated concentrations of
401 pesticides including dichlorodiphenyltrichloroethane (DDT) metabolites (e.g. p,p'
402 dichlorodiphenyldichloroethylene (DDE))[29], endosulfan and lindane[55] were detected
403 in the tissues of roach sampled at this location 20 years after a pesticide formulation
404 factory next to this site closed in 1982.[29] The p,p'DDE concentrations equated to
405 those known to affect the early life stages of fish (gene expression and gonadal
406 intersex) and approaching reported effect concentrations for adult fish.[29] Several DDT
407 metabolites are anti-androgenic and some are also estrogenic.[56-57] and alter
408 expression of estrogen receptors in fish.[58]
409 It is also possible that the shifts in allele frequency at the androgen receptor do not
410 relate to adaptation to pollution as one *ar* SNP is also highly differentiated in the
411 population from an isolated stretch of the Grantham Canal (GraBas) with no known
412 WwTW inputs (see Additional file 1 Table S1). We cannot exclude selection from
413 other EDC pollutant from an unidentified source or in the neighboring polluted river
414 Trent[60] before the separation of these waterways approximately 50 years ago. Thus
415 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.[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

437 selection due to differences in oxygen availability; average water temperatures are
438 approximately 2^{o} C higher in the more southerly Thames Catchment[64] and rivers in
439 both catchments would have suffered from nutrient-rich pollution e.g. from fertilizers
440 and poorly treated sewage. High differentiation at coagulation factor IXb (*f9b*) may
441 relate to adaptation against blood pathogens; the coagulation system has been
442 under strong selective pressure in primates, possibly for this resason.[65]

443 Analysis of the population from the estrogenic River Lee (LeeWhe) using RAD-seq
444 provided no evidence for adaptation to estrogen pollution, as genes involved in
445 estrogen response were not overrepresented among loci with elevated F_{ST} values in
446 comparisons with populations from clean sites. Indeed the only gene that was found
447 to be related to endocrine disruption under directional selection in the LeeWhe
448 population was oxysterol binding protein 7. Interestingly 3 other oxysterol binding
449 proteins were also identified in the LeeWhe-CufBro comparison but the direction of
450 selection was not determined. Oxysterols modify estrogen receptor function and can
451 bind to, and modulate, the activity of ERα and ERβ.[66] Expression of genes for these
452 proteins is modified by estrogen[67] and lindane[68] found at elevated concentrations in
453 tissues from roach from this River Lee location[29] The LeeWhe-CufBro comparison
454 identified *ar*, confirming the result from the targeted gene analysis. Nevertheless,
455 *cyp1A* was not identified using this method, despite the large allele shift in the
456 CufBro population. Thus a resequencing approach[21] would enable a more complete
457 and detailed analysis of genes under selection.

458 Limitations of this study include that the full history of roach within these rivers is
459 not known. Each population will have had different levels of immigration, most
460 restocking events are undocumented and the success of this restocking is unknown.
461 Levels of estrogen contamination will have varied over time with changes in waste-
462 water treatments processes and changes in industry chemical use. For instance the
463 concentration of nonylphenol, responsible for a major part of the estrogenicity in the
464 River Aire,[27] decreased during the 1990s.[70] Levels of other EDC pollutants have not
465 been recorded; the high levels of DDT metabolites for fish in the River Lee were only
466 discovered accidently.[29] For further information on history of fish in these rivers see
467 Additional file 1. Another limitation is that false positives in F_{ST} outlier tests can arise
468 from historic demographic events such as recent range expansions.{Hoban, 2016
469 #2416} Here, average F_{ST} is low, and all populations have high genetic diversity,
470 reducing false positives.{Hoban, 2016 #2416} The key role of *cyp1A* in adaptation in
471 other fish species, and the near certainty of the occurrence similar pollutants in some
472 of these rivers adds confidence that this has resulted from natural selection.
473 Likewise, the occurrence of independent large allele shifts in *ar* in both catchments
474 adds confidence that at least one of these shifts results from natural selection.
475 Irrespective of the cause of the highly differentiated loci observed in this study, our
476 results caution against extrapolating effects from fish derived from only one
477 population for assessing the impacts of endocrine disrupting chemicals on the health
478 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,[71] 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

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

564 References

565

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

569 2. Hinck, J. E.; Blazer, V. S.; Schmitt, C. J.; Papoulias, D. M.; Tillitt, D. E., Widespread
570 occurrence of intersex in black basses (*Micropterus* spp.) from US rivers, 1995-2004. *Aquat. Toxicol.*
571 **2009,** *95*, (1), 60-70

