Supplemental information

Strong genome-wide divergence between sympatric European river and brook lampreys

Catarina S. Mateus, Madlen Stange, Daniel Berner, Marius Roesti, Bernardo R. Quintella, M. Judite Alves, Pedro R. Almeida, and Walter Salzburger

Supplemental Inventory

Supplemental Figures and Tables

Figure S1. Analysis of genomic divergence between males and females based on sex-specific read coverage across RAD loci in the lamprey *L. planeri* (A) and in threespine stickleback (*Gasterosteus aculeatus*) (B).
The existence of a relatively large genomic region highly differentiated between males and females will cause RAD loci within these regions to show sex-biased read coverage (details in [S1]). In a male-heterogametic system, for instance, read coverage for X-linked loci will be twofold higher in females than males as compared to autosomal loci for which read coverage between the sexes should be equal. The reason is that Y-linked sequences align poorly to their X-counterpart. Exactly this situation is found in stickleback: while most data points lie within the region predicted for autosomal loci (shown as yellow line in the plot), an additional cluster is visible along the line predicted for X-linked loci (green line; the expectation for W-linked loci in a female-heterogametic system is shown as blue line). By contrast, no deviation from the autosomal expectation is evident in *L. planeri*, indicating the absence of physically extensive genomic differentiation between males and females. Hence, if sex determination in this lamprey species is genetically based, the underlying system evolved without major chromosome divergence. Alternatively, sex determination might be under strong environmental influence, as generally assumed to occur in lampreys [S2–S4].

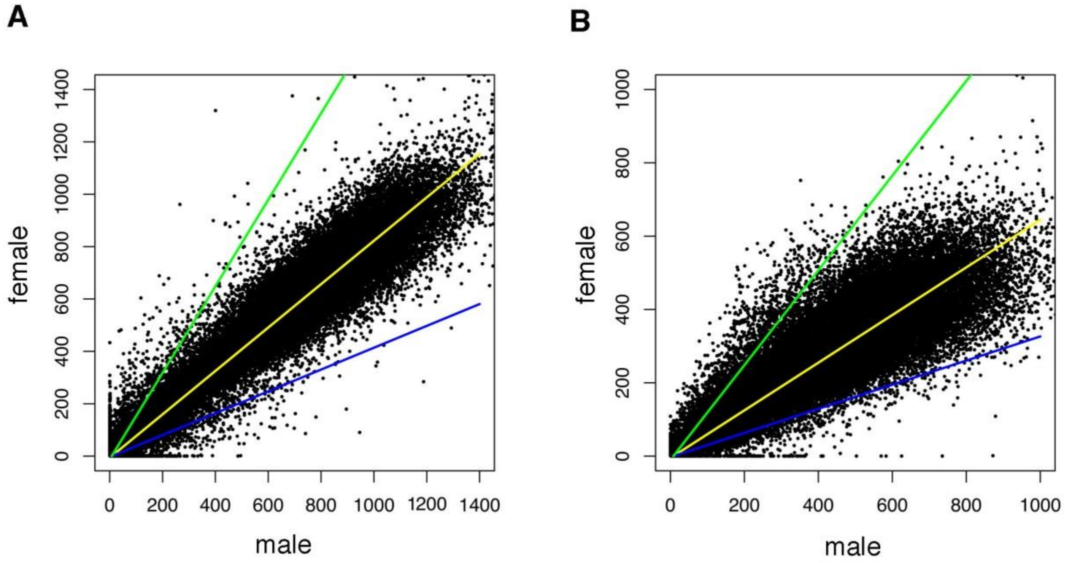

Table S1. Genes and gene families attained after BLAST of the SNPs with F_{ST} = 1.

Gene/Gene family	Function	References
Neurohypophysial gene (vasotocin)	Osmoregulation	[S5-S8]
Gonadotrophin-releasing hormone 2 precursor (GnRH2)	Gonadal maturation and migratory behavior	[S8-S11]
Pineal gland-specific opsin gene (P opsin)	Photoreception	[S12-S17]
Mannose-binding lectin-associated serine protease-1 (MASP-1) gene	Immunity	[S18, S19]
Ikaros-like genes (IKLF2)	Immunity	[S20, S21]
Variable lymphocyte receptor (VLR) gene	Immunity	[S22-S25]
CD45 gene (PTPRC, Protein tyrosine phosphatase, receptor type C)	Immunity	[S26, S27]
Homeobox genes (HoxW10a, Hox7, Emx)	Axial patterning and segmental identity	[S28-S31]
Voltage-gated sodium channel gene	Conduction of electrical signaling in nerves and muscles	[S32, S33]
Protein tyrosine phosphatase receptor type A precursor (PTPRA) gene	Regulation of cellular processes	[S5, S34]

