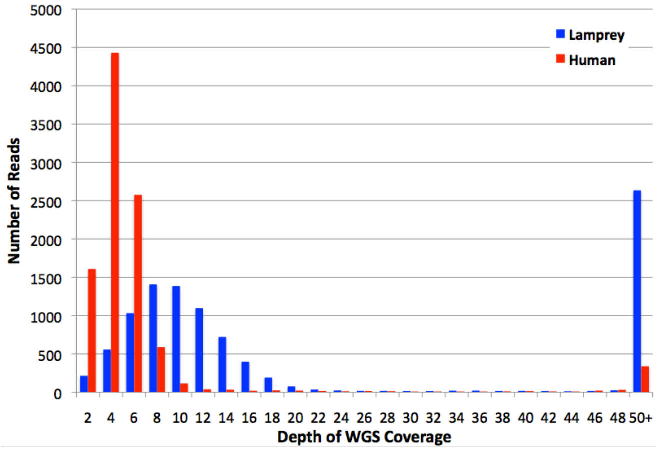
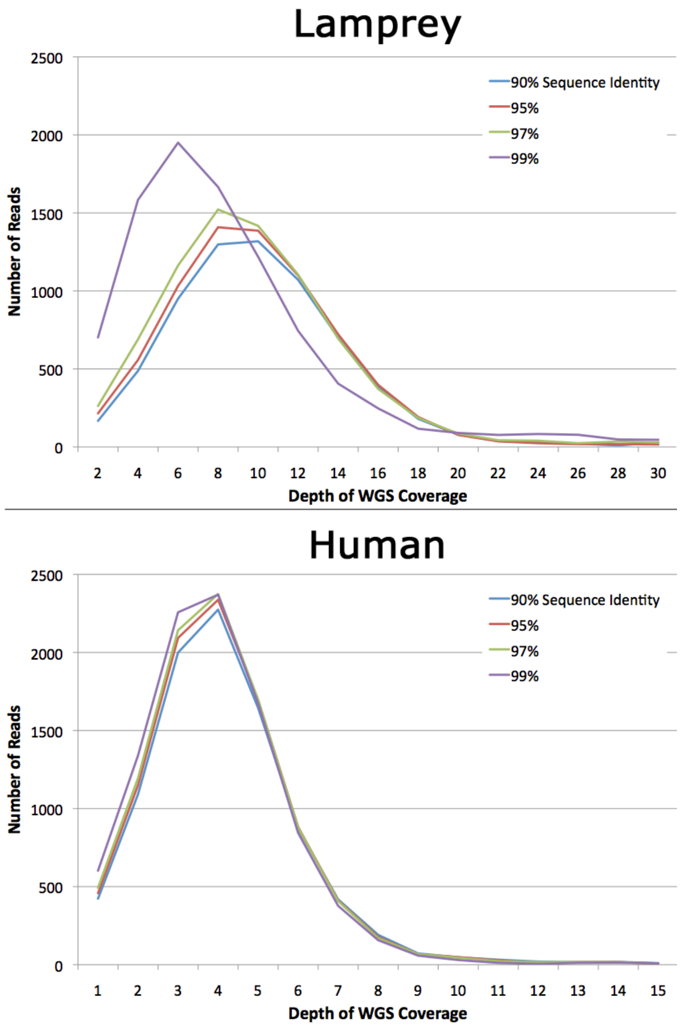
Supplementary Figures

Supplementary Figure 1 – Histogram of read coverage depths for lamprey and human WGS sequencing projects.

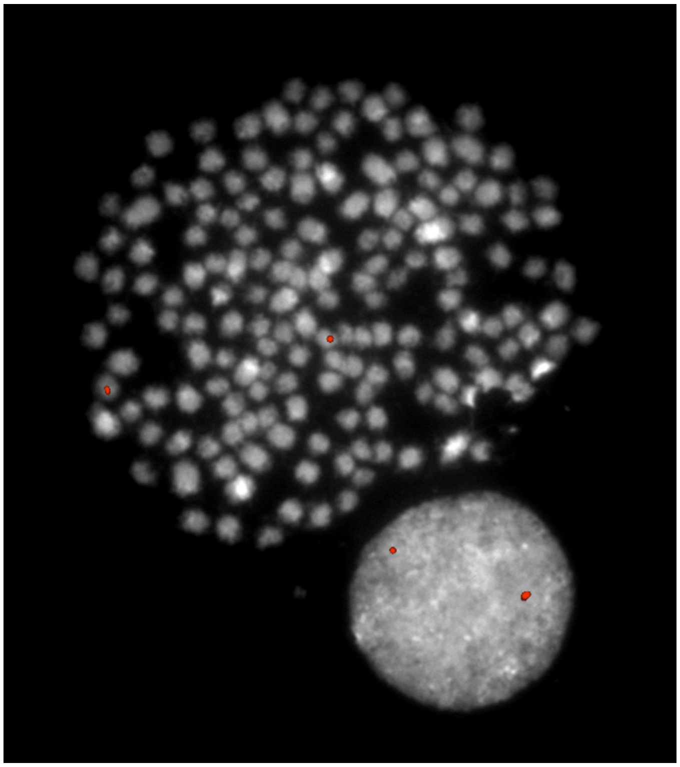


Nature Genetics: doi:10.1038/ng.2568

Supplementary Figure 2 – Distribution of read coverage depths for lamprey and human genomes, considering various thresholds of sequence identity.

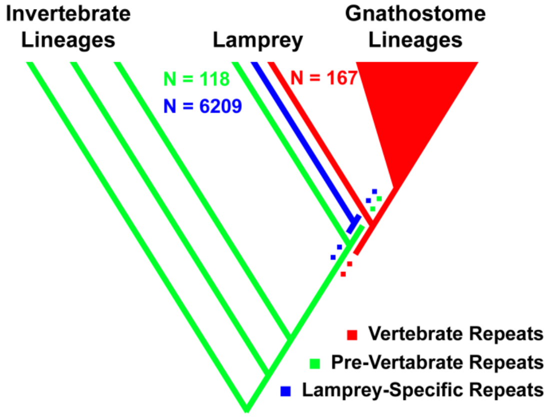


Supplementary Figure 3 – Hybridization of a predicted single-copy BAC (PMAY-25E18) to lamprey somatic metaphase and interphase nuclei. Hybridization patterns are consistent with this ~100 kb region being present at a diploid copy state.

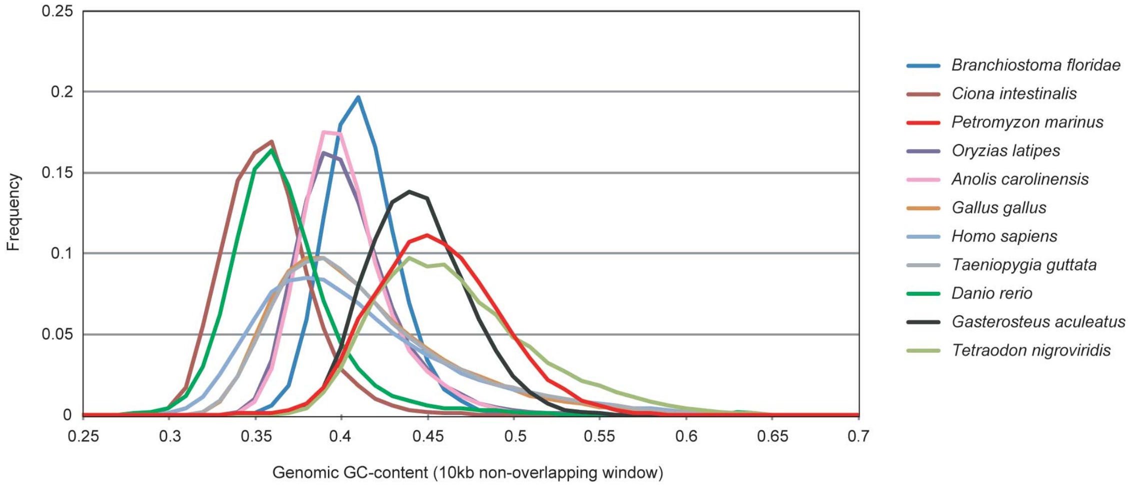


Nature Genetics: doi:10.1038/ng.2568

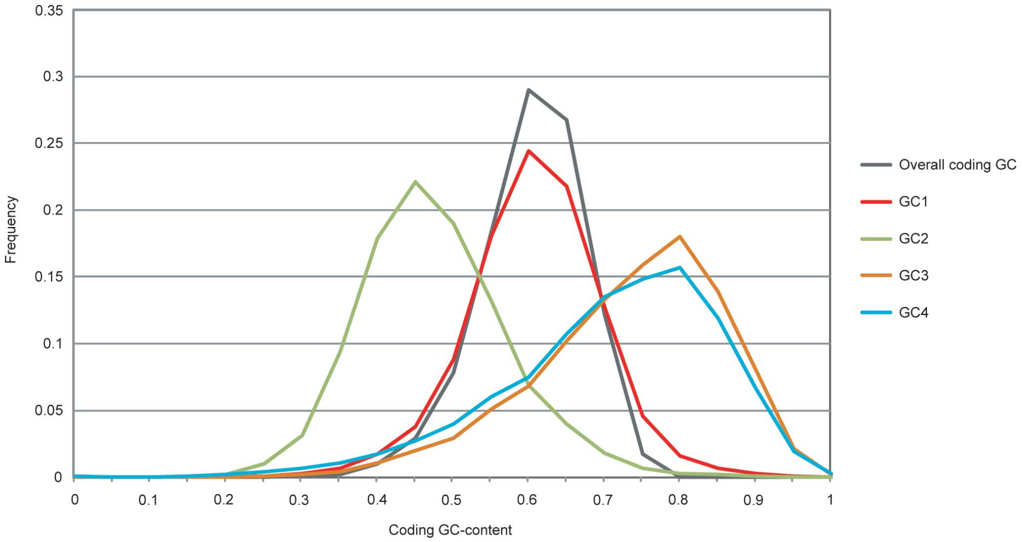
Supplementary Figure 4 - Phylogenetic distribution of lamprey repetitive elements. The repetitive fraction of the lamprey genome is a composite of vertebrate-specific elements, more anciently derived elements that have not been identified in other vertebrate lineages and multiple elements that are apparently unique to the lamprey lineage. The phylogenetic distribution of repetitive elements reflects the ancient shared ancestry of lamprey and gnathostome lineages and a billion years of independent evolution (2x 500 million years) subsequent to the lamprey/gnathostome split.



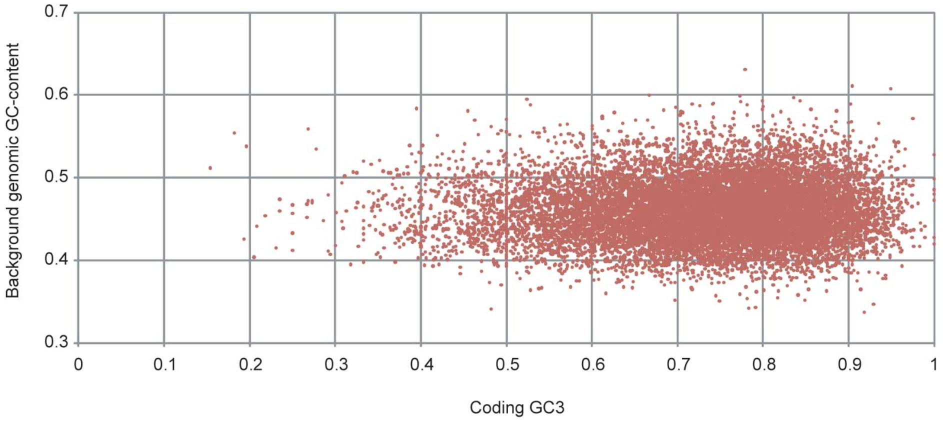
Supplementary Figure 5 - Global genomic GC-content of the lamprey and other chordates. Non-masked genome sequences, including coding regions, were cut into 10 kb fragments. In our analysis, because of shorter scaffold lengths, we employed 10 kb windows instead of 20 kb often employed in previous reports [1,2]. Global GC-content of individual genomic fragments was computed and shown in the histogram. Fragments less than 10 kb and those with more than 10% of 'N' in length were discarded. No distinction was made between coding (only 2%) and non-coding regions.



Supplementary Figure 6 - GC-content of different codon positions of lamprey protein-coding genes. All predicted genes were included.

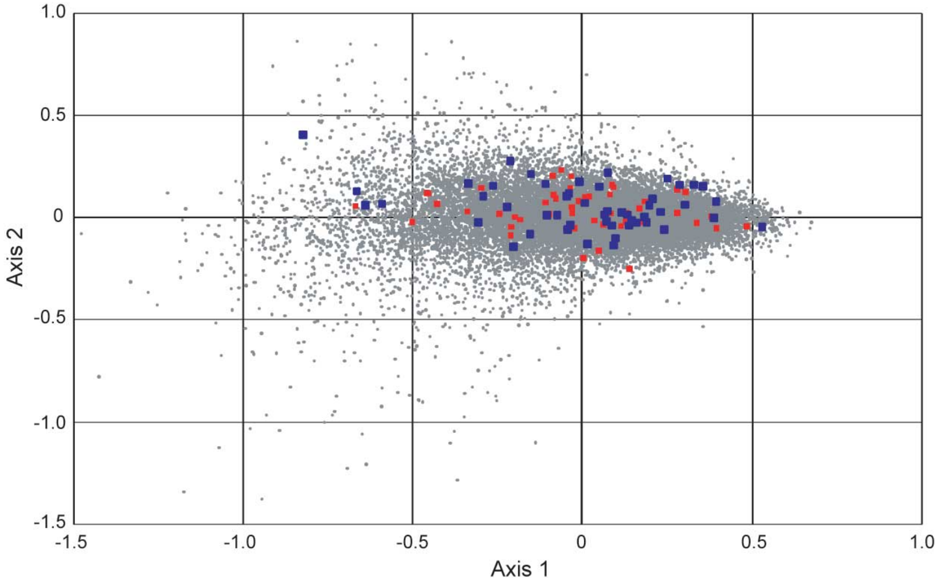


Supplementary Figure 7 - Gene-by-gene comparison of coding GC3 and background genomic GC-content in the lamprey genome. Among the 10 kb genomic sequences prepared for Supplementary Figure 5 (n = 49,918), those containing at least one protein-coding gene were selected. For each protein-coding gene, GC3 and genomic GC-content of the 10 kb fragment harboring the gene(s) are computed and plotted. The distribution of genomic GC-content for selected scaffolds (n = 12,250) did not differ significantly from that of the entire assembly (data not shown).

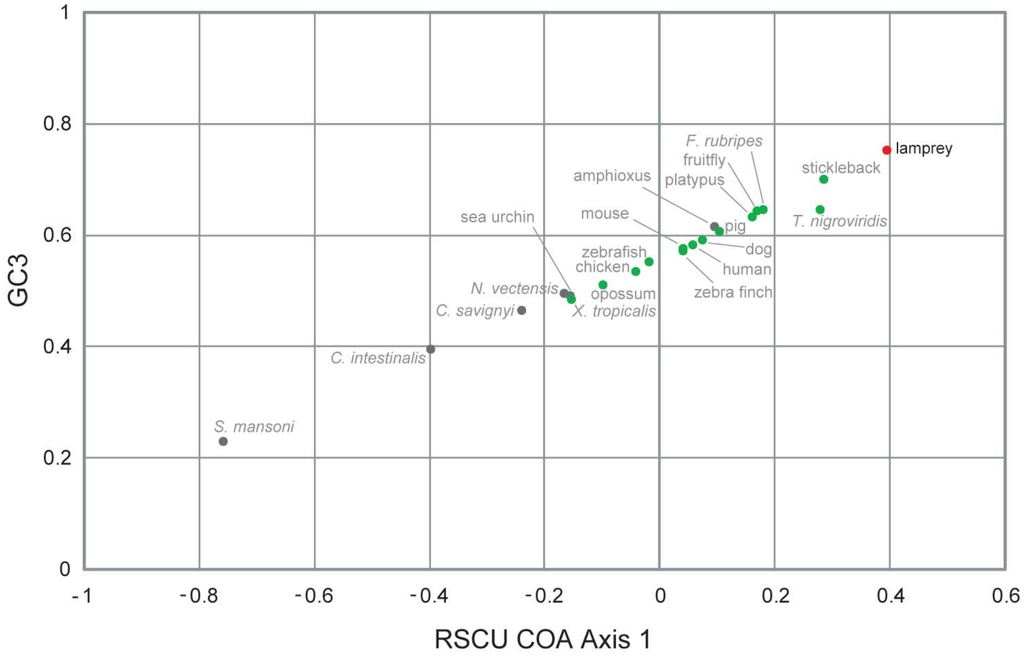


Nature Genetics: doi:10.1038/ng.2568

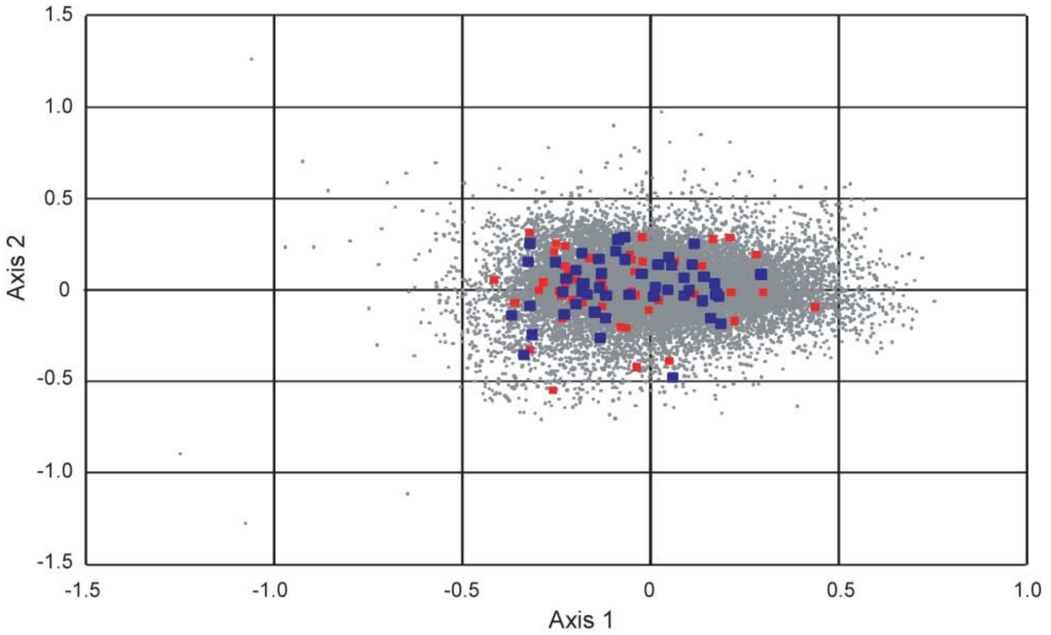
Supplementary Figure 8 - Variation of relative synonymous codon usage (RSCU) among lamprey genes. Individual protein-coding genes are plotted along the first two principal axes generated by correspondence analysis (COA) on RSCU values. The 50 highest expressed and 50 lowest expressed genes are shown as red and blue squares, respectively.