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

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

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

578 6. Jobling, S.; Williams, R.; Johnson, A.; Taylor, A.; Gross-Sorokin, M.; Nolan, M.; Tyler, C.
579 R.; van Aerle, R.; Santos, E.; Brighty, G., Predicted exposures to steroid estrogens in UK rivers
580 correlate with widespread sexual disruption in wild fish populations. *Environ. Health Persp.* **2006,**
581 *114*, 32-39

582 7. Jobling, S.; Coey, S.; Whitmore, J. G.; Kime, D. E.; Van Look, K. J. W.; McAllister, B. G.;
583 Beresford, N.; Henshaw, A. C.; Brighty, G.; Tyler, C. R.; Sumpter, J. P., Wild intersex roach (*Rutilus*
584 *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

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 exposure of the Chinese rare minnow (*Gobiocypris rarus*). *Toxicol. Appl. Pharm.* **2008,** *226*, (3), 298-308

15. Kidd, K. A.; Blanchfield, P. J.; Mills, K. H.; Palace, V. P.; Evans, R. E.; Lazorchak, J. M.; Flick, R. W., Collapse of a fish population after exposure to a synthetic estrogen. *P. Natl. Acad. Sci. USA* **2007,** *104*, (21), 8897-8901

16. Blanchfield, P. J.; Kidd, K. A.; Docker, M. F.; Palace, V. P.; Park, B. J.; Postma, L. D., Recovery of a wild fish population from whole-lake additions of a synthetic estrogen. *Environ. Sci. Technol.* **2015,** *49*, (5), 3136-3144

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

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

19. Oziolor, E. M.; Bigorgne, E.; Aguilar, L.; Usenko, S.; Matson, C. W., Evolved resistance to PCB- and PAH-induced cardiac teratogenesis, and reduced CYP1A activity in Gulf killifish (*Fundulus grandis*) populations from the Houston Ship Channel, Texas. *Aquat. Toxicol.* **2014,** *150*, (0), 210-219.

20. Williams, L. M.; Oleksiak, M. F., Ecologically and evolutionarily important SNPs identified in natural populations. *Mol. Biol. Evol.* **2011,** *28*, (6), 1817-1826

21. Reid, N. M.; Proestou, D. A.; Clark, B. W.; Warren, W. C.; Colbourne, J. K.; Shaw, J. R.; Karchner, S. I.; Hahn, M. E.; Nacci, D.; Oleksiak, M. F.; Crawford, D. L.; Whitehead, A., The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. *Science* **2016,** *354*, (6317), 1305-1308

631 22. Brazzola, G.; Chèvre, N.; Wedekind, C., Additive genetic variation for tolerance to estrogen
632 pollution in natural populations of Alpine whitefish (*Coregonus* sp., Salmonidae). *Evol Appl* **2014,** *7*,
633 (9), 1084-1093

634 23. Shi, B.; Wen, H. S.; He, F.; Dong, S. L.; Ma, S.; Chen, C. F.; Chen, X. Y.; Zhang, J. R.; Jin,
635 G. X., Single nucleotide polymorphisms within the estrogen receptor beta gene are linked with
636 reproductive indices in Japanese flounder, *Paralichthys olivaceus*. *Comp. Biochem. Phys. B.* **2009,**
637 *154*, (1), 62-67

638 24. Yong , E. L.; Loy, C. J.; Sim, K. S., Androgen receptor gene and male infertility. *Human*
639 *Reproduction Update* **2003,** *9*, (1), 1-7

640 25. Foll, M.; Gaggiotti, Oscar E.; Daub, Josephine T.; Vatsiou, A.; Excoffier, L., Widespread
641 signals of convergent adaptation to high altitude in Asia and America. *Am J Hum Genet* **2014,** *95*, (4),
642 394-407

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

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

649 28. Williams, R. J.; Keller, V. D. J.; Johnson, A. C.; Young, A. R.; Holmes, M. G. R.; Wells, C.;
650 Gross-Sorokin, M.; Benstead, R., A national risk assessment for intersex in fish arising from steroid
651 estrogens. *Environ. Toxicol. Chem.* **2009,** *28*, (1), 220-230

