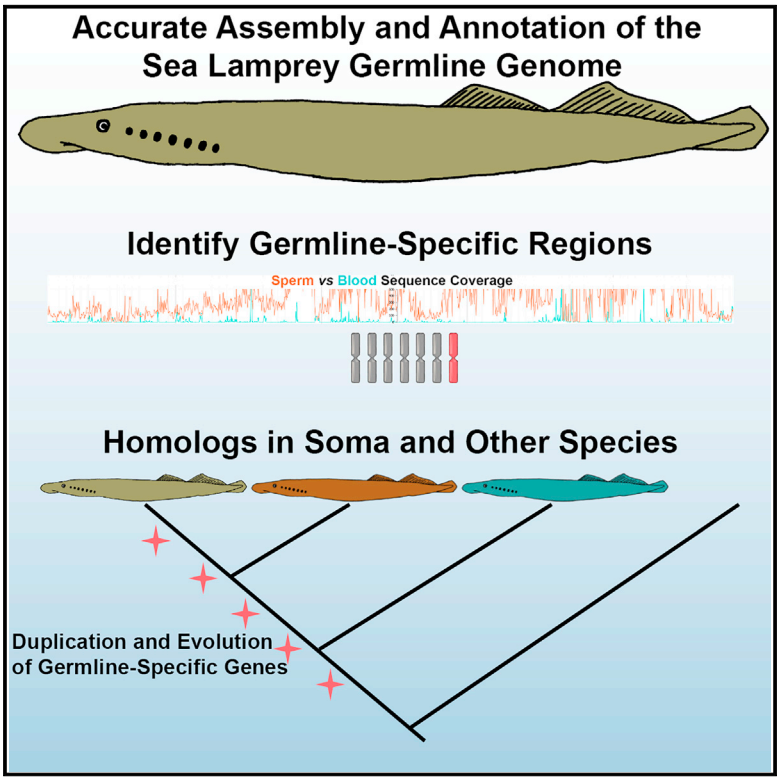
Cell Reports

Article

# An improved germline genome assembly for the sea lamprey *Petromyzon marinus* illuminates the evolution of germline-specific chromosomes

## Graphical abstract




## Authors

Nataliya Timoshevskaya, Kaan İ. Eşkut, Vladimir A. Timoshevskiy, ..., Olivier Fedrigo, Erich D. Jarvis, Jeramiah J. Smith

## Correspondence

jjsmit3@uky.edu

## In brief

Timoshevskaya et al. report an improved genome assembly for sea lamprey that aids in resolving the structure and evolution of chromosomes that are programmatically eliminated during development (germline-specific chromosomes). Using data from other species, these analyses indicate major roles of duplication and selection in the long-term evolution of germline-specific chromosomes.

## Highlights

- We report an improved assembly of the sea lamprey (*Petromyzon marinus*) genome
- The assembly resolves at least one germline-specific chromosome
- Many germline-specific genes have somatic paralogs in the sea lamprey genome
- Data from other species provide insight into the timing of duplication events

Timoshevskaya et al., 2023, Cell Reports *42*, 112263
March 28, 2023 
https://doi.org/10.1016/j.celrep.2023.112263

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

# An improved germline genome assembly for the sea lamprey *Petromyzon marinus* illuminates the evolution of germline-specific chromosomes

Nataliya Timoshevskaya,[1] Kaan İ. Eşkut,[1] Vladimir A. Timoshevskiy,[1] Sofia M.C. Robb,[2] Carson Holt,[3] Jon E. Hess,[4] Hugo J. Parker,[2] Cindy F. Baker,[5] Allison K. Miller,[6] Cody Saraceno,[1] Mark Yandell,[3] Robb Krumlauf,[2,7] Shawn R. Narum,[8] Ralph T. Lampman,[9] Neil J. Gemmell,[6] Jacquelyn Mountcastle,[10] Bettina Haase,[10] Jennifer R. Balacco,[10] Giulio Formenti,[10,11] Sarah Pelan,[12] Ying Sims,[12] Kerstin Howe,[12] Olivier Fedrigo,[10] Erich D. Jarvis,[10,11,13] and Jeramiah J. Smith[1,14,*]

[1]Department of Biology, University of Kentucky, Lexington, KY 40506, USA
[2]Stowers Institute for Medical Research, Kansas City, MO 64110, USA
[3]Department of Human Genetics, University of Utah, Salt Lake City, UT 84112, USA
[4]Columbia River Inter-Tribal Fish Commission, Portland, OR 97232, USA
[5]National Institute of Water and Atmospheric Research Limited (NIWA), Hamilton, Waikato 3261, New Zealand
[6]Department of Anatomy, School of Biomedical Sciences, University of Otago, Dunedin, Otago 9054, New Zealand
[7]Department of Anatomy & Cell Biology, The University of Kansas School of Medicine, Kansas City, KS 66160, USA
[8]Columbia River Inter-Tribal Fish Commission, Hagerman, ID 83332, USA
[9]Yakama Nation Fisheries Resource Management Program, Pacific Lamprey Project, Toppenish, WA 98948, USA
[10]Vertebrate Genome Lab, The Rockefeller University, New York, NY 10065, USA
[11]Laboratory of Neurogenetics of Language, The Rockefeller University, New York, NY 10065, USA
[12]Tree of Life, Wellcome Sanger Institute, Cambridge CB10 1SA, UK
[13]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
[14]Lead contact
*Correspondence: jjsmit3@uky.edu
https://doi.org/10.1016/j.celrep.2023.112263

## SUMMARY

Programmed DNA loss is a gene silencing mechanism that is employed by several vertebrate and nonvertebrate lineages, including all living jawless vertebrates and songbirds. Reconstructing the evolution of somatically eliminated (germline-specific) sequences in these species has proven challenging due to a high content of repeats and gene duplications in eliminated sequences and a corresponding lack of highly accurate and contiguous assemblies for these regions. Here, we present an improved assembly of the sea lamprey (*Petromyzon marinus*) genome that was generated using recently standardized methods that increase the contiguity and accuracy of vertebrate genome assemblies. This assembly resolves highly contiguous, somatically retained chromosomes and at least one germline-specific chromosome, permitting new analyses that reconstruct the timing, mode, and repercussions of recruitment of genes to the germline-specific fraction. These analyses reveal major roles of interchromosomal segmental duplication, intrachromosomal duplication, and positive selection for germline functions in the long-term evolution of germline-specific chromosomes.

## INTRODUCTION

The sea lamprey (*Petromyzon marinus*) is one of a growing number of animal,[1–6] plant,[7] and protist[8–10] species that are known to possess a collection of genes in their germ cells that are not found in any other cell type.[11,12] These germline-specific genes are lost from most somatic cell lineages early in development (starting at the early blastula stage) but are retained in the germline.[1–3,13,14] Evidence from sequencing, embryology, and karyotyping studies indicates that several other lamprey and hagfish species (apparently all jawless vertebrates) experience similar patterns of DNA loss during embryogenesis, likely involving the selective removal of whole chromosomes or chromosome fragments during early embryogenesis.[2–5] Additionally work in birds indicates that the ancestor of all songbirds likely evolved a similar genome biology, resulting in the evolution of a single germline-specific chromosome in the Passerine lineage.[15,16] While the somatic loss of germline-specific chromosomes has only been observed during embryogenesis in two vertebrate species (sea lamprey and Pacific lamprey [*Entosphenus tridentatus*]),[3,13,17] the discrete localization of germline-specific sequences to the adult germline is taken as evidence that elimination events take place during the earliest stages of development in all species that undergo programmed DNA loss.

The presence of germline-specific chromosomes in these species provides insights into the biological function of their germline-specific genes: all of these genes are expressed exclusively by the germline and presumably have functions in the germline that are sufficiently important as to permit their retention over evolutionary time. The long-term maintenance of programmed DNA elimination also implies a selective advantage to permanently silence some germline-specific genes within somatic tissues, presumably due to deleterious effects that could arise from their misexpression. For some eliminated genes, paralogs are also present in somatically retained fractions of the genome, which in theory could release germline-specific paralogs from pleiotropic effects imposed by the soma, thereby permitting a more rapid response to selection.[15,16] Studies of programmatically eliminated chromosomes in diverse taxa, including lampreys, songbirds, roundworms, and plants, support these general expectations.[7,14,16,18–20] Recent studies in songbirds indicate that the structure and content of germline-specific chromosomes (also known as germline-restricted chromosomes [GRCs]) may evolve rapidly over evolutionary time and likely harbor large numbers of genes that have been duplicated from somatically retained chromosomes.[15,16] As programmed elimination appears to have evolved several times among deeply diverged eukaryotic lineages, and might generally evolve rapidly in these lineages, the diverse set of species that are known to undergo programmed DNA loss provides a unique platform for studying the selective, regulatory, and developmental constraints that drive the evolution of germ cells and germline genes.

The sea lamprey also has served as an essential comparative model for studying several biomedically important aspects of vertebrate development, evolution, regeneration, and immunology.[21,22] Moreover, sea lampreys pose a significant ecological threat to the Great Lakes basin, having invaded the system in the 1930s, resulting in the decimation of several commercial fish populations.[23–26] All of these factors have made the sea lamprey an attractive target for the development of genomic resources, including genome assemblies.[19,27] In addition to the sea lamprey, chromosome-level genome assemblies have also been developed for species from the genera *Entosphenus*[28] and *Lethenteron*.[29,30] These species are part of a recently diverged clade that last shared a common ancestor approximately 12–13 million years ago (mya; roughly equivalent to human/orangutan divergence)[31–33] and that shared a common ancestor with *Petromyzon* approximately 30 mya. Other lamprey species that have not been fully assembled thus far include members of the genera *Geotria* (a draft assembly is available)[34] and *Mordacia*, which are found in the Southern hemisphere and diverged from the common ancestor of Petromyzon and other Northern hemisphere species ~200 mya.[21,31]

Lamprey genomes in general, and the sea lamprey genome in particular, present notable challenges to assembly, including exceptionally high GC content,[27] large numbers of chromosomes (1N = 96),[3] duplications of varying ages,[19,29] and the presence of numerous high copy satellite elements, which are particularly enriched within the germline-specific chromosomes.[3] In addition, the sea lamprey genome appears to have undergone a recent expansion in size, apparently due to the accumulation of sequences from recently active transposable elements; as such, the sea lamprey genome is ~0.5–0.9 Gb larger than other closely related lamprey species within the Northern Hemisphere clade (sea lamprey germline genome: ~2.3 Gb; sea lamprey somatic genome 1.8 Gb; other species: ~1.29–1.42 Gb as estimated via Feulgen densitometry and flow cytometry using somatic tissues).[27,35] Improvements in sequencing methods and assembly algorithms have solved several of the aforementioned issues that previously impeded assembly of germline-specific chromosomes[36] and show promise for resolving the larger-scale structure and evolutionary history of germline-specific chromosomes.

Here, we present an improved assembly of the sea lamprey genome and use this assembly to resolve several open questions regarding the evolution of germline-specific chromosomes related to the timing and mode of recruitment of germline-specific genes, to the depth of ancestry of programmed DNA loss in sea lamprey, and to the evolutionary consequences of recruitment to the germline-specific fraction of the genome. These analyses are empowered by the dramatically improved contiguity and scaffolding of germline-specific regions, including a highly contiguous assembly of one germline-specific chromosome, and by the increased accuracy of the assembly compared with a previous version that was based on older long-read data. This assembly allows multiple analyses aimed at the discovery of duplication events, both ancient and modern, that shaped the content of germline chromosomes and the partitioning of ancestral pleiotropic gene functions.

## RESULTS AND DISCUSSION

### Assembly and annotation

The sea lamprey genome was assembled from meiotic testes of a single individual (distinct from individuals used in previous assemblies) using a combination of PacBio continuous long reads (CLRs), 10X Genomics linked reads, BioNano Genomics optical maps, and Hi-C data, the combination of which was assembled using the VGP 1.6 pipeline, including manual curation of any errors found.[36] The resulting assembly has a contig N50 of 2.54 vs. 0.17 Mb for the previous assembly, which is a 150-fold improvement in contiguity. Most chromosomes in the improved assembly have less than 20 gaps, some less than 7. It has a scaffold N50 length of 13 Mb, with 33 scaffolds being larger than N50 (fewer than half the number of chromosome pairs: 1N = 96)[3] and a total of 1,434 scaffolds. This reflects an 88.7% reduction in the number of scaffolds that are not fully integrated into distinct chromosomes (unlocalized and orphans: from 11,967 to 1,349). The overall structure of the assembly is similar to the previous version (with notable exceptions of previous chromosomal scaffolds 1, 2, and 11) and is consistent with patterns of chromatin contacts inferred from our Hi-C library (Figure S1), improving on the previous version (Figure S2). Various sequence-based estimates of error rate and assembly completeness indicate that this improved sea lamprey genome improves upon the previously published lamprey assemblies (Figure 1; Table S1). The assembly contains 92% of predicted universal Metazoan homologs (BUSCO Metazoa)[37,38] and 91% of predicted universal vertebrate homologs (core vertebrate genes [CVGs]).[39] Notably, of the seven CVGs that are missing

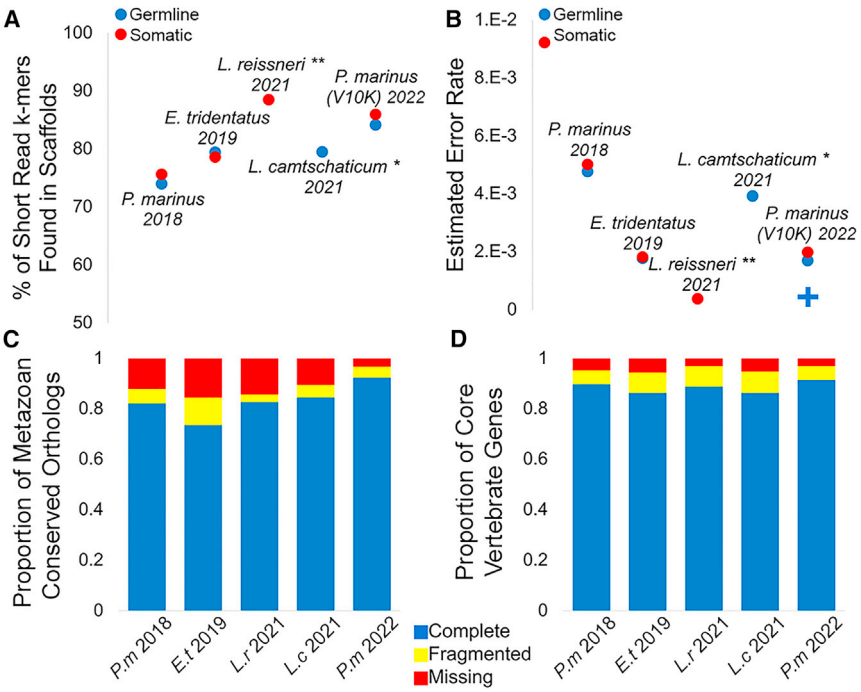



**Figure 1. Assembly quality metrics for the VGP lamprey genome and other lamprey genomes**

(A) Percent of short-read k-mers found in the assembly. This statistic reflects the degree to which the assembly incorporates sequences that are sampled from the organism. Analyses were performed using germline (blue) and somatic (red) reads.

(B) Estimated error rate reflects the degree to which the assembly bases are consistent with large samples of reads. This statistic can be impacted by intraspecific polymorphism when assembly and evaluation read sets are generated from different tissues or individuals: the blue plus (+) symbol shows the value for this estimate when short-read data are from the same animal used for the assembly.

(C) Percentage of predicted conserved metazoan orthologs detected.

(D) Percentage of predicted conserved vertebrate orthologs detected.

*, somatic reads not available; **, germline reads not available; *P.m*, *P. marinus*; *E.t*, *E. tridentatus*; *L.c*, *L. camtschaticum*; *L.r*, *L. reissneri*. Datasets and detailed numbers provided in Table S1.

from our assembly (homologs of LRRC34, TBCD, TMEM43, SORL1, NEXN, YEATS2, and CEP192), only one is found in any other lamprey assembly (TMEM43 is detected in *E. tridentatus* and *L. camtschaticum*), suggesting that these genes have been lost or were never present in the lamprey lineage. Direct comparisons of assembly quality with other lamprey assemblies are challenging due to variable sampling across projects, including a lack of paired somatic/germline sequence data for both *Lethenteron* species; the absence of definitive germline sequence data for others (sequence data from the *L. reissneri* assembly are reported as having been sampled from muscle, PRJNA558325, and are reported as germline elsewhere[30]); and intraspecific variation within sequenced animals. Overall, continuous improvement of sequencing and assembly methods has led to the development of increasingly accurate and informative lamprey genome assemblies.

## Identification of germline-specific/-enriched intervals

Germline-specific intervals were identified using new high coverage resequencing data from a separate male *P. marinus* (~52× coverage in germline [sperm] reads and ~95× coverage in somatic [blood] reads). These analyses identified 29.1 Mb germline-specific/-enriched sequences that could be anchored to one or more high-confidence nonrepetitive (approximately single copy) intervals (Table S2). Nonrepetitive and moderate copy-number (up to ~30 copies) intervals contain a total of 483 annotated genes, 373 of which correspond to known homologs in other vertebrate species (Table S3). Cross-referencing previously published PCR validation studies confirms the programmatic elimination of 37 predicted germline-specific scaffolds/regions and accounts for 259 annotated genes[19] (Table S3). Consistent with previous studies,[14,18,19] genes encoded in the germline-specific fraction of the genome are highly enriched for ontologies associated with several functions that are relevant to germ cell development and maturation including meiotic cell division, recombination, cell migration/adhesion, and WNT signaling (Table S4).

## Resolving the large-scale structure of a germline-specific chromosome

Analysis of germline vs. somatic sequence coverage revealed that one of the large chromosomal scaffolds that was reconstructed by the assembly pipeline represents a germline-specific chromosome (Figures 2 and S3). We refer to this chromosome as chromosome G1 (ChrG1; originally named Chr81) in reference to the fact that it is germline specific and one of 12 germline-specific chromosomes that are known to exist in the sea lamprey genome.[3] The availability of a highly contiguous assembly for one vertebrate germline-specific chromosome provides new insights into the content and evolutionary history of this chromosome, which are likely relevant to understanding the evolution of germline-specific chromosomes in general. First, the chromosome contains a large number of interspersed repetitive elements, as well as 85 annotated genes, 71 of which are homologs of 27 distinct vertebrate genes (Table S3). On ChrG1, homologs of these 27 genes range in copy number (distinct annotated copies) from one to seven, with homologs of HYKK (hydroxylysine kinase) being the most abundant. Six HYKK gene copies are located within a 360 kb interval and are interspersed with copies of SPOP (speckle type BTB/POZ protein), suggesting that some copies of SPOP, beyond the initial integration, trace their origin to the same duplication events that amplified HYKK (Figures 2 and S3). In general, annotation and analysis of ChrG1 is consistent

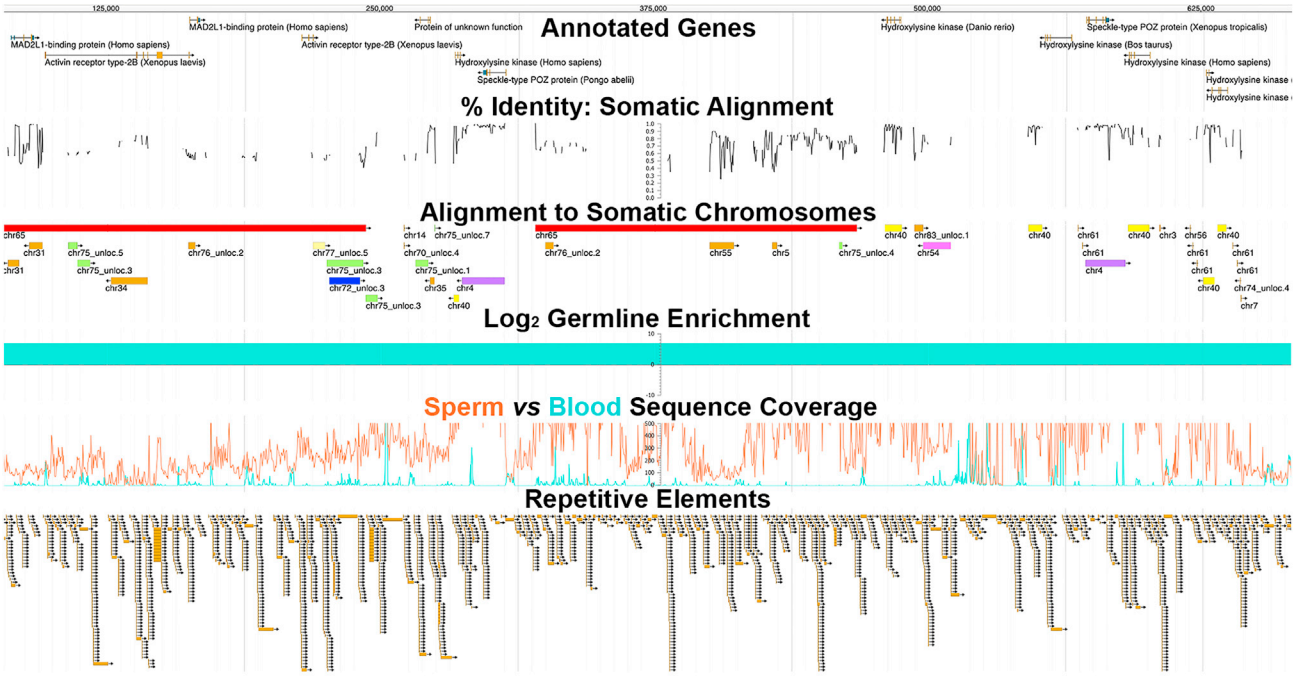



**Figure 2. Annotation and analysis of the germline-specific chromosome G1**

A browser view showing a 600 kb subregion of ChrG1. This region shows strong segmental homology to at least seven somatically retained chromosomes (highlighted by coloration of the corresponding segments) and contains 13 gene models, including 6 paralogs of HYKK. Segments homologous to somatic chromosomes vary with respect to their length and the degree of sequence identity between ChrG1 and the corresponding somatic chromosome across nonrepetitive intervals. Browser view is from SIMRbase (https://simrbase.stowers.org) and displays a subset of tracks that are most informative to assessing DNA elimination and the evolution of the sequence content of this chromosome. A view of the entire chromosome is shown in Figure S3.

with studies of germline-specific paralogs in birds[16,40] that suggest a major role for intrachromosomal duplication in shaping the gene content of germline-specific chromosomes across divergent taxa.

The most in-depth studies of songbird germline-specific chromosomes have been performed using the zebra finch and have proposed that a large portion of the finch GRC traces its origin to duplications of somatic chromosomal segments.[16,40] To assess the degree to which interchromosomal duplication (including somatic to germline) has shaped the evolution of sea lamprey ChrG1, we aligned this chromosome to all other assembled sea lamprey chromosomes to search for homologous regions. These searches yielded alignment to several somatic chromosomes, including 13 somatic chromosomes that each covered more than 20 Kb of ChrG1. Among these, alignments to Chr65 and Chr82 (both somatic) covered the largest fraction of ChrG1, corresponding to 337 and 234 kb, respectively. Duplication of Chr65 accounts for the origins of MAD2L1 and ACVR2B homologs, which have undergone secondary duplications, as well as an apparent pseudogene of myosin heavy chain. Duplications from Chr82 account for the origins of four ChrG1 genes: CDC20, NCAM, COP1, and PKP4. The presence of syntenic duplications and the fact that all annotated genes possess intron/exon structures indicate that a majority of germline and somatic paralogs trace their origins to large segmental duplications of portions of somatic chromosomes (Figure 2).

## Reconstructing the evolutionary origins and history of germline-specific genes

To resolve the timing of recruitment of genes to the germline-specific chromosomes, we generated phylogenetic trees from five species that provided resolution as to the timing and location of germline-specific paralogs. Datasets used in this study included all annotated sea lamprey genes; a draft germline assembly from *E. tridentatus* (Pacific lamprey); a germline/embryo transcriptome from *Geotria australis* (known as kanakana, piharau, or pouched lamprey); and genes from two gnathostomes, spotted gar (*Lepisosteus oculatus*) and human (*Homo sapiens*). Following automated construction of genome-wide gene trees via OrthoFinder,[41] we identified 70 orthogroups that provided insight into the evolutionary history of the germline-specific chromosomes (all germline-specific scaffolds). A majority of orthogroups (42 groups containing 163 genes) had no definable somatic homolog in sea lamprey, suggesting that these genes have either resided on the germline-specific chromosomes since their origin and the ancestral somatic copy was lost over evolutionary time or that the somatic homolog was otherwise not identifiable in the assembly (including cases where the germline homologs have diverged in sequence to the point that they are no longer recognizable as such). Another 28 germline-specific gene lineages (encompassing 113 germline-specific genes) were grouped with their paralogous somatic genes (Figure S4; Table S5). Because the species used vary in their divergence time with sea lamprey (Pacific lamprey

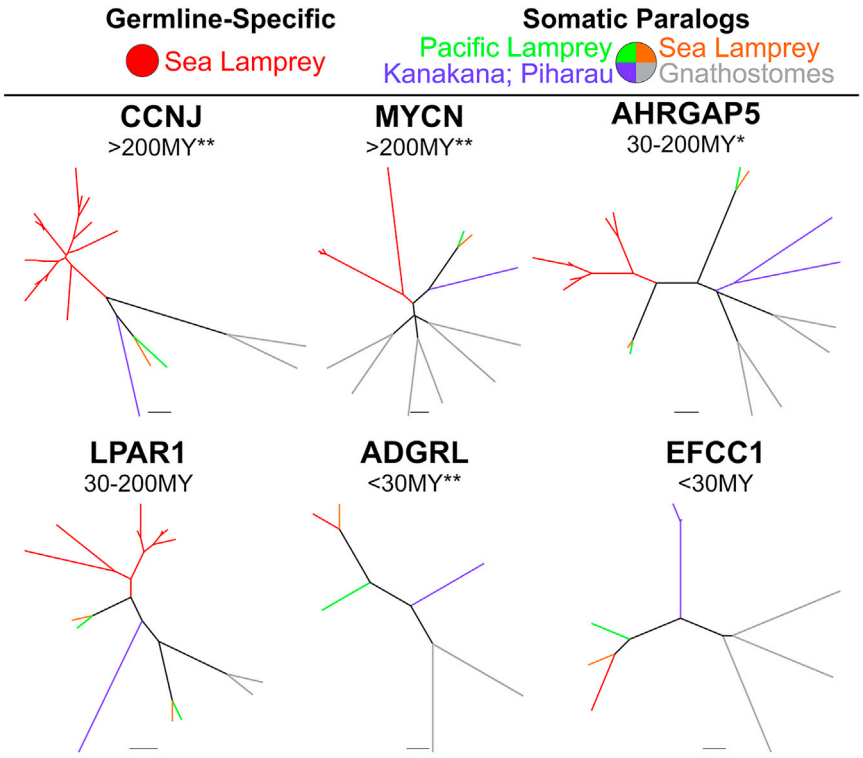



**Figure 3. Example gene trees for germline-specific genes with well-defined somatic homologs in sea lamprey**

A complete set of trees is shown in Figure S4. Germline-specific genes are highlighted in red. Lamprey somatic lineages are highlighted in orange (sea lamprey), green (Pacific lamprey), and purple (kanakana/piharau/pouched lamprey). Gnathostome lineages are highlighted in gray. Divergence dates are relative to three major divergence events within the lamprey lineage. **, gene trees with germline-specific clades with *dN*/*dS* >1 and higher than somatically retained clades; *, *dN*/*dS* not determined due to a high frequency of gap characters in amino acid alignments. Scale bars: 0.02 substitutions per amino acid site.

~30 my, kanakana ~200 my, gnathostomes ~550 my), these trees provided an estimate of the timing of recruitment of genes to the germline chromosomes. Gene tree reconstruction yielded somatic gene trees that were generally consistent with the true evolutionary relationships among taxa in this phylogeny, and for 23 orthogroups, it was possible to resolve the approximate timing of divergence of germline vs. somatic gene lineages (Figures 3 and S4). These duplication events were distributed across all three age classes: younger than 30 my (7 genes), 30–200 my (7 genes), and older than 200 my (9 genes). This wide distribution of ages suggests that programmed DNA loss traces its origins to the deep ancestry of the lamprey (or deeper stem) lineages and that recruitment of germline-specific genes to the sea lamprey germline-specific chromosomes has occurred through a roughly continuous process that includes both ancient and very recent integration events.

The construction of these gene trees also provides an opportunity to assess how genes evolve following recruitment to the germline-specific fraction of the genome. We asked whether clades of germline-specific genes differed in rates of synonymous vs. nonsynonymous mutation relative to their somatic counterparts. Among the 28 orthogroups with identifiable somatic and germline paralogs, PAML was able to perform analyses of evolutionary rates for 23, whereas for five others (AHRGAP2, CDC20, CDH2, IRS1/4, and YTHDC2), a combination of large numbers of insertions or deletions (indels) and large numbers of divergent paralogs reduced the number of gap-free sites such that *dN*/*dS* ratios could not be estimated (Figures 3 and S4). Among these 23 trees, eight germline-specific gene lineages showed strong evidence of increased positive selection following recruitment to the germline-specific fraction of the genome ($p < 0.01$; Table S5; Figures 3 and S4), and another six were either too diverse to calculate *dN*/*dS* statistics or had *dN*/*dS* >1 but were not statistically different from somatic orthologs. In this context, it is worth noting that the use of *dN*/*dS* to detect selection should be most effective when sufficient time has passed to permit accumulation of mutations and when genes in a clade show consistent patterns.[37] As such, this measure may underestimate selective forces acting on young lineages, which may explain why few young germline-specific gene lineages show evidence of positive selection compared with relatively older germline paralogs that arose >30 mya.

The presence of several older duplicates on sea lamprey germline chromosomes seems to have improved our ability to detect signatures of selection associated with programmed DNA loss. Overall, these analyses show that germline-specific genes often accumulate amino acid substitutions at a higher rate than their somatic homologs and that these changes may take place over the course of tens to hundreds of millions of years after landing on the germline chromosomes. Presumably this increased rate of protein evolution reflects the fact that these genes have been released from stabilizing selection in the context of somatic cells. While not every gene or gene variant that contributes positively to germline development would be expected to show the sorts of antagonistic pleiotropic effects that would select for permanent somatic silencing (loss), the collection of genes and substitutions that are retained in the germline-specific chromosomes of lamprey and other eliminating species may be particularly relevant to understanding evolutionary genetic trade-offs between germline and soma.

## Varying roles of germline-specific genes in reproduction and embryogenesis

Given evidence for widespread response to selection following recruitment of genes to the germline-specific chromosomes,

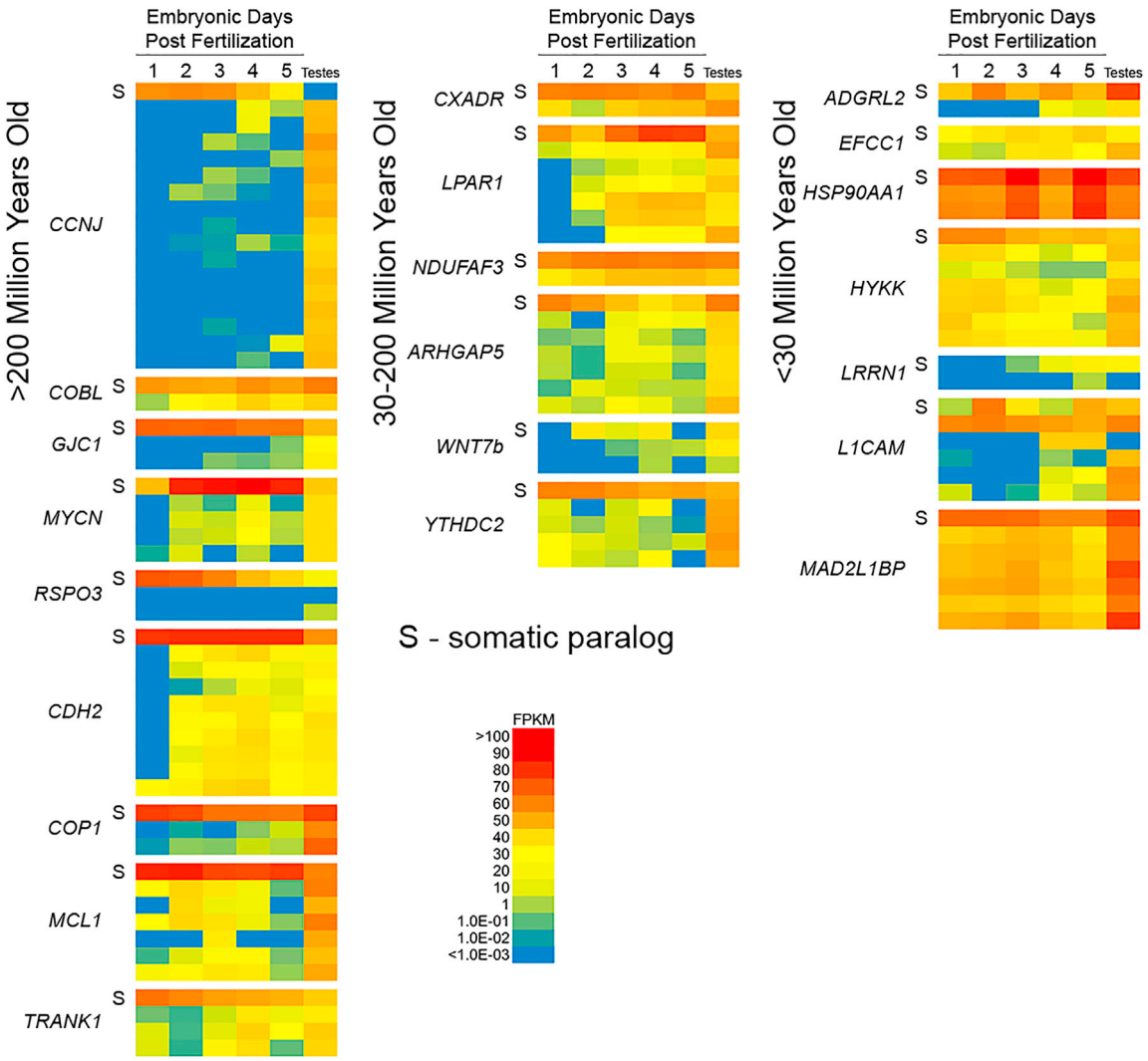



**Figure 4. Expression of germline-specific genes and their somatic homologs during early embryogenesis and spermatogenesis**

Groups of paralogs are separated based on the predicted timing of divergence for somatic- (S) vs. germline-specific copies. Expression is presented relative to FKPM (fragments per kilobase per million reads) to aid in comparisons between homologs of varying length. Colors depicting varying levels of gene expression are scaled relative to the $\log_{10}$(FKPM) to permit visualization of expression metrics spanning several orders of magnitude (inset scale).

we sought to better understand how germline-specific genes have integrated into early embryonic development and spermatogenesis after diverging from their somatic counterparts. Focusing on the set of germline genes with identifiable somatic homologs (Figure 3), we estimated transcript abundance across early embryonic development (morula though neurula) and within spermatogenic/meiotic germline cells using publicly available RNA sequencing (RNA-seq) datasets[14,18] (Figure 4). First, we found that expression of germline-specific genes is generally highest in meiotic testes. This may not be surprising, as germ cells comprise a large fraction of the testes and are highly transcriptionally active at this stage of meiosis,[14] whereas early embryos possess only a small number of primordial germ cells. Notable exceptions to this general pattern were observed for individual paralogs of L1CAM, LRRN1, and LPAR1, which show relatively higher expression during embryogenesis, suggesting that some germline-specific paralogs may preferentially contribute to earlier (or later) stages of germ cell development or maturation.

Second, we found that somatic homologs are generally highly expressed during early embryogenesis (Figure 4). Somatic homologs are also highly expressed in meiotic testes, with the exception of somatic CCNJ, which appears to have evolved highly specific expression in the earliest stages of embryogenesis, relative to spermatogenesis. Notably, CCNJ is among the oldest and most diverse of germline-specific genes (Figures 3 and S4), and somatic paralogs of other genes in this age class also show reduced expression in the germline relative to early embryogenesis, suggesting they may be on a similar evolutionary trajectory (Figure 4). Taken together, patterns of

expression suggest that somatically retained genes often retain functions relevant to the germline, but accumulation of changes over evolutionary time can reduce or eventually replace germline functions of somatically retained copies.

## Concluding remarks

The availability of a highly accurate and contiguous assembly for the sea lamprey, in the context of deeply diverging lamprey lineages, sheds new light on the tempo and mode of evolution in germline-specific chromosomes. These analyses illustrate the importance of large segmental duplications in shaping the gene content of somatically eliminated chromosomes and reveal extensive evolutionary changes that are associated with germline specificity and the corresponding release from constraints imposed by somatic antagonistic pleiotropy.

## Data use

Embargoes on the use of this improved lamprey genome assembly are lifted upon publication of this study. While this study was under consideration, a study by Yasmin et al.[42] used the current genome assembly to report genome-wide predictions of germline-specific scaffolds and chromosomes, as well as the impacts of our improved assembly on the annotation of germline-specific genes.

## Limitations of the study

This study reports a genome assembly for a single individual and uses somatic and germline resequencing data from a second individual to identify germline-specific portions of the genome (those listed by programmatic elimination). These analyses do not fully resolve the entire structure or gene content of eliminated regions or account for the potential for population-level variation in the complement of germline-specific genes or duplicate copies. The relatively young age of many of the identified duplicates suggests that some are likely to vary among individual sea lampreys.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Organisms used as source material
- METHOD DETAILS
  - Genome sequencing and assembly
  - Assembly quality metrics for lamprey genomes
  - Gene annotation
  - Identification of germline-specific regions
  - Alignment of G1 to somatic chromosomes
  - Geotria transcriptome assembly
  - Construction and analysis of gene trees
  - Reanalysis of RNAseq data
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2023.112263.

## ACKNOWLEDGMENTS

We thank Jane Kitson for facilitating discussions to ensure the appropriate representation of Maori iwi interests in relation to the collection and presentation of data from kanakana/piharau. The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of NIH or NSF. Partial computational support was provided by The University of Kentucky High Performance Computing complex. The support and resources from the Center for High Performance Computing at the University of Utah are gratefully acknowledged. This work was funded by grants from the National Institutes of Health (NIH) (R35GM130349) and the National Science Foundation (NSF) (MCB1818012) to J.J.S., a grant from the Great Lakes Fishery Commission to J.J.S. and E.D.J., HHMI funds to E.D.J., and New Zealand Ministry of Business Innovation and Employment (MBIE) contract C01X1615 to C.F.B. A.K.M. and N.J.G. are supported by the University of Otago and a subcontract from C01X1615.

## AUTHOR CONTRIBUTIONS

Conceptualization, J.J.S. and E.D.J.; investigation, N.T., K.İ.E., V.A.T., S.M.C.R., C.H., J.E.H., H.J.P., C.F.B., A.K.M., C.S., S.R.N., R.T.L., J.M., B.H., J.R.B., G.F., S.P., Y.S., K.H., and O.F.; writing – original draft, J.J.S.; writing – review & editing, N.T. and E.D.J.; funding acquisition, J.J.S., E.D.J., and C.F.B.; resources, M.Y., N.J.G., and R.K.; supervision, J.J.S. and E.D.J.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

## INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in their field of research or within their geographical location. One or more of the authors of this paper received support from a program designed to increase minority representation in their field of research. We avoided "helicopter science" practices by including the participating local contributors from the region where we conducted the research as authors on the paper.

Received: June 7, 2022
Revised: October 17, 2022
Accepted: February 28, 2023
Published: March 15, 2023

## REFERENCES

1. Boveri, T. (1887). Uber Differenzierung der Zellkerne wlhrend der Furchung des Eies van Ascarims egalocephala. Anat. Anz. *2*, 288–693.
2. Smith, J.J., Antonacci, F., Eichler, E.E., and Amemiya, C.T. (2009). Programmed loss of millions of base pairs from a vertebrate genome. Proc. Natl. Acad. Sci. USA *106*, 11212–11217. https://doi.org/10.1073/pnas.0902358106.
3. Timoshevskiy, V.A., Timoshevskaya, N.Y., and Smith, J.J. (2019). Germline-specific repetitive elements in programmatically eliminated chromosomes of the sea lamprey (Petromyzon marinus). Genes *10*, 832. https://doi.org/10.3390/genes10100832.
4. Kohno, S., Nakai, Y., Satoh, S., Yoshida, M., and Kobayashi, H. (1986). Chromosome elimination in the Japanese hagfish, Eptatretus burgeri

(Agnatha, Cyclostomata). Cytogenet. Cell Genet. *41*, 209–214. https://doi.org/10.1159/000132231.

5. Kojima, N.F., Kojima, K.K., Kobayakawa, S., Higashide, N., Hamanaka, C., Nitta, A., Koeda, I., Yamaguchi, T., Shichiri, M., Kohno, S.i., and Kubota, S. (2010). Whole chromosome elimination and chromosome terminus elimination both contribute to somatic differentiation in Taiwanese hagfish Paramyxine sheni. Chromosome Res. *18*, 383–400. https://doi.org/10.1007/s10577-010-9122-2.
6. Pigozzi, M.I., and Solari, A.J. (1998). Germ cell restriction and regular transmission of an accessory chromosome that mimics a sex body in the zebra finch, Taeniopygia guttata. Chromosome Res. *6*, 105–113. https://doi.org/10.1023/a:1009234912307.
7. Ruban, A., Schmutzer, T., Wu, D.D., Fuchs, J., Boudichevskaia, A., Rubtsova, M., Pistrick, K., Melzer, M., Himmelbach, A., Schubert, V., et al. (2020). Supernumerary B chromosomes of Aegilops speltoides undergo precise elimination in roots early in embryo development. Nat. Commun. *11*, 2764. https://doi.org/10.1038/s41467-020-16594-x.
8. Yao, M.C., and Gall, J.G. (1979). Alteration of the Tetrahymena genome during nuclear differentiation. J. Protozool. *26*, 10–13.
9. Yao, M.C., and Gorovsky, M.A. (1974). Comparison of the sequences of macro- and micronuclear DNA of Tetrahymena pyriformis. Chromosoma *48*, 1–18. https://doi.org/10.1007/bf00284863.
10. Klobutcher, L.A. (1987). Micronuclear organization of macronuclear genes in the hypotrichous ciliate Oxytricha nova. J. Protozool. *34*, 424–428. https://doi.org/10.1111/j.1550-7408.1987.tb03206.x.
11. Smith, J.J., Timoshevskiy, V.A., and Saraceno, C. (2021). Programmed DNA elimination in vertebrates. Annu. Rev. Anim. Biosci. *9*, 173–201. https://doi.org/10.1146/annurev-animal-061220-023220.
12. Wang, J., and Davis, R.E. (2014). Programmed DNA elimination in multicellular organisms. Curr. Opin. Genet. Dev. *27*, 26–34. https://doi.org/10.1016/j.gde.2014.03.012.
13. Timoshevskiy, V.A., Herdy, J.R., Keinath, M.C., and Smith, J.J. (2016). Cellular and molecular features of developmentally programmed genome rearrangement in a vertebrate (sea lamprey: Petromyzon marinus). PLoS Genet. *12*, e1006103. https://doi.org/10.1371/journal.pgen.1006103.
14. Bryant, S.A., Herdy, J.R., Amemiya, C.T., and Smith, J.J. (2016). Characterization of somatically-eliminated genes during development of the sea lamprey (Petromyzon marinus). Mol. Biol. Evol. *33*, 2337–2344. https://doi.org/10.1093/molbev/msw104.
15. Torgasheva, A.A., Malinovskaya, L.P., Zadesenets, K.S., Karamysheva, T.V., Kizilova, E.A., Akberdina, E.A., Pristyazhnyuk, I.E., Shnaider, E.P., Volodkina, V.A., Saifitdinova, A.F., et al. (2019). Germline-restricted chromosome (GRC) is widespread among songbirds. Proc. Natl. Acad. Sci. USA *116*, 11845–11850. https://doi.org/10.1073/pnas.1817373116.
16. Kinsella, C.M., Ruiz-Ruano, F.J., Dion-Côté, A.M., Charles, A.J., Gossmann, T.I., Cabrero, J., Kappei, D., Hemmings, N., Simons, M.J.P., Camacho, J.P.M., et al. (2019). Programmed DNA elimination of germline development genes in songbirds. Nat. Commun. *10*, 5468. https://doi.org/10.1038/s41467-019-13427-4.
17. Timoshevskiy, V.A., Lampman, R.T., Hess, J.E., Porter, L.L., and Smith, J.J. (2017). Deep ancestry of programmed genome rearrangement in lampreys. Dev. Biol. *429*, 31–34. https://doi.org/10.1016/j.ydbio.2017.06.032.
18. Smith, J.J., Baker, C., Eichler, E.E., and Amemiya, C.T. (2012). Genetic consequences of programmed genome rearrangement. Curr. Biol. *22*, 1524–1529. https://doi.org/10.1016/j.cub.2012.06.028.
19. Smith, J.J., Timoshevskaya, N., Ye, C., Holt, C., Keinath, M.C., Parker, H.J., Cook, M.E., Hess, J.E., Narum, S.R., Lamanna, F., et al. (2018). The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. Nat. Genet. *50*, 270–277. https://doi.org/10.1038/s41588-017-0036-1.
20. Wang, J., Gao, S., Mostovoy, Y., Kang, Y., Zagoskin, M., Sun, Y., Zhang, B., White, L.K., Easton, A., Nutman, T.B., et al. (2017). Comparative genome analysis of programmed DNA elimination in nematodes. Genome Res. *27*, 2001–2014. https://doi.org/10.1101/gr.225730.117.
21. McCauley, D.W., Docker, M.F., Whyard, S., and Li, W. (2015). Lampreys as diverse model organisms in the genomics era. Bioscience *65*, 1046–1056. https://doi.org/10.1093/biosci/biv139.
22. Green, S.A., and Bronner, M.E. (2014). The lamprey: a jawless vertebrate model system for examining origin of the neural crest and other vertebrate traits. Differentiation *87*, 44–51. https://doi.org/10.1016/j.diff.2014.02.001.
23. Dymond, J.R. (1922). A provisional list of fishes of Lake Erie. Univ. Toronto Stud. Biol. Ser. Pub. Ont. Fish Res. Lab. *4*, 57–73.
24. Applegate, V.C. (1950). Natural history of the sea lamprey, Petromyzon marinus, in Michigan. Special Scientific Report No. 55 (US Fish and Wildlife Services).
25. Smith, B.R., and Tibbles, J.J. (1980). Sea lamprey (Petromyzon marinus) in Lakes Huron, Michigan, and Superior: history of invasion and control, 1936–78. Can. J. Fish. Aquat. Sci. *37*, 1780–1801.
26. Wingfield, J., Brant, C., Eshenroder, R., Gaden, M., Miehls, A., and Siefkes, M. (2021). 100 years of sea lampreys above Niagara Falls: a reflection on what happened and what we learned. J. Great Lakes Res. *47*, 1844–1848. https://doi.org/10.1016/j.jglr.2021.10.013.
27. 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 insights into vertebrate evolution. Nat. Genet. *45*, 415–421. https://doi.org/10.1038/ng.2568.
28. Hess, J.E., Smith, J.J., Timoshevskaya, N., Baker, C., Caudill, C.C., Graves, D., Keefer, M.L., Kinziger, A.P., Moser, M.L., Porter, L.L., et al. (2020). Genomic islands of divergence infer a phenotypic landscape in Pacific lamprey. Mol. Ecol. *29*, 3841–3856. https://doi.org/10.1111/mec.15605.
29. Nakatani, Y., Shingate, P., Ravi, V., Pillai, N.E., Prasad, A., McLysaght, A., and Venkatesh, B. (2021). Reconstruction of proto-vertebrate, proto-cyclostome and proto-gnathostome genomes provides new insights into early vertebrate evolution. Nat. Commun. *12*, 4489. https://doi.org/10.1038/s41467-021-24573-z.
30. Zhu, T., Li, Y., Pang, Y., Han, Y., Li, J., Wang, Z., Liu, X., Li, H., Hua, Y., Jiang, H., et al. (2021). Chromosome-level genome assembly of Lethenteron reissneri provides insights into lamprey evolution. Mol. Ecol. Resour. *21*, 448–463. https://doi.org/10.1111/1755-0998.13279.
31. Kuraku, S., and Kuratani, S. (2006). Time scale for cyclostome evolution inferred with a phylogenetic diagnosis of hagfish and lamprey cDNA sequences. Zoolog. Sci. *23*, 1053–1064. https://doi.org/10.2108/zsj.23.1053.
32. Genner, M.J., Hillman, R., McHugh, M., Hawkins, S.J., and Lucas, M.C. (2012). Contrasting demographic histories of European and North American sea lamprey (Petromyzon marinus) populations inferred from mitochondrial DNA sequence variation. Mar. Freshw. Res. *63*, 827–833.
33. Artamonova, V.S., Kucheryavyy, A.V., and Pavlov, D.S. (2011). Nucleotide sequences of the mitochondrial cytochrome oxidase subunit I (co1) gene of lamprey classified with Lethenteron camtschaticum and the Lethenteron reissneri complex show no species-level differences. Dokl. Biol. Sci. *437*, 113–118.
34. Miller, A.K., Timoshevskaya, N., Smith, J.J., Gillum, J., Sharif, S., Clarke, S., Baker, C., Kitson, J., Gemmell, N.J., and Alexander, A. (2022). Population genomics of New Zealand pouched lamprey (kanakana; piharau; Geotria australis). J. Hered. *113*, 380–397. https://doi.org/10.1093/jhered/esac014.
35. Gregory, T.R. (2005). Animal genome size database. http://www.genomesize.com.
36. Rhie, A., McCarthy, S.A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. Nature *592*, 737–746. https://doi.org/10.1038/s41586-021-03451-0.
37. Seppey, M., Manni, M., and Zdobnov, E.M. (2019). BUSCO: assessing genome assembly and annotation completeness. Methods Mol. Biol. *1962*, 227–245. https://doi.org/10.1007/978-1-4939-9173-0_14.

38. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

39. Hara, Y., Tatsumi, K., Yoshida, M., Kajikawa, E., Kiyonari, H., and Kuraku, S. (2015). Optimizing and benchmarking de novo transcriptome sequencing: from library preparation to assembly evaluation. BMC Genomics *16*, 977. https://doi.org/10.1186/s12864-015-2007-1.

40. Asalone, K.C., Takkar, A.K., Saldanha, C.J., and Bracht, J.R. (2021). A transcriptomic pipeline adapted for genomic sequence discovery of germline-restricted sequence in zebra finch, Taeniopygia guttata. Genome Biol. Evol. *13*, evab088. https://doi.org/10.1093/gbe/evab088.

41. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. *20*, 238. https://doi.org/10.1186/s13059-019-1832-y.

42. Yasmin, T., Grayson, P., Docker, M.F., and Good, S.V. (2022). Pervasive male-biased expression throughout the germline-specific regions of the sea lamprey genome supports key roles in sex differentiation and spermatogenesis. Commun Biol *5*, 434. https://doi.org/10.1038/s42003-022-03375-z.

43. Chin, C.S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods *10*, 563–569. https://doi.org/10.1038/nmeth.2474.

44. Chin, C.S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods *13*, 1050–1054. https://doi.org/10.1038/nmeth.4035.

45. Guan, D., McCarthy, S.A., Wood, J., Howe, K., Wang, Y., and Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics *36*, 2896–2898. https://doi.org/10.1093/bioinformatics/btaa025.

46. Ghurye, J., Pop, M., Koren, S., Bickhart, D., and Chin, C.S. (2017). Scaffolding of long read assemblies using long range contact information. BMC Genomics *18*, 527. https://doi.org/10.1186/s12864-017-3879-z.

47. Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. Preprint at arXiv. https://doi.org/10.48550/arXiv.1207.3907.

48. Chow, W., Brugger, K., Caccamo, M., Sealy, I., Torrance, J., and Howe, K. (2016). gEVAL - a web-based browser for evaluating genome assemblies. Bioinformatics *32*, 2508–2510. https://doi.org/10.1093/bioinformatics/btw159.

49. Miller, J.R., Delcher, A.L., Koren, S., Venter, E., Walenz, B.P., Brownley, A., Johnson, J., Li, K., Mobarry, C., and Sutton, G. (2008). Aggressive assembly of pyrosequencing reads with mates. Bioinformatics *24*, 2818–2824. https://doi.org/10.1093/bioinformatics/btn548.

50. Rhie, A., Walenz, B.P., Koren, S., and Phillippy, A.M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. *21*, 245. https://doi.org/10.1186/s13059-020-02134-9.

51. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

52. Cabanettes, F., and Klopp, C. (2018). D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ *6*, e4958. https://doi.org/10.7717/peerj.4958.

53. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

54. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

55. Campbell, M.S., Holt, C., Moore, B., and Yandell, M. (2014). Genome annotation and curation using MAKER and MAKER-P. Curr. Protoc. Bioinformatics *48*, 4.11.1–4.11.39. https://doi.org/10.1002/0471250953.bi0411s48.

56. Stanke, M., Schöffmann, O., Morgenstern, B., and Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinf. *7*, 62. https://doi.org/10.1186/1471-2105-7-62.

57. Harris, R.S. (2007). Improved Pairwise Alignment of Genomic DNA. Ph.D. Thesis (The Pennsylvania State University).

58. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006. https://doi.org/10.1101/gr.229102.

59. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652. https://doi.org/10.1038/nbt.1883.

60. Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. *16*, 157. https://doi.org/10.1186/s13059-015-0721-2.

61. Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-Coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. *302*, 205–217. https://doi.org/10.1006/jmbi.2000.4042.

62. Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. *34*, W609–W612. https://doi.org/10.1093/nar/gkl315.

63. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591. https://doi.org/10.1093/molbev/msm088.

64. Rambaut, A. (2018). FigTree, version 1.4.4. http://tree.bio.ed.ac.uk/software/figtree/.

65. Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. *37*, 907–915. https://doi.org/10.1038/s41587-019-0201-4.

66. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295. https://doi.org/10.1038/nbt.3122.

67. Howe, K., Chow, W., Collins, J., Pelan, S., Pointon, D.L., Sims, Y., Torrance, J., Tracey, A., and Wood, J. (2021). Significantly improving the quality of genome assemblies through curation. GigaScience *10*, giaa153. https://doi.org/10.1093/gigascience/giaa153.

68. Tahara, Y. (1988). Normal stages of development in the lamprey, Lampetra reissued (dybowski). Zool. Sci. *5*, 109–118.

69. Chang, J.M., Di Tommaso, P., Taly, J.F., and Notredame, C. (2012). Accurate multiple sequence alignment of transmembrane proteins with PSI-Coffee. BMC Bioinformatics *13* (*Suppl 4*), S1. https://doi.org/10.1186/1471-2105-13-S4-S1.

70. Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. *13*, 555–556. https://doi.org/10.1093/bioinformatics/13.5.555.

71. Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360. https://doi.org/10.1038/nmeth.3317.

72. Pertea, M., Kim, D., Pertea, G.M., Leek, J.T., and Salzberg, S.L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat. Protoc. *11*, 1650–1667. https://doi.org/10.1038/nprot.2016.095.

## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| **Biological samples** | | |
| Sea Lamprey male | kPetMar1 | SAMN12629506 |
| Sea Lamprey male testes | Male7219 | SAMN23067069 |
| Sea Lamprey male blood | Male7219 | SAMN23067060 |
| *Geotria australis* testes | TAONGA-AGDR00015 | N1M |
| *Geotria australis* embryos | TAONGA-AGDR00015 | N1 |
| Pacific Lamprey testes | EtrAdultCRITFC18_GenomeStudy-0004 | SAMN23426542 |
| Pacific Lamprey blood | EtrAdultCRITFC18_GenomeStudy-0004 | SAMN23426543 |
| TruSeq Stranded Total RNA library prep | Illumina | Cat# 20020599 |
| **Critical commercial assays** | | |
| PacBio long Reads (Sequel I) | Pacific Biosciences | Pacific Biosciences Sequel I |
| Bionano optical maps (Saphyr) | BioNano | BioNano Saphyr |
| 10X Genomics linked reads | 10X Genomics | N/A |
| Arima Genomics Hi-C linked reads | Arima | N/A |
| Illumina NovaSeq | Illumina | Illumina NovaSeq |
| **Deposited data** | | |
| Sea lamprey germline genome assembly | This paper | BioProject PRJNA562011 |
| Sea lamprey germline genomic sequence reads | This paper | https://genomeark.github.io/genomeark-all/Petromyzon_marinus |
| Comparative sequencing datasets for sea lamprey | This paper | BioProject PRJNA779416 |
| Comparative sequencing datasets for pacific lamprey | This paper | BioProject PRJNA784541 |
| RNA sequencing data for *G. australis* | This paper | Aotearoa Genomic Data Repository (AGDR) under accession number TAONGA-AGDR00015 (https://data.agdr.org.nz/study-viewer/project/AGDR00015). |
| **Software and algorithms** | | |
| FALCON v. DNANexus 1.9.0 | Chin et al.[43] | http://www.dnanexus.com |
| FALCON-Unzip v. DNANexus 1.0.6 | Chin et al.[44] | http://www.dnanexus.com |
| purge_dups v. github ca23030ccf4254dfd2d3a5ea90d0eed41c24f88b | Guan et al.[45] | https://github.com/dfguan/purge_dups |
| scaff10x v. 4.1.0 | High Performance Algorithms Group at the Wellcome Sanger Institute | https://github.com/wtsi-hpag/Scaff10X |
| Bionano Solve DLS v. 3.2.1 | Bionano Genomics | https://s3.amazonaws.com/www.bnxinstall.com/access/tools/access.tools.tgz |
| Salsa v. 2.2 | Ghurye et al.[46] | https://github.com/marbl/SALSA |
| Arrow smrtanalysis v. smrtlink_6.0.0.47841 | PacBio | https://downloads.pacbcloud.com/public/software/installers/smrtlink_6.0.0.47841.zip |
| longranger align v. 2.2.2 | 10X Genomics | https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines |
| freebayes Illumina polishing v. 1.3.1 | Garrison and Marth[47] | https://github.com/freebayes/freebayes |
| gEVAL v. 2019–12–09 | Guan et al.[48] | https://geval.sanger.ac.uk |
| Meryl v. 1.1 | Miller et al.[49] | https://github.com/marbl/meryl |
| Merqury v. 2020-01-29 | Rhie et al.[50] | https://github.com/marbl/merqury |

(*Continued on next page*)

*Continued*

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| BUSCO v. 5.1.3 | Simao et al.[38] | https://gitlab.com/ezlab/busco |
| minimap2 v 2.17 | Li[51] | https://github.com/lh3/minimap2 |
| D-Genies | Cabanettes et al.[52] | https://dgenies.toulouse.inra.fr |
| bwa v. 0.7.17 | Li and Durbin[53] | https://github.com/lh3/bwa |
| samtools v. 1.14 | Li et al.[54] | https://github.com/samtools/samtools |
| PretextMap v. 0.1.8 | | https://github.com/wtsi-hpag/PretextMap |
| PretextView v. 0.2.4 | | https://github.com/wtsi-hpag/PretextMap |
| MAKER v. 3.01 | Campbell et al.[55] | https://www.yandell-lab.org/software/maker.html |
| Augustus v. 3.3 | Stanke et al.[56] | https://bioinf.uni-greifswald.de/augustus/downloads/ |
| DifCover v. 3.0.1 | Smith et al.[19] | https://github.com/timnat/DifCover https://doi.org/10.5281/zenodo.7574262 |
| LastZ v. 1.04.15 | Harris[57] | https://github.com/lastz/lastz |
| ChainNet v. 302.1 | Kent et al.[58] | https://github.com/ucscGenomeBrowser/kent |
| trinityrnaseq v. 2.11.0 | Grabherr et al.[59] | https://github.com/trinityrnaseq/trinityrnaseq |
| Orthofinder v. 2.5.2 | Emms and Kelly[60] | https://github.com/davidemms/OrthoFinder |
| T-Coffee v. 13.45.0.4846264 | Notredame et al.[61] | https://tcoffee.org/Projects/tcoffee |
| Pal2Nal v. 14 | Suyama et al.[62] | http://www.bork.embl.de/pal2nal |
| PAML v. 4.9j | Yang[63] | http://abacus.gene.ucl.ac.uk/software/paml.html |
| figTree v. 1.4.4 | Rambaut[64] | http://tree.bio.ed.ac.uk/software/figtree |
| hisat2 v. 2.2.0 | Kim et al.[65] | http://daehwankimlab.github.io/hisat2 |
| StringTie v. 2.1.5 | Perta et al.[66] | https://ccb.jhu.edu/software/stringtie |
| Other | | |
| *Drosophila melanogaster* protein sequences | FlyBase r6.30 | https://doi.org/10.1126/science.287.5461.2185 |
| *Homo sapiens* protein sequences | NCBI release 109.20190905 | https://doi.org/10.1038/35057062, https://doi.org/10.1126/science.1058040 |
| *Mus musculus* protein sequences | NCBI release 108 | https://doi.org/10.1038/nature01262 |
| *Callorhinchus milii* protein sequences | NCBI release 100 | https://doi.org/10.1038/nature12826 |
| *Danio rerio* protein sequences | NCBI release 106 | https://doi.org/10.1038/nature12111 |
| *Hydra vulgaris* protein sequences | NCBI release 102 | https://doi.org/10.1038/nature08830 |
| *Lottia gigantea* protein sequences | JGI v1.0 | https://doi.org/10.1038/nature11696 |
| *Ciona intestinalis* protein sequences | JGI v2.0 | https://doi.org/10.1126/science.1080049 |
| *Branchiostoma floridae* protein sequences | JGI v1.0 | https://doi.org/10.1038/nature06967 |
| *Nematostella vectensis* protein sequences | JGI v1.0 | https://doi.org/10.1126/science.1139158 |
| *Takifugu rubripes* protein sequences | JGI v4.0 | https://doi.org/10.1126/science.1072104 |
| *Xenopus tropicalis* protein sequences | JGI v4.1 | https://doi.org/10.1126/science.1183670 |
| *Trichoplax adhaerens* protein sequences | JGI v1.0 | https://doi.org/10.1038/nature07191 |
| UniProt/Swiss-Prot protein sequences | UniProt/Swiss-Prot | https://doi.org/10.1093/nar/gkaa1100 |
| RepBase repeat database | RepBase | https://doi.org/10.1186/s13100-015-0041-9 |
| *P. marinus* species specific repeats | | https://doi.org/10.1038/s41588-017-0036-1 |

## RESOURCE AVAILABILITY

### Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Jeramiah Smith (jjsmit3@uky.edu).

### Materials availability

This study did not generate new unique reagents.

### Data and code availability

- Data: Sequence data for *Petromyzon marinus* and *Entosphenus tridentatus* have been deposited at NCBI and GenomeArk (https://genomeark.github.io/genomeark-all/Petromyzon_marinus) and are publicly available as of the date of publication. The assembly of the *P. marinus* germline genome is deposited under BioProject PRJNA562011 in NCBI (accession number GCF_010993605.1 for the primary assembly and GCA_010993595.1 for the alternate contig only assembly). Comparative sequencing datasets for *P. marinus* are deposited under BioProject PRJNA779416. Comparative sequencing datasets for *E. tridentatus* are deposited under BioProject PRJNA784541. RNA Sequence data for *Geotria australis* have been deposited at the Aotearoa Genomic Data Repository (AGDR) and are publicly available as of the date of publication under accession number TAONGA-AGDR00015 (https://data.agdr.org.nz/study-viewer/project/AGDR00015). Accession numbers are also listed in the key resources table. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.
- Code: All original code (an updated version of DifCover) has been deposited at GitHub and is publicly available as of the date of publication. The DOI is listed in the key resources table.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

## EXPERIMENTAL MODEL AND SUBJECT DETAILS

### Organisms used as source material

Petromyzon marinus, adult, male, kPetMar1, sampled under protocol number 2011-0848 (University of Kentucky Institutional Animal Care and Use Committee).

Petromyzon marinus, adult, male, Male7219, sampled under protocol number 2011-0848 (University of Kentucky Institutional Animal Care and Use Committee).

*Entosphenus tridentatus*, adult, male, EtrAdultCRITFC18_GenomeStudy-0004, sampled under a 2018 Yakama Nation scientific collector's permit to RL.

*Geotria australis*, adult, male, TAONGA-AGDR00015M1M, sampled under the Ministry for Primary Industries, Fisheries New Zealand Special Permit 666/2.

## METHOD DETAILS

### Genome sequencing and assembly

The genome was sequenced and assembled using the VGP 1.6 pipeline.[36] An adult male sea lamprey was captured from the Great Lakes, and testis was dissected and immediately flash frozen, and stored at the University of Kentucky (BioSample SAMN12629506). Spermatogenic (meiotic) testis was chosen for this project because it permitted the preparation of high molecular weight DNA and Hi-C libraries from the same individual under the anticipation meiotic germline is likely to yield more informative patterns of chromatin contact in comparison to highly condensed spermatocyte nuclei. Ultralong DNA molecules were isolated (>300Kb), and sequencing conducted with PacBio continuous long reads (CLR) on a Sequel I at (62.36X coverage, ~40kb insert size), Bionano optical maps on a Saphyr (538.18X coverage), and 10X Genomics linked reads (67.01X coverage) and Arima Genomics Hi-C linked reads (70X coverage) on an Illumina NovaSeq at the Rockefeller University Vertebrate Genomes Lab.

Following initial assembly and haplotype purging using FALCON (v. DNANexus 1.9.0),[43] FALCON-Unzip (v. DNANexus 1.0.6)[44] and purge_dups (v. github ca23030ccf4254dfd2d3a5ea90d0eed41c24f88b\),[45] the primary (longer) set of contigs were scaffolded sequentially using 10x linked reads with scaff10x (v. 4.1.0; https://github.com/wtsi-hpag/Scaff10X), Bionano cmaps with Bionano Solve DLS (v. 3.2.1), and Hi-C linked reads with Salsa (v. 2.2).[46] The primary assembly base calls were then error corrected (polished) and scaffolds gap-filled with using the original CLR with Arrow smrtanalysis (v. smrtlink_6.0.0.47841), and further polished with the 10x short reads and longranger align (v. 2.2.2) and freebayes Illumina polishing (v. 1.3.1).[47] Manual curation including decontamination was conducted at the Sanger Institute using gEVAL (v. 2019–12–09)[48] as previously described.[67] Manual curation issued 440 structural changes leading to an increase of scaffold N50 by 16%, a reduction in scaffold number by 13% and a decrease of assembly

size by 3% due to removal of retained haplotypic duplication. Of the resulting assembly, 92.3% could be assigned to 85 identified chromosomes. The assembly was submitted to the public NCBI archives, and assigned to a BioProject (PRJNA562011).

### Assembly quality metrics for lamprey genomes

Meryl v1.1 (https://github.com/marbl/meryl)[49] and Merqury (2020-01-29)[50] were used for error rate estimation and calculation of percentage of short reads k-mers, k=21, found in the assemblies (Table S1). To assess assembly completeness, we searched for single copy orthologs that are conserved across all metazoans (lineage dataset metazoan_odb10, n = 954) and core vertebrates (n = 233) using BUSCO pipeline v. 5.1.3 in mode ''genome'' with gene predictor ''metaeuk''. Assessment of largescale chromosome structure used chromatin contact data and a second sea lamprey genome assembly that was generated from independent sequence/mapping datasets with no overlap to the animal or datasets used to generate this assembly.[19] The assemblies were aligned to one another using minimap2 (v 2.17)[51] and alignments were displayed using D-Genies.[52] Chromatin contact maps were generated by first aligning Hi-C libraries from the same meiotic testes (NCBI BioProject PRJNA562011) to the current and previous genome assemblies with bwa (v. 0.7.17 )[53] and filtered using samtools (v. 1.14)[54] to include alignment scores greater than or equal to ten. Chromatin contact densities were calculated and summarized using PretextMap (v. 0.1.8: https://github.com/wtsi-hpag/PretextMap) and visualized using PretextView (v. 0.2.4: https://github.com/wtsi-hpag/PretextMap).

### Gene annotation

NCBI annotation with 78 RNAseq data sets from various tissues (brain, olfactory organ, liver, and whole embryos) yielded 22,167 genes (17,580 protein coding and 4,361 non-coding) contributing to 43,324 transcripts (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Petromyzon_marinus/100/). An additional custom annotation of *P. marinus* and *E. tridentatus* germline assemblies was performed using the MAKER genome annotation pipeline using published MAKER annotation protocols.[55] (Basic Protocols 1 and 5 as well as Support Protocols 1, 2, 3, and 4). MAKER was configured to use Augustus for gene prediction, RepBase supplemented with a species-specific library for repeat masking, assembled mRNA-seq datasets for transcriptome evidence, and whole proteomes of multiple animal species and all of UniProt/Swiss-Prot for protein evidence. Transcriptome evidence from lamprey (SIMRbase: https://simrbase.stowers.org) and protein datasets from multiple organisms (Key Resources) and repeat annotations (Key Resources) were used as evidence for *de novo* annotation following published MAKER annotation protocols.[55] Augustus[56] was trained using MAKER generated alignments of the Swiss-Prot protein dataset against each assembly followed by a single round of bootstrap training (MAKER Support Protocol 1). Gene predictions that were rejected by MAKER were added to the final annotation set if they contained identifiable InterPro Protein domains (MAKER Basic Protocol 5). For *P. marinus*, new annotations were matched to previous PMZ_v3.1 annotations by first mapping the earlier annotation set to the new assembly (MAKER Support Protocol 4) and then identifying model overlap to new genome annotations.

### Identification of germline-specific regions

Germline specific regions of sea lamprey assembly were identified using ~52X coverage in germline (sperm) reads (413 million 150 bp Illumina NovaSeq 6000 read pairs) and ~96X coverage in somatic (blood) reads (486 million 150 bp Illumina read pairs). Sequences were aligned to the genome assembly using BWA-mem (v 0.7.17)[53] with option -a and filtered by samtools view[54] with option -F2308, such that only primary alignments were retained for further analysis. The resulting files were processed using DifCover (v 3.0.1)[19] to calculate the degree of germline enrichment across all discontiguous 500bp intervals of low-copy sequence using modal coverages for sperm and blood, low coverage masking of regions with read depth <1/3X in both samples and high coverage masking of sequences with read depth >3X modal coverage in both samples. To identify germline-specific genes that are present at higher copy number, we ran DifCover using low coverage masking with read depth <10X in both samples and high coverage masking of sequences with read depth >30X modal coverage.

Short read sequences of *E. tridentatus* sperm (567 million 150 bp Illumina read pairs) and blood (463 million 150 bp Illumina read pairs) DNA were aligned to the assembly (GCA_014656915.2) with BWA-mem and filtered by samtools view with option -F2308 to retain only primary alignments. Read pairs duplicates were removed with samtools markduplicates and properly paired reads (344 million pairs in sperm and 274 million pairs in blood) were selected with samtools view -f2 yielding modal coverages for sperm and blood of 106X and 77X respectively. Relative coverage of DNA sequence from sperm and blood (standardized $\log_2$ ratio) was estimated using DifCover as described above for sea lamprey. All tracks are publicly accessible as a browser track labeled ''Germline-specific Regions'' on SIMRbase.

### Alignment of G1 to somatic chromosomes

To identify regions of homology between chromosome G1 and somatic chromosomes, ChrG1 scaffolds were first aligned to all other chromosomal scaffolds using LastZ (v. 1.04.15).[57] Alignment scoring files and alignment parameters match those used for the generation of high-divergence vertebrate genome alignments hosted at the UCSC browser (–scores=scoring_file –inner=2000 –hspthresh=2200 –gappedthresh=6000 –ydrop=3400 –masking=50 –notransition –chain –gapped). Generation of alignment nets via ChainNet (v. 302.1) also followed the methods used to generate UCSC deep vertebrate alignment tracks (-linearGap=loose) except that chain score cutoffs were not implemented after examining the impact of chain score cutoffs of 2500 and 5000[58] (these have minimal impact on alignment tracks but omit some small extensions of syntenic chains).

### Geotria transcriptome assembly

A transcriptome for *G. australis* was generated using RNAseq data from the testes of a nest guarding male and 40 embryos (similar to Tahara stage 21[68]). Animals and embryos used for this aim were collected under the Ministry for Primary Industries, Fisheries New Zealand Special Permit 666/2. RNA was extracted from RNAlater-preserved testes tissue and whole embryos using a Zymo Research Direct-zol RNA Miniprep Plus kit following the kit protocol (v.1.0.1) with some modifications. Approximately 50mg of testes (minced with a sterile razor) and intact (whole) embryos were used as starting tissue material. Excess RNAlater was removed from the tissues prior to the tissue lysis step by blotting (Kimwipe) and then rinsing in ice-cold PBS solution for ~5 seconds. Tissue lysis methods varied by tissue sample and included dry ice freezing with crushing (mortar and pestle), handheld homogenizing, and bead bashing (20 freq for 2–6 minutes). An optional DNAase 1 treatment was performed following the kit protocol. DNAase/RNA-Free Water was used to elute 26–50 ml of RNA. The extracted RNA was tested for quality and quantity with a NanoDrop spectrophotometer and Qubit Fluorometer and stored at -80 prior to library preparation. Subsampling, tissue lysis, and RNA purification steps were performed on ice and in a chilled centrifuge (~4° C). In addition, all surfaces, hood-space, and centrifuges were wiped with ethanol and RNaseZap before, and during, each extraction batch.

RNA extractions were evaluated using a bioanalyzer to assess the quality of the samples. Samples were considered partially degraded (RNA integrity number [RIN] < 7); however, a preliminary sequencing analyses demonstrated success for lamprey embryos of this level of reported RIN and additional external quality control was performed during the downstream analyses. Library preparation was performed by the Otago Genomics Facility using the Illumina TruSeq Stranded Total RNA library and TruSeq Stranded Total RNA Gold rRNA depletion to remove downstream inhibiting excess eukaryotic cytoplasmic rRNA, mitochondrial rRNA, and globin mRNA. Sequencing was performed by the Roy J Carver Biotechnology Center at the University of Illinois using S4 a NovaSeq 6000 (S4) to generate paired-end reads (2x150nt). The resulting reads were assembled using trinityrnaseq-v2.11.0.[59]

### Construction and analysis of gene trees

Gene trees were built from a set of species chosen on the basis of several factors that dictate their utility in dissecting the evolution of germline-specific genes. First, in comparison to the *P. marinus* genome, sequence data from *E. tridentatus* and *G. australis* span key ancestral nodes within the lamprey phylogeny.[21,31] Second, available germline and somatic sequence datasets for *P. marinus* and *E. tridentatus* provide the information necessary to assign sequences to germline-specific vs somatic compartments, whereas these data are absent for closely related *Lethenteron* species necessitating their exclusion from this analysis. Finally, two (presumable non-rearranging) gnathostome outgroups (human and gar) were used to aid in the definition of ancestral states in the lamprey lineage. Datasets used for these species included annotated genes for *P. marinus* (https://genomes.stowers.org/sealamprey),[19] parallel gene annotations for *E. tridentatus* (https://genomes.stowers.org/pacificlamprey), a nonredundant set of transcripts for *G. australis* (the highest expressed isoforms from the RNAseq assembly above), human gene annotations (GRCh38.p13 Ensembl genebuild V104.38) and spotted gar gene annotations (LepOcu1, Ensembl genebuild V104.1). Protein sequences from these genes were used to identify orthology groups using Orthofinder (v. 2.5.2).[41,60] Trees containing germline-specific genes were extracted for manual curation (collection of missing orthologs and pruning of excessively long branches due to misannotations) and further analysis. Orthology groups were realigned using the PSI-Coffee module of T-Coffee (v. 13.45.0.4846264)[61,69] then integrated with transcript sequences to generate codon alignments in PAML format using Pal2Nal (v 14).[62] The resulting alignments and trees were further analyzed using PAML[63,70] to estimate substitutional rates and likelihood statistics for three models: model 0 (Model =0 NSsites=0), model 1a (Model =1 NSsites= 0), and clade model D (Model = 3 NSsites =3) that was used to test whether rates in each clade of germline-specific genes differed their somatically retained homologs. P-values for the test of significance of the clade model (indicating differences in substitutional rates for germline vs somatic genes) used the convention that sampling probabilities for two times the difference in the likelihood ratio statistics between model 1a and model D are approximated by the $\chi^2$ distribution. Tree visualization and labeling was performed using figTree v1.4.4.[64] For several trees it is not possible to justify a specific root, even ignoring issues of cyclostome/gnathostome duplication history (e.g. MYCN / AHRGAP5) although it is still possible to resolve the relative branching patterns within lamprey lineages on the gene tree. As such, we present unrooted trees to avoid misleading the reader with respect to gnathostome/lamprey relationships.

### Reanalysis of RNAseq data

Published RNAseq datasets (PRJNA306044, SRP009181) were aligned to the genome assembly using hisat2-2.2.0[65,71] and the resulting sam files were filtered to extract single best alignments using samtools view (v. 1.11)[54] then converted to bam using samtools sort. Filtered and sorted alignments were processed using StringTie v2.1.5[66,72] to generate FPKM estimates.

## QUANTIFICATION AND STATISTICAL ANALYSIS

Meryl v1.1 (https://github.com/marbl/meryl)[49] and Merqury (2020-01-29)[50] were used for error rate estimation and calculation of percentage of short reads and k-mers, k=21, found in the assemblies (Figure 1, Table S1).

DifCover (v 3.0.1)[19] was used to calculate normalized enrichment statistics for germline sequence data relative to somatic sequence data, accounting for differences in sequence modal coverage (Figure 2, Table S3).

PAML[63,70] was used to estimate substitutional rates and likelihood statistics for substitution models and P-values were calculated as recommended in PAML user documentation (Figure 3, Table S4).

StringTie v2.1.5[66,72] to generate fragments per kilobase of exon per million mapped fragments (FPKM) estimates (Figure 4).

## ADDITIONAL RESOURCES

SIMRbase sea lamprey genome browser: https://simrbase.stowers.org/sealamprey.

SIMRbase Pacific lamprey genome browser: https://simrbase.stowers.org/pacificlamprey.