Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/5/7/eaaw7006/DC1

Supplementary Materials for

Stress, novel sex genes, and epigenetic reprogramming orchestrate socially controlled sex change

Erica V. Todd*, Oscar Ortega-Recalde*, Hui Liu, Melissa S. Lamm, Kim M. Rutherford, Hugh Cross, Michael A. Black, Olga Kardailsky, Jennifer A. Marshall Graves, Timothy A. Hore, John R. Godwin, Neil J. Gemmell*

*Corresponding author. Email: ericavtodd@gmail.com (E.V.T.); ojavieror@gmail.com (O.O.-R.); neil.gemmell@otago.ac.nz (N.J.G.)

Published 10 July 2019, *Sci. Adv.* **5**, eaaw7006 (2019)
DOI: 10.1126/sciadv.aaw7006

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/5/7/eaaw7006/DC1)

Supplementary Materials and Methods

Bluehead wrasse draft genome assembly, scaffolding and annotation

To provide a genomic reference for our methylation analyses, we constructed the first genome assembly for the bluehead wrasse. Ovary tissue of a single adult female (sex verified by histological analysis), from which methylome data was also derived, was used to provide DNA for sequencing. Genomic DNA was isolated using a lithium-chloride protocol (*48*), including RNAse digestion, followed by column purification through a DNA Clean & Concentrator kit (Zymo Research). Two TruSeq PCR-free libraries (350 and 550 bp inserts) and two Nextera mate-pair libraries (5 and 8 kb inserts) were constructed and sequenced together on 2 lanes of an Illumina Hi-Seq Rapid V2 (2x250 bp PE) at the Otago Genomics and Bioinformatics Facility at the University of Otago. Sequencing yielded a total of 1.84×10^{11} bases of data: 271.1, 182.4, 139.9 and 143.6 million paired reads from the 350 bp and 550 bp TruSeq, and 5 kb and 6 kb Nextera Mate Pair libraries, respectively. Based on the unassembled data, genome size was estimated at 0.76 Gb using SGA preqc (*64*).

Separate assemblies were performed for each TrueSeq library using the DISCOVAR de novo assembler (*65*) with default parameters. Based on the higher number of input reads and post-assembly statistics, the 350 bp insert assembly was used as a substrate for scaffolding. Prior to scaffolding, reads were trimmed of sequencing adapters and low-quality bases in Trimmomatic v0.35 (*66*), using the parameters 'TRAILING:26' and 'MINLEN:20'.

Scaffolding was performed in SSPACE v3.0 (*67*) using trimmed data from all four sequencing libraries. First, reads from both TruSeq libraries were mapped to the DISCOVAR contigs in

BWA (*68*). Mate Pair libraries were pre-processed with NextClip v1.3.1 (*69*) and all category A, B, and C mapping files were used for scaffolding. Resulting bam files were sorted by sequence name and converted to tab-delimited SSPACE files using the in-built script 'sam_bam2tab.pl'. SSPACE was run several times for parameter optimization, with the following parameters used for final scaffolding: 'minimum number of links to consider read pair (-k)', 3; 'maximum ratio between best two contig pairs (-a)', 0.7; 'minimum overlap between contigs to merge (-n)', 15; and contig extension: 'minimum number of reads needed to call a base during extension (-o)', 10; 'minimum number of overlapping bases during overhang consensus buildup (-m)', 50; 'minimal base ratio to accept overhang consensus base (-r)', 0.8.

GapFiller v1-11 (*70*) was used to close gaps in scaffolds. Three iterations were run, using data from all four libraries and the following parameters: 'minimum number of overlapping bases (-m)', 50; 'minimum number of reads to call a base (-o)', 2; 'minimal base ratio to accept overhang consensus (-r)', 0.7; 'minimum overlap to merge two sequences (-n)', 10; 'number of nucleotides trimmed at sequence edges of the gap (-t)', 10.

Annotation was performed following the Trinotate version 3.1.1 (*71*) and PASA version 2.2 (*72*) pipelines. To first improve gene models from previous transcriptome annotations, and determine gene positions across the genome, transcripts from our published transcriptome assembly for bluehead wrasse (*19*) were aligned to the genome assembly using GMAP version 2018-03-11 (*73*) and BLAT version 3.5 (*74*), and these results were then fed into PASA. The program seqclean version x86_64 (https://sourceforge.net/projects/seqclean/files/) was used to validate the transcript sequences and trim unwanted sequences (e.g., vectors, adaptors, polyA tails).