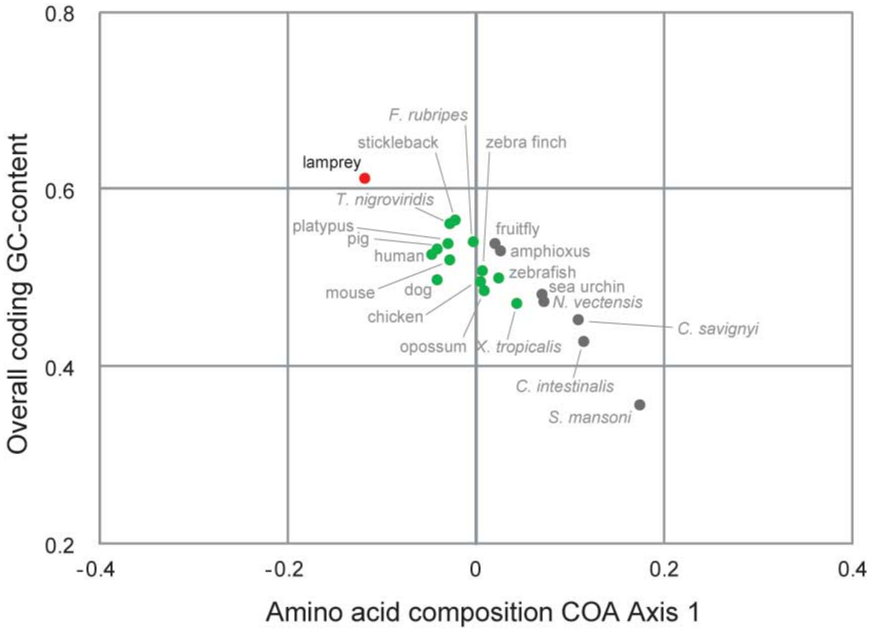
Supplementary Figure 9 - Plot of diverse animals along axis1 in RSCU correspondence analysis and GC3. Axis1 in this plot is identical to Axis1 in Figure 2, Panel A. Red: lamprey. Grey: invertebrates. Green: jawed vertebrates.



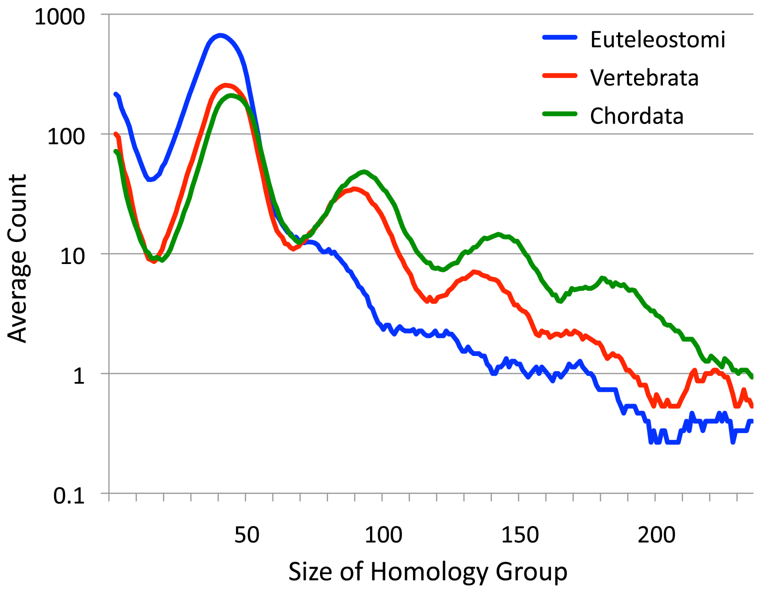
Supplementary Figure 10 - Variation of amino acid compositions among lamprey genes analyzed with COA. Individual protein-coding genes are shown as dots along axes 1 and 2. The 50 most highly expressed genes are shown in red, and the 50 most lowly expressed genes are in blue.



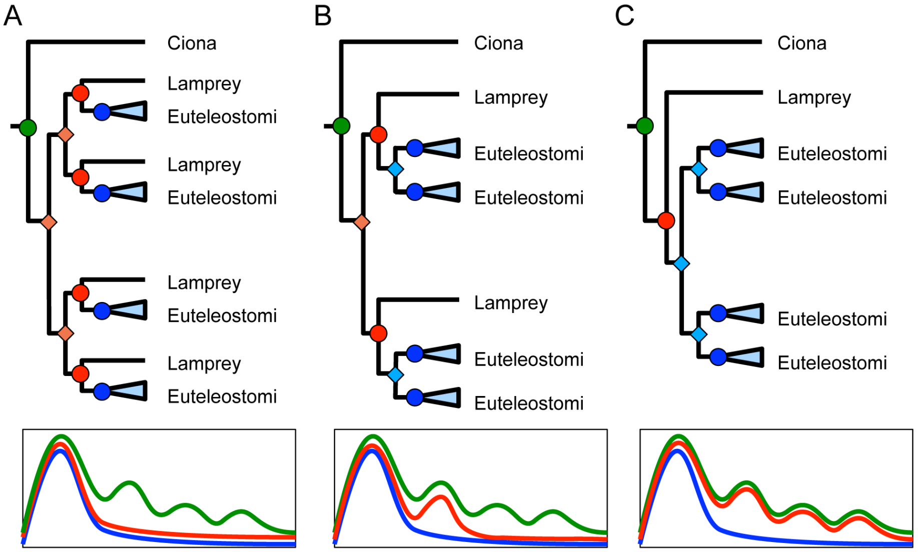
Supplementary Figure 11 - Plot of Axis1 in the correspondence analysis of amino acid composition against overall protein-coding GC-content. Axis1 in this plot is identical to Axis1 in Figure 2, Panel B. Red: lamprey. Grey: invertebrates. Green: jawed vertebrates.



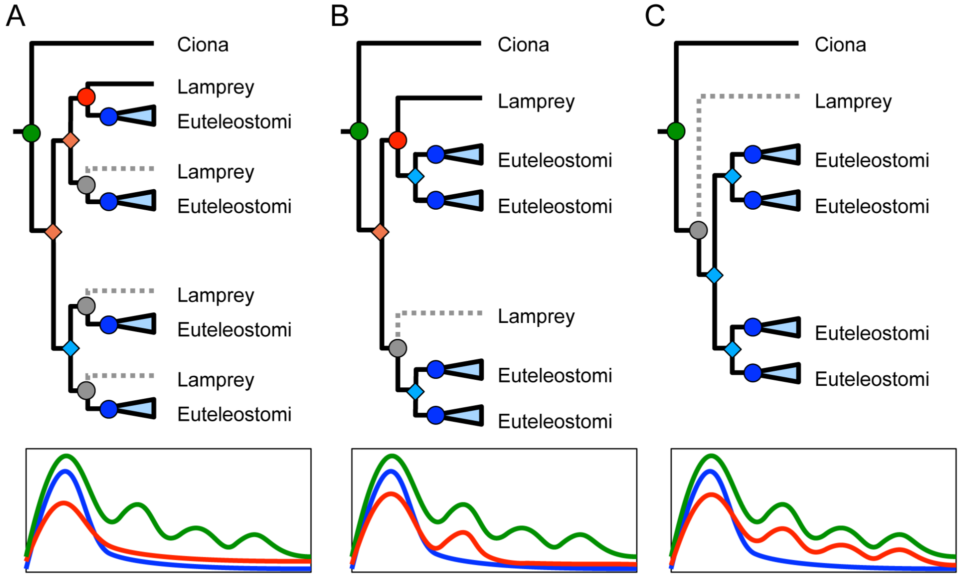
Supplementary Figure 12 – Observed distribution of sizes of homology groups that share a most recent common ancestor within ancestral taxonomic groups. Quadrimodal distributions are indicative of two rounds of whole genome duplication.



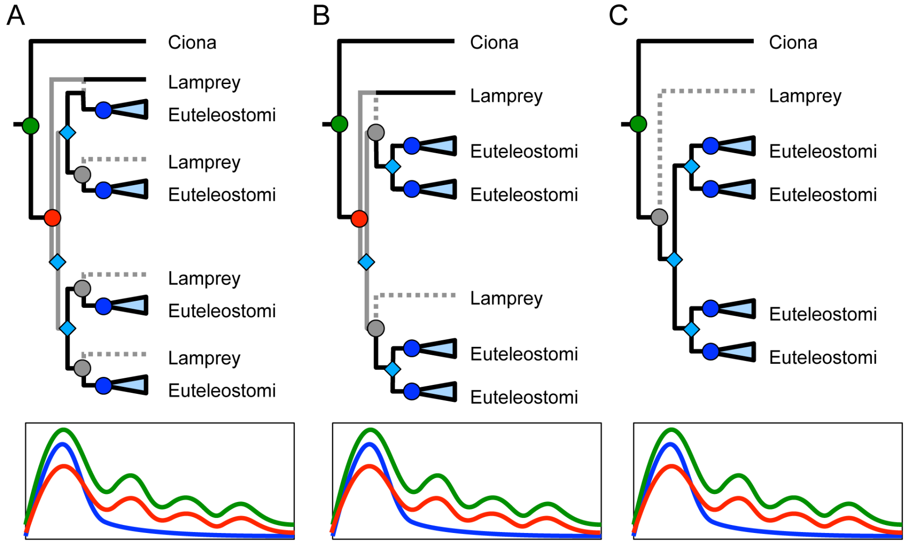
Supplementary Figure 13 – Expected distributions of the sizes of the homology groups, depending on the position of the gnathostome/lamprey split (A: after 2R, B: between 1R and 2R, C: before 2R). The distribution at a given node after 2R should exhibit with mode nE (the number of species in the Euteleostomi clade) (blue curve). If the node is between the 2R, a second mode at 2·nE should be present (red curve, in B). If the node is before the 2R, two additional modes (3·nE and 4·nE) should be visible (green curve).



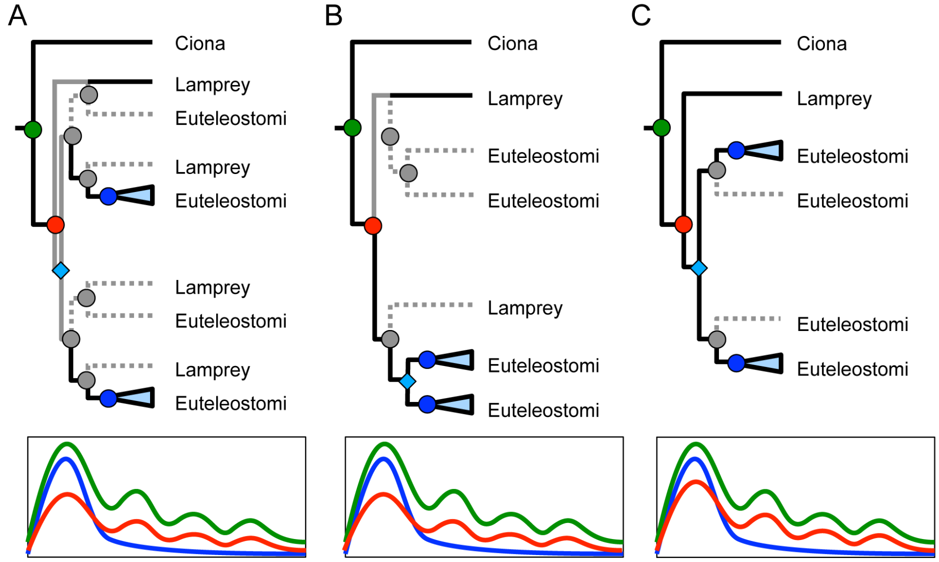
Supplementary Figure 14 - Expected distributions of the sizes of the homology groups, taking into account missing annotations in the lamprey genome. The only difference is for the Vertebrata curve (in red): the levels are expected to be lower (less Vertebrata speciation nodes) but the shape of the distribution would remain identical to the default one (Supplementary Figure 13).



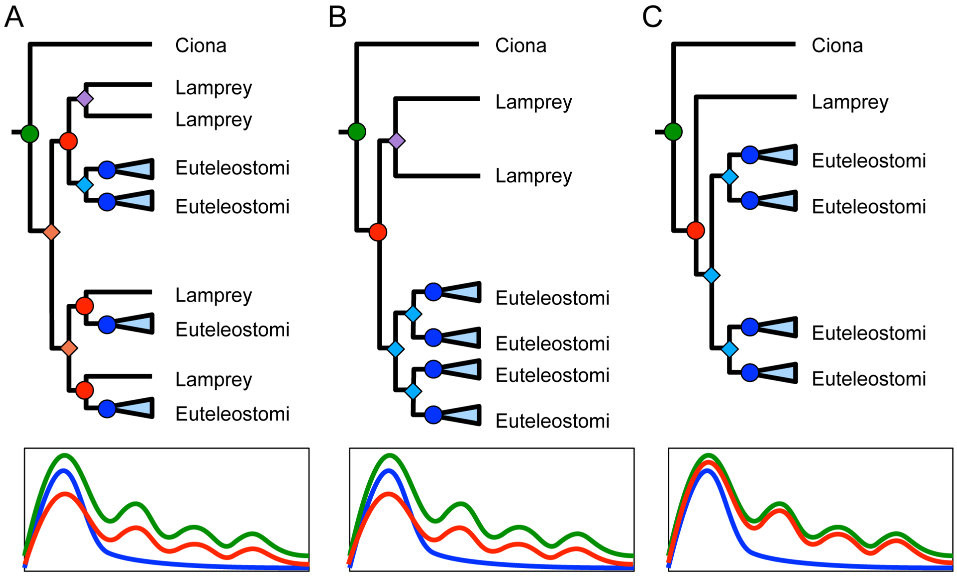
Supplementary Figure 15 - Expected distributions of the sizes of the homology groups, taking into account missing annotations in the lamprey genome, and TreeBeST reconstructions. TreeBeST will favor topologies with the lamprey genes as outgroups of Euteleostomi-specific duplications. The lamprey would be virtually positioned as an outgroup to the 2R event, similar to *Ciona*, regardless of the true position of the lamprey / gnathostome split. The Vertebrata curve (in red) would then be similar to the Chordata curve (in green) in all three scenarios.



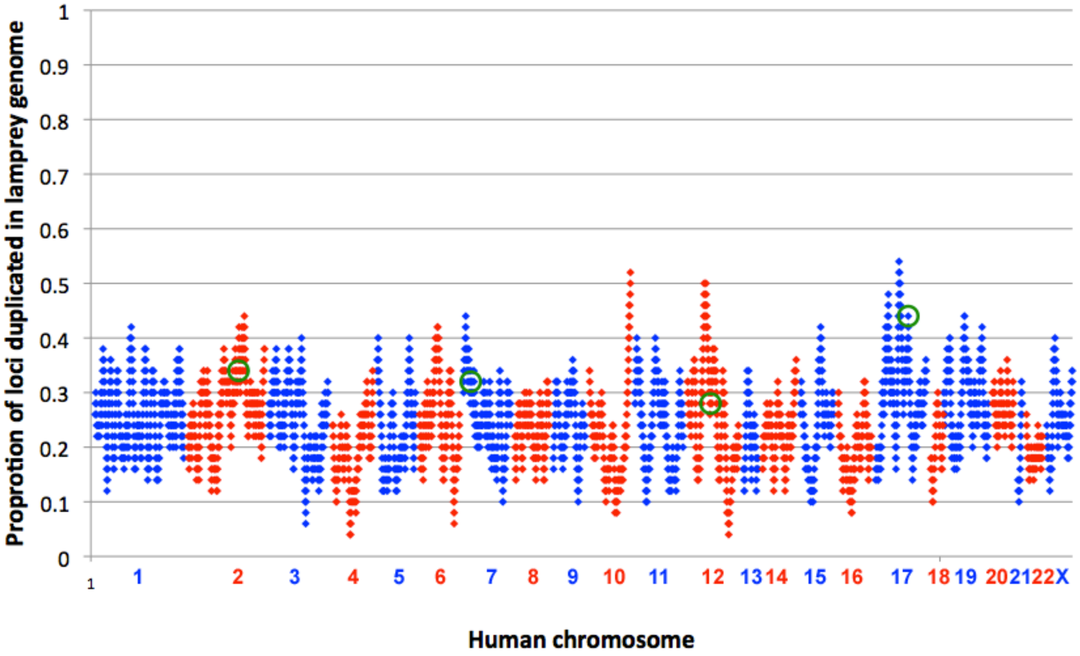
Supplementary Figure 16 – Expected distributions of the sizes of the homology groups, taking into account differential gene losses, and TreeBeST reconstructions. The effect is very similar to those shown in Supplementary Figure 15: TreeBeST will group the remaining Euteleostomi families on one branch, and lamprey genes on the other. The Vertebrata homology groups would then encompass all the paralogous Euteleostomi families (all three scenarios).