652 29. Jürgens, M. D.; Crosse, J.; Hamilton, P. B.; Johnson, A. C.; Jones, K. C., The long shadow of
653 our chemical past – High DDT concentrations in fish near a former agrochemicals factory in England.
654 *Chemosphere* **2016,** *162*, 333-344

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

31. Grabherr, M. G.; Haas, B. J.; Yassour, M.; Levin, J. Z.; Thompson, D. A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; Chen, Z.; Mauceli, E.; Hacohen, N.; Gnirke, A.; Rhind, N.; di Palma, F.; Birren, B. W.; Nusbaum, C.; Lindblad-Toh, K.; Friedman, N.; Regev, A., Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotech* **2011,** *29*, (7), 644-652

32. Li, H.; Durbin, R., Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **2009,** *25*, (14), 1754-1760

33. Frichot, E.; François, O., LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution* **2015,** *6*, (8), 925-929

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

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

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

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

38. Excoffier, L.; Lischer, H. E. L., Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* **2010,** *10*, (3), 564-567

677 39. Etter, P.; Bassham, S.; Hohenlohe, P.; Johnson, E.; Cresko, W., SNP Discovery and
678 Genotyping for Evolutionary Genetics Using RAD Sequencing. In *Molecular Methods for*
679 *Evolutionary Genetics*, Orgogozo, V.; Rockman, M. V., Eds. Humana Press: 2011; Vol. 772, pp 157-
680 178.

681 40. Catchen, J.; Hohenlohe, P. A.; Bassham, S.; Amores, A.; Cresko, W. A., Stacks: an analysis
682 tool set for population genomics. *Mol. Ecol.* **2013,** *22*, (11), 3124-314

683 41. Catchen, J. M.; Amores, A.; Hohenlohe, P.; Cresko, W.; Postlethwait, J. H., Stacks: building
684 and genotyping loci de novo from short-read sequences. *G3-Genes Genom. Genet.* **2011,** *1*, (3), 171-
685 182

686 42. Fraser, B. A.; Künstner, A.; Reznick, D. N.; Dreyer, C.; Weigel, D., Population genomics of
687 natural and experimental populations of guppies (*Poecilia reticulata*). *Mol. Ecol.* **2015,** *24*, (2), 389-
688 408

689 43. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., Basic local alignment
690 search tool. *J. Mol. Biol.* **1990,** *215*, (3), 403-410

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

699 45. Huang, D. W.; Sherman, B. T.; Lempicki, R. A., Systematic and integrative analysis of large
700 gene lists using DAVID bioinformatics resources. *Nat Protoc* **2009,** *4*, (1), 44-57

46. Gottlieb, B.; Beitel, L. K.; Nadarajah, A.; Paliouras, M.; Trifiro, M., The androgen receptor gene mutations database: 2012 update. *Human Mutation* **2012,** *33*, (5), 887-894

47. Grummer, J. A.; Beheregaray, L. B.; Bernatchez, L.; Hand, B. K.; Luikart, G.; Narum, S. R.; Taylor, E. B., Aquatic landscape genomics and environmental effects on genetic variation. *Trends Ecol. Evol.* **2019,** *34*, (7), 641-654

48. Kuang, Y.-Y.; Zheng, X.-H.; Li, C.-Y.; Li, X.-M.; Cao, D.-C.; Tong, G.-X.; Lv, W.-H.; Xu, W.; Zhou, Y.; Zhang, X.-F.; Sun, Z.-P.; Mahboob, S.; Al-Ghanim, K. A.; Li, J.-T.; Sun, X.-W., The genetic map of goldfish (*Carassius auratus*) provided insights to the divergent genome evolutions in the Cyprinidae family. *Scientific Reports* **2016,** *6*, (1), 34849

49. Wang, Z.; Li, T.; Xing, X.; Gao, X.; Zhang, X.; You, L.; Zhao, H.; Ma, J.; Chen, Z.-J., Replication study of RAD54B and GREB1 polymorphisms and risk of PCOS in Han Chinese. *Reprod. Biomed. Online* **2013,** *27*, (3), 316-321

50. Borg, B., Androgens in teleost fishes. *Comp. Biochem. Physiol. C-Toxicol. Pharmacol.* **1994,** *109*, (3), 219-245

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

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

53. 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

54. 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

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.