Using both the cleaned and original transcripts, several rounds of PASA were performed, incorporating the genome alignments, and our previous transcriptome annotations (*19*).

The Trinotate pipeline was used to annotate the mapped assemblies from the PASA output. First, TransDecoder version 5.0.2 (https://github.com/TransDecoder) was used to predict coding regions. Predicted peptide sequences and transcripts were then used as queries to search multiple protein and nucleotide databases. The protein database SwissProt was queried using BLAST (*74*), using *blastp* for peptide sequences and *blastx* for transcripts. Additionally, protein databases of the Zebrafish and Tilapia genomes were searched. The program HMMER 3.1b2 (http://hmmer.org/) was used to identify protein domains in the peptide sequences using the Pfam database. Search results were consolidated into a Trinotate sqlite database to produce an annotation report.

Custom Python and R scripts were used to extract annotations from the Trinotate report and link these to the mapped location of transcripts on the genome (https://github.com/hughcross/bluehead_methylome_bioinformatics). A custom Python script was also used to create separate mapping files (gff3 format) for the three gene references used (sprot, zebrafish, tilapia), using the best match annotation as the gene description. Mapping files were used to visualize gene annotations in SeqMonk.

The scaffolded assembly was more complete and more contiguous than the DISCOVAR assembly and was used in methylation analyses. The scaffolded genome included 379,332 scaffolds, with a scaffold and contig N50 of 15.6 and 12.5 kb, respectively, and total length of

1095.9 Mb. According to a BUSCO analysis (*75*), this assembly is relatively complete (96.7% complete, 1.4% fragmented orthologues), but with a duplication level of 13.2%. Although the large number of scaffolds indicates a fragmented assembly, over 91% of the total genome length is represented within large scaffolds (29,971 over 1 kb in length; 10,270 over 10 kb in length). Therefore, although scaffolds less than 1 kb in length were numerous (349,361; 92.1%) these comprise less than 9% of the genome length and, from initial surveys, largely represent repetitive regions. Furthermore, almost no genes mapped to these small scaffolds. Therefore, only scaffolds larger than 1 kb were used for bisulfite mapping and methylation analyses.

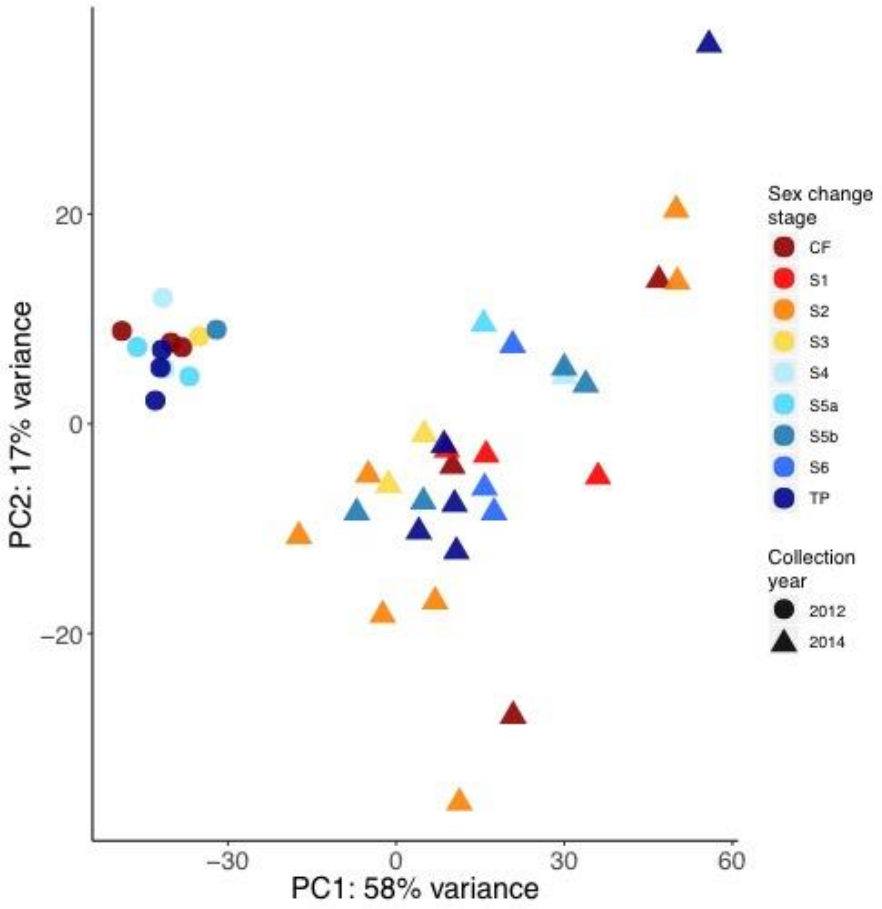

Fig. S1. Lack of sex-biased variation in global gene expression in the bluehead wrasse forebrain. Principal component analysis (PCA) of forebrain samples across sex change (10,000 most variable transcripts) reveals samples do not cluster by sex or transitional stage. Variation across collection years is evident, and samples collected in 2012 and 2014 were sequenced separately (Materials and Methods). CF, control female; S1-6, stage 1 to 6; TP, terminal phase male.

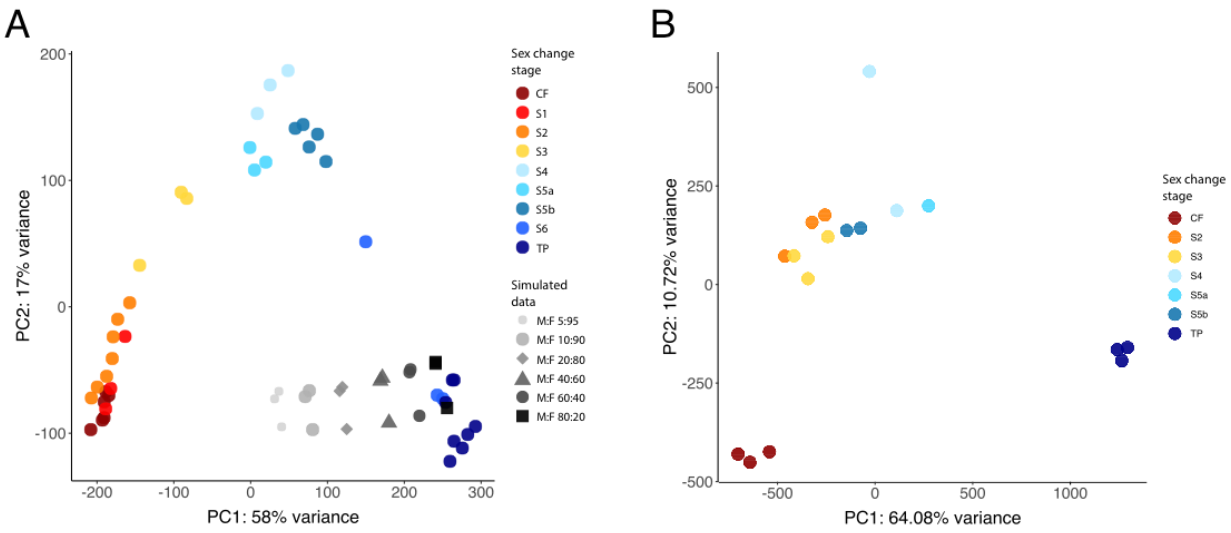

Fig. S2. Intermediate gonads are molecularly distinct from ovary and testis. (A) Principal component analysis (PCA) of gonadal gene expression (10,000 most variable transcripts) across sex change, plus simulated 'mixed' samples (grey) containing female (F) and male (M) reads combined in varying ratios (subsampled from three CF ovary and three TP testis libraries). The transition from ovary to testis is captured along PC1 (58% variance), whereas PC2 (17% variance) delineates fully differentiated gonads of control females (bottom left) and TP males (bottom right) from those of transitioning fish. Distribution of simulated 'mixed' samples along PC1 represents the expected location of intermediate samples should sex change occur via simple proportional change from female to male cells. (B) PCA of gonadal methylomes across sex change. Samples were combined by sex change stage and only probes of 10 kb with more than 100 CpG calls were included. CF, control female; S1-6, stage 1 to 6; TP, terminal phase male.