Supplementary Figure 17 - Possible distributions of the sizes of the homology groups, taking into account a possible long-branch-attraction effect of the lamprey genes. This would lead to lamprey genes being grouped together, as outgroups of Euteleostomi-specific duplications. Again, the Vertebrata curve (in red) would be similar to the Chordata curve (in green) in the three scenarios.

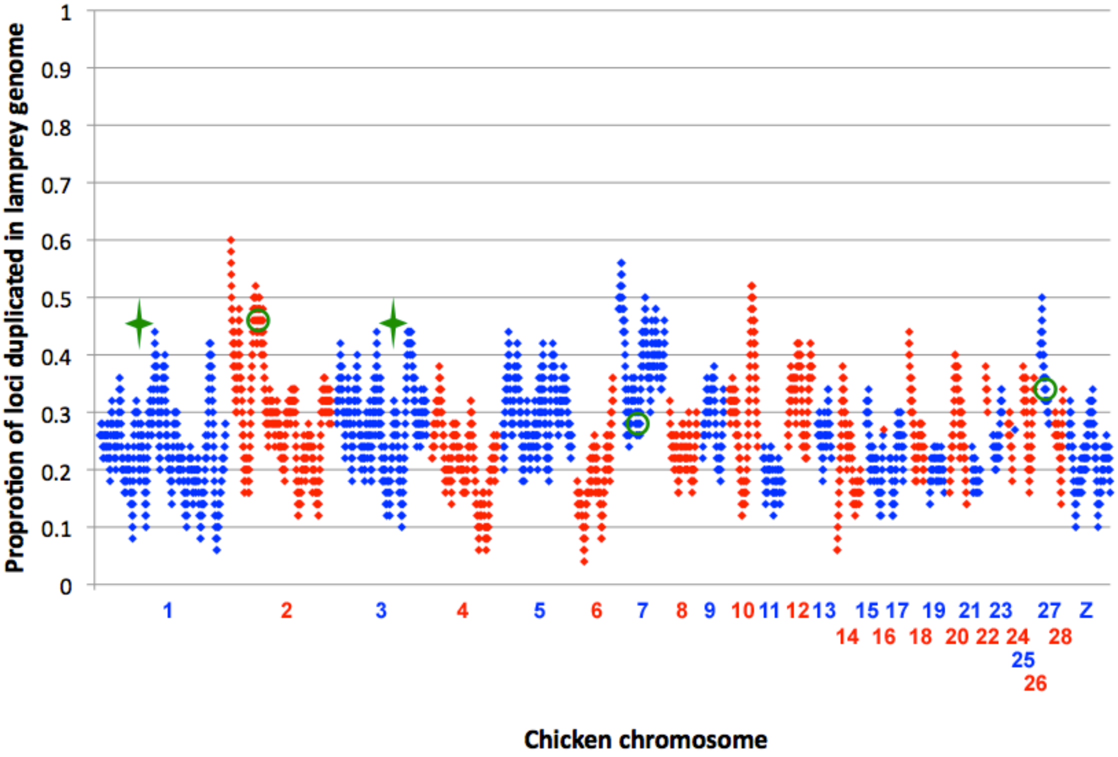


Supplementary Figure 18 – Plot of the frequency of lamprey paralogous duplications, relative to their orthologous locations in the human genome. Individual points show the frequency of duplicated genes within sliding windows of 50 orthologous loci. Green circles show the position of Hox clusters.

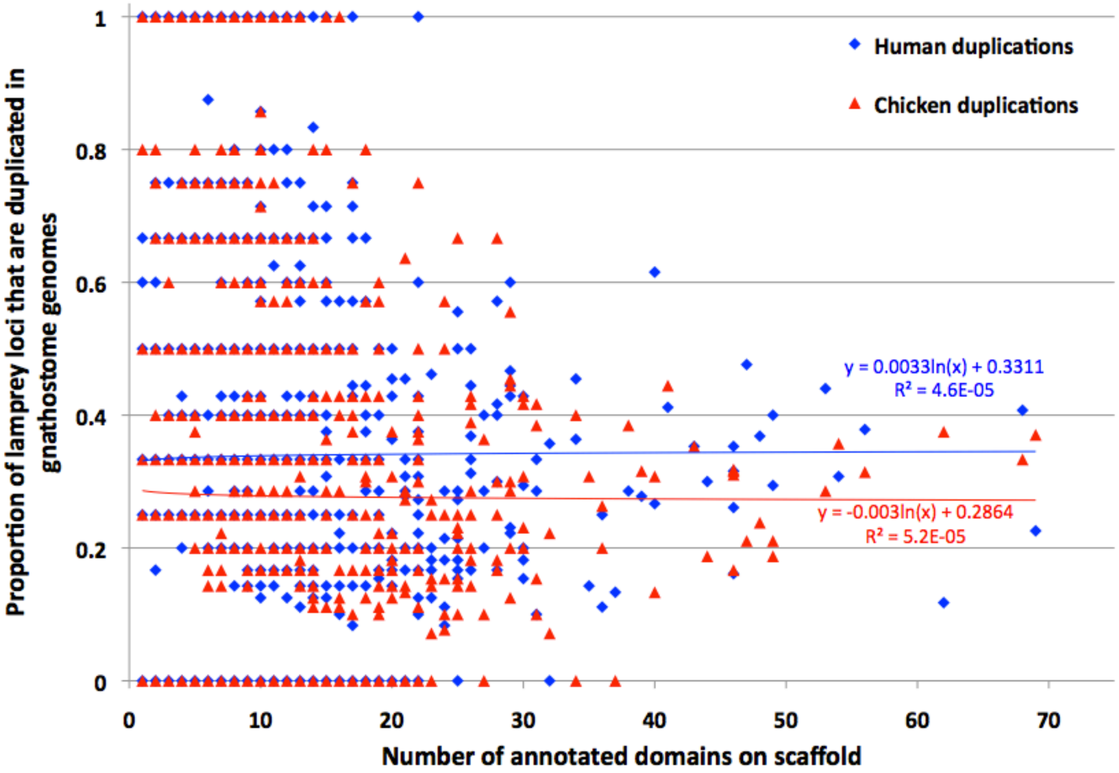


Nature Genetics: doi:10.1038/ng.2568

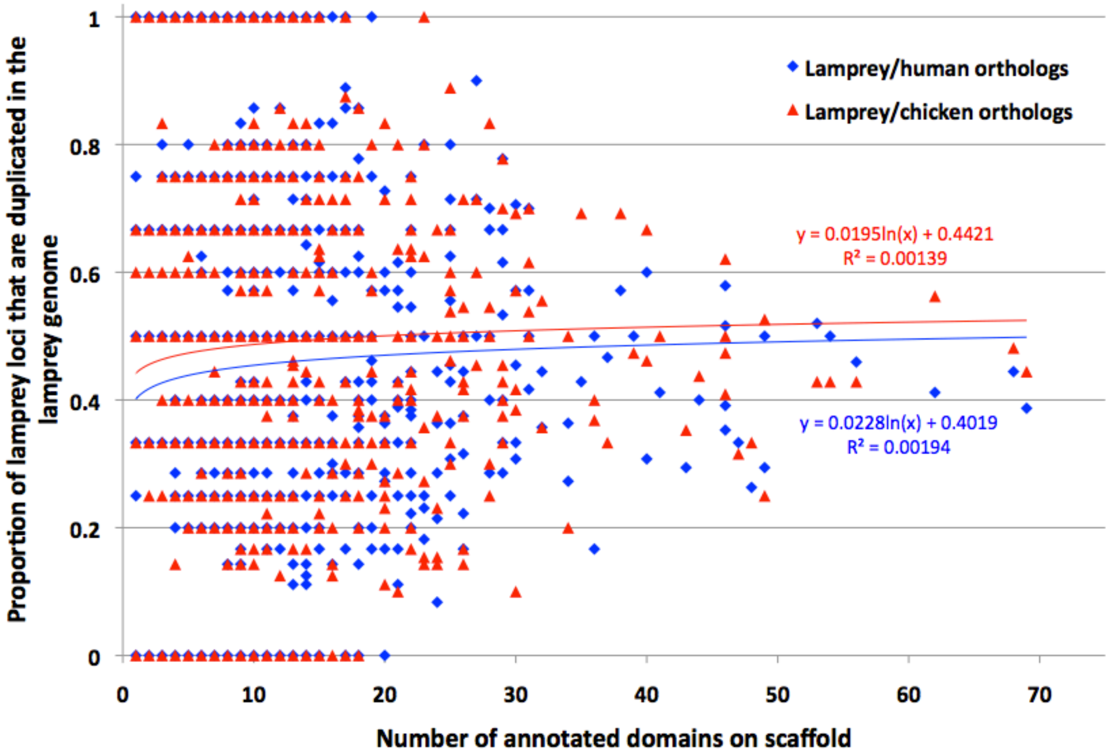
Supplementary Figure 19 - Plot of the frequency of lamprey paralogous duplications, relative to their orthologous locations in the chicken genome. Individual points show the frequency of duplicated genes within sliding windows of 50 orthologous loci. Green circles show the position of assembled Hox clusters, green crosses show approximate locations of Hox clusters that have not yet been assembled.



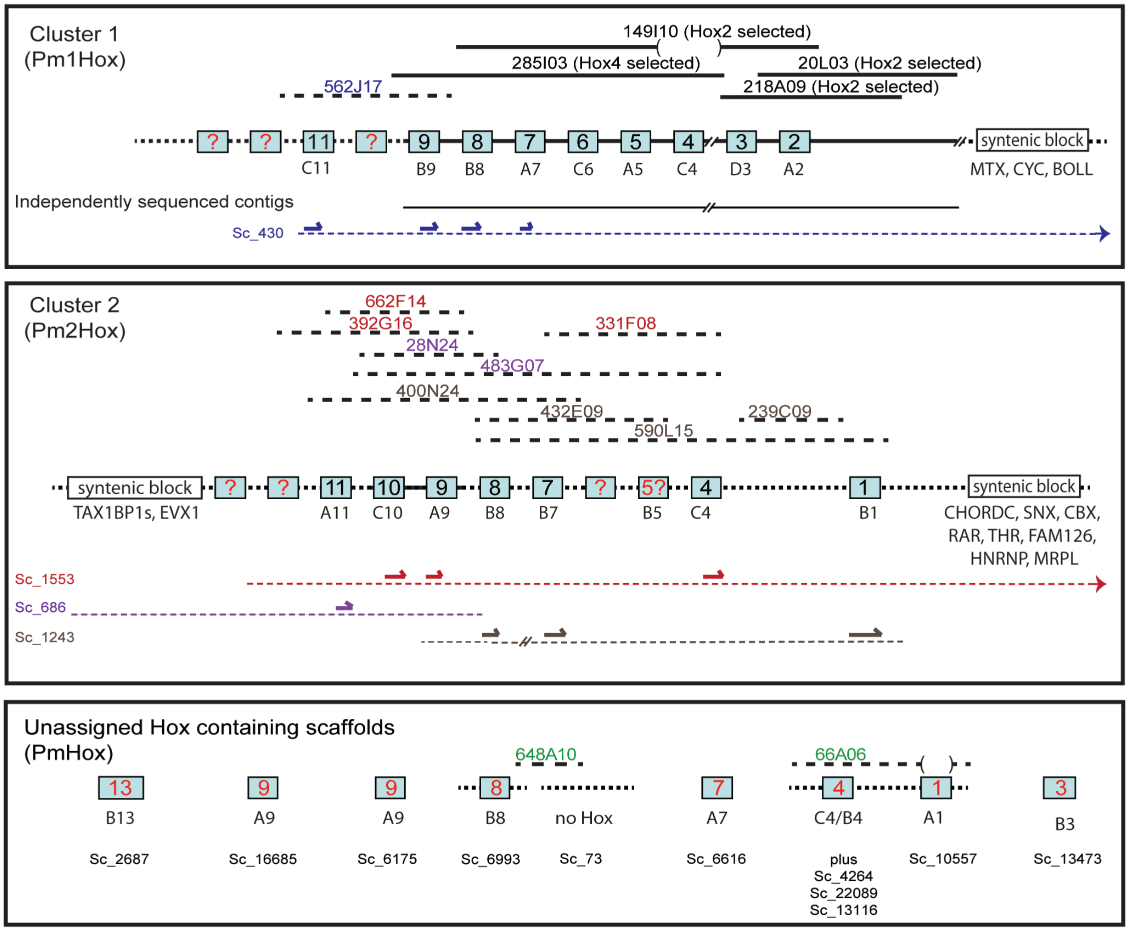
Supplementary Figure 20 – Relationship between detection of gnathostome paralogous duplicates and scaffold information content. Regression lines are included only to show general trends.



Supplementary Figure 21 - Relationship between detection of lamprey paralogous duplicates and scaffold information content. Regression lines are included only to show general trends.

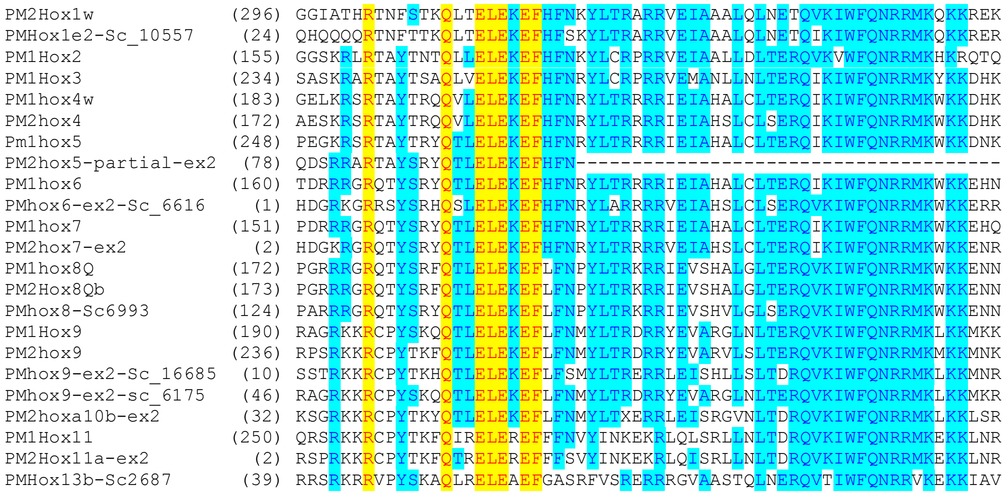


Supplementary Figure 22 - Summary of Hox genes and clusters identified in the lamprey genome. To supplement the lamprey genomic scaffolds, we combined a variety of additional sequence sources targeting these loci, including conventional sequencing via a shot-gun approach and subsequent walking to bridge gaps of a series of BACs. Selected representative BAC names forming the backbone sequence are shown, those in black were extensively sequenced to close gaps, those dotted were sequence by Illumina. Additional sequence contigs from 454 sequenced BACs were used in the assembly of the map. These clusters also contain genes syntenic with Hox clusters in other species, albeit with some paralogs being differentially retained in lamprey vs. gnathostome lineages. In addition, the best human Hox paralog match is indicated for each gene as established translation and blastx searches. Some BACs contain deletions, for example 149I10 has a deletion encompassing Hox4, which may represent a somatic deletion as previously documented in lamprey development [3,4] or a rearranged BAC generated during library construction.



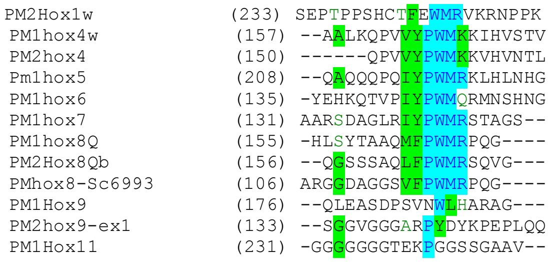
Supplementary Figure 23 - Alignments of the homeodomain and hexapeptide regions of the predicted Lamprey Hox genes. Alignments were produced with default values in Vector NTI's AlignX program. Hox homology group assignments were established by manual curation of the predicted coding exons of each gene following blastx searches and GeneMaker analysis. Predicted proteins were compared by blastp, against NCBI RefSeq (limited by "Hox" Entrez term) and aligned to the best-matched human paralogs. In some cases, exons were evaluated individually when putative genes were not located on a single contig and significantly large gaps remained to make the firm association between the two exons unclear (Cluster 2 Hox4 and Hox9). In all cases, introns and flanking regions confirm the distinct genomic origin of the predicted paralogs.

Homeodomain containing region of exon 2 Vector NTI AlignX comparisons



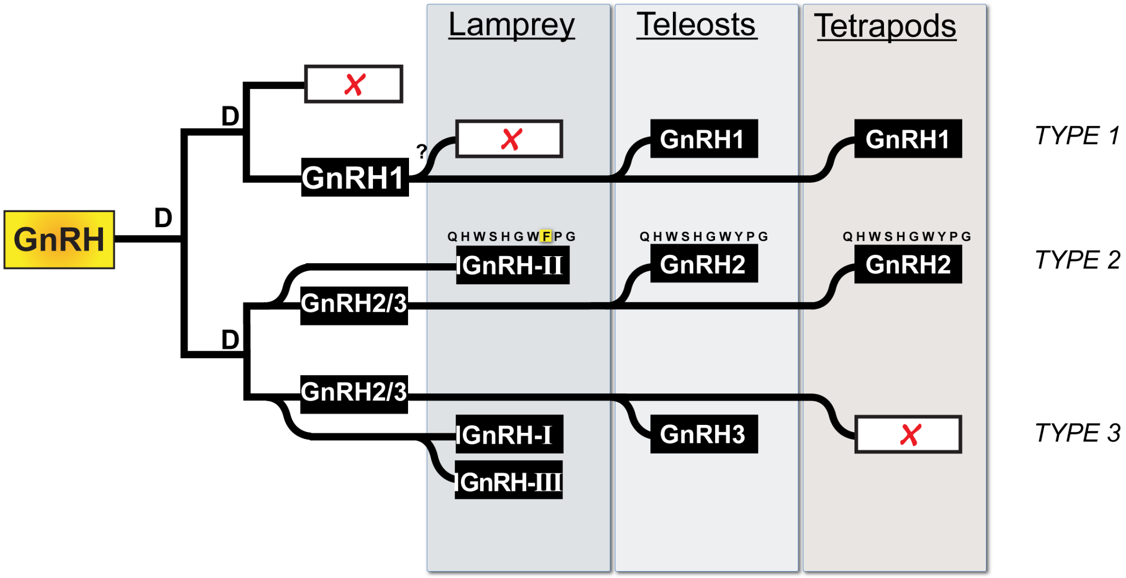
```
PM2Hox1w             (296) GGIATHRTNFSTKQLTELEKEFHFNKYLTRARRVEIAALQLNETQVKIWFQNRRMKQKKREK
PMHox1e2-Sc_10557     (24) QHQQQQRTNFTTKQLTELEKEFHFSKYLTRARRVEIAALQLNETQIKIWFQNRRMKQKKRER
PM1Hox2              (155) GGSKRLRTAYTNTQLLELEKEFHFNKYLCRPRRVEIAALLDLTERQVKVWFQNRRMKHKRQTQ
PM1Hox3              (234) SASKRARTAYTSAQLVELEKEFHFNRYLCRPRRVEMANLLNLTERQIKIWFQNRRMKYKKDHK
PM1hox4w             (183) GELKRSRTAYTRQQVLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKDHK
PM2hox4              (172) AESKRSRTAYTRQQVLELEKEFHFNRYLTRRRRIEIAHSLCLSERQIKIWFQNRRMKWKKDHK
Pm1hox5              (248) PEGKRSRTAYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKDNK
PM2hox5-partial-ex2   (78) QDSRRARTAYSRYQTLELEKEFHFN-------------------------------------
PM1hox6              (160) TDRRRGRQTYSRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEHN
PMhox6-ex2-Sc_6616     (1) HDGRKGRRSYSRHQSLELEKEFHFNRYLARRRRVEIAHSLCLSERQVKIWFQNRRMKWKKERR
PM1hox7              (151) PDRRRGRQTYSRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEHQ
PM2hox7-ex2            (2) HDGKRGRQTYSRYQTLELEKEFHFNRYLTRRRRVEIAHSLCLTERQIKIWFQNRRMKWKKENR
PM1hox8Q             (172) PGRRRGRQTYSRFQTLELEKEFLFNPYLTRKRRIEVSHALGLTERQVKIWFQNRRMKWKKENN
PM2Hox8Qb            (173) PGRRRGRQTYSRFQTLELEKEFLFNPYLTRKRRIEVSHALGLTERQVKIWFQNRRMKWKKENN
PMhox8-Sc6993        (124) PARRRGRQTYSRYQTLELEKEFLFNPYLTRKRRIEVSHVLGLSERQVKIWFQNRRMKWKKENN
PM1Hox9              (190) RAGRKKRCPYSKQQTLELEKEFLFNMYLTRDRRYEVARGLNLTERQVKIWFQNRRMKLKKMKK
PM2hox9              (236) RPSRKKRCPYTKFQTLELEKEFLFNMYLTRDRRYEVARVLSLTERQVKIWFQNRRMKMKKMNK
PMhox9-ex2-Sc_16685   (10) SSTRKKRCPYTKHQTLELEKEFLFSMYLTRERRLEISHLLSLTDRQVKIWFQNRRMKLKKMNR
PMhox9-ex2-sc_6175    (46) RAGRKKRCPYSKQQTLELEKEFLFNMYLTRDRRYEVARGLNLTERQVKIWFQNRRMKLKKMKR
PM2hoxa10b-ex2        (32) KSGRKKRCPYTKYQTLELEKEFLFNMYLTXERRLEISRGVNLTDRQVKIWFQNRRMKLKKLSR
PM1Hox11             (250) QRSRKKRCPYTKFQIRELEREFFFNVYINKEKRLQLSRLLNLTDRQVKIWFQNRRMKEKKLNR
PM2Hox11a-ex2          (2) RSPRKKRCPYTKFQTRELEREFFFSVYINKEKRLQISRLLNLTDRQVKIWFQNRRMKEKKLNR
PMHox13b-Sc2687       (39) RRSRKRRVPYSKAQLRELEAEFGASRFVSRERRRGVAASTQLNERQVTIWFQNRRVKEKKIAV
```

Hexapeptide region in Exon 1 Vector NTI AlignX comparisons

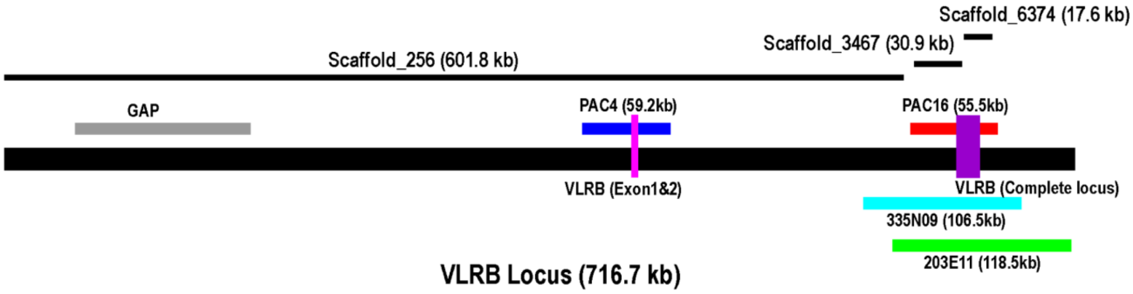


```
PM2Hox1w             (233) SEPTPPSHCTFEWMRVKRNPPK
 PM1hox4w            (157) --AALKQPVVYPWMKKIHVSTV
 PM2hox4             (150) ------QPVVYPWMKKVHVNTL
 Pm1hox5             (208) --QAQQQPQIYPWMRKLHLNHG
 PM1hox6             (135) -YEHKQTVPIYPWMQRMNSHNG
 PM1hox7             (131) AARSDAGLRIYPWMRSTAGS--
 PM1hox8Q            (155) -HLSYTAAQMFPWMRPQG----
 PM2Hox8Qb           (156) --QGSSSAQLFPWMRSQVG---
 PMhox8-Sc6993       (106) ARGGDAGGSVFPWMRPQG----
 PM1Hox9             (176) --QLEASDPSVNWLHARAG---
 PM2hox9-ex1         (133) --SGGVGGGARPYDYKPEPLQQ
 PM1Hox11            (231) -GGGGGGTEKPGGSSGAAV--
```

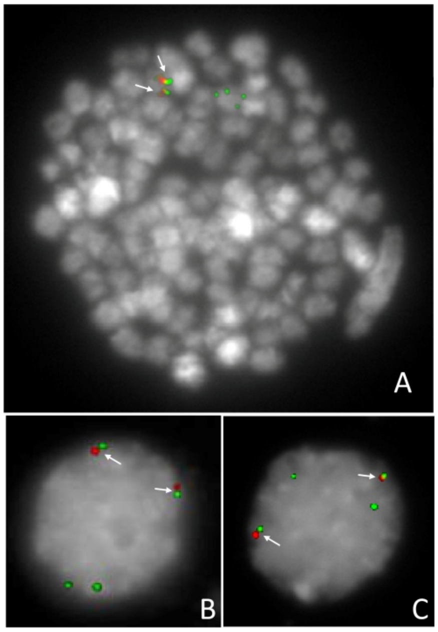
Supplementary Figure 24 – Proposed evolution of the GnRH gene family in vertebrates. The deduced amino acids corresponding to the lamprey GnRH-II and GnRH2 (zebrafish, human) decapeptides are indicated above the representative blocks with the single amino acid difference in lamprey highlighted. 'D' represents whole-genome duplication events. Importantly, although we have drawn the scenario where lamprey GnRH-II is orthologous to GnRH2, a scenario wherein lamprey GnRH-II and gnathostome GnRH3 share most-recent common ancestry (post duplication) is also plausible.



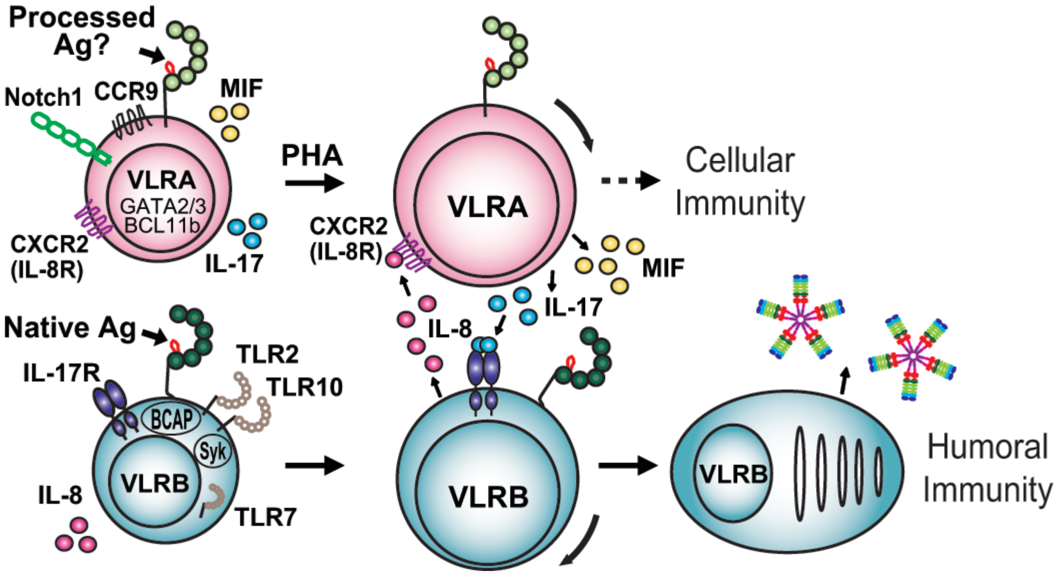
Supplementary Figure 25 - Map of the composite VLRB scaffold that was stitched together with the aid of sequenced BAC clones.



Supplementary Figure 26 - Fluorescence *in situ* hybridization of lamprey chromosomes and interphase nuclei using fluorescently labeled PAC4 (red) and PAC16 (green). (A) meiotic spread from testis in diakinesis/metaphase I; (B) interphase nuclei from gill; (C) interphase nuclei from kidney. Data were kindly obtained by Francesca Antonacci (Genome Sciences, Univ. Washington). Red and green signals are adjacent to each other (arrows) indicating their physical proximity in the genome. The additional hybridization site for PAC16 is likely due to a repetitive tract as opposed to VLRB homology based on *BLASTN* searches of the lamprey genomic assembly and traces.

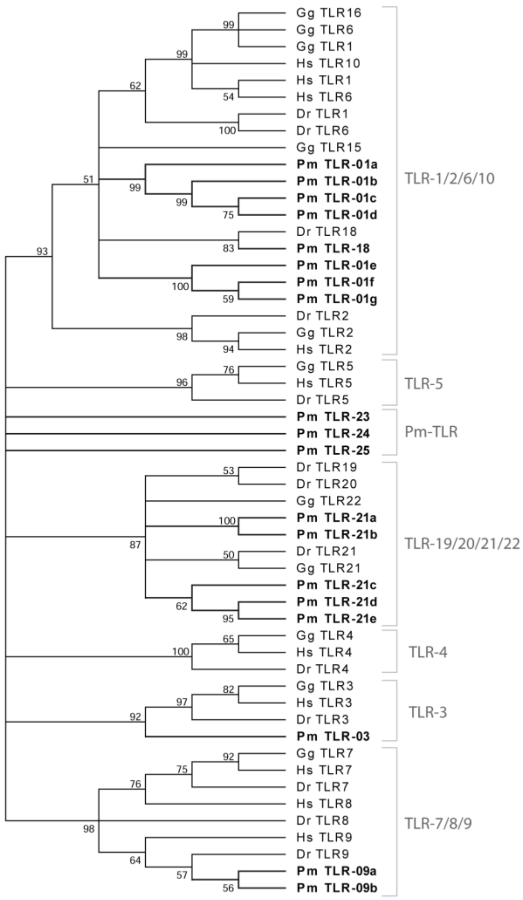


Supplementary Figure 27 - T-like and B-like lymphocyte subsets in lampreys. Antigens (Ag) induce lymphoblastoid transformation of VLRA and VLRB cells. Ag-stimulated VLRB cells differentiate into VLRB-secreting plasmacytes. VLRA is not secreted, but activated VLRA cells produce the proinflammatory cytokines, IL-17, and macrophage migration inhibitory factor (MIF). VLRA and VLRB cells express transcripts that encode orthologs for several genes essential for respective T cell and B cell development in jawed vertebrates: GATA binding protein 2/3 (GATA2/3), B cell lymphoma/leukemia 11b (BCL11b), C-C chemokine receptor 9 (CCR9), Notch1, C-X-C chemokine receptor 2 (CXCR2, IL-8 receptor), Syk, B cell adaptor protein (BCAP), IL-8, IL-17 receptor (IL-17R), and TLR orthologs TLR2, TLR7, and TLR10. The reciprocal expression of cytokines (IL-17 in VLRA and IL-8 in VLRB cells) and their receptors (IL-17R in VLRB and IL-8R in VLRA cells) suggest functional interaction between the two types of lymphocytes. PHA: phytohemagglutinin.

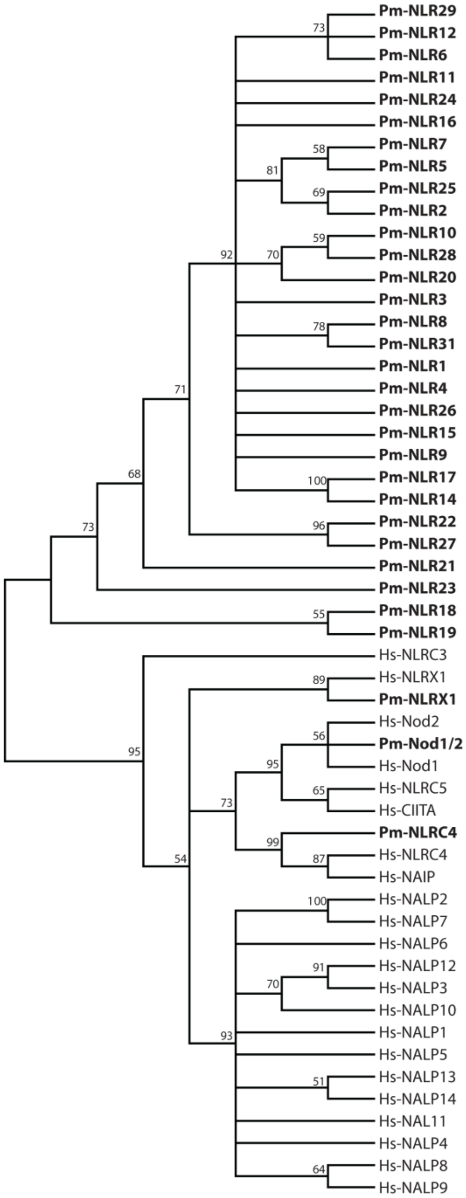


Nature Genetics: doi:10.1038/ng.2568

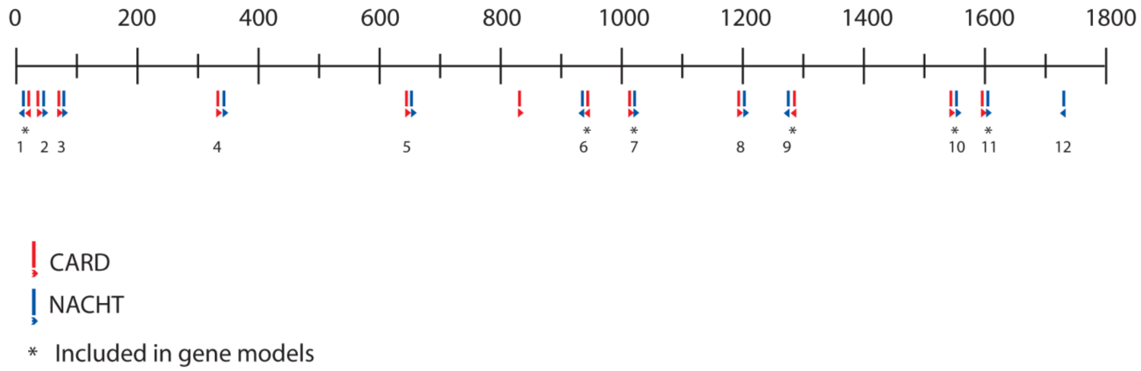
Supplementary Figure 28 - Phylogenetic analysis of the lamprey Toll-like receptors. To understand the relationships between the lamprey TLR genes and the characterized vertebrate TLRs, a neighbor-joining tree was constructed in MEGA5 [5] using complete gap deletion. The values shown at the nodes are the bootstrap values based on 1000 replicates. TLR sequences were collected from NCBI from human (Hs), chicken (Gg) and zebrafish (Dr). The lamprey sequences (Pm) are shown in bold. Accession numbers for the sequences used to build the tree are as follows: Hs-TLR1, NP_003254.2; Hs-TLR2, NP_003255.2; Hs-TLR3, NP_003256.1; Hs-TLR4, NP_612564.1; Hs-TLR5, NP_003259.2; Hs-TLR6, NP_006059.2; Hs-TLR7, NP_057646.1; Hs-TLR8, NP_619542.1; Hs-TLR9, NP_059138.1|; Hs-TLR10, NP_001182036.1; Gg-TLR1, BAD67422.1; Gg-TLR2, NP_989609.1; Gg-TLR3, NP_001011691.3; Gg-TLR4, NP_001025864.1; Gg-TLR5, NP_001019757.1; Gg-TLR6, NP_001075178.2; Gg-TLR7, NP_001011688.1; Gg-TLR15, NP_001032924.1; Gg-TLR16, ABQ85926.1; Gg-TLR21, NP_001025729.1; Dr-TLR1, AAQ91305.1; Dr-TLR2, NP_997977.1; Dr-TLR3, NP_001013287.2; Dr-TLR4, AAQ90475.1; Dr-TLR5, NP_001124067.1; Dr-TLR6, NP_001124065.1; Dr-TLR7, XP_003199309.1; Dr-TLR8, XP_003199440.1; Dr-TLR9, NP_001124066.1; Dr-TLR18, NP_001082819.1; Dr-TLR19, XP_002664892.2; Dr-TLR20, AAI63786.1; Dr-TLR21, NP_001186264.1; Dr-TLR22, NP_001122147.1.



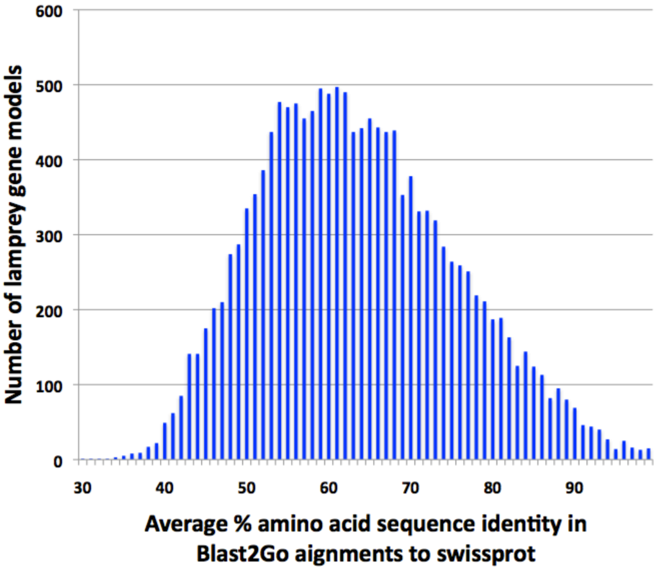
Supplementary Figure 29 - Phylogenetic analysis of the lamprey NLRs. A neighbor-joining tree was constructed using the conserved NACHT domains of the lamprey NLR genes and the human NLRs in MEGA5 [5]. The values shown at the nodes are the bootstrap values based on 1000 replicates. The lamprey sequences (Pm) are shown in bold. Accession numbers for the remaining NLR sequences are as follows: Hs-Nod1, NP_006083.1; Hs-Nod2, NP_071445.1; Hs-NLRC3, ACP40993.1; Hs-NLRC4, AAH31555.1; Hs-NLRC5, NP_115582.3; Hs-NAIP, AAI36274.1; Hs-NALP1, Q9C000.1; Hs-NALP2, Q9NX02.1; Hs-NALP3, Q96P20.3; Hs-NALP4, Q96MN2.3; Hs-NALP5, NP_703148.4; Hs-NALP6, NP_612202.1; Hs-NALP7, Q8WX94.1; Hs-NALP8, Q86W28.2; Hs-NALP9, Q7RTR0.1; Hs-NALP10, NP_789791.1; Hs-NALP11, P59045.2; Hs-NALP12, NP_653288.1; Hs-NALP13, NP_789780.2; Hs-NALP14, NP_789792.1; Hs-CIITA, P33076.3; Hs-NLRX1, AAI10891.1.



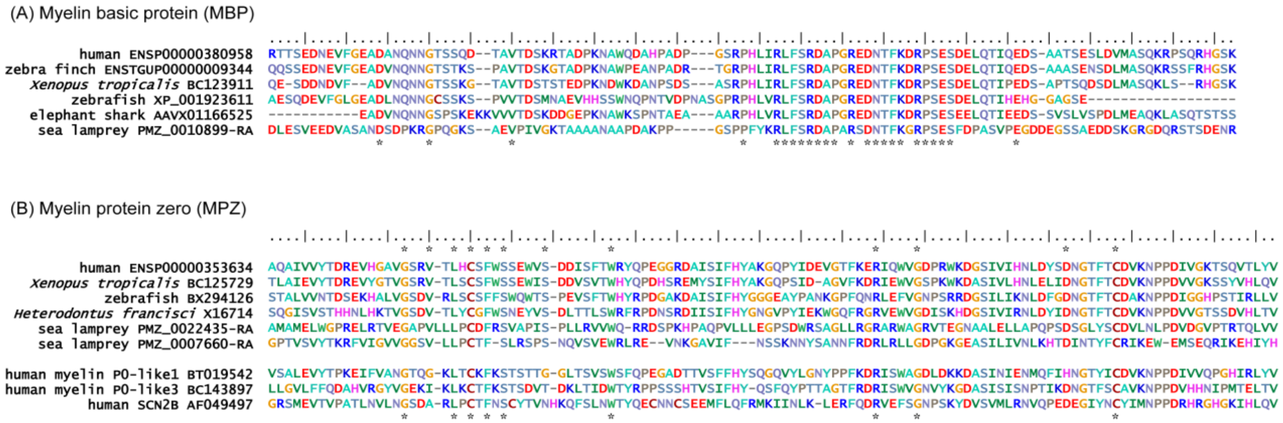
Supplementary Figure 30 - Twelve of the lamprey NLRs are organized on a single scaffold. The NACHT (blue boxes) and CARD domains (red boxes) identified on scaffold_357 are shown. Arrows underneath the boxes indicate the direction of the coding sequence. The sequence is shown to scale (numbers indicate kb). The NLR genes are numbered according to Supplementary Table 21. Asterisks are shown below NACHT domains that were captured by gene models. No effector domains were identified near the NACHT domain for Pm-NLR12.



Supplementary Figure 31 – Distribution of alignment statistics used in assigning ontologies to lamprey gene models.



Supplementary Figure 32 - Amino acid alignments of myelin-related proteins. (A) Myelin basic proteins (MBP) of jawed vertebrates and their possible lamprey ortholog. Asterisks at the bottom indicate amino acid residues conserved among these sequences. The alignment covers the region from amino acid position 37 to 145 of the human sequence ENSP00000380958. (B) Myelin protein zero (MPZ) sequences of jawed vertebrates and their possible lamprey orthologs. Three human proteins are added at the bottom as distant paralogs in the same protein family. Asterisks at the top indicate amino acid residues conserved among the MPZ proteins including the lamprey sequences, while those at the bottom indicate residues conserved throughout all the sequences included. The alignment covers the region from amino acid position 70 to 156 of the human sequence ENSP00000353634. Amino acid alignments were constructed using the program MAFFT [6] with default settings.



(A) Myelin basic protein (MBP)

```
                               ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|..
         human ENSP00000380958 RTTSEDNEVFGEADANQNNGTSSQD--TAVTDSKRTADPKNAWQDAHPADP---GSRPHLIRLFSRDAPGREDNTFKDRPSESDELQTIQEDS-AATSESLDVMASQKRPSQRHGSK
zebra finch ENSTGUP00000009344 QQSSEDNEVFGEADVNQNNGTSTKS--PAVTDSKGTADPKNAWPEANPADR---TGRPHLIRLFSRDAPGREDNTFKDRPSESDELQTIQEDS-AAASENSDLMASQKRSSFRHGSK
   Xenopus tropicalis BC123911 QE-SDDNDVF--ADVNQNNGTSSKG--TAVTDSTSTEDPKNDWKDANPSDS---ASRPHLIRLFSRDAPGREDNTFKDRPSESDELQTIPEDS-APTSQDSDLMASQKLS--RHGSK
        zebrafish XP_001923611 AESQDEVFGLGEADLNQNNGCSSKS--PVVTDSMNAEVHHSSWNQPNTVDPNASGPRPHLVRLFSRDAPGREDNTFKDRPSESDELQTIHEHG-GAGSE------------------
   elephant shark AAVX01166525 -----------EADVNQNNGSPSKEKKVVVTDSKDDGEPKNAWKSPNTAEA---AARPHLVRLFSRDAPGREDNTFKDRPSESEELQTIEEDS-SVSLVSPDLMEAQKLASQTSTSS
    sea lamprey PMZ_0010899-RA DLESVEEDVASANDSDPKRGPQGKS--AEVPIVGKTAAAANAAPDAKPP-----GSPPFYKRLFSRDAPARSDNTFKGRPSESFDPASVPEGDDEGSSAEDDSKGRGDQRSTSDENR
                                            *     *         *                           *   ******** * ***** *****       *
```

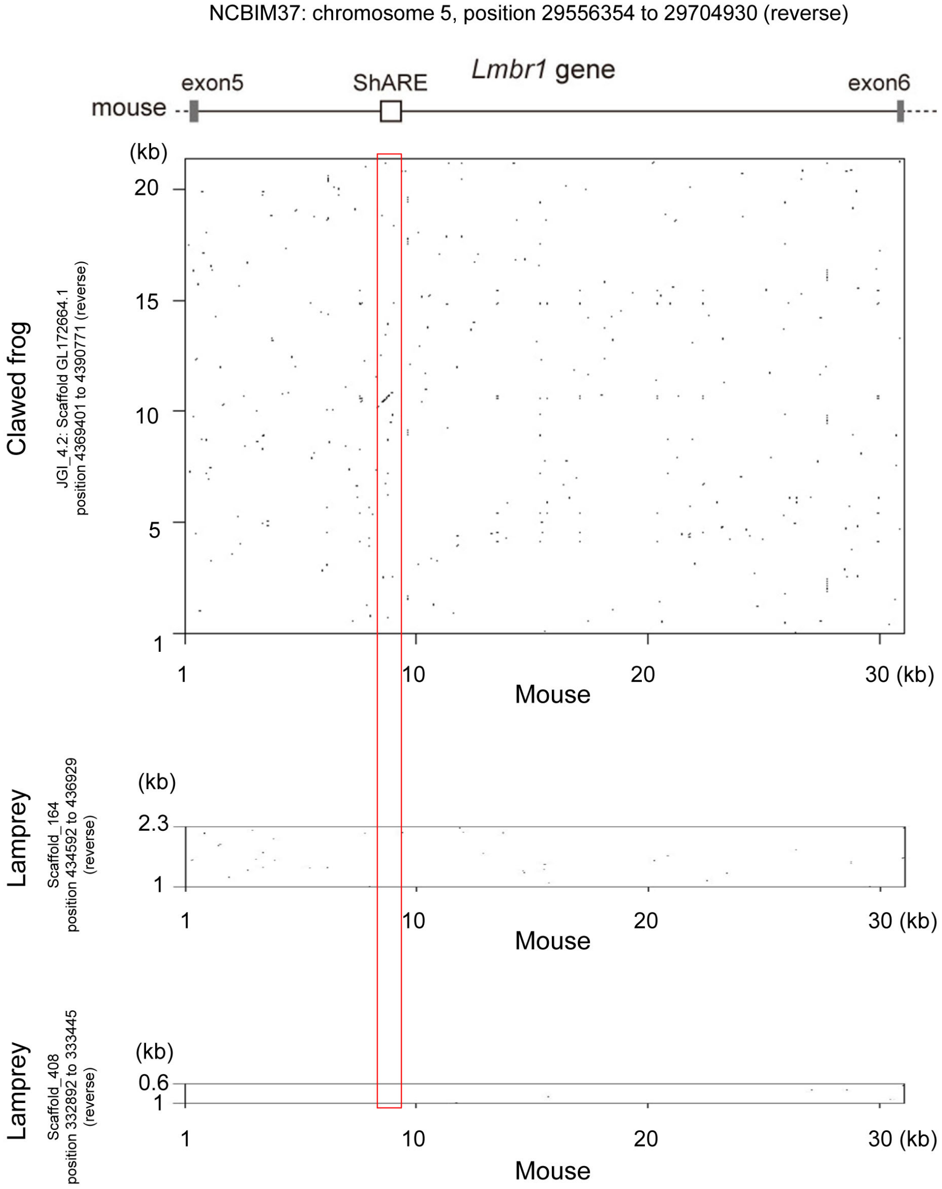
(B) Myelin protein zero (MPZ)

```
                               ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|..
                                               *  *  * * * *    *       *                               *    *                 *     *

         human ENSP00000353634 AQAIVVYTDREVHGAVGSRV-TLHCSFWSSEWVS-DDISFTWRYQPEGGRDAISIFHYAKGQPYIDEVGTFKERIQWVGDPRWKDGSIVIHNLDYSDNGTFTCDVKNPPDIVGKTSQVTLYV
   Xenopus tropicalis BC125729 TLAIEVYTDREVYGTVGSRV-TLSCSFWSSEWIS-DDVSVTWHYQPDHSREMYSIFHYAKGQPSID-AGVFKDRIEWVGSPKWKDASIVLHNLELIDNGTFTCDVKNPPDVVGKSSYVHLQV
            zebrafish BX294126 STALVVNTDSEKHALVGSDV-RLSCSFFSWQWTS-PEVSFTWHYRPDGAKDAISIFHYGGGEAYPANKGPFQNRLEFVGNPSRRDGSILIKNLDFGDNGTFTCDAKNPPDIGGHPSTIRLLV
 Heterodontus francisci X16714 SQGISVSTHHNLHKTVGSDV-TLYCGFWSNEYVS-DLTTLSWRFRPDNSRDIISIFHYGNGVPYIEKWGQFRGRVEWVGDISKHDGSIVIRNLDYIDNGTFTCDVKNPPDVVGTSSDVHLTV
    sea lamprey PMZ_0022435-RA AMAMELWGPRELRTVEGAPVLLLPCDFRSVAPIS-PLLRVVWQ-RRDSPKHPAQPVLLLEGPSDWRSAGLLRGRARWAGRVTEGNAALELLAPQPSDSGLYSCDVLNLPDVDGVPTRTQLVV
    sea lamprey PMZ_0007660-RA GPTVSVYTKRFVIGVVGGSV-LLPCTF-SLRSPS-NQVSVEWRLRE--VNKGAVIF---NSSKNNYSANNFRDRLRLLGDPGKGEASILIVNLKHTDINTYFCRIKEW-EMSEQRIKEHIYH

human myelin P0-like1 BT019542 VSALEVYTPKEIFVANGTQG-KLTCKFKSTSTTG-GLTSVSWSFQPEGADTTVSFFHYSQGQVYLGNYPPFKDRISWAGDLDKKDASINIENMQFIHNGTYICDVKNPPDIVVQPGHIRLYV
human myelin P0-like3 BC143897 LLGVLFFQDAHVRGYVGEKI-KLKCTFKSTSDVT-DKLTIDWTYRPPSSSHTVSIFHY-QSFQYPTTAGTFRDRISWVGNVYKGDASISISNPTIKDNGTFSCAVKNPPDVHHNIPMTELTV
          human SCN2B AF049497 GRSMEVTVPATLNVLNGSDA-RLPCTFNSCYTVNHKQFSLNWTYQECNNCSEEMFLQFRMKIINLK-LERFQDRVEFSGNPSKYDVSVMLRNVQPEDEGIYNCYIMNPPDRHRGHGKIHLQV
                                               *     * * * *            *                               *    *                       *
```

Supplementary Figure 33 – Dot matrix plot for sequence comparison of an intron of *Lmbr1* homologs (homologous to mouse intron 5). The plot was produced by the Dottup web server (http://emboss.bioinformatics.nl/cgi-bin/emboss/dottup) developed as part of the EMBOSS package, with word size set at 11. Input sequences include an upstream exon and a downstream exon as well as the intron harboring the ShARE in the mouse *Lmbr1* gene.



Supplementary Tables

Supplementary Table 2 - Identification of CEGs in the lamprey genome.

CEG group	# in assembly	% of group
1	49	74.2
2	42	75.0
3	48	78.7
4	44	67.7
All	183	73.8

Supplementary Table 3 - Features used for curation of transposable elements.

	Super family	Terminal repeat	TSD	Other
Class I	LINE	Direct/inverted/none	Variable	Poly A signal at 3'end
	SINE	None	Variable	Poly A signal at 3'end
	LTR elements	Direct	4-5 bps	
Class II	Chapaev	Inverted	4 bps	
	hAT	Inverted	8 bps	
	Helitron	None	None	Starts with "TC" and ends with "CTAG"
	PIF/Harbinger/ Tourist	Inverted	3 bps	
	Tc1/Mariner	Inverted	2 bps	

Supplementary Table 4 - Composition of transposable elements in the lamprey genome.

Class of element	Curated	Non Curated	Total genomic fraction (%)
SINEs	14	35	7.2
LINEs	47	312	12.6
LTR elements	42	135	2.7
DNA elements	39	196	5.7
Unknown		6720	19.2

Supplementary Table 5 - SRA Identifiers for RNAseq experiments.

SRA Experiment Identifier	Sequencing Technology	Tissue Sequenced
SRX109761.3	GS20	parasitic sea lamprey olfactory epithelium
SRX109762.3	GS20	adult sea lamprey olfactory epithelium
SRX109764.3	GS-FLX	adult sea lamprey brain
SRX109765.3	GS-FLX	larval/parasitic sea lamprey brain
SRX109766.3	GS-FLX	larval sea lamprey liver
SRX109767.3	GS-FLX	parasitic sea lamprey liver
SRX109768.3	Illumina GA2 75bp reads	adult sea lamprey brain
SRX109769.3	Illumina GA2 75bp reads	parasitic sea lamprey liver
SRX109770.3	Illumina GA2 75bp reads	larval sea lamprey intestine
SRX110023.2	Illumina GA2 75bp reads	larval sea lamprey kidney
SRX110024.2	Illumina GA2 75bp reads	small parasitic sea lamprey kidney
SRX110025.2	Illumina GA2 75bp reads	small parasitic sea lamprey proximal intestine
SRX110026.2	Illumina GA2 75bp reads	small parasitic sea lamprey distal intestine
SRX110027.2	Illumina GA2 75bp reads	adult sea lamprey intestine
SRX110028.2	Illumina GA2 75bp reads	adult sea lamprey kidney
SRX110029.2	Illumina GA2 100bp reads	sea lamprey late blastula embryo (stage 18)
SRX110030.2	Illumina GA2 100bp reads	sea lamprey gastrula embryo (stage 20)
SRX110031.2	Illumina GA2 100bp reads	sea lamprey neurula embryo (stage 22a)
SRX110032.2	Illumina GA2 100bp reads	sea lamprey neurula embryo (stage 22b)
SRX110033.2	Illumina GA2 100bp reads	sea lamprey embryo: neural crest migration (stage 23)
SRX110034.2	Illumina GA2 100bp reads	sea lamprey embryo: neural crest migration (stage 24c1)
SRX110035.2	Illumina GA2	sea lamprey embryo: neural crest migration

	100bp reads	stage 24c2)

Supplementary Table 7 - Numbers of tRNA isotypes in the lamprey genome. tRNA isotype genes assigned for different anticodons were identified using *tRNAScan*[7] and *RNAFOLD*[8] software. Numbers of pseudogenes are in parentheses.

Anticodon	Codon	Amino acid	tRNA count
AAA	TTT	Phe/F	3
GAA	TTC		60(11)
TAA	TTA	Leu/L	5 (2)
CAA	TTG		50
AAG	CTT		126
GAG	CTC		0 (1)
TAG	CTA		44 (4)
CAG	CTG		75 (1)
AAT	ATT	Ile/I	115 (12)
GAT	ATC		10
TAT	ATA		10
CAT	ATG	Met/M	222
AAC	GTT	Val/V	186
GAC	GTC		1
TAC	GTA		32
CAC	GTG		74
AGA	TCT	Ser/S	22
GGA	TCC		1
TGA	TCA		20
CGA	TCG		38
ACT	AGT		0
GCT	AGC		70 (1)
AGG	CCT	Pro/P	23 (3)
GGG	CCC		2
TGG	CCA		30
CGG	CCG		16
AGT	ACT	Thr/T	72 (3)
GGT	ACC		150 (81)
TGT	ACA		33 (2)
CGT	ACG		27
TTA	TAA	Stop	
CTA	TAG		

Anticodon	Codon	Amino acid	tRNA count
AGC	GCT	Ala/A	144 (7)
GGC	GCC		2
TGC	GCA		30 (8)
CGC	GCG		24
ATA	TAT	Tyr/Y	0
GTA	TAC		93 (5)
ATG	CAT	His/H	2 (7)
GTG	CAC		34 (1)
TTG	CAA	Gln/Q	47
CTG	CAG		156 (1)
ATT	AAT	Asn/N	0 (1)
GTT	AAC		177 (7)
TTT	AAA	Lys/K	24
CTT	AAG		21 (6)
ATC	GAT	Asp/D	2
GTC	GAC		135
TTC	GAA	Glu/E	15 (1)
CTC	GAG		12
ACA	TGT	Cys/C	0
GCA	TGC		71 (5)
TCA	TGA	Stop(SelCys)	3
CCA	TGG	Trp/W	11
ACG	CGT	Arg/R	83 (11)
GCG	CGC		2 (1)
TCG	CGA		28 (2)
CCG	CGG		10 (3)
TCT	AGA		18
CCT	AGG		52 (8)
ACC	GGT	Gly/G	0 (1)
GCC	GGC		60 (2)
TCC	GGA		20
CCC	GGG		6

Supplementary Table 11 – Counts of single copy genes and retained duplicates in lamprey and chicken genomes. Numbers correspond to counts of ancestral (pre-duplication) loci for gene families with less than six members in both species.

		Chicken		
		Single Copy	**Retained Duplicate**	**Total**
Lamprey	**Single Copy**	5123	928	6051
	Retained Duplicate	1008	1246	2254
	Total	6131	2174	8305

Supplementary Table 12 – Counts of single copy genes and retained duplicates in lamprey and human genomes. Numbers correspond to counts of ancestral (pre-duplication) loci for gene families with less than six members in both species.

		Human		
		Single Copy	**Retained Duplicate**	**Total**
Lamprey	**Single Copy**	5181	1370	6551
	Retained Duplicate	917	1364	2281
	Total	6098	2734	8832

Supplementary Table 13 – Counts of presumptive ancestral vertebrate genes for single-copy versus duplication states lamprey/chicken interdigitated syntenic blocks.

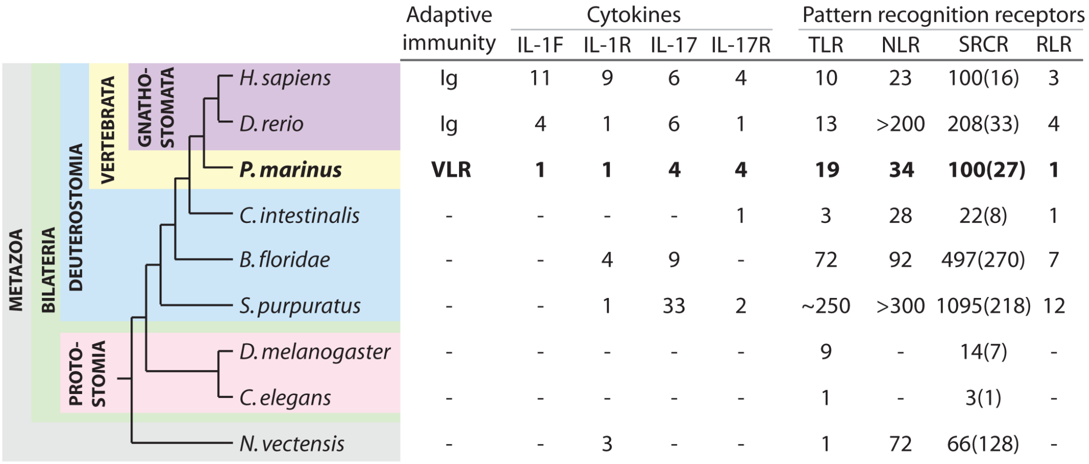
		Chicken		
		Single Copy	**Retained Duplicate**	**Total**
Lamprey	**Single Copy**	330	77	407
	Retained Duplicate	173	131	304
	Total	503	208	711

Supplementary Table 14 – Proposed nomenclature shift for GnRH genes. Modified from previously published material[9,10].

Old Groupings of Paralogs	Representative members
GnRH1	mammal GnRH in mouse, human, sheep, pig, eel, newt, frog; seabream GnRH in *goldfish*, *salmon*, *catfish*; chicken GnRH-I in *chicken*, *lizard;* catfish GnRH in *catfish*; guinea pig GnRH in *guinea pig*; medaka GnRH in *medaka*
GnRH2	chicken GnRH-II in *mouse*, *primate*, *human*, *chicken*, *lizard*, *frog*, *newt*, *eel*, *goldfish*, *catfish*, *salmon*, *medaka*, *red seabream*, *tilapia*, *ratfish*; lamprey GnRH-II in *lamprey*
GnRH3	salmon GnRH in *medaka*, *red seabream*, *tilapia, Atlantic salmon, brook trout, chinook salmon, zebrafish*
GnRH4 (IV)	lamprey GnRH-I and lamprey GnRH-III in *lamprey*

New Groupings of Paralogs	Representative members
GnRH1	mammal GnRH in mouse, human, sheep, pig, eel, newt, frog; seabream GnRH in *goldfish*, *salmon*, *catfish*; chicken GnRH-I in *chicken*, *lizard;* catfish GnRH in *catfish*; guinea pig GnRH in *guinea pig*; medaka GnRH in *medaka*
GnRH2	chicken GnRH-II in *mouse*, *primate*, *human*, *chicken*, *lizard*, *frog*, *newt*, *eel*, *goldfish*, *catfish*, *salmon*, *medaka*, *red seabream*, *tilapia*, *ratfish*; *lamprey GnRH-II in lamprey*
GnRH3	salmon GnRH in *medaka*, *red seabream*, *tilapia*, *Atlantic salmon*, *brook trout*, *chinook salmon*, *zebrafish;* lamprey GnRH-I and lamprey GnRH-III in *lamprey*
GnRH4	Lost

Supplementary Table 17 - The lamprey innate immune system resembles that of jawed vertebrates. Traits of adaptive and innate immune systems for representative metazoan species are shown. Lampreys possess a distinct, VLR-based adaptive immune system that differs from the immunoglobulin-based system of jawed vertebrates both in the domain structure of the receptor molecules and also in the mechanism of somatic diversification. The lamprey innate immune system, however, is more similar to those of jawed vertebrates than invertebrate deuterostomes, particularly with respect to the multiplicity of the gene families that encode pattern recognition receptors. Numbers of genes encoding toll-like receptors (TLR), Nod-like receptors (NLR), and Rig-I like receptors (RLR) for each of the species is shown. The total number of scavenger receptor cysteine rich (SRCR) domains is shown, with the number of genes in parentheses. In contrast to the expanded gene families of sea urchin and amphioxus, the lamprey genome contains similar numbers of TLR, NLR, SRCR, and RLR genes as Divergent homologs of cytokines have also been identified in the *P. marinus* genome, including IL-1 and IL-17.



	Adaptive immunity	Cytokines				Pattern recognition receptors			
		IL-1F	IL-1R	IL-17	IL-17R	TLR	NLR	SRCR	RLR
H. sapiens	Ig	11	9	6	4	10	23	100(16)	3
D. rerio	Ig	4	1	6	1	13	>200	208(33)	4
P. marinus	**VLR**	**1**	**1**	**4**	**4**	**19**	**34**	**100(27)**	**1**
C. intestinalis	-	-	-	-	1	3	28	22(8)	1
B. floridae	-	-	4	9	-	72	92	497(270)	7
S. purpuratus	-	-	1	33	2	~250	>300	1095(218)	12
D. melanogaster	-	-	-	-	-	9	-	14(7)	-
C. elegans	-	-	-	-	-	1	-	3(1)	-
N. vectensis	-	-	3	-	-	1	72	66(128)	-

Supplementary Table 18 - Scaffolds of VLR loci in the sea lamprey genome.

Locus	Scaffolds	Size (bp)	Comments
VLRA	Scaffold 1054	322,821	Contains entire genomic locus. Contains two big gaps (106,601 bp and 124,240 bp).
VLRB	Scaffold 256	617,513	Contains exons 1 and 2 and includes the PAC4 type locus. Contains a 116,201 bp gap.
	Scaffold 3467	31,119	Contains exons 1 and 2 and part of LRRNT, includes part of PAC16 type locus.
	Scaffold 6374	17,510	Contains LRRCT+3' UTR and downstream LRR modules. Includes part of PAC16 type locus
VLRC	Scaffold 92	1,070,441	Contains entire germline locus. Contains gaps totaling ~120 kb.

Supplementary Table 19 - Numbers of immune-related domains in deuterostome genomes.

	H. sapiens		*C. milii*		*P. marinus*		*C. intestinalis*		*B. floridae*		*S. purpuratus*	
	< 10^{-5}	< 1	< 10^{-5}	< 1	< 10^{-5}	< 1	< 10^{-5}	< 1	< 10^{-5}	< 1	< 10^{-5}	< 1
TIR	22	32	18	28	17	20	4	11	103	144	248	318
NACHT	25	43	58	82	32	37	54	121	91	160	360	422
SRCR	100	113	459	523	99	112	6	15	389	454	1658	1857
CARD	23	42	11	31	8	38	3	8	135	198	13	36
DEAD	65	112	31	86	18	58	43	78	49	207	60	253
DED	7	9	0	0	1	4	3	6	136	181	8	22
Death	18	32	12	23	8	18	2	19	326	604	81	767
Pyrin	23	24	0	1	0	0	0	1	1	7	0	0
IL17	8	11	4	4	4	4	2	6	11	14	21	46
MACPF	11	17	6	7	1	2	11	16	136	207	0	15
Ig	248	1193	225	1433	184	604	35	267	362	1341	250	1314
C1-set	104	144	312	476	1	19	0	7	0	47	1	72
V-set	535	1109	1054	1883	106	549	7	184	215	998	350	1214

Supplementary Table 20 - Lamprey TIR domain containing proteins.

	Gene model or sequence	**Name**	**Domains**
TLR-1/6/10 like	PMZ_0000753-RA	Pm_TLR-01a	TIR
	PMZ_0025017 *	Pm_TLR-01b	LRR(12) – LRRCT – TIR
	PMZ_0018923-RA	Pm_TLR-01c	LRR(6) – TIR
	Scaffold_5807 (unannotated similarity)	Pm_TLR-01d	LRR(12) – TIR
	PMZ_0007133-RA	Pm_TLR-01e	TIR
	PMZ_0008529-RA	Pm_TLR-01f	LRR(2) – TIR
	PMZ_0025012 *	Pm_TLR-01g	LRR(14) – TIR
TLR-18 like	PMZ_0022340-RA	Pm_TLR-18	LRR(7) – TIR
TLR-3 like	PMZ_0020961-RA	Pm_TLR-03	LRR(8) – TIR
TLR-7/8/9 like	PMZ_0014254-RA	Pm_TLR-09a	LRR(26) – LRRNT – TIR
	Scaffold_16 (unannotated similarity)	Pm_TLR-09b	LRR(14) – TIR
TLR-21/22 like	PMZ_0011018-RA	Pm_TLR-21a	LRR(8) – TIR
	PMZ_0025013*	Pm_TLR-21b	TIR
	PMZ_0017468-RA	Pm_TLR-21c	LRR(21) – TIR
	PMZ_0012373-RA	Pm_TLR-21d	LRR(23) – TIR
	PMZ_0017038-RA	Pm_TLR-21e	LRR(18) – TIR
Unique lamprey TLRs	PMZ_0000773-RA	Pm_TLR-23	TIR
	Scaffold_15 (unannotated similarity)	Pm_TLR-24	TIR
	PMZ_0005952-RB	Pm_TLR-25	TIR
TLR adaptors	PMZ_0002660-RA	Pm-MyD88	DEATH –TIR
	PMZ_0009064-RA	Pm-TICAM	TIR
	PMZ_0009066-RA	Pm-TICAM	TIR
	PMZ_0020456-RA	Pm-Sarm	SAM(2) – TIR
	PMZ_0010925-RA	Pm-IL1R	Ig(2) – TIR

* Manually annotated

Supplementary Table 21 - NLRs and NLR adaptors in the lamprey genome.

Name	Gene model or scaffold [start – stop]	Domains
Pm-Nod1/2	PMZ_0014569-RA	NACHT
Pm-NLRX1	PMZ_0009055-RA	CARD-NACHT
Pm-NLRC4	PMZ_0025030*	NACHT
Pm-NLR1	PMZ_0000296-RA	CARD-NACHT
Pm-NLR2	PMZ_0025001*	CARD-NACHT
Pm-NLR3	PMZ_0025002*	CARD-NACHT
Pm-NLR4	PMZ_0000325-RA	CARD-NACHT
Pm-NLR5	PMZ_0025003*	CARD-NACHT
Pm-NLR6	PMZ_0025004*	CARD-NACHT
Pm-NLR7	PMZ_0000316-RA	CARD-NACHT
Pm-NLR8	PMZ_0000317-RA	CARD-NACHT
Pm-NLR9	PMZ_0025005*	CARD-NACHT
Pm-NLR10	PMZ_0000336-RA	CARD-NACHT
Pm-NLR11	PMZ_0000339-RA	CARD-NACHT
Pm-NLR12	PMZ_0025006*	NACHT
Pm-NLR13	PMZ_0004181-RA	CARD-NACHT
Pm-NLR14	PMZ_0025010*	NACHT
Pm-NLR15	PMZ_0025009*	CARD-NACHT
Pm-NLR16	PMZ_0014048*	CARD-NACHT
Pm-NLR17	PMZ_0025019*	NACHT
Pm-NLR18	PMZ_0007535-RA	NACHT
Pm-NLR19	PMZ_0011734-RA	NACHT
Pm-NLR20	PMZ_0002760-RA	CARD-NACHT
Pm-NLR21	PMZ_0010190-RA	CARD-NACHT
Pm-NLR22	PMZ_0005721-RA	CARD-NACHT
Pm-NLR23	PMZ_0020511-RA	NACHT
Pm-NLR24	PMZ_0021096-RA	CARD-NACHT
Pm-NLR25	PMZ_0025025*	NACHT
Pm-NLR26	PMZ_0005662*	CARD-NACHT
Pm-NLR27	PMZ_0025008*	CARD-NACHT
Pm-NLR28	PMZ_0025015*	NACHT
Pm-NLR29	PMZ_0025028*	CARD-NACHT
Pm-NLR30	PMZ_0025011*	NACHT
Pm-NLR31	PMZ_0025024*	CARD-NACHT

* Manually annotated

Supplementary Table 22 - Cytokines and cytokine receptors in the lamprey genome.

Cytokine system	Molecule	Gene model or location
IL-8	IL-8	PMZ_0016236
	IL-8 receptor A	Transcriptome
Mif	Mif	PMZ_002555
IL-1	IL-1	PMZ_0025029*
	IL-1R	PMZ_0010925-RA
IL-6	IL-6	PMZ_0007799
	IL-6	PMZ_0005494
IL-17	IL-17B-like	PMZ_0004037
	IL-17C-like	PMZ_0025018*
	IL-17D-like	PMZ_0014858
	IL-17D-like	PMZ_0025014*
	IL-17R	PMZ_0011633
		PMZ_0011631
		PMZ_0012207
		PMZ_0012673
	Act1/CIKS/TRAF3IP2	PMZ_0005656

* Manually annotated

Supplementary Table 23 – Myelin associated proteins present in the lamprey genome. These genes were found in genome by the GO category myelin and by manual curation of the genome using the gene accession numbers from mammalian (mouse) sequences provided.

Similar to: Gene Name	Gene Symbol	GenBank ID	Lamprey Gene ID
Peripheral myelin protein 22	PMP22	NP_032911.1	PMZ_0004596, PMZ_0025020
Myelin protein zero	MPZ	AAI41227.1	PMZ_0023820, PMZ_0007660
Myelin and lymphocyte protein	MAL	CAA68907.1	PMZ_0016745
Myelin and lymphocyte protein 2	MAL2	NP_849251.1	PMZ_0010355, PMZ_0011570
Myelin transcription factor-1 like protein	Myt1l	NP_001087245.1	PMZ_0022251, PMZ_0020295, PMZ_0001265
Proteolipid protein 1	PLP1	CAA98191.1	PMZ_0025021
CNP	CNP	NP_034053.2	PMZ_0022232
Myelin basic protein	MBP	NP_001020422.1	PMZ_0010899

Supplementary Table 24 – Proteins associated with neurodegenerative diseases present in the lamprey genome. For illustration, listed below is a subset of genes found in the lamprey genome with known associations with diseases or abnormalities in the human central nervous system. These genes were found by manual curation of the genome, using the gene accession numbers provided below from mammalian sequences.
*indicates previously and independently cloned in lamprey.

Disease Relevance	Similar to: Gene Name	Gene Symbol	GeneBank No.	Lamprey Genome ID
Alzheimer's	APP amyloid beta A4 protein isoform 2 precursor	APP	NP_031497.2	PMZ_0007078 PMZ_0007079 PMZ_0001045 PMZ_0001046 PMZ_0019398
	presenilins	PSEN1, PSEN2	NP_032969.1 NP_001122077	PMZ_0002493
	nicastrin	NCSTN	AF240469_1	PMZ_0019510
	gamma-secretase subunit APH-1A isoform 1	APH1A	NP_666216.1	PMZ_0021437
	gamma-secretase subunit PEN-2	PSENEN	NP_079774.1	PMZ_0007928
Parkinson's	protein DJ-1	DJ-1	NP_065594.2	PMZ_0008498
	leucine-rich repeat serine/threonine-protein kinase 2	LRRK2	NP_080006.3	PMZ_0010096 PMZ_0011162
	alpha-synuclein	SNCA	NP_001035916.1	PMZ_0006175
Huntington's	huntingtin	HTT	NP_034544.1	PMZ_0003465 PMZ_0003466 PMZ_0003467 PMZ_0003468 PMZ_0003469 PMZ_0003470 PMZ_0003473 PMZ_0008637
Autism	Neurexin 1	NRXN1	NP_068535.2	PMZ_0015049 PMZ_0016531 PMZ_0017004 PMZ_0003243 PMZ_0009982 PMZ_0008983
	Neuroligin 3	NLGN3	AAA97871.1	PMZ_0002122 PMZ_0002123
	Neuroligin 4	NLGN4	ABV59297.1	PMZ_0003442

Disease Relevance	Similar to: Gene Name	Gene Symbol	GeneBank No.	Lamprey Genome ID
axon regeneration	Semaphorin 3a*	SEMA3A	NP_033178.2 *AAU94361.1	PMZ_0007070 PMZ_0010392 PMZ_0018042 PMZ_0016843 PMZ_0013836 PMZ_0000260 PMZ_0023212
	Netrin family members*	NTN	CAI25793.1 *ABI54137.1	PMZ_0005518 PMZ_0021834 PMZ_0008460 PMZ_0006688
	Reticulon family-partial coverage	Nogo/Rtn4	NP_918943.1	PMZ_0006144 PMZ_0022679 PMZ_0018632
	aggrecan	ACAN	AAA21000.1	PMZ_0006402 PMZ_0015047 PMZ_0021265
	chondroitin sulfate proteoglycan 4	CSPG4	EDL25876.1	PMZ_0022167 PMZ_0017660 PMZ_0018635 PMZ_0018636 PMZ_0014661
	Neurofilament light polypeptide*	NEFL	NP_113971.1 *ABI29893.1	PMZ_0006465 PMZ_0016713
	Vimentin*	VIM	NP_112402.1 *ADN06664.1	PMZ_0023009 PMZ_0010749
	Synapsins*	SYN	AF192747_1 AF192749_1 AF192748_1 AF192750_1 *AAF08808.1 *AAF08807.1 *AAF08806.1 *AAF08805.1	PMZ_0022482 PMZ_0015921 PMZ_0007547 PMZ_0007549 PMZ_0007550

Supplementary Note

Background: natural history, relationship to other organisms and early evolution of vertebrates

Vertebrates arose ~550 million years ago in the Cambrian period [11], and began exploiting terrestrial habitats ~350 million years ago [12]. In exploiting diverse aquatic habitats, and through the transition from aquatic to terrestrial environments, they have undergone substantial morphological evolution, giving rise to a diversity of forms [11,12]. Despite this vast radiation there are many basic structural features common to all vertebrates, though some are only visible in the embryonic stages [11,12]. Some of these features trace their evolutionary origin to structures that were present in invertebrate ancestors. The **respiratory system** develops from the pharynx and in the adult shows diverse forms from gills to lungs. The **single heart**, a multichambered organ pumping the blood around the body, is the major transport system of metabolic substances throughout the tissues. The **liver**, an organ common to all vertebrates, is important in the utilization of the products of digestion [12]. Other features appear to have arisen more recently, since the origin of the vertebrate lineage. Vertebrate bodies are structurally supported by a **dorsal vertebral column** (a set of jointed vertebrae composed of bone or cartilage) and possess **anal fins** that are supported by fin rays. The dorsal nervous system lies close to the vertebral column and expands anteriorly to form a tripartite **brain** with an optic tectum. The **craniate head** consists of a **brain**, **pituitary**, sense organs including **eyes**, and a **skull**. Vertebrates also possess paired **nasal sacs** and **prenasal sinus** as components of their chemosensory system. Other diverse sensory systems, including **lateral line**, **electroreceptors** and the **labyrinth** of canals in the inner ear appear to share common embryological and evolutionary origins [11]. Elaboration of the brain, **neural crest** and **neurogenic placodes** within the embryo seem to associate with increasing complexity of visual, acoustic and lateral-line system within vertebrates [11,13,14].

Lamprey Anatomy in Relation to Other Vertebrates

Lampreys have a simple eel-like body plan, supported by a notochord and a cartilaginous skeleton. They are distinguished from the Gnathostomata by the absence of both jaws and paired fins, the possession of only two semicircular canals in the labyrinth, and a branchial skeleton consisting of joined branchial arches, situated externally to the branchial arteries and nerves and to the trunk arteries [11,15-18]. Other features such as the absence of internal ossification, scales and paired fins; and possession of a single nostril with paired olfactory nerves running into the brain, proterocercal tail, pore-liked gill openings, ceratotrichial fin rays, multicuspid lingual laminae and monomeric hemoglobin are common among extinct and extant gnathostome outgroups [11,15,16]. The lamprey is the only gnathostome outgroup known to possess electrosensory organs [19-21], which share some anatomical similarity with the ampullary organs in some jawed vertebrates [22].

Phylogenetic Relationships among Lampreys

The extant lampreys are separated into three major lineages, which are currently recognized as distinct families [16,23]. Lampreys exhibit an antitropical distribution, with two families endemic to the Southern Hemisphere and the third but largest family (Petromyzontidae) restricted to the Northern Hemisphere [13,15]. Among the Southern Hemisphere lampreys, the family Geotriidae contains a single species, *Geotria australis*, which occurs throughout New Zealand and the southern regions of Australia (including Tasmania) and South America. The family Mordaciidae is represented by two species in Australia, *Mordacia mordax* and *M. praecox*, and one in Chile, *M. lapicida* [24]. Among the family Petromyzontidae, the genera *Entosphenus* (western North America and eastern Asia), *Eudontomyzon* (central/eastern Europe and eastern Asia), *Lampetra* (western Eurasia and western North America), *Lethenteron* (North America, Asia, and southern Europe), and *Petromyzon* (eastern North America and Europe) are widespread. On the other hand, *Caspiomyzon* (Caspian Sea basin), *Ichthyomyzon* (eastern North America), and *Tetrapleurodon* (central Mexico) have more restricted distributions [23,24].

Life History

The sea lamprey (*Petromyzon marinus*) is anadromous and semelparous; adults live at sea as ectoparasites on several fish species and return to fresh water to spawn at the end of their life cycle. The sea lamprey has rarely been caught in the sea, where they may live at depths of up to 500 m. Ocean going sea lampreys may attain lengths up to approximately 1 meter and weights up to 2.5 kg. In spring, the adults migrate upriver to spawn in streams with strong currents and sand or gravel beds. They build nests by removing stones with their suckers and piling them on the downstream side to form a depression in the bed of the stream. Spawning typically begins in May or June when the water temperature reaches approximately 15°C. The female extrudes eggs, which are fertilized externally by the male and drift to the nest edge and remain amongst the stones. This procedure is repeated until the adults are spent, and die. On average 170,000 small eggs are laid by each female, and they hatch in 10-12 days. About 20 days after hatching the larvae drift to quieter waters where they remain in a burrow until metamorphosis. Metamorphosis typically begins in autumn when the animals are about 5.5 years of age, 13 to 16 cm long, and is completed by mid-winter. During the metamorphic period, the lampreys develop eyes, fins, and a tooth-bearing oral disk [25] and initiate their downstream migration. The sea lampreys used in this study are the land-locked population, which invaded the Great Lakes of North America in the 1800s. This population does not return to sea, completing its adult phase in the Great Lakes and spawning in adjacent tributaries.

Impact and Relevance

The 19th century invasion by sea lampreys of the upper Great Lakes devastated one of the greatest inland fisheries in North America. Partly as a consequence of their massive proliferation and the ensuing attempts to eradicate the species, the sea lamprey (among all other lampreys) has captured the attention of biologists, both as a target of biocontrol and as a model for comparative studies. Mature male sea lampreys release the sex pheromone 3 keto-petromyzonol sulfate (3kPZS) [26] through the gills at the onset of spermiation [27]. Ovulatory females are attracted to this novel bile acid pheromone and can thus be lured into traps [26,28-31]. The research in sea lamprey pheromones represents the first attempt to use vertebrate pheromones as part of an integrated pest management strategy [26,28-30,32,33]. For decades, the sea lamprey has been the primary model for spontaneous regeneration of injured spinal cords, and a prominent model for understanding locomotive control. Its giant neurons in the central nervous system are unique *in situ* cellular models for studying neurodegenerative diseases. Studies of its immune system have shed light on the evolution and molecular logics of adaptive immunity. In addition, its liver, which loses the entire biliary tree during metamorphosis, offers a unique opportunity to study human diseases such as biliary atresia, cholestasis and jaundice.

By virtue of its phylogenetic position, the lamprey genome is uniquely poised to provide insight into the ancestry of vertebrate genomes and fundamentals of vertebrate biology. The sequencing of the sea lamprey genome enables researchers to dissect the molecular mechanisms that led to landmark events during early vertebrate evolution such as genome duplication and rediploidization, the expansion of neuroendocrine signaling, myelination of axons, development of appendages and evolutionary diversification of the adaptive immune system. The lamprey genome provides an important resource for reconstructing vertebrate origins and subsequent evolutionary events that have shaped extant organisms and their genomes.

Sea lamprey (*Petromyzon marinus*) genome sequencing

Sea lamprey DNA for whole genome shotgun (WGS) sequencing, fosmid and BAC libraries was derived from liver dissected from a single female lamprey captured from the Great Lakes, USA and processed in the laboratory of Marianne Bronner, California Institute of Technology, where remaining samples are deposited. Production of BAC library CHORI-303 was described previously [34]. Other libraries were cloned into bacterial vectors (POTW13 for WGS, PBACGK1.1 for BACs) and clones were arrayed individually into the wells of growth trays.

Sequencing was performed as previously described [1,35-37]. Briefly, cells (0.7 ul) from bacterial growth trays were automatically transferred to cycle trays for DNA extraction. 2 ul of DI water was added to each tray well, and cycle trays were centrifuged for 30 sec at 1500 rpm to bring the liquid mixture to the bottom of the wells. Cycle trays were then placed on a heat block at 95ºC for 7 minutes to lyse the cells. Subsequently, lysed cycle trays were automatically filled with 1.5ul of water, followed by addition of 1.5ul of ABI sequencing master mix (containing the appropriate primer and Big Dye Terminator) to each well. Sequencing trays were then covered, centrifuged (30 sec, 1500rpm), and thermocycled, using the following conditions: 95ºC for 30 seconds, 50ºC for 15 seconds, 60ºC for 2 minutes for a total of 35 total cycles.

Once thermal cycling was complete, 12ul of isopropanol was added to each well, using a Thermo Scientific Matrix WellMate Microplate Dispenser. Trays were sealed then centrifuged for 30 minutes at 3500rpm. Trays were then removed from the centrifuge, drained, and subjected to an inverted centrifugation step (500rpm for 30 seconds) to remove residual isopropanol. After isopropanol precipitation, pellets were washed by adding 12ul of 70% ethanol, sealing and centrifuging for 15 minutes at 3500rpm. Trays were removed from the centrifuge, drained, and subjected to an inverted centrifugation step (500rpm for 30 seconds) to remove residual ethanol. Finally, the trays were placed in a speed vacuum and dried for 15-30 minutes to permit evaporation of all remaining ethanol.

The prepped plates were re-hydrated and sequenced on ABI 3730 robots, according to manufacturer's directions.

This sequencing effort yielded approximately 19 million sequence reads [18,562,580 short insert WGS end reads (insert size ~4kb), 19,728 fosmid end reads (insert size ~40kb) and 379,929 BAC end reads (insert size ~150kb].

Genome assembly

Several analyses were performed prior to initiating the assembly, these provided insight as to the selection of the assembler (*Arachne* [38]), parameters for assembly, and the expected fraction of the genome that can be represented in an assembly. These analyses revealed that: 1) ~30% of the genome consists of high-identity repeats that are capable of disrupting the assembly, 2) the genome is apparently highly polymorphic and 3) the genome is high in GC-content (analyses of GC-content are covered in Supplement 4).

WGS vs. WGS depth of coverage analysis

Repeat Content: We performed a pre-assembly analysis of the WGS dataset in order to gain insight into 1) genomic sampling obtained through the WGS sequencing effort and 2) the repetitive content of the lamprey genome. This analysis was performed by selecting a subset of 10,000 high quality shotgun sequence reads (>500 bp at Q20) and aligning these to the complete dataset of 18.5 million WGS reads (Q20 trimmed). A complementary analysis was also performed by aligning 10,000 trimmed WGS reads from a single human genome [39] to a complete dataset 12.1 million WGS reads (Q20 trimmed). All reads were downloaded from NCBI Trace Archives in .scf format and processed with *phred* [40-42] to generate base calls and quality scores. Alignments to human and lamprey WGS sequence datasets were performed using Megablast [43].

Alignments were summarized as the average depth of alignment over each of the 10,000 selected reads. Supplementary Figure 1 shows the distribution of coverage depth estimates, considering all alignments >400bp in length and >95% sequence identity. Both lamprey and human datasets have unimodal distributions, with a large tail to the right of the distribution. The humped portion of the distribution corresponds to alignments involving low-copy DNA, with the mode of the distribution approximating the expected depth of coverage across the entire genome. Aside from differences in sequence depth, there are two notable differences in these distributions. First, the low copy distribution for the lamprey WGS project shows substantially higher dispersion than the low copy distribution for the human WGS project. This suggests that cloning biases may play a stronger role in determining sampling probabilities for various regions of the lamprey genome, relative to the human genome. Second, the proportion of reads falling to the right of the low copy distribution is significantly higher in lamprey, relative to human. This is taken as evidence that "assembly-relevant" repetitive DNA comprises a much larger fraction of the lamprey genome. Specifically, ~30% of the genome consists of repeats that are sufficiently long and sufficiently similar that they could potentially disrupt linear assembly (400bp intervals that have more than one >95% repeat somewhere else in the genome). By comparison, the same analysis in human indicates that <7% of the human genome consists of similar "assembly-relevant" repeats (Supplementary Figure 1). Taken at face value, this might seem to indicate that repeats should have an extremely disruptive effect on assembly. However, performing a similar analysis using paired-end BAC ends revealed that repeats are strongly clustered at the subchromosomal (~100 kb) scale [4].

Evidence for High Allelic Divergence: To gain insight into the potential influence of allelic polymorphism, we estimated depth of coverage as described above, with varying thresholds for percent nucleotide identity between aligning sequences. Distributions of coverage depth were estimated using sequence identity thresholds of 90, 95, 97, and 99%. Comparison of these distributions revealed a large shift in coverage depth as the required percent identity for alignment was increased from 97% to 99% (Supplementary Figure 2). A similar shift was not observed for human, although increasing the threshold always shifted the distribution to the left (Supplementary Figure 2). This abrupt shift in modal coverage depth is consistent with high sequence divergence (~2%) between alleles at most lamprey loci. Alternately, these patterns could be explained by recent duplication of the lamprey genome, however preliminary FISH studies yield no evidence for recent broad-scale duplication (Supplementary Figure 3).

Assembly

Assembly of the lamprey genome was performed using a ~19 million sequence reads [18,562,580 short insert WGS end reads (insert size ~4kb), 19,728 fosmid end reads (insert size ~40kb) and 379,929 BAC end reads (insert size ~150kb] with *Arachne* [38] parameterized for assembly of a outbred diploid genome. Following assembly by the *Assemblez* module, contigs corresponding to divergent haplotypes were assembled together using the *Rebuilder* module parameterized with liberal settings that permit merger of divergent haplotypes (http://www.broadinstitute.org/crd/wiki/index.php/Arachne_Main_Page) and then haplotypes were joined using linking information from end read mapping information. End mapping information was incorporated via the *ExtendHaploSupers* module in a series of steps that prioritized the number of end reads supporting linkages between contigs and the source of end mapping information (Shotgun reads vs. large-insert clones). Specifically, paired end mapping information was incorporated in the following steps, where subsequent linkages may not supplant linkages that have been previously identified at a more stringent threshold: at least four paired end linkages from large-insert clones, at least four paired end linkages from large-insert clones or WGS clones, three paired end linkages from large-insert clones, three paired end linkages from large-insert clones or WGS clones, two paired end linkages from large-insert clones, two paired end linkages from large-insert clones or WGS clones, a single paired end linkage from a large-insert clone and finally, a single paired end linkage from a WGS clone.

Analyses were performed with and without data from long insert clones (BACs/fosmids) because these clones derived from a tissue source that was not identical to the WGS reads (WGS clones are from liver and long-insert clones are from blood), and there is evidence that tissues may differ in terms of DNA content [3]. The highest N50 contig size and smallest number of contigs was obtained when the base-run included BACs and fosmids. This implies that tissue variation (if any) is less disruptive than the exclusion of these data. The final draft assembly is 0.816 Gb and is distributed across 25,073 contigs. The size of the assembled fraction is within the range of expectation, given the content and distribution of repetitive reads within the genome. Contiguity of the assembly permits the identification of multiple genes per scaffold across a majority of the assembly: half of the assembly is in 1214 scaffolds of 173 kb or longer and the longest scaffold is 2.4 Mb. The assembly was screened for contamination and submitted to the NCBI assembly archive (NCBI accession number AEFG00000000).

Unassembled Reads: Exclusion of disruptive reads is a critical aspect of genome assembly. Reads may be excluded for a number of reasons including: 1) low sequence quality, 2) exceedingly high copy number, 3) likely contaminant sequence and 4) likely read chimerism. Notably, apparent read chimerism could potentially result from either cloning artifacts or bona fide structural variation within the sequenced animal's genome. In total, 7,249,772 sequence reads were excluded from the genome assembly. A complete list of excluded reads and justification for their exclusion is provided in Supplementary Table 1 (included as a separate file).

Fraction of the Somatic Genome Represented: we used read-coverage statistics to estimate the fraction of the (low copy) genome that was represented by Q20 shotgun reads. The distribution of roughly single-copy coverage-depth statistics for lamprey (≤30) is unimodal with a mean of 4.24 and standard deviation of 8.92. For a normal distribution with these values, 2.4% of the distribution is expected to fall below copy number 0.5. Thus, we estimate that 97.6% of (samplable) low copy sequence was captured at least once through our sequencing efforts.

Another generalized method for evaluating "completeness" of genome assembly is based on representation of a set of core eukaryotic genes (CEGs)[44,45]. Based on representation of CEGs, one would estimate that ~75% of expected protein coding sequences can be readily identified the current assembly by standard homology searches (Supplementary Table 2). However, we noted that the numbers of incomplete and complete CEGs identified varied depending on the exact algorithm used to search the genome, moreover, additional CEGs were identified in the genome by searching our transcriptome assembly for CEGs and realigning these to the genome (>98% identity). It is possible that divergence over the last 1 billion years of independent evolution has prevented the detection of homologs for some fraction of CEGs. In this context, it may be worth noting that the pattern of conservation of CEGs is not consistent with either evolutionary divergence or genome completeness, as outlined by the authors of the program CEGMA[45]. Specifically, CEGs from group 4 are expected to show the highest level of conservation and group 1 the lowest level of conservation. For taxa that are evolutionarily distant from the species used to develop model CEGs one might expect to identify more "group 4" CEGs and fewer from other groups. With respect to estimates of completeness, it is also important to note that it has been estimated that ~20% of the lamprey germline genome is deleted during embryonic development. It is currently unclear how this deletion will affect estimates of "completeness" based on conserved protein-coding genes, but we expect that it should result in some reduction of completeness statistics. Overall, we expect that the assembly will provide reasonable coverage of genes that are present in the vast majority of lamprey cell types, but recommend an abundance of caution in interpreting the perceived absence of homologs for any specific vertebrate gene.

Collection and identification of repetitive sequence in sea lamprey

Repetitive sequences were collected with *RECON* (version 1.06, http://www.repeatmasker.org/) [46], with a cutoff of 10 copies. This resulted in a total of 9,880 repetitive sequences. After filtering putative gene families (sequences matching non-transposase proteins), 8,790 repetitive sequences remained. Thereafter, a subset of the repetitive sequences was manually curated to verify their identity, individuality and 5'/3' boundaries. First, the relevant sequence was used to search the sea lamprey genomic sequences and at least 10 hits (*BLASTN* [47], E< 10^{-10}) plus 100 bp of 3' and 5' flanking sequence were recovered. Recovered sequences were then aligned using "*dialign 2*" [48], with the resulting output examined for the presence of possible boundaries between putative elements and their flanking sequences. A boundary was defined as the position to which sequence homology is conserved over more than half of the aligned sequences (e.g. six of ten sequences). Sequences flanking the boundary of the putative element were compared with those of known transposable elements (TE) and were examined for the possible presence of target site duplication, which is created during insertion of most DNA transposons. Each transposon family has unique terminal sequences and target site duplication, which can aid in the identification of a specific transposon class [49]. For some large transposable elements, fragmented sequences identified by *RECON* were joined to derive a compete sequence. If a particular lamprey sequence is similar to a known transposon at the nucleotide or protein level (*BLASTX* or *BLASTN* E< 10^{-5}, RepBase14.12), it was assigned to that repeat class. Finally, the putative terminal sequence was aligned (directly and inversely) using "gap" in GCG to detect possible inverted or direct repeats. This classification scheme is summarized in Supplementary Table 3.

Manually-curated sequences were compared to the remaining repetitive sequences using *RepeatMasker (RepeatMasker-open-3-2-7)* [50]. Lamprey repeats matching the curated sequences were considered to belong to the same repeat family and excluded from the repeat library. The criteria for exclusion were based on previously published guidelines [49]. Specifically, if two elements share 80% or higher identity over 80% of their element length, they are considered to be the same family. If a repetitive sequence matches the curated sequences without reaching the "80% identity in 80% length" criteria, this sequence is retained and is considered to belong to a new family in the same superfamily. This process led to the exclusion of 1,234 repetitive sequences from original *RECON* output. As a result, the current lamprey repeat library is composed of a total of 7,556 repetitive elements. The remaining repeats were then searched for homology to known repeat classes in *RepBase* 14.12 (www.girinst.org/repbase/) [51], using RepeatMasker and BLAST (*BLASTX* E< 10^{-5}), to identify elements similar to known other transposable elements. If an uncurated sequence matched any known TE, it was considered to be homologous to that element. The number

of families assigned to different classes of transposable elements in the curated and uncurated subsets are provided in Supplementary Table 4.
We inferred the phylogenetic distributions of repeats searching "vertebrate" and "invertebrate" subsets of *RepBase* using the program *RepeatMasker* (Supplementary Figure 4). After removing low complexity (N = 210), simple sequence repeats (N = 210) and predicted repeats that could not be directly aligned to the assembly by *RepeatMasker* (N = 634), sequences with matches to known vertebrate elements were considered to have evolved prior to the gnathostome/lamprey split and to have been retained in both lineages. These were assigned to the "Vertebrate Repeats" category (N = 167). Remaining lamprey repeats were searched for a match to other Repbase sequences. These were assigned to the "Invertebrate Repeats" category (N = 118). Remaining sequences that did not produce a match to either database were assigned to the "Lamprey-specific Repeats" category (N = 6209). This last category of repeats corresponds to repetitive elements that fall into at least three categories: 1) repeats that were uniquely derived within the lamprey lineage, 2) repeats that were present in the ancestral "invertebrate" lineage but have not-yet been sampled from invertebrate taxa, or 3) repeats that were present in the ancestral vertebrate lineage but have not-yet been sampled from invertebrate taxa.

Sea lamprey transcriptome sequencing

EST sequencing

In order to add experimental support to the *in silico* gene predictions for the lamprey genome, we produced cDNA sequence data from several lamprey tissues: 1) whole brain from adult male and female; 2) olfactory tissue from adult male and female; 3) pooled adult liver, muscle, testis, skin, gill from male and female; 4) pooled embryos: 10 stages from fertilization to completion of digestive tract; 5) adult muscle from male and female; and 6) embryo stages 2-12 days post fertilization. The sequencing protocol was as follows.

1. A magnetic bead preparation was used to purify DNA from arrayed EST libraries. Growth archive trays, and prepping solutions were arranged on a Biomek Laboratory Automation Workstation. Cycle trays, placed on magnets, were also added to the deck. 7.5ul of archived bacterial culture was added to each cycle tray well, followed by 15ul of Homogenation Bead Solution. Tip washes occurred between every growth archive tray and bead addition. Cycle trays were placed on the magnets and formed a ring shaped pellet in each well, composed of magnetic particles (Seradyn Indianapolis, IN 46268 USA; Lot#200935 Part#44152105050450) and plasmid DNA. After removing supernatant, beads were washed by adding 25ul of 85% EtOH.

2. 3ul Sequencing Reactions: The deck of a Biomek Laboratory Automation Workstation was filled with prepped cycle trays (containing purified plasmid

DNA).1.5ul of water, followed by 1.5ul of ABI 3730 master mix (appropriate primer and Big Dye Terminator) was added to each well. Tip washes occurred between the water and master mix additions. Sequencing trays were then were thermocycled, using the following conditions: 95ºC for 30 seconds, 50ºC for 15 seconds, 60ºC for 2 minutes for a total of 35 total cycles.

3. Cleanup and Sequencing: 15ul of 100%EtOH/3M Sodium Acetate (100:1; ph5.2) precipitation solution was added to each cycle well. Trays were then centrifuged at 3500rpm for 30 minutes to pellet sequencing products. An inverted centrifugation was then performed at 500rpm for 30 seconds to remove the precipitation solution. Pellets were then washed by the addition of 15ul of 70% EtOH and centrifugation at 3500 rpm for 15 minutes. To remove the excess EtOH, a final inverted centrifugation step was performed at 500rpm for 30 seconds. The trays were then dried in a speed vacuum for 5-20 minutes. Finally, sequencing trays were loaded on ABI 3730 sequencers for automated sequencing.

mRNA-Seq sequencing

Lamprey RNA for 454 GS20, 454 GS FLX, and Illumina mRNA-Seq sequencing was extracted from twenty samples (Supplementary Table 5).

454 GS20 sequencing: Sea-lamprey olfactory epithelium from both adult life stages was sequenced using Roche 454 GS20 Life Sciences technology. Tissues were flash-frozen in liquid nitrogen and stored at -80°C until extraction. Total RNA was extracted using the VersaGene RNA Tissue Kit (Gentra, Inc), quality-checked by gel electrophoresis, and quantified using a NanoDrop ND-1000 spectrophotometer. Total RNA was used as template for first-strand cDNA synthesis using the SMART cDNA Synthesis Kit (Clonetech Laboratories, Inc.). Single-strand cDNA was amplified in 13 cycles of LD-PCR using the Advantage® 2 PCR Kit (Clonetech Laboratories, Inc.) according to manufacturer's instructions. PCR products were purified using QIAquick PCR Purification Kit (Qiagen, Inc) and concentrated on Millipore YM-30 (MWCO 30,000) columns. cDNA were submitted to Michigan State University Research Technology Support Facility (MSU RTSF) for 454 GS20 sequencing. One complete 454 run was performed for cDNAs from each of the two life stages. The TIGR SeqClean (http://compbio.dfci.harvard.edu/tgi/software/) sequence trimming pipeline was used to remove low quality, low complexity, polyA and adapter sequences from the cDNA sequences. 409,174 raw reads were identified with an average read length of 106 nucleotides, and 373,391 high-quality reads with an average read length of 93 nucleotides. This resulted in a total raw read length of 43,438,497 nucleotides, and high-quality read length of 35,035,388 nucleotides (81%).

454 GS-FLX sequencing: 454 GS-FLX was then used to sequence four normalized samples of sea-lamprey larva/parasite brain, larva liver, parasite liver, and adult brain. Collected tissues were flash-frozen in liquid nitrogen, homogenized in extraction buffer while frozen and RNA was extracted by the PerfectPure RNA method (5 Prime Inc., Gaithersburg, MD). RNA was treated with 2 rounds of DNase I while immobilized on the column to assure complete removal of contaminating genomic DNA. RNA quality was evaluated on agarose gels by size distribution and relative intensity of 28s and 18s ribosomal RNA bands. Elimination of genomic DNA was verified using 40 cycles of no-RT PCR amplification for 40s and 60s ribosomal RNA (40s: 40sF 5'-ACCTACGCAGGAACAGCTATGAC-3', 40sR 5'-CGACGAATTCCACCACATTG-3', 60s: 60sF 5'-CGCATCCGCGCAATG-3', 60sR 5'-GTCGGGTATGTCCACGATCTG-3'). In this quality assurance step, RNA samples, in the absence of reserve transcription, were subjected to PCR using single-exon and intron-nested primers. Full-length cDNA was generated using the SMART cDNA synthesis kit (Clontech Laboratories Inc., Mountain View CA). SMART- amplified cDNA was normalized with the Evrogen (Moscow, Russia) *Trimmer* duplex specific nuclease (DSN) protocol (Zhulidov et al. 2004). Agarose gel electrophoresis and q-RT-PCR were used to assess the effectiveness of the SMART cDNA synthesis and Evrogen normalization protocols. Full length enriched cDNA flanked by known primers were reliably produced. Normalized cDNA ranged between 0.5 and 4.5 kb in sea lamprey, and appeared on agarose gels as a distribution of cDNA of the same size as non-normalized cDNA but lacking predominant banding pattern seen in non-normalized cDNA. cDNA were submitted to the MSU RTSF for 454 GS-FLX sequencing. TIGR SeqClean was used for trimming and quality-filtering. A total of 592,194 raw reads were identified with average read length of 184 nucleotides, and a total of 403,472 high-quality reads averaging 186 nucleotides, resulting in a total of 108,690,455 raw nucleotides, and 75,219,931 clean nucleotides (69%).

454 GS-FLX sequencing was also used to generate reads from PCR subtracted cDNA of myeloid cells. RNA was extracted using the RNeasy Mini Kit (Qiagen), treated with Turbo DNase (Ambion) and measured for quantity and quality via spectrophotometry and gel electrophoresis. Library construction was performed using a variation of the Clontech SMART system in which the 5' and 3' PCR adapters (5' sequence; 3' sequence) contain type IIs restriction enzyme sites (*MmeI*). Post-amplification of the library was performed with a single PCR primer (manufacturer proprietary sequence) that maintained the MmeI site at the 5' and 3' termini. This product was titrated for the optimum number of PCR cycles, to avoid over-cycling of the product.

Optimally-cycled products were then normalized using a duplex-specific nuclease (DSN) that preferentially digests double-stranded DNA in the presence of single-stranded DNA (Trimmer; Evrogen). Briefly, the cDNA library DNA is boiled and allowed to re-anneal for approximately 5 hours in a buffered salt solution. During this time, high-copy molecules re-anneal while the low-copy molecules maintain a single-stranded state. Following incubation, duplex-specific nuclease (DSN), mentioned above, is added to degrade dsDNA molecules, leaving single-stranded sequences as template for re-amplification using the single primer discussed above. At this step, a second PCR cycle titration was performed at this stage to prevent over-cycling. Optimally-cycled (2nd stage) products were then purified by binding biotin labeled adapters to M280 streptavidin beads (Invitrogen). The inclusion of MmeI restriction sites allowed for cleavage of the poly-A tail from the 3' end of cDNAs and removal of both 5' and 3' adapter sequences prior to sequencing. Normalized and purified products were then sequenced according to the standard 454-FLX library protocol and resulting reads were deposited in GenBank Trace Archives (www.ncbi.nlm.nih.gov/Traces/trace.cgi) and assembled using Newbler (Roche).

Illumina mRNA-Seq: Total RNA from the four previous GS-FLX sequencing runs was combined into two samples of larva/parasite liver and larva/parasite/adult brain. These were submitted to WUGSC for sequencing. Samples were poly-A selected, fragmented, and randomly primed for reverse transcription to cDNA using the Illumina mRNA-Seq-8 Sample Prep Kit RNA (Illumina, Inc.), and sequenced in two separate runs using an Illumina GA2 sequencer. A total of 119,412,170 reads of 50-nucleotide length were produced, yielding 5,970,608,500 nucleotides (6.0 Gb).

In a second Illumina sequencing round, sea lamprey tissues were sampled individually from larval intestine, larval kidney, small parasite kidney, small parasite proximal intestine, small parasite distal intestine, adult intestine, and adult kidney. Samples were flash-frozen in liquid nitrogen and stored at -80C until RNA extraction. Total RNA was extracted from tissue samples using TRIzol Reagent (Life Sciences Corp.) according to manufacturer's instruction. Total RNA samples were submitted to the MSU RTSF for subsequent processing and sequencing using the Illumina mRNA-Seq-8 Sample Prep Kit RNA (Illumina, Inc.). Samples were poly-A selected, fragmented, and randomly primed for reverse transcription to cDNA, and then sequenced by an Illumina GA2 machine yielding approximately 40 million 75 base reads per sample for a total of 21,246,929,250 nucleotides (21 Gb) of mRNA-Seq sequence. Sequence was quality-filtered using the Illumina Genome Analyzer Pipeline, yielding 14,516,488,350 nucleotides (15 Gb) of pass-filter sequence (68%).

Finally, sea-lamprey tissues were sampled from seven stages of embryo development: late blastula (stage 18), gastrula (stage 20), neurula (stages 22a and 22b), neural-crest migration (stages 23, 24c1, and 24c2)[52]. Samples were processed as above, with mRNA-Seq read length increased to 100 nucleotides. Sequencing yielding 41-45 million 100-base reads per sample totaling 29,046,625,200 (29 Gb) of raw mRNA-Seq sequence, and 17,856,991,800 nucleotides (18 Gb) of pass-filter sequence (61%).

All samples were submitted to the National Center for Biotechnology Information's Sequence Read Archive (NCBI SRA) under SRA048296.1 (Supplementary Table 5).

Gene annotation

Annotations for the lamprey genome assembