The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution

Jeremiah J. Smith, Nataliya Timoshevskaya, Chengxi Ye, Carson Holt, Melissa C. Keinath, Hugo J. Parker, Malcolm E. Cook, Jon E. Hess, Shawn R. Narum, Francesco Lamanna, Henrik Kaessmann, Vladimir A. Timoshevskiy, Courtney K. M. Waterbury, Cody Saraceno, Leanne M. Wiedemann, Sofia M. C. Robb, Carl Baker, Evan E. Eichler, Carin Yee, Tatjana Sauka-Spengler, Mark Yandell, Robb Krumlauf, Greg Elgar, and Chris T. Amemiya.

The sea lamprey (*Petromyzon marinus*) serves as a comparative model for reconstructing vertebrate evolution. To enable more informed analyses, we developed a new assembly of the lamprey germline genome that integrates several complementary data sets. Analysis of this highly contiguous (chromosome-scale) assembly shows that both chromosomal and whole-genome duplications have played significant roles in the evolution of ancestral vertebrate and lamprey genomes, including chromosomes that carry the six lamprey HOX clusters. The assembly also contains several hundred genes that are reproducibly eliminated from somatic cells during early development in lamprey. Comparative analyses show that gnathostome (mouse) homologs of these genes are frequently marked by polycomb repressive complexes (PRCs) in embryonic stem cells, suggesting overlaps in the regulatory logic of somatic DNA elimination and bivalent states that are regulated by early embryonic PRCs. This new assembly will enhance diverse studies that are informed by lampreys’ unique biology and evolutionary/comparative perspective.

The sea lamprey is a member of an ancient lineage that diverged from the vertebrate stem approximately 550 million years ago (MYA). By virtue of this deep evolutionary perspective, lampreys have served as a critical model for understanding the evolution of several conserved and derived features that are relevant to broad fields of biology and biomedicine. Studies have used lampreys to provide perspective on the evolution of developmental pathways that define vertebrate embryogenesis, vertebrate nervous and neuroendocrine systems, genome structure, immunity, clotting, and other features. These studies show aspects of vertebrate biology that have been conserved over deep evolutionary time and identify evolutionary modifications that gave rise to novel features that emerged within the jawed vertebrate lineage (gnathostomes). Lampreys also possess several features that are not observed in gnathostomes, which could represent either aspects of ancestral vertebrate biology that have not been conserved in the gnathostomes or features that arose since the divergence of the ancestral lineages that gave rise to lampreys and gnathostomes. These include the ability to achieve full functional recovery after complete spinal cord transaction, the deployment of evolutionarily independent yet functionally equivalent adaptive immune receptors, and the physical restructuring of the genome during development known as programmed genome rearrangement (PGR).

PGR results in the physical elimination of ~0.5 Gb of DNA from the organism’s ~2.3-Gb genome. The elimination events that mediate PGR are initiated at the 7th embryonic cell division and are essentially complete by 3 days post fertilization. As a result, lampreys are effectively chimeric, with germ cells possessing a full complement of genes and all other cell types possessing a smaller, reproducible fraction of the germline genome. Previous analyses support the idea that the somatic genome lacks several genes that contribute to the development and maintenance of germ cells but are potentially deleterious if misexpressed in somatic lineages. However, our understanding of the mechanisms and consequences of PGR remains incomplete, as only a smaller, reproducible fraction of the genome lacking 0.5 Gb of sequence that is invariably specific to the germline has been sequenced to date.

In contrast to the germline genome, the somatically retained portions of the genome are relatively well characterized. Because it was not known until 2009 that lampreys were subject to PGR, sequencing efforts focused on somatic tissues from which DNA or intact nuclei could be readily obtained (e.g., blood and liver). Sequencing
of the sea lamprey somatic genome followed an approach that had proven successful for other vertebrate genomes before the advent of next-generation sequencing technologies (Sanger sequencing of clone ends, fosmid ends and BAC ends). Because of the abundance of highly identical interspersed repetitive elements and moderately high levels of polymorphism (approaching 1%), assembly of the somatic genome resulted in a consensus sequence that was substantially more fragmentary than other Sanger-based vertebrate assemblies11. Nonetheless, this initial assembly yielded significant improvements in our understanding of the evolution of vertebrate genomes and fundamental aspects of vertebrate neurobiology, immunity and development1–7.

Here we present the first assembly of the sea lamprey germline genome. Through extensive optimization of assembly pipelines, we identified a computational solution that allowed us to generate an assembly from next-generation sequence data (Illumina and Pacific Biosciences reads) that surpasses the existing Sanger-based somatic assembly. Analysis of the resulting assembly identifies several hundred genes that are eliminated from somatic tissues by PGR and sheds new light on the evolution of genes and functional elements in the wake of ancient large-scale duplication events.

Results
Assembly and annotation of the sea lamprey genome. Several shotgun-sequencing and scaffolding data sets were generated in order to permit assembly of the sea lamprey germline genome (>100× sequence coverage in Illumina paired-end reads, >300× physical coverage in 4-kb Illumina mate pairs and >600× physical coverage in 40-kb Illumina mate pairs). Previous analyses demonstrated that the lamprey genome is highly repetitive, and initial analysis of Illumina shotgun sequence data confirmed that the repeat content of lamprey (~60% high-identity repeats) is substantially higher than that of human (Fig. 1). To enable the development of a highly contiguous assembly, we also generated ~17× genome coverage in single-molecule long-read data (Pacific Biosciences XL/C2 chemistry, N50 read length = 5,424) and performed hybrid assembly using DBG2OLC15. This approach yielded an assembly with contiguity statistics (23,286 contigs, N50 = 164,585 bp) that rivaled those of a previously published Sanger-based assembly of the lamprey somatic genome4. To further improve the large-scale structure of this assembly, we integrated scaffolding data (~56× coverage in BioNano optical mapping; >150 kb molecules, and 325 million Chicago (Dovetail) linked read pairs: 2×132 bp), as well as published meiotic mapping data1. Linkages identified through these three independent data sets were cross-validated and integrated using AllMaps (Fig. 2)16. This integrated scaffolding approach allowed us to further increase the contiguity of the assembly (12,077 contigs, N50 = 12 Mb, N60 contig number = 34). In total, 74.8% of the current germline genome assembly is anchored to one of 94 previously defined linkage groups4, and >80% of the assembly is present in super-scaffolds that are 1 Mb or longer. Given that the sea lamprey has 99 pairs of chromosomes in its germline, this integrated assembly appears to approach chromosome-scale contiguity.