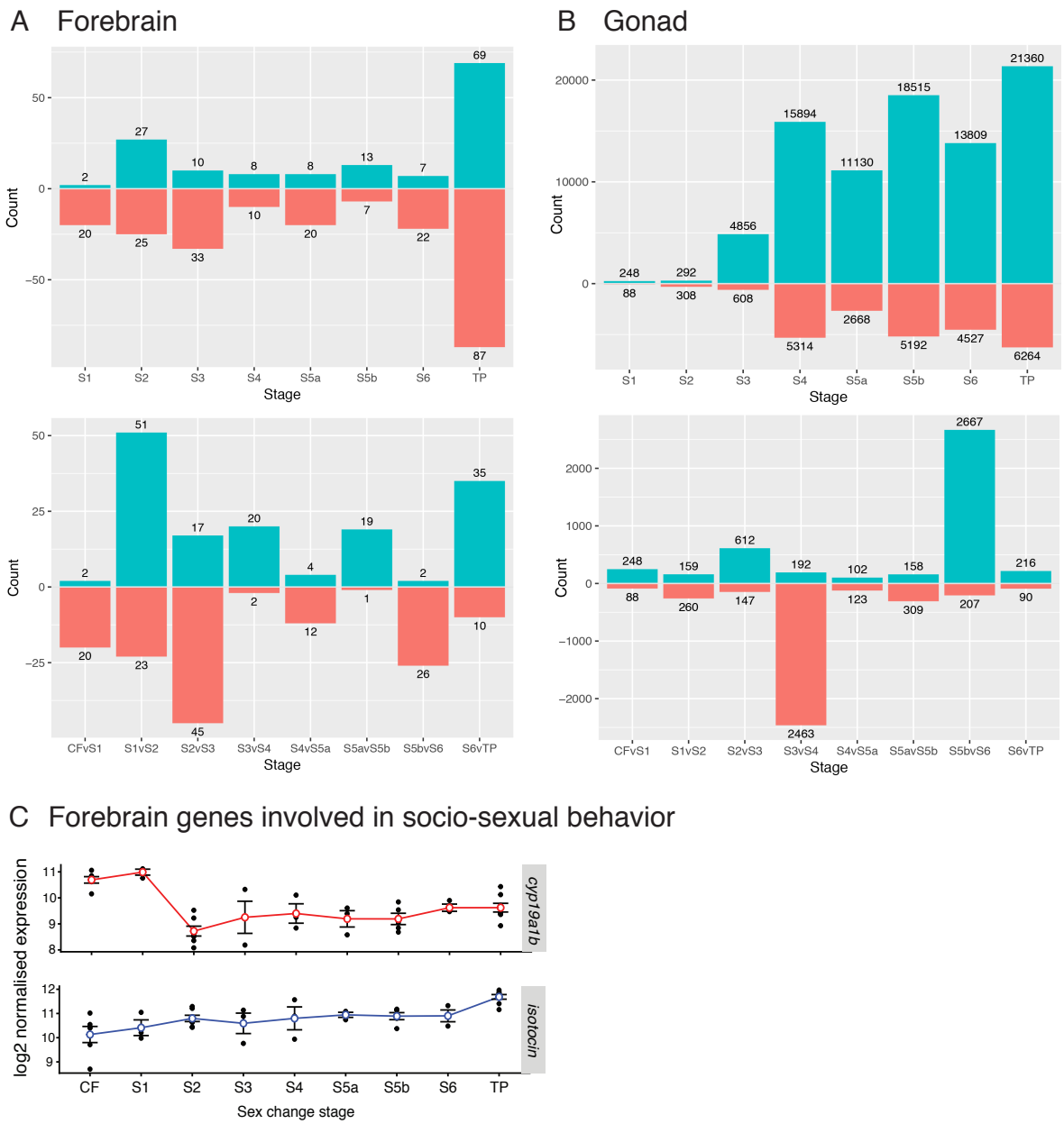

Fig. S3. Differential transcript expression in forebrain and gonad of bluehead wrasses across sex change. For (A) forebrain and (B) gonad, numbers of differentially expressed transcripts are plotted for pairwise comparisons between control females and sex-change stages (top) and between neighboring stages (bottom). Cut-off, gonad: adjusted p-value <0.05, fold-change >2. Cut-off, brain: adjusted p-value <0.05. Up/down-regulation refers to the second stage in each comparison. C) Normalized forebrain expression of selected key genes involved in teleost socio-sexual behavior. Brain aromatase *cyp19a1b* shows female biased expression and is sharply downregulated from stage 2, whereas *isotocin* expression is male biased and increases steadily across sex change. CF, control female; S1-6, stage 1 to 6; TP, terminal phase male.

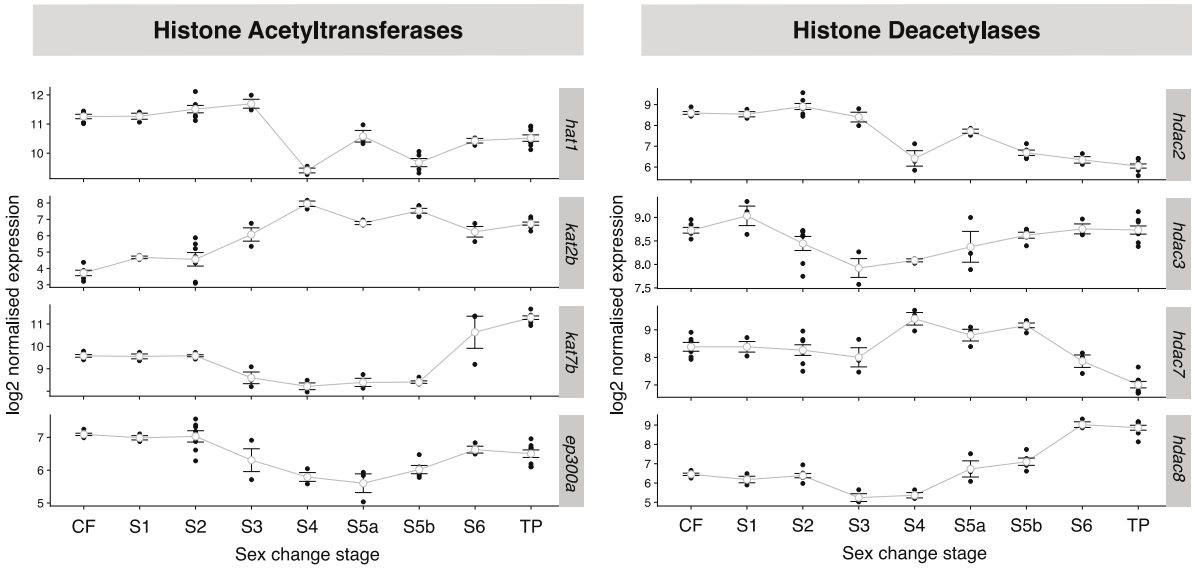

Fig. S4. Writers and erasers of histone acetylation are dynamically expressed across sex change. Normalized gonadal expression of genes encoding histone acetyltransferases (writers) and deacetylases (erasers) across sex change: *hat1* (histone acetyltransferase 1), *kat2b* (K(lysine) acetyltransferase 2b), *kat7* (K(lysine) acetyltransferase 7), *ep300a* (E1A binding protein p300a acetyltransferase), *hdac2* (histone deacetylase 2), *hdac3* (histone deacetylase 3), *hdac7* (histone deacetylase 7), *hdac8* (histone deacetylase 8).

Table S1. Genes enriched in the JAK-STAT signaling pathway are up-regulated across sex change. Genes upregulated in sex-changers (stage S2 to S6) against control females (CF) and enriched in the Jak-STAT signaling pathway. *Indicates significant enrichment (p-value <0.05) in a DAVID functional enrichment analysis.

Comparison	CF v S2	CF v S3*	CF v S4*	CF v S5a*	CF v S5b*	CF v S6
Number of genes	2	20	41	32	37	27
Fold enrichment	na	1.9	1.7	1.7	1.5	1.4
P-value	na	7.60E-03	1.90E-04	2.20E-03	7.80E-03	8.10E-02
Upregulated genes enriched in Jak-STAT signaling pathway	*cish*	*ccnd2a* *cish* *cntfr* *csf2rb* *csf3r* *epor* *il11ra* *il13ra2* *il21r.1* *il2rga* *il6st*	*akt3b* *ccnd2a* *cish* *cntfr* *crfb4* *csf2rb* *csf3r* *epor* *ghra* *grb2b* *ifngr1l* *il10* *il11ra* *il13ra1* *il13ra2* *il21r.1* *il22ra2* *il2rga* *il6* *il6st*	*akt3b* *ccnd2a* *cish* *cntfr* *crfb4* *csf2rb* *csf3r* *epor* *ghra* *il11ra* *il13ra2* *il21r.1* *il2rga* *il6* *il6st*	*akt3b* *ccnd2a* *cish* *cntfr* *crfb4* *csf2rb* *csf3r* *epor* *ghra* *ifngr1l* *il10* *il11ra* *il13ra1* *il13ra2* *il21r.1* *il2rga* *il6st*	*akt3b* *bcl2l1* *ccnd2a* *crfb4* *csf2rb* *csf3r* *epor* *il11ra* *il13ra2* *il21r.1* *il2rga*