56. 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

57. 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

58. 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

59. Agency, U. E. *Upper Lee (Main River) Fisheries Report 2009*; Environment Agency: 2010.

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

61. Williams, L. M.; Oleksiak, M. F., Evolutionary and functional analyses of cytochrome P4501A promoter polymorphisms in natural populations. *Mol. Ecol.* **2011,** *20*, (24), 5236-5247

62. Wirgin, I.; Waldman, J. R., Resistance to contaminants in North American fish populations. *Mutat. Res.-Fundam. Mol. Mech. Mutagen.* **2004,** *552*, (1-2), 73-100

63. Wang, D.; Yang, X.; Shen, B., A novel erythroid differentiation related gene EDRF2 inhibited α-globin gene expression in K562 cells. *Chin. Sci. Bull.* **2002,** *47*, (5), 398-402

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

65. Rallapalli, P. M.; Orengo, C. A.; Studer, R. A.; Perkins, S. J., Positive selection during the evolution of the blood coagulation factors in the context of their disease-causing mutations. *Mol. Biol. Evol.* **2014,** *31*, (11), 3040-3056

66. Nelson, E. R.; DuSell, C. D.; Wang, X.; Howe, M. K.; Evans, G.; Michalek, R. D.; Umetani, M.; Rathmell, J. C.; Khosla, S.; Gesty-Palmer, D.; McDonnell, D. P., The oxysterol, 27-hydroxycholesterol, links cholesterol metabolism to bone homeostasis through its actions on the estrogen and liver x receptors. *Endocrinology* **2011,** *152*, (12), 4691-4705

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

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

69. Martyniuk, C. J.; Bissegger, S.; Langlois, V. S., Current perspectives on the androgen 5 alpha-dihydrotestosterone (DHT) and 5 alpha-reductases in teleost fishes and amphibians. *Gen. Comp. Endocr.* **2013,** *194*, 264-274

70. 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

770 in the concentration of industrially derived surfactants. *Environ. Toxicol. Chem.* **2002,** *21*, (3), 515-
771 519

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

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

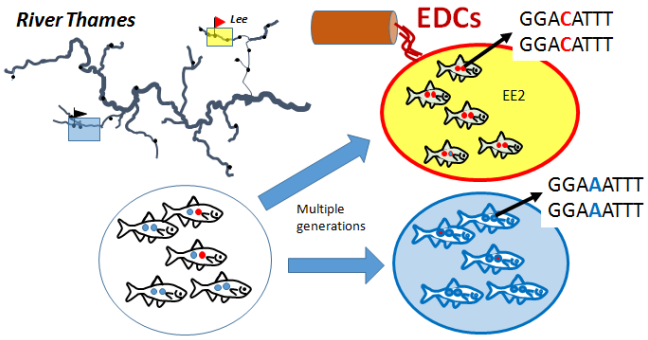
777

778

779
780

781 GRAPHICAL ABSTRACT

782



783

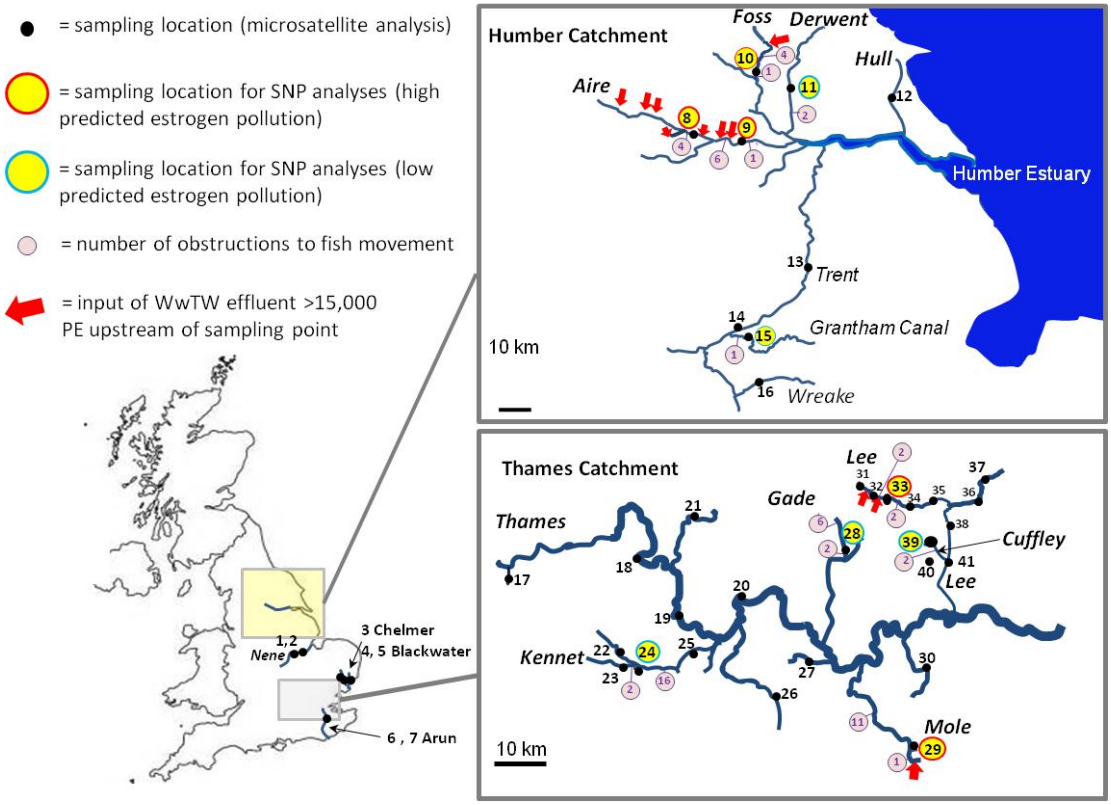
784

785

786

787

788 FIGURES



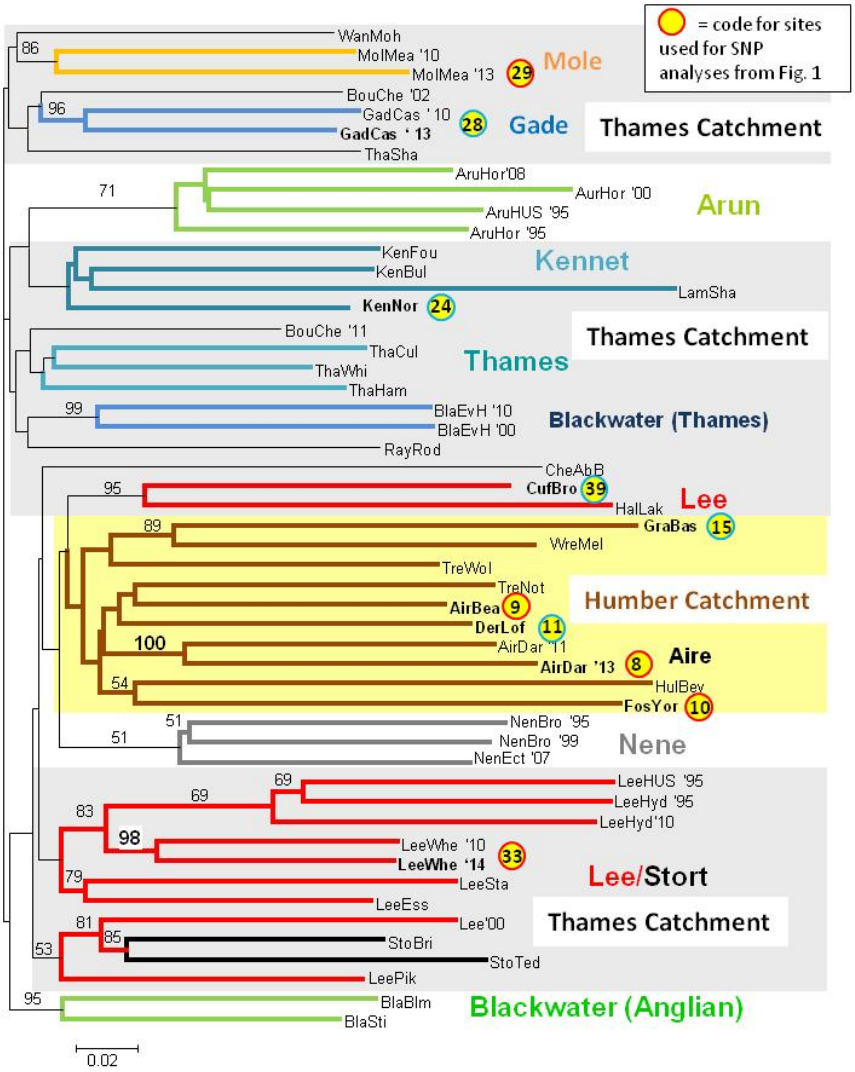
789

790 Figure 1. Locations of river sample sites in England genotyped in this study. Sample
791 codes: 1. NenBro; 2. NenEct; 3. CheAbB; 4. BlaBlM; 5. BlaSti, 6. AruHor; 7. AruHUS; 8.
792 AirDar; 9. AirBea; 10. FosYor; 11. DerLof; 12. HulBev; 13. TreWol; 14. TreNot; 15
793 GraBas; 16 WreMel; 17. RayRod; 18. ThaCul; 19. ThaWhi; 20. ThaHam; 21. ThaSha 22.
794 LamSha 23. KenBul 24. KenNor; 25. KenFou; 26. BlaEvH; 27. BouChe; 28. GadCas; 29.
795 MolMea; 30. WanMoh; 31. LeeHUS; 32. LeeHyd; 33. LeeWhe; 34. LeeSta; 35. LeeEss;
796 36. StoBri; 37. StoTed; 38. Lee'00 (exact location uncertain); 39. CufBro; 40. HalLak; 41.
797 LeePik. Details of newly sampled locations are given in Additional file 1, Table S1. For
798 the River Aire locations there are 9 and 15 WwTWs with a population served greater
799 than 15,000 upstream of AirDar and AirBea respectively. Further details on sample

800 sites and obstructions to fish movement (locks and weirs) in the Thames Catchment

801 are given in the map figure in Hamilton *et al*.[17]

802



803

804

805 **Figure 2.** Neighbor-joining tree for roach population samples produced from data

806 from 2369 roach from 41 sample sites. Several locations were sampled in different

807 years, producing a total of 51 'samples'. The tree is based on the data from 14

808 microsatellite loci using Cavalli-Sforza and Edwards' chord distance measure, D_C[72].

809 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).

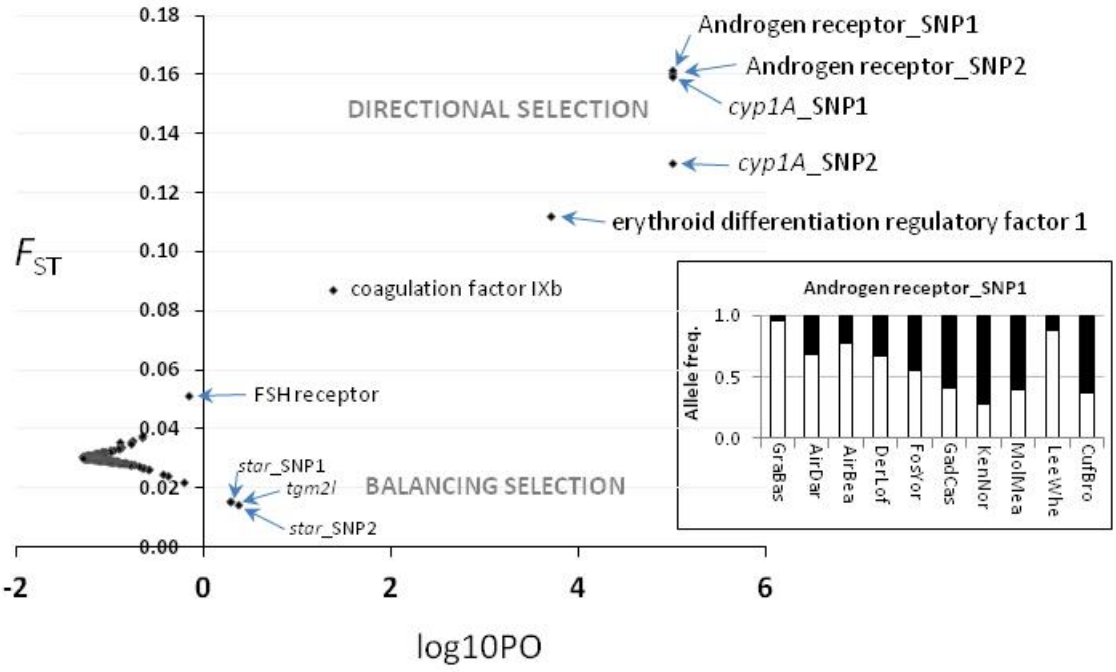


Figure 3. Identification of F_{ST} outlier loci potentially subject to differential selection constructed using data from 217 SNPs loci and 10 sample sites using BayeScan. The x axis represents Log transformed Bayes factors and the y axis represents locus specific F_{ST} from BayeScan. Loci with a posterior 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.