Our long-range scaffolding approach used three independent methods that extend and cross-validate one another (Fig. 2), and we consider strong agreement among these three methods as evidence that the large-scale structure of the assembly accurately reflects the structure of P. marinus chromosomes. For many vertebrates, it is possible to independently assess long-range contiguity by measuring conservation of gene orders with closely related species. Highly contiguous assemblies are not yet available for any other jawless vertebrate, although an anchorench draft assembly does exist for the Arctic lamprey (Lethenteron camtschaticum: syn. Lethenteron japonicum)17. To provide perspective on the chromosomal structure of a closely related species, we developed a meiotic map for the Pacific lamprey (Entosphenus tridentatus). The species is a representative of a clade of lampreys (genera Entosphenus, Lethenteron and Lampetra) that diverged from the lineage represented by Petromyzon ~40 MYA18, and embryos of known parentage are available through ongoing hatchery efforts aimed at restoring the species to its native waterways in the US Pacific Northwest19. Meiotic mapping was performed using restriction site–associated DNA (RAD) sequencing of 94 F1 siblings generated from a controlled cross between two wild-captured individuals. The resulting meiotic map provides dense coverage of the genome and represents 83 linkage groups, covering 9,956 cM with an average intermarker distance of 3.4 cM (Supplementary Table 1). Alignment of RAD markers to the sea lamprey genome identified 1,733 homologous sequences, which show strong conservation of synteny and gene order (Fig. 3, Supplementary Table 1). This broad conservation of gene order is considered strong evidence that the sea lamprey assembly and Pacific lamprey meiotic map accurately reflect the chromosomal structure of their respective species.

The repetitive nature of the lamprey genome presents challenges not only to its assembly but also the identification of genes within assembled contigs. This is largely attributable to the interspersion of transposable coding sequences within and among the coding
sequences of low-copy genes. To circumvent these issues, we used a two-tiered approach to gene prediction. Annotation and identification of repetitive elements was performed using RepeatModeler and RepeatMasker\(^23,24\). The entire set of annotated repeats, published gene models and transcriptomic data sets\(^10,13\) were integrated to generate a conservative set of 18,205 gene predictions using MAKER\(^22\). After generating initial gene calls, a second round of gene predictions was generated that permitted extraction of gene models that include low-copy repetitive sequences, yielding another 2,745 gene models, for a total of 20,950 MAKER gene models. In total, Maker was able to assign 18,367 of these gene models to a likely vertebrate homolog on the basis of multispecies BLAST alignments, which included the vast majority of single-copy orthologs expected for lamprey (Supplementary Note)\(^23,24\). An additional 2,583 genes (12%) could not be immediately assigned a homolog on the basis of multispecies alignments. Although these may represent lamprey-specific genes, careful manual curation is likely to be necessary to define their precise evolutionary origins. Such efforts will be enabled through the publicly available genome browser (see URLs). This annotation set was subsequently used to identify the location of 35,382 long noncoding RNA (lncRNA) transcripts in 18,857 lncRNA gene bodies (Supplementary Note, Supplementary Table 1 and Supplementary Fig. 1). These and other annotation sets, including RNA sequencing and genome re-sequencing tracks, are available through SIMRbase (see URLs).

Vertebrate genome evolution. Lamprey occupies a critical phylogenetic position with respect to reconstructing ancestral karyotypes and inferring the timing and mode of duplication events that occurred in ancestral vertebrate and gnathostome lineages. Alignment to the chicken\(^15\) and gar\(^26\) genomes (Supplementary Tables 3–5) permits reconstruction of ancestral orthology groups that are highly consistent with previous reconstructions based on the lamprey meiotic map\(^4\). Because these comparisons require resolution of homologies that are the product of duplication (i.e., 1:1 orthology is not expected) our operational definition of “orthology groups” is expanded to include higher-order relationships (see ref. \(^4\) for more detail). Inclusion of comparative mapping data from the recently published gar genome assembly provides further support for the observation that the majority of ancestral vertebrate chromosomes experienced a single large-scale duplication event in the ancestral vertebrate lineage (Fig. 4, Supplementary Fig. 2). Most ancestral orthology groups correspond to two derived chicken chromosomes (6/11 chicken–lamprey orthology groups identified here). Three other orthology groups possess four derived chromosomes, suggesting that these groups have experienced an additional large-scale duplication: these include well-defined fourfold orthologies that occurred in ancestral vertebrate and gnathostome lineages.

Fig. 2 | Long-range scaffolding and assessment of long-range contiguity of lamprey super-scaffolds. Data from three independent strategies were used to place contigs on larger chromosomal structures. Data from meiotic maps (Smith and Keinath, 2015; blue), Dovetail maps (red) and optical maps (green) complement and extend one another. a, Information used to generate super-scaffold 5. b, Ordering of anchors along super-scaffold 5. c, Information used to generate super-scaffold 21. d, Ordering of anchors along super-scaffold 21. $p =$ Pearson correlation coefficient based on the following numbers of markers; a (top to bottom), $n =$ 18, 28, 14, 10, 34, 156, 78 and 162 independent scaffolding anchors; b (top to bottom), $n =$ 10, 22, 36, 196 and 79 independent scaffolding anchors.

Fig. 3 | Alignment of the Pacific lamprey (E. tridentatus) meiotic map to assembled sea lamprey (P. marinus) super-scaffolds. The relative position of homologous sequences is shown for sea lamprey (y axis) and Pacific lamprey (x axis). A single homologous site (aligning RAD-seq read, Supplementary Table 1) is marked by a single dot. Chromosomes and linkage groups (LGs) are ordered from longest to shortest within species, and individual chromosomes and LGs are highlighted by alternating dark and light shading. Groups of adjacent dots (regions showing conservation of synteny and gene order) appear as diagonal lines.
that chicken chromosome 26 and a portion of chicken chromosome 1 were likely fused in the bony vertebrate (Euteleostome) ancestor approximately 450 MYA and subsequently experienced a derived fission in the chicken lineage. Other deviations from 1:2 or 1:4 are interpreted as the product of derived fission/fusion events that occurred during the first 150 MY following divergence of basal lamprey and gnathostome lineages, derived fission/fusion events in the lamprey lineage, or misassembled regions of the lamprey genome. Although it is possible that the observed genome-wide patterns of conserved synteny could have arisen through two whole-genome duplication events (the 2R hypothesis)\(^27,28\) accompanied by large numbers of chromosome losses\(^29,30\), a previously proposed alternative scenario involving one whole-genome duplication preceded by three distinct chromosome-scale duplication events requires fewer evolutionary steps and is consistent with the data underlying all previous reconstructions\(^1\).