			il7r	*il7r*	*il7r*	
			irf9	*irf9*	*irf9*	
		jak1	*jak1*	*jak1*	*jak1*	*jak1*
			jak2b		*jak2b*	
		lifra	*lifra*	*lifra*	*lifra*	*lifra*
		m17	*m17*	*m17*	*m17*	*m17*
						pik3ca
			pik3cd	*pik3cd*	*pik3cd*	*pik3cd*
			pik3cg			*pik3cg*
			pik3r1	*pik3r1*	*pik3r1*	
			pik3r2	*pik3r2*	*pik3r2*	*pik3r2*
		pik3r5				
			pik3r3b	*pik3r3b*	*pik3r3b*	*pik3r3b*
			pik3r5	*pik3r5*	*pik3r5*	*pik3r5*
			pim1	*pim1*	*pim1*	
						pim2
					ptpn11b	*ptpn11b*
		ptpn6	*ptpn6*	*ptpn6*	*ptpn6*	*ptpn6*
		si:dkey-13m1.2	*si:dkey-13m1.2*	*si:dkey-13m1.2*	*si:dkey-13m1.2*	*si:dkey-13m1.2*
		socs1a	*socs1a*	*socs1a*	*socs1a*	*socs1a*
		socs3b	*socs3b*	*socs3b*	*socs3b*	
			socs7		*socs7*	
			stam2			*stam2*
	stat1b	*stat1b*	*stat1b*	*stat1b*	*stat1b*	*stat1b*
			stat4	*stat4*	*stat4*	

Additional data files:

Data S1. GO enrichment detailed results. Gene Ontology (GO) enrichment was performed on transcripts found uniquely assigned to each of four spatial regions identified in the PCA of gonadal gene expression: 'Female', 'Male', 'Transitionary' and 'Differentiated' (Fig. 2B). For each region, enriched GO terms and associated gene lists are provided in separate tabs.

Data S2. Differential expression statistical results. Pairwise comparisons of (A) forebrain of control females against sex change stages, (B) forebrain among neighboring sex change stages, (C) gonad of control females against sex change stages, and (D) gonad among neighboring sex change stages of bluehead wrasse. Statistical results are reported in separate tabs for each pairwise comparison, and for upregulated (up, blue tabs) versus downregulated (down, pink tabs) transcripts. Up/down-regulation refers to the second stage in each comparison. In each tab, transcripts with statistically significant differential expression are colored. Cut-off, gonad: adjusted p-value <0.05, fold-change >2. Cut-off, forebrain: adjusted p-value <0.05. CF, control female; S1-S6, sex change stages 1 through 6; TP, Terminal Phase male. Column headers: contigs, contig name; baseMean, mean normalized count value; log2FoldChange, effect size estimate; lfcSE, log2 fold change standard error; stat, Wald statistic; pvalue, uncorrected p-value; padj, FDR adjusted p-value; Zfish_name, Ensembl zebrafish gene annotation; Zfish_description, Ensembl zebrafish gene description; Sprot_name, SWISS-PROT protein annotation; Sprot_source, SWISS-PROT annotation source species. NA, not available.

Data S3. RNA-seq metadata for bluehead wrasse brain and gonad samples. The table lists the number of trimmed reads and average quality 'Q' value obtained for each RNA-seq library as used in downstream analyses, and the corresponding NCBI BioSample and SRA accession number/s. *RIN values for ovarian samples are not representative of RNA quality as the large quantity of small RNAs interferes with calculations. N/A not calculated. ^Number of read pairs after quality trimming and filtering, used in expression analyses.

Data S4. WGBS metadata for bluehead wrasse gonads. The table lists the number of cytosine calls at either symmetric CG dinucleotides ('CG') or in other sequence contexts ('non-CG'), mapped against the draft bluehead wrasse genome. Number of calls are following deduplication. The frequency of non-CG methylation indicates the maximum rate of non-conversion during the bisulfite treatment step; by this measure, all libraries had a bisulfite conversion efficiency of at least 98.94%. na, not applicable.