Supplemental Experimental Procedures

Sampling

We collected (by electric fishing) 17 juvenile specimens of the anadromous *L. fluviatilis* at the start of their downstream trophic migration in January of two consecutive years (2009 and 2010), and 18 adult specimens of the resident *L. planeri* during the breeding season between late November 2009 and January 2010. All samples were collected in the Sorraia River, a tributary of the left bank of the Tagus River basin, where both species occur in sympatry. On the Iberian Peninsula, Tagus is the only river where the anadromous *L. fluviatilis* is known to occur, and it represents the southern range limit of both species [S35]. Tissue samples were preserved in 100% ethanol and deposited in the zoological collection 'Museu Bocage' of the Museu Nacional de História Natural e da Ciência (MUHNAC) (Lisbon, Portugal). Sampling was performed under the permission of the Instituto da Conservação da Natureza e das Florestas.

Restriction-site associated DNA (RAD) library preparation

RAD library preparation followed the protocol of Baird *et al*. [S36] and further modifications [S37, S38]. Briefly, DNA was extracted with the 'DNeasy Blood & Tissue Kit' (Qiagen) following the manufacturer's protocol. Genomic DNA from each individual was digested with the *Sbf*1 restriction enzyme. Each digest was then 5-mer barcoded for sample identification, and the 35 total samples were multiplexed into a single library. Final PCR enrichment was performed in 8 separate reactions to reduce amplification bias. Finally, the library was single-end sequenced with 100 cycles in a single Illumina HiSeq 2000 genome analyzer lane at D-BSSE Basel. Illumina reads are available from the Sequence Read Archive (SRA) at NCBI under the accession number PRJNA206554.

Marker generation

The reads were first quality-filtered and demultiplexed according to the individual barcodes. Using sequence data from the one individual with the highest read number, the reads were clustered by tolerating a maximum of two mismatches. For each cluster (representing a RAD locus), the consensus sequence was derived, and the unique consensus sequences were concatenated to form a 3.79 Mb pseudo-reference genome. These steps were carried out using Stacks v0.9996 [S39]. Next, data from each of the 35 individuals were aligned against the pseudo-reference genome using Novoalign v2.08.03 (http://www.novocraft.com), tolerating approximately six high-quality mismatches (-t 'flag' = 180). We enforced unique alignment, thereby avoiding that distinct loci in the pseudo-genome actually derived from the same locus in the true genome because of substantial polymorphisms. The alignments were then converted to *bam* format using Samtools v0.1.18 [S40]. Next, each RAD locus was genotyped at the whole-haplotype level. We here called a homozygous genotype when the dominant haplotype occurred in at least 18 copies and the second most frequent haplotype occurred less than six times. A heterozygote was called when the two most frequent haplotypes occurred in at least 18 copies each. A locus not matching these criteria received a haploid genotype based on the dominant haplotype if that haplotype occurred in at least six copies, or were scored as missing data otherwise. As genotyping used fixed coverage thresholds, loci with excessive read coverage were down-sampled at random to 70x before genotyping (average read coverage per individual and RAD locus was 114.2, sd = 59.8). Finally, we combined the consensus sequences of all individuals to screen each RAD locus for SNPs. To exclude polymorphisms with low information content and technical artifacts [S41], SNPs displaying a minor allele frequency of 0.06 or lower were excluded from the data set. The resulting SNP panel for analysis included 34,267 SNPs. Genotyping and SNP calling was carried out using the R language [S42], benefiting from the bioconductor packages Biostrings and Rsamtools.

Population genetic and phylogenetic analyses

Prior to the analyses of genetic differentiation we eliminated SNPs with insufficient representation across individuals (threshold: 15 nucleotide calls from each population). The SNPs were used to calculate the haplotype-based fixation index (F_{ST}) [see S38] between the two samples. We then used Structure 2.3.4 [S43] to determine the number of genetic clusters (K) in our dataset and to estimate, for each individual, the assignment probability to these clusters. First, structure was run for 100,000 generations, with a burnin of 10,000 generation, and applying the admixture model for $K = 1$ to $K = 5$ and three independent replicates for each K. Using Structure Harvester [S44], we found that the most likely number of K was 2. We then repeated the Structure analysis for $K = 2$, running it for 500,000 generations (Figure 1C) and applying a burnin of 50,000. PAUP* [S45] was used to perform a phylogenetic analysis with the SNP dataset under maximum parsimony applying a heuristic search (stepwise addition and TBR branch swapping and allowing polymorphisms). Confidence assessment was performed with a bootstrap analysis and 1000 replicates. The resulting tree (Figure 1D) had a length of 22,632 steps. We also performed a neighbor-joining tree search (not shown), which produced a highly similar topology.

Screening fixed polymorphisms for candidate genes

For the 166 SNPs fixed for different alleles ($F_{ST} = 1$) between the samples, a homology search was first completed by performing a BLAST [S46] search on the NCBI public database. BLAST hits were then mapped to annotated genes in the Ensembl database [S47] making use of the recently released genome of the sea lamprey (*Petromyzon marinus*) [S48]. The hits were then confirmed by a reciprocal BLAST search, i.e., blasting the respective sea lamprey contig against all RAD tags. In total, we could link twelve RAD loci to annotated genes (Table S1). We found fixed differences in vasotocin, which is involved in many aspects of fish physiology and behavior, including circadian and seasonal biology, metabolism, reproduction and osmoregulation [S5-S8]; in the gonadotropin-releasing hormone 2 (GnRH2), a key gene in gonadal development and differentiation, and regulation of the reproductive and migratory behavior, by controlling secretion of pituitary hormones [S8-S11]; in the non-visual pineal gland-specific opsin gene (P opsin), which is key in photoreception in lamprey larvae, controlling the changes in body coloration and metamorphosis, and in adults through control of sexual maturation [S12-S17]. We found four genes implicated with immune functions: a mannose-binding lectin-associated serine protease (MASP), the ikaros factor-like 2 gene (IKFL2), variable lymphocyte receptor (VLR), and the protein tyrosine phosphatase receptor type C (PTPRC or CD45) [see S18-S27]. We also found hits with three homeobox genes (HoxW10a, Hox7, Emx), which are known to be involved in the specification and patterning of different regions along the body axes [S28-S31]. In particular, Emx is known to play a major role in forebrain development. Hits were also found with the voltage-gated sodium channel gene, known to play an essential role in physiology through the initiation and propagation of action potentials in neurons and other electrically excitable cells such as myocytes and endocrine cells [S32, S33], and finally, in the protein tyrosine phosphatase receptor type A precursor (PTPRA). The protein encoded by PTPRA is a member of the protein tyrosine phosphatase (PTPase) family. PTPases are involved in a variety of cellular processes including cell activation, growth and differentiation, mitotic cycle, and oncogenic transformation [S5, S34].

Genomic screen for large sex-specific regions

We here used a subsample of five females and seven males from the resident species *L. planeri*. This included all lamprey individuals for which sex was known (note that *L. fluviatilis* were sampled as migrating juveniles, precluding the phenotypic identification of sex). The full alignments of these 12 individuals were used to screen visually for the presence of a major sex-linked genomic region (Figure S1). For this, the total number of reads was counted separately across all males and all females at each of the 38,308 total RAD loci. For each locus, the total female count was then plotted against the total male count. The rationale was that RAD loci in sex-specific regions should exhibit systematic read coverage bias between males and females relative to loci in autosomal regions, because of differential alignment success to the reference sequence [for details see S1]. This approach should thus allow detecting at least large-scale differentiation between males and females visually. For comparison, we performed an analogous investigation with exactly the same sample size using RAD data from threespine stickleback [S1], a species with a major XY chromosomal system [S48].

Supplemental References

S1. Roesti, M., Moser, D., and Berner, D. (2013). Recombination in the threespine stickleback genome - patterns and consequences. Mol. Ecol. *22*, 3014-3027.

S2. Beamish, F. W. H. (1993). Environmental sex determination in southern brook lamprey, *Ichthyomyzon gagei*. Can. J. Fish. Aquat. Sci. *50*, 1299–1307.

S3. Lowartz, S. M., and Beamish, F. W. H. (2000). Novel perspectives in sexual lability through gonadal biopsy in larval sea lampreys. J. Fish Biol. *56*, 743–757.

S4. Docker, M. F., and Beamish, F. W. H. (1994). Age, growth, and sex ratio among populations of least brook lamprey, *Lampetra aepyptera*, larvae: an argument for environmental sex determination. Enf. Biol. Fishes *41*, 191–205.

S5. Gwee, P.-C., Tay, B.-H., Brenner, S., and Venkatesh, B. (2009). Characterization of the neurohypophysial hormone gene loci in elephant shark and the Japanese lamprey: origin of the vertebrate neurohypophysial hormone genes. BMC Evol. Biol. *9*, 47.

S6. Balment, R. J., Warne, J. M., Tierney, M., and Hazon, N. (1993). Arginine vasotocin and fish osmoregulation. Fish Physiol. Biochem. *11*, 189–194.

S7. Balment, R. J., Lu, W., Weybourne, E., and Warne, J. M. (2006). Arginine vasotocin a key hormone in fish physiology and behaviour: a review with insights from mammalian models. Gen. Compar. Endocrinol. *147*, 9–16.

S8. Sower, S. A., and Kawauchi, H. (2001). Update: brain and pituitary hormones of lampreys. Comp. Biochem. Physiol. B *129*, 291–302.

S9. Onuma, T., Higa, M., Ando, H., Ban, M., and Urano, A. (2005). Elevation of gene expression for salmon gonadotropin-releasing hormone in discrete brain loci of prespawning chum salmon during upstream migration. J. of Neurobiol. *63*, 126–145.

S10. Gazourian, L., Deragon, K. L., Chase, C. F., Pati, D., Habibi, H. R., and Sower, S. A. (1997). Characteristics of GnRH binding in the gonads and effects of lamprey GnRH-I and -III on reproduction in the adult sea lamprey. Gen. Compar. Endocrinol. *108*, 327–339.

S11. Rissman, E. F. (1996). Behavioral regulation of gonadotropin-releasing hormone. Biol. of Reproduct. *54*, 413–419.

S12. Joss, J. M. P. (1973). Pineal-gonad relationships in the lamprey *Lampetra fluviatilis*. Gen. Compar. Endocrinol. *21*, 118–122.

S13. Joss, J. M. P. (1973). The pineal complex, melatonin, and color change in the lamprey Lampetra. Gen. Compar. Endocrinol. *21*, 188–195.

S14. Cole, W. C., and Youson, J. H. (1981). The effect of pinealectomy, continuous light, and continuous darkness on metamorphosis of anadromous sea lampreys, *Petromyzon marinus* L. J. Exp. Zool. *218*, 397–404.

S15. Pu, G. A., and Dowling, J. E. (1981). Anatomical and physiological characteristics of pineal photoreceptor cell in the larval lamprey, Petromyzon marinus. J. Neurophysiol. *46*, 1018–1038.

S16. Tamotsu, S., and Morita, Y. (1986). Photoreception in pineal organs of larval and adult lampreys, *Lampetra japonica*. J. Comp. Physiol. A *159*, 1–5.

S17. Yokoyama, S., and Zhang, H. (1997). Cloning and characterization of the pineal gland-specific opsin gene of marine lamprey (*Petromyzon marinus*). Gene *202*, 89–93.

S18. Endo, Y., Takahashi, M., Nakao, M., Saiga, H., Sekine, H., Matsushita, M., Nonaka, M., and Fujita, T. (1998). Two lineages of mannose-binding lectin-associated serine protease (MASP) in vertebrates. J. Immunol. *161*, 4924–4930.

S19. Endo, Y., Nonaka, M., Saiga, H., Kakinuma, Y., Matsushita, A., Takahashi, M., Matsushita, M., and Fujita, T. (2003). Origin of mannose-binding lectin-associated serine protease (MASP)-1 and MASP-3 involved in the lectin complement pathway traced back to the invertebrate, amphioxus. J. Immunol. *170*, 4701–4707.

S20. Mayer, W. E., Huigin, C. O., Tichy, H., Terzic, J., and Saraga-Babic, M. (2002). Identification of two Ikaros-like transcription factors in lamprey. Scan. J. Immunol. *55*, 162–170.

S21. Haire, R. N., Miracle, A. L., Rast, J. P., and Litman, G. W. (2000). Members of the Ikaros gene family are present in early representative vertebrates. J. Immunol. *165*, 306–312.

S22. Cooper, M. D., and Alder, M. N. (2006). The evolution of adaptive immune systems. Cell *124*, 815–822.

S23. Pancer, Z., Saha, N. R., Kasamatsu, J., Suzuki, T., Amemiya, C. T., Kasahara, M., and Cooper, M. D. (2005). Variable lymphocyte receptors in hagfish. Proc. Natl. Acad. Sci. USA *102*, 9224–9229.

S24. Alder, M. N., Rogozin, I. B., Iyer, L. M., Glazko, G. V, Cooper, M. D., and Pancer, Z. (2005). Diversity and function of adaptive immune receptors in a jawless vertebrate. Science *310*, 1970–1973.

S25. Boehm, T., McCurley, N., Sutoh, Y., Schorpp, M., Kasahara, M., and Cooper, M. D. (2012). VLR-based adaptive immunity. Ann. Rev. Immunol. *30*, 203–220.

S26. Uinuk-ool, T., Nikolaidis, N., Sato, A., Mayer, W. E., and Klein, J. (2005). Organization, alternative splicing, polymorphism, and phylogenetic position of lamprey CD45 gene. Immunogenetics *57*, 607–617.

S27. Uinuk-ool, T., Mayer, W. E., Sato, A., Dongak, R., Cooper, M. D., and Klein, J. (2002). Lamprey lymphocyte-like cells express homologs of genes involved in immunologically relevant activities of mammalian lymphocytes. Proc. Natl. Acad. Sci. USA *99*, 14356–14361.

S28. Irvine, S. Q., Carr, J. L., Bailey, W. J., Kawasaki, K., Shimizu, N., and Amemiya, C. T. (2002). Genomic analysis of Hox clusters in the sea lamprey *Petromyzon marinus*. J. Exp. Zool. B *294*, 47–62.

S29. Pendleton, J. W., Nagai, B. K., Murtha, M. T., and Ruddle, F. H. (1993). Expansion of the Hox gene family and the evolution of chordates. Proc. Natl. Acad. Sci. USA *90*, 6300–6304.

S30. Boncinelli, E. (1999). Otx and Emx Homeobox Genes in Brain Development. The Neuroscientist *5*, 164–172.

S31. Tank, E. M., Dekker, R. G., Beauchamp, K., Wilson, K. A., Boehmke, A. E., and Langeland, J. A. (2009). Patterns and consequences of vertebrate Emx gene duplications. Evol. Devel. *11*, 343–353.

S32. Yu, F. H., and Catterall, W. A. (2003). Overview of the voltage-gated sodium channel family. Genome Biol. *4*, 207.

S33. Novak, A. E., Jost, M. C., Lu, Y., Taylor, A. D., Zakon, H. H., and Ribera, A. B. (2006). Gene duplications and evolution of vertebrate voltage-gated sodium channels. J. Mol. Evol. *63*, 208–21.

S34. Saadat, M., Nakamura, K., Mizuno, Y., Kikuchi, K., and Yoshida, M. C. (1995). Regional localization of rat and mouse protein-tyrosine phosphatase

PTPα/LRP gene (Ptpra) by fluorescence *in situ* hybridization. Japan. J. Genetics *70*, 669–674.

S35. Mateus, C. S., Rodríguez-Muñoz, R., Quintella, B. R., Alves, M. J., and Almeida, P. R. (2012). Lampreys of the Iberian Peninsula: distribution, population status and conservation. Endangered Species Research *16*, 183–198.

S36. Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Zachary, A., Selker, E. U., Cresko, W. A., and Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE *3*, e3376.

S37. Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., and Cresko, W. A. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genetics *6*, e1000862.

S38. Roesti, M., Hendry, A. P., Salzburger, W., and Berner, D. (2012). Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. Mol. Ecol. *21*, 2852–62.

S39. Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., and Postlethwait, J. H. (2011). Stacks: building and genotyping Loci de novo from short-read sequences. G3 *1*, 171–182.

S40. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) *25*, 2078–2079.

S41. Roesti, M., Salzburger, W., and Berner, D. (2012). Uninformative polymorphisms bias genome scans for signatures of selection. BMC Evol. Biol. *12*, 94.

S42. R Development Core Team (2010). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).

S43. Pritchard, J. K., Stephens, M., and Donelly, P. (2000). Inference of population structure using multicolus genotype data. Genetics *155*, 945-959.

S44. Earl, D. A. and vonHoldt, B, M. (2012). Structure harvester: a website and program for visualizing Structure output and implementing the Evanno method. Cons. Genetic Res. *4*, 359-361.

S45. Swofford, D, L. (2003). PAUP* - Phylogenetic Analyses Using Parsimony and other methods, version 4.0. Sinauer, Sunderland, MA.

S46. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic Local Alignment Search Tool. J. Mol. Biol. *215*, 403–410.

S47. Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., et al. (2013). Ensembl 2013. Nucl. Acids Res. *41*, D48–D55.

S48. Smith, J. J., Kuraku, S., Holt, C., Sauka-Spengler, T., Jiang, N., Campbell, M. S., Yandell, M. D., Manousaki, T., Meyer, A., Bloom, O. E., et al. (2013). Sequencing of the sea lamprey (Petromyzon marinus) genome provides

S49. Peichel, C. L., Ross, J. A., Matson, C. K., Dickson, M., Grimwood, J., Schmutz, J., Myers, R. M., Mori, S., Schluter, D., and Kingsley, D. M. (2004). The master sex-determination locus in threespine sticklebacks is on a nascent Y chromosome. Curr. Biol. *14*, 1416–1424.