**Lamprey HOX clusters: duplication and divergence.** Historically, descriptions of genome duplications have relied heavily on the HOX gene clusters. This is partially due to their highly conserved organization with respect to gene order and orientation, which contributes to the generation of coordinated patterns of axial expression (collinearity) associated with their roles in embryonic development. Assembly of the Arctic lamprey genome led to the tentative prediction of (at least) six, and possibly eight, HOX clusters, suggesting that the duplication history of at least the lamprey HOX-bearing chromosomes differs from that in the jawed vertebrates\(^17\). We identify 42 HOX genes in the sea lamprey, which all fall within six HOX clusters that are highly similar in content to the HOX clusters predicted in the Arctic lamprey (Fig. 5a, Supplementary Figs. 3 and 4).

In principle, a number of duplication scenarios could potentially explain the existence of six paralogous HOX-bearing chromosomes. These include: (1) whole-genome duplication followed by triplication, or vice versa; (2) a gnathostome-like duplication history (either 2R accompanied by large numbers of chromosome losses\(^29,30\) or one whole-genome duplication preceded by three chromosome-scale duplication events\(^1\)) followed by a further round of whole-genome duplication (yielding eight ancestral HOX clusters) and loss of two entire paralogous chromosomes; (3) a gnathostome-like duplication history followed by duplication of two individual chromosomes. Initial synteny comparisons between lamprey and gnathostome HOX loci showed no clear orthology relationships, but show substantial similarities in the gene content of lamprey HOX-\(\epsilon\) and HOX-\(\beta\) clusters. Notably, phylogenetic analyses of paralogy groups with \(\geq 4\) retained copies (HOX4, HOX8, HOX9, HOX11 and HOX13) also show no clear orthology between lamprey and gnathostome clusters, but they reproducibly place members of HOX-\(\epsilon\) and HOX-\(\beta\) clusters in sister clades with high bootstrap support (Fig. 5b, Supplementary Figs. 5–9). Taken at face value, this would seem to suggest that the \(\epsilon\) and \(\beta\) clusters diverged from one another more recently than other paralogous clusters, apparently lending support to scenario 3. Alternatively, this might also reflect greater functional constraint with respect to the membership of these clusters.

To gain further perspective on the duplication history of lamprey HOX clusters, we extended the analyses to compare the...
chromosome-wide distribution of two-copy paralogs on all HOX-bearing chromosomes. Because post-duplication patterns of conserved synteny are strongly driven by paralog loss, we reasoned that more recent duplication events should yield pairs of chromosomes that share more two-copy duplications, exclusive of all other paralogous chromosomes (the latter of which would have experienced more extensive loss of redundant paralogs over time). Two pairs of chromosomes were observed to share more duplicates relative to all other pairwise combinations of HOX-bearing chromosomes. The strongest enrichment of two-copy paralogs was observed between super-scaffolds 5 and 16 ($\chi^2 = 14.22, P = 1.6 \times 10^{-4}, \text{d.f.} = 1$, Fig. 5, Supplementary Table 6), which carry the HOX-ε and HOX-β clusters. In conjunction with the internal structure of HOX clusters and consistent phylogenetic clustering of ε and β HOX members, we interpret this as indicating that the ε- and β-bearing chromosomes trace their ancestry to a chromosome-scale duplication event that occurred substantially more recently than the genome- and chromosome-scale duplication events that define all other pairwise contrasts, perhaps within the last 200–300 MY. Only one other pair of chromosomes shows significant enrichment of two-copy paralogs relative to all other contrasts. The chromosomes bearing HOX-α and HOX-δ clusters are enriched in shared two-copy paralogs ($\chi^2 = 8.41, P = 3.7 \times 10^{-4}, \text{d.f.} = 1$, Fig. 5, Supplementary Table 6), although α and δ HOX members show no consistent pattern of clustering within gene trees. This difference could be interpreted as indicating that these two chromosomes are the product of a slightly older duplication event, or alternatively it might reflect differential constraints relative to the retention of duplicates by individual pairs of paralogous chromosomes. However, it is unclear what processes might constrain the evolution of one pair of paralogous chromosomes relative to all others.

Programmed genome rearrangement. Identification of eliminated DNA. In lampreys approximately 20% of zygotically inherited DNA is eliminated from somatic cell lineages during early embryogenesis, being retained only by the germline. To identify germline-enriched (i.e., somatically eliminated) regions, we generated whole-genome shotgun sequence data for both sperm (73x coverage) and blood (80x coverage) DNA that were isolated from the same individual. Analysis of read counts identified 1,077 super-scaffolds with enrichment scores (log2[(standardized sperm coverage/blood coverage)]) exceeding 2, over more than 80% of the scaffold (Fig. 6, Supplementary Table 7). These presumptively germline-specific regions cover ~13 Mb of the genome assembly and contain 356 annotated protein coding genes. The distribution of enrichment scores also suggests that other regions with lower enrichment scores are likely to be affected by PGR. To further evaluate our predictions, we designed primers for the 96 longest super-scaffolds with enrichment scores of 2 or higher. In total, primers from 90 (94%) of these scaffolds yielded specific amplification in one or more experiments in C2. Overall, comparisons with ChIP experiments, apparently indicating that that these genes may be marked by bivalent promoters in embryonic stem cells (ESCs) and then presumably released from silencing in germline at later developmental stages. To test this idea, we more closely examined a cluster of genes (denoted GS1) that was highly enriched within C1 ChIP experiments. Notably, all of these genes were previously found to be marked by bivalent (poised) promoters in murine ESCs and primordial germ cells (bivalent in ESCs: 16/16, $\chi^2 = 77.0, P = 8.8 \times 10^{-18}, \text{d.f.} = 1$; bivalent in primordial germ cells (PGCs): 15/16, $\chi^2 = 47.3, P = 3.1 \times 10^{-12}, \text{d.f.} = 1$). A second cluster of eliminated genes (denoted GS51) also showed strong enrichment for these two functional categories (bivalent in ESCs: 14/22, $\chi^2 = 34.6, P = 2.0 \times 10^{-4}, \text{d.f.} = 1$; bivalent in PGCs: 14/22, $\chi^2 = 23.2, P = 7.5 \times 10^{-7}, \text{d.f.} = 1$).

Other enriched ChIP experiments (C2) correspond primarily to the binding targets of transcriptional modifiers in embryonic stem cells ($N = 7$), embryonic progenitor lineages ($N = 7$) and transcriptional activators in cancer ($N = 15$; Fig. 7). Notably, all but one (PCDHGB5) of the genes detected in C1 are present in one or more experiments in C2. Overall, comparisons with ChIP analyses performed in non-eliminating species lends further support to the idea that PGR acts to prevent misexpression of ‘germline’ genes and suggests that misexpression of orthologous genes may directly contribute to oncogenesis in a diverse range of cancers. Moreover, these comparative analyses provide new insight into the regulatory functions of PGR by finding overlap between early gene-silencing events that are achieved by PGR and those that are mediated by the PRC during differentiation of germline and soma.

Discussion
The lamprey genome presents an interesting target for sequencing because of its phylogenetic position and unique genome biology, yet a particularly challenging target given its high repeat content and divergence from other species with highly contiguous assemblies. In an attempt to resolve this complexity, we leveraged several complementary technologies to generate a highly contiguous assembly that approaches the scale of entire chromosomes. Moreover, we were
able to validate the chromosome-scale contiguity of our assembly by generating a dense meiotic map for a related species. The high contiguity of our assembly provides critical context for understanding the evolution of gene content and genome structure in vertebrates. Here we highlighted the utility of this assembly in addressing fundamental questions related to understanding changes in large-scale structure of vertebrate genomes, specifically reconstructing the deep evolutionary origins of vertebrate chromosomes and understanding how PGR mediates genetic conflicts between germline and somatic tissues.

Our improved assembly permits robust resolution of a complement of ancestral chromosomes that existed before the divergence of ancestral gnathostome and agnathan lineages and prior to whole-genome duplication(s) within the shared ancestral lineage of all extant vertebrates. These reconstructions largely validate previous analyses that were performed using meiotic mapping data, but they provide improved resolution of ancestral homology groups. Analyses also lend further support to the idea that chromosome-scale duplication events may have been more common over the course of vertebrate ancestry than has been appreciated from the

---

**Fig. 5 | Structure and evolution of HOX clusters.** a, Six HOX clusters can be identified within the sea lamprey genome assembly. Lamprey cluster designations α through ζ follow the convention of Mehta et al. HOX genes are represented as boxes, with the direction of their transcription indicated by the black arrow. Flanking non-HOX genes are depicted as arrowheads, which indicate their direction of transcription. The positions of known microRNAs are indicated. The four human HOX loci and the inferred ancestral vertebrate HOX locus are shown for comparison. The white arrow downstream of the black arrow. Flanking non-HOX genes are depicted as arrowheads, which indicate their direction of transcription. The positions of known microRNAs are indicated. The four human HOX loci and the inferred ancestral vertebrate HOX locus are shown for comparison. The white arrow downstream of the black arrow. Flanking non-HOX genes are depicted as arrowheads, which indicate their direction of transcription. The positions of known microRNAs are indicated. The four human HOX loci and the inferred ancestral vertebrate HOX locus are shown for comparison.
analysis of bony vertebrate genomes. Parallel lines of evidence supporting a relatively recent duplication having given rise to lamprey HOX-ε- and HOX-β-bearing chromosomes further highlights the potential for large-scale duplication outside the context of whole-genome duplication. It appears that two features of lamprey biology might favor the fixation of chromosomal duplications. First, lampreys possess a large number of small chromosomes, and consequently chromosomal duplications will generally impact fewer genes than similar events in human. Duplication events (in addition to a single presumptive whole-genome duplication) appear to have affected other groups of lamprey chromosomes, though not all (Supplementary Fig. 11). Second, individuals are highly fecund (~100,000 eggs per female), and therefore a single mutant can introduce thousands of carriers (including stable carriers) into a population13–15. While it is likely that the reproductive biology and distribution of chromosome sizes has fluctuated over the course of vertebrate evolution, available evidence suggests that lampreys have possessed similar karyotypes and reproductive biologies for hundreds of millions of years. As such, extant lampreys may represent a better model for conceptualizing phases of evolution during which ancestral vertebrates were characterized by higher fecundity and larger numbers of relatively gene-poor microchromosomes, in addition to providing phylogenetic perspective on early stages of vertebrate genome evolution.

The assembly also identifies a large number of genes that are reproducibly eliminated via PGR. Enrichment analyses reveal a strong overlap in the targets of PGR-mediated elimination and the targets of silencing via PRC proteins in embryonic stem cells. The PRC is a deeply conserved complex that plays roles in gene silencing related to the maintenance of stem cell identity, silencing of oncogene expression and X-chromosome inactivation, among other functions18,39. These well-defined functions of the PRC mirror several aspects of PGR, particularly in that both act to achieve strong transcriptional silencing and both appear to target an overlapping subset of proto-oncogenes. It is interesting to speculate that the overlapping targets of PGR and the PRC may indicate that these two modes of silencing share common underlying mechanisms.

However, it is notable that PRC repression is strongly associated with the deposition and binding to trimethylated lysine 27 of histone H3 (H3K27me3), whereas previous studies have shown that this mark is absent prior to the onset of PGR in lamprey embryos41. It therefore appears that PGR acts to (in part) regulate a subset of germline-expressed targets of the PRC and that it may work upstream of the PRC in lamprey embryos.

The analyses presented here address a focused set of topics that are specifically related to understanding the evolution and development of genome structure in lamprey and other vertebrates. We anticipate that this assembly will substantially improve our ability to use lamprey as a comparative evolutionary model. Because sequences are anchored to their broader chromosomal structure, the current assembly should enhance the ability to reconstruct the deep evolutionary history of the vast majority of genes within vertebrate genomes and perform robust tests of hypotheses related to historical patterns of duplication and divergence. Moreover, the availability of a highly contiguous assembly for an agnathan species should aid in the development and analysis of other genome assemblies from this highly informative vertebrate lineage.

**URLs.** SIMRbase/Lamprey Genome Browser, https://genomes.stowers.org/organism/Petromyzon/marinus; DilCover, https://github.com/tinmat/DilCover; RepeatMasker, http://www.repeatmasker.org. Original data pertaining to the Chicago assembly (Dovetail) and HOX cluster curation can be accessed from the Stowers Original Data Repository at http://www.stowers.org/research/publications/LIBPB-1215.
Methods
Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-017-0036-1.

Received: 14 August 2017; Accepted: 15 December 2017; Published online: 22 January 2018

References
1. Parker, H. J., Bronner, M. E. & Krumlauf, R. A Hox regulatory network of hindbrain segmentation is conserved to the base of vertebrates. Nature 514, 490–493 (2014).
2. Green, S. A., Simone-Costa, M. & Bronner, M. E. Evolution of vertebrates as viewed from the crest. Nature 520, 474–482 (2015).
3. Sower, S. A. et al. Emergence of an ancestral glycoprotein hormone in the pituitary of the sea lamprey, a basal vertebrate. Endocrinology 156, 3026–3037 (2015).
4. Smith, J. J. & Keinan, M. C. The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications. Genome Res. 25, 1081–1090 (2015).
5. Das, S. et al. Evolution of two prototypic T cell lineages. Cell. Immunol. 296, 87–93 (2015).
6. Doolittle, R. F. Bioinformatic characterization of genes and proteins involved in blood clotting in lampreys. J. Mol. Evol. 81, 121–130 (2015).
7. McCauley, D. W., Dobson, M. F., Whyard, S. & Li, W. Lampeyes as diverse model organisms in the genomics era. Bioscience 65, 1046–1056 (2015).
8. Smith, J. J., Antonacci, F., Eicher, E. E. & Amemiya, C. T. Programmed loss of millions of base pairs from a vertebrate genome. Proc. Natl. Acad. Sci. USA 106, 11212–11217 (2009).
9. Smith, J. J., Stuart, A. B., Sauka-Spengler, T., Clifton, S. W. & Amemiya, C. T. Development and analysis of a germine BAC resource for the sea lamprey, a vertebrate that undergoes substantial chromatin diminution. Chromosoma 119, 381–389 (2010).
10. Smith, J. J., Baker, C., Eicher, E. E. & Amemiya, C. T. Genetic consequences of programmed genome rearrangement. Curr. Biol. 22, 1524–1529 (2012).
11. Timosheskiv, V. A., Herdy, J. R., Keinan, M. C. & Smith, J. J. Cellular and molecular features of developmentally programmed genome rearrangement in a vertebrate (sea lamprey: Petromyzon marinus). PLoS Genet. 12, e1006103 (2016).
12. Bryant, S. A., Herdy, J. R., Amemiya, C. T. & Smith, J. J. Characterization of somatically-eliminated genes during development: lamprey (Petromyzon marinus). Mol. Biol. Evol. 33, 2337–2344 (2016).
13. Smith, J. J. et al. Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. Nat. Genet. 45, 415–421 (2013).
14. Seip, M. L. et al. The UCSC Genome Browser database: 2016 update. Nucleic Acids Res. 44 D1, D717–D725 (2016).
15. Ye, C., Hill, C. M., Wu, S., Ruan, J. & Ma, Z. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. Sci. Rep. 6, 31900 (2016).
16. Tang, H. et al. ALLMAPS: robust scaffold ordering based on multiple maps. Genome Biol. 16, 3 (2015).
17. Melba, T. K. et al. Evidence for at least six Hox clusters in the Japanese lamprey (Lethenteron japonicum). Proc. Natl. Acad. Sci. USA 110, 16044–16049 (2013).
18. Kuraku, S. & Kuratani, S. Time scale for cyolocyte evolution inferred with a phylogenetic diagnosis of hagfish and lamprey CENF sequences. Zool. Sci. 23, 1053–1064 (2006).
19. Lampman, R. et al. In Jawless Fishes of the World Vol. 2 (eds. A. Orlov & R. Beamish) Ch. 22, 160–195 (Cambridge Scholars Publishing, Newcastle-upon-Tyne, UK, 2016).
20. Smith, A. F. A. & Hubley, R. RepeatModeler-1.0. (2015).
21. Smith, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0 (2015).
22. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12, 491 (2011).
23. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212 (2015).
24. Hera, Y. et al. Optimizing and benchmarking de novo transcriptome sequencing: from library preparation to assembly evaluation. BMC Genomics 16, 977 (2015).
25. Warren, W. C. et al. A new chicken genome assembly provides insights into avian genome structure. G3 (Bethesda) 7, 109–117 (2017).
26. Braschler, L. et al. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nat. Genet. 48, 427–437 (2016).
27. Ohno, S. Evolution by Gene Duplication (Springer-Verlag, New York, 1970).

Acknowledgements
Research reported in this publication was supported by the National Institute of General Medical Sciences of the US National Institutes of Health under award number R01GM104123 to J.J.S., the Stowers Institute under award number SIMR 1001 to H.J.P., M.E.C., L.M.W., S.M.C.R. and R.K., and the Bonneville Power Administration to J.E.H. and S.R.N. E.E.F. is an investigator of the Howard Hughes Medical Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Support and resources from the Center for High Performance Computing at the University of Utah are gratefully acknowledged. Additional computational support was provided by The University of Kentucky High Performance Computing complex.

Author contributions
J.J.S., R.K., C.T.A. and G.E. conceived of the study. J.J.S., N.T., C.Y., C.H., M.C.K., H.J.P., M.E.C., J.E.H., S.R.N., V.A.T., C.K.M.W., C.S., H.K., F.L., L.M.W., S.M.C.R., C.B., E.E., D.H., T.S.-S., M.Y. and R.K. contributed analyses. J.J.S., N.T., M.C.K., H.J.P. and R.K. wrote the manuscript.

Competing interests
E.E.F. is on the scientific advisory board (SAB) of DNAExpress, Inc.

Additional information
Supplementary information accompanies this paper at https://doi.org/10.1038/s41588-017-0036-1.
Reprints and permissions information is available at www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.J.S.
Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/
Methods

Research animals. This study complied with all relevant ethical guidelines and was performed under protocol number 2011-0848 (University of Kentucky Institutional Animal Care and Use Committee).

Sequencing. Fragment libraries were prepared by Covaris shearing of sperm genomic DNA isolated from a single individual and size selected to achieve average insert sizes of ~205 and 231 bp. These libraries were sequenced on the Illumina HiSeq2000 platform. Two separate 4-kb mate pair libraries were generated. One 4-kb library was prepared and sequenced by the Genomic Services Laboratory at HudsonAlpha (Huntsville, AL) and another was prepared and sequenced using the standard Illumina mate-pair kit. Two 4-kb libraries were prepared and sequenced by Lucigen (Middleton, WI). Long reads were prepared by the University of Florida Interdisciplinary Center for Biotechnology Research (Gainesville, FL) and sequenced using Pacific Biosciences (Menlo Park, CA) XL/C2 chemistry on a Single Molecule, Real-Time (SMRT) Sequencing platform.

Hybrid assembly. Hybrid assembly of Illumina fragment reads and Pacific Biosciences single-molecule reads was performed using the programs SparseAssembler2 and DBG2OLC.1 The first 159 Gb of the high-quality paired-end reads were used to construct a short but accurate de Bruijn graph containing the program SparseAssembler with k-mer size 51 and a skip length of 15. The program DBG2OLC was then used to map short contigs to PacBio SMRT sequencing reads and generate a hybrid assembly. Each PacBio read was compressed using high-quality short-read contigs and aligned to all other reads for structural scaffolding wherein chimeric reads are split and trimmed. A read-overlap-based assembly graph was generated and unbranched linear regions of the graph were output as the initial assembly backbones. Consensus sequences for the backbones were generated by joining overlapped raw sequencing reads and short-read contigs. In practice, many regions of the initial consensus sequences can be erroneous due to the high error rates of the PacBio reads. In order to polish each backbone, all related PacBio reads and contigs are first collected and realigned using Sparc43 to calculate the most likely consensus sequence for the genome.

Scaffolding. Scaffolding of the hybrid assembly was performed using SSPACE2.0 to incorporate mate pair data, followed by ALLMAPS version 0.5.3 to incorporate optical mapping (BioNano), linked-read (Dovetail) and previously published meiotic mapping data. Scaffolding by SSPACE imposed a stringent scaffolding threshold requiring 5 or more consistent linkages to support scaffolding of any pair of contigs. Scaffolding via ALLMAPS was implemented with default parameters and with equal weights assigned to all three types of mapping data with initial anchoring to meiotic maps. For scaffolds without linkage mapping data, additional ALLMAPS runs were performed using the remaining data sets. Conflicts among the three mapping methods were resolved by majority rule or by manually breaking contigs that could not be placed by majority rule.

Meiotic mapping of E. tridentatus. A meiotic map was generated for E. tridentatus using a single outbred adult pair collected from Willamette Falls (Oregon City, OR, USA) and from which larvae were artificially propagated in May 2013 at the USGS Animal Care and Use Committee). This study complied with all relevant ethical guidelines and was performed under protocol number 2011-0848 (University of Kentucky Institutional Animal Care and Use Committee).

Identification of coding sequences. Genome annotations were produced using the MAKER2 genome annotation pipeline, which supports re-annotation using pre-existing gene models as input. Previous Petrozomyc marinus gene models (WUGSC.7.0/petMar2 assembly)39 were mapped against the new genome assembly into GFF3 format and were used as prior model input to MAKER for re-annotation. Snap39 and Augustus50 were also used with MAKER and were trained using the pre-existing lamprey gene models. Additional input to MAKER included previously published mRNA-seq reads derived from lamprey embryos and testes16,31 and assembled using Trinity32, as well as mRNA-seq reads (HiSeq 75–100bp paired-end) that were derived from whole embryos and dissected heads at Tahara stage 20 and dissected embryonic dorsal neural tubes at Tahara stage 18, 20 and 21. The following protein data sets were also used: Ciona intestinalis (sea squirt)33, Lottia gigantea (limpet)34, Nematosatella vectensis (sea anemone)35, Takifugu rubripes (pufferfish)36, Branchiostoma floridae (lancelet)37, Carcharhinus melanopterus (fruit fly)38, Homo sapiens (human)39, Mus musculus (mouse)40, Danio rerio (zebrafish)41, Hydra magnipapillata42, Trichoplax adhaerens43, and the Uniprot/Swiss-Prot protein database44. Protein domains were identified in final gene models using the InterProScan domain identification pipeline51,52, and putative gene functions were assigned using BLASTP53 identified homology to the Uniprot/Swiss-Prot protein database.

Intron RNA annotation. Putative IncRNAs were predicted from RNA-seq reads obtained from brain, heart, kidney, and ovary/testis sampled from two ripe adult individuals (one female, one male). In total, 8 libraries were produced using the Illumina stranded TruSeq mRNA kit (Illumina Inc.). Sequencing (single-end, directional 100 bp) was performed on a HiSeq 2000. The resulting reads were mapped to the genome assembly using GSNAP (v2017-04-24)34, the resulting bam files were then assembled into transcript models using StringTie (v1.3.6b3). The following parameters were optimized in order to maximize the number of correctly predicted IncRNAs and reduce the number of assembly artifacts: (1) minimum isoform abundance of the predicted transcripts as a fraction of the most abundant transcript assembled at a given locus: lower-abundance transcripts are often artifacts of incompletely spliced precursor of processed transcripts; (2) minimum read coverage allowed for the predicted transcripts; (3) minimum locus gap separation value: reads that are mapped closer than 10 bp distance are merged together in the same processing bundle; (4) smallest anchor length: junctions that do not have spliced reads that align across them with at least 10 bases on both sides are filtered out; (5) minimum length allowed for the predicted transcripts (200 bp); (6) minimum number of spliced reads that align across a junction (i.e. junction coverage); (7) removal of monoexonic transcripts. The resulting transcriptomes from each library were then merged into a single GTF file (–merge option in StringTie). Transcript overlapping (in sense) exons of the protein coding annotated genes were removed using the script FEELnc_filter.pl54. The filtered gene models file was then used to compute the Coding Potential Score (CPS) for each of the candidate non-coding transcript with the script FEELnc_codpot_pl54. In FEELnc two reads with the same adapters were considered a species-specific IncRNA set, as is the case for P. marinus, the implemented machine-learning strategy requires to simulate non-coding RNA sequences to train the model by shuffling the set of mRNAs while preserving their 7-mer frequencies. This approach is based on the hypothesis that at least some IncRNAs are derived from "debris" of protein-coding genes. The simulated data were then used to calculate the CPS cutoff separating coding (mRNAs) from non-coding (IncRNAs) using 10-fold cross-validation on the input training files in order to extract the CPS that maximizes both sensitivity and specificity.

Analysis of conserved synteny. Analyses of conserved synteny were performed as previously described.18 Briefly, predicted protein sequences from the lamprey genome were aligned to proteins from the Gar (LepOcu1: GCA_000242695.1) and Chicken (Gallgal4: GCA_000002315.2) genome assemblies45. All alignments with bitscore ≥100 and ≥90% of the best match (within a species) were considered putative orthologs of each lamprey, chicken or gar gene. Groups of orthologs were removed to those with more than 6 members in any given group. Enrichment of orthologs on chromosomes or chromosomal segments was assessed using χ² tests, incorporating Yates’ correction for continuity and Bonferroni corrections for multiple testing as previously described.18

Identification and characterization of germline-specific/enriched sequences. Single-copy genes. To identify germline-specific regions, we separately aligned paired-end reads from blood and sperm DNA to the genome assembly using BWA-MEM (v0.7.10)55 with default parameters and filtered to exclude unmapped reads and supplementary alignments (samtools v.1.2 with option: -v -F23080). Initial coverage analyses was implemented using bedtools v2.23.056 and revealed that the modal coverage of reads from sperm DNA was slightly lower than the coverage of reads from blood, ~73K and ~80K, respectively, but contained a larger amount of low-copy DNA (Supplementary Fig. 12). To identify germine-enriched intervals, data were filtered to remove regions with coverage both from sperm and blood of <10 (underrepresented regions: computed with genomcov-bga, bedtools v2.23.0) and also regions with coverage exceeding three
times the modal value in sperm or blood (high-copy regions). The remaining data were processed to generate coverage ratios for discrete intervals containing 1,000 bp (or >500 bp at contig ends) of approximately single-copy sequence. Identification of contiguous intervals and re-estimation of coverage ratios was performed using DNAc copy version 1.42.0 after removing trailing windows that were <500 bp in length. Ontology analyses used naming assignments that were generated using multispecies BLAST alignments via MAKER2/Ago and were performed using Enrichr8.

Repetitive sequences. High-identity repetitive elements were assembled de novo from k-mers (k = 31) that were abundant in sperm and blood reads, with k-mer counting via Jellyfish version 2.2.4 and assembly using Velvet version 1.2.10. Copy-number thresholds for abundant k-mers set at 3x modal copy numbers for 31-mers: 165 for sperm and 180 for blood. Abundant k-mers from sperm and blood were combined and assembled as a single simple-end read for Velvet running with 29mers. These analyses resulted in a de novo repeat library with 130,632 sequences (overall length ~11 Mb with individual contigs lengths range from 57 bases to 15.5kb). These repeats were annotated using RepeatMasker version open-4.0.51 (see URLs) and repeat libraries generated for the germline assembly and from RepeatMasker repeats (2014/03/11).

For downstream analyses we used a set of model repeats representing the union of de novo repeats, those identified within assembled genomic sequences via RepeatModeler and an updated assembly of the previously identified Germl1 element1. Enrichment analyses were performed by separately aligning paired-end reads from blood and sperm DNA to the repeat data set. As with single copy sequence, alignments were pre-filtered to exclude unmapped reads and supplementary alignments. The remaining data were processed to generate average coverage ratios for intervals of ~100bp.

Manual curation of HOX clusters. Manual curation of gene models was carried out using Agp2 implemented in Integrative7. Indels in the assembly were identified and corrected by comparison with RNA-seq and genomic DNA re-sequencing data. Gene predictions from Maker were refined based on whole-embryo RNA-seq data from multiple developmental stages and homology with gene sequences from other vertebrates. In addition to the 42 clustered HOX genes in the genome assembly, 6 further HOX genes were predicted that did not fall within the 6 HOX clusters. To investigate these genes further, the genomic scaffolds harboring these gene loci were extracted and used as queries for alignment against the assembly by BLASTN. Five of these gene loci (homologs of hoxA3, D8, C9, B13 and B13a) were found to align with high sequence similarity (≥97% identity) across long stretches of their sequence (>4kb, containing predicted HOX coding sequence and flanking, non-coding sequence) to loci of individual members of the 42 clustered lamprey HOX genes (Supplementary Table 13). These loci either could represent recent duplications of HOX loci or could be assembly artifacts arising from the relatively high heterozygosity of the lamprey genome. Based on their exceptionally high levels of coding and non-coding sequence similarity to clustered HOX loci, we infer that these 5 loci are assembly artifacts due to polymorphism and that they do not represent additional singleton HOX genes in the lamprey genome. The 6th predicted singleton HOX gene shows equal levels of homology to ANTP-class homeobox genes of both HOX and non-HOX families, suggesting it is a derived ANTP-class homeobox gene and not necessarily a HOX gene.

Phylogenetic analysis of HOX genes. Phylogenetic analysis was performed on HOX paralog groups with 4 or more members in sea lamprey; groups 4, 8, 9, 11 and 13. For each paralog group, predicted sea lamprey HOX protein sequences were aligned against homologs from other vertebrate species and amphioxus, retrieved from GenBank. Our approach was informed by the experiences detailed by Simakov et al.13, Qiu et al.3, Mehta et al.3 and Manousaki et al.5. In selecting jawed vertebrate taxa for these analyses, we avoided tetost fish and Xenopus laevis as these lineages have undergone additional genome duplication events, which can lead to their co-orthologous genes/proteins being more derived than those from non-duplicated lineages. Thus, we focused on medaka (Oryzias latipes), zebrafish (Danio rerio) and coelacanth (L. menadoensis) as these lineages have undergone additional genome duplication events, which can lead to their co-orthologous genes/proteins being more derived than those from non-duplicated lineages. Thus, we opted for elephant shark (Heterocercus milii) and axolotl (Ambystoma mexicanum) which can lead to their co-orthologous genes/proteins being more derived than those from non-duplicated lineages. The Rad1-Rad10 complex promotes the production of gross chromosomal rearrangements from spontaneous DNA damage in Saccharomyces cerevisiae. Genetics 169, 1927–1937 (2005).

Korf, I. Gene finding in novel genomes. BMC Bioinformatics 5, 59 (2004).

Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19 (Suppl. 2), ii215–ii225 (2003).

Stanke, M., Duhans, M., Baertsch, R. & Haussler, D. Using native and sytemically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24, 637–644 (2008).

Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652 (2011).

Dohal, P. et al. The draft genome of Goniobrama microlepis: insights into choroid and vertebrate origins. Science 339, 2157–2162 (2012).

Simakov, O. et al. Insights into bilaterian evolution from three spiralian genomes. Nature 493, 526–531 (2013).

Putnam, N. H. et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317, 86–94 (2007).

Aparicio, S. et al. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297, 1301–1310 (2002).

Putnam, N. H. et al. The amphioxus genome and the evolution of the chordate karyotype. Nature 453, 1064–1071 (2008).

Venkatesh, B. et al. Elephant shark genome provides unique insights into gnathostome evolution. Nature 505, 174–179 (2014).

Hellesten, U. et al. The genome of the Western clawed frog Xenopus tropicalis. Science 328, 633–636 (2010).

Adams, M. D. et al. The genome sequence of Drosophila melanogaster. Science 287, 1825–1829 (2000).

Venter, J. C. et al. The sequence of the human genome. Nature 449, 210–215 (2007).
70. The UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res. 39, D214–D219 (2011).
71. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TREMBL in 2000. Nucleic Acids Res. 28, 45–48 (2000).
72. Mulder, N. J. et al. InterPro: an integrated documentation resource for protein families, domains and functional sites. Brief. Bioinform. 3, 225–235 (2002).
73. Jones, P. et al. InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236–1240 (2014).
74. Quevillon, E. et al. InterProScan: protein domains identifier. Nucleic Acids Res. 33, W116–W120 (2005).
75. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990).
76. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26, 873–881 (2010).
77. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33, 290–295 (2015).
78. Wucher, V. et al. FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. Nucleic Acids Res. 45, e57 (2017).
79. Duret, L., Chureau, C., Samain, S., Weissenbach, J. & Avner, P. The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. Science 312, 1653–1655 (2006).
80. Kuraku, S., Meyer, A. & Kuratani, S. Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after? Mol. Biol. Evol. 26, 47–59 (2009).
81. Manousaki, T. et al. in Jawless Fishes of the World Vol. 1 (eds. Orlov, A. & Beamish, R.) 2–16 (Cambridge Scholars Publishing, Newcastle-upon-Tyne, UK, 2016).
82. Camacho, C. et al. BLAST+: architecture and applications. BMC Bioinformatics 10, 421 (2009).
83. Skinner, M. E., Uzilov, A. V., Stein, L. D., Mungall, C. J. & Holmes, I. H. JBrowse: a next-generation genome browser. Genome Res. 19, 1630–1638 (2009).
84. Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23, 657–663 (2007).
85. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27, 764–770 (2011).
86. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829 (2008).
87. Quevillon, E. et al. Web Apollo: a web-based genomic annotation editing platform. Genome Biol. 14, R93 (2013).
88. Quevillon, E. et al. BLAST+: architecture and applications. BMC Bioinformatics 10, 421 (2009).
89. Manousaki, T. et al. in Jawless Fishes of the World Vol. 1 (eds. Orlov, A. & Beamish, R.) 2–16 (Cambridge Scholars Publishing, Newcastle-upon-Tyne, UK, 2016).
90. Ravi, V. et al. Elephant shark (Callorhinchus milii) provides insights into the evolution of Hox gene clusters in gnathostomes. Proc. Natl. Acad. Sci. USA 106, 16327–16332 (2009).
91. Amerini, C. T. et al. Complete HOX cluster characterization of the coelacanth provides further evidence for slow evolution of its genome. Proc. Natl. Acad. Sci. USA 107, 3622–3627 (2010).
Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

### Experimental design

1. **Sample size**
   - Describe how sample size was determined.
   - **no sample size calculation was performed**

2. **Data exclusions**
   - Describe any data exclusions.
   - In Supplementary Figure 8, amphioxus Hox11 was not used as an outgroup due to its relatively short sequence impinging on the alignment. Hox-11 from the Japanese lamprey was omitted from the alignment as full-length sequence is not available.

3. **Replication**
   - Describe whether the experimental findings were reliably reproduced.
   - **all attempts at replication were successful except where noted in the manuscript, with respect to a few predicted deletions.**

4. **Randomization**
   - Describe how samples/organisms/participants were allocated into experimental groups.
   - The study does not include experimental groups. Animals used for sequencing were captured from nature and selected based on the fact that they were producing sperm.

5. **Blinding**
   - Describe whether the investigators were blinded to group allocation during data collection and/or analysis.
   - **No, no experimental groups were allocated in this study.**

   *Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.*

6. **Statistical parameters**
   - For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

   **n/a** Confirmed

   - [x] The **exact sample size** \((n)\) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
   - [x] A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
   - [x] A statement indicating how many times each experiment was replicated
   - [x] The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
   - [x] A description of any assumptions or corrections, such as an adjustment for multiple comparisons
   - [x] The test results (e.g. \(P\) values) given as exact values whenever possible and with confidence intervals noted
   - [x] A clear description of statistics including **central tendency** (e.g. median, mean) and **variation** (e.g. standard deviation, interquartile range)
   - [x] Clearly defined error bars

*See the web collection on statistics for biologists for further resources and guidance.*
### Software

Describe the software used to analyze the data in this study.

- SparseAssembler - Genome assembler
- DBG2OLC - Genome assembler
- SSPACE 2.0 - Genome scaffold
- ALLMAPS v0.5.3 - Genome scaffold
- JoinMap v4.1 - Linkage analysis
- Blast - Sequence alignment
- RepeatModeler 1.0.9 - Repeat identification
- RepeatMasker v4.0.5 - Repeat identification/Annotation
- Enrichr - Ontology enrichment analysis
- MAKER2 - Gene annotation
- Snap v2013-11-29 - Gene annotation
- Augustus v2.5.5 - Gene annotation
- GSNAP v2017-04-24 - Noncoding gene annotation
- StringTie v1.3.3b - Transcript reconstruction
- FEELnc_filter.pl - Noncoding gene annotation
- BWA-MEM v.0.7.10 - Short read alignment
- samtools v.1.2 - Alignment processing
- bedtools v2.23.0 - Alignment processing
- DifCover - Alignment processing
- DNAcopy v1.46.0 - Alignment processing
- Jellyfish v2.1.3 - k-mer counting
- Apollo v2.0.8 - Manual annotation
- Jbrowse v1.12.3 - Manual annotation
- MEGA7 - Phylogenetic analysis
- MUSCLE - Sequence multiple alignment

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

### Materials and reagents

#### 8. Materials availability
Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

- no unique materials were used

#### 9. Antibodies
Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

- no antibodies were used

#### 10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

- no cell lines were used

b. Describe the method of cell line authentication used.

- no eukaryotic cell lines were used
c. Report whether the cell lines were tested for mycoplasma contamination.

- no eukaryotic cell lines were used
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

- no eukaryotic cell lines were used

### Animals and human research participants

#### 11. Description of research animals
Provide details on animals and/or animal-derived materials used in the study.

- Two different male lampreys (wild captured spawning adults) were used.
12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

the study did not involve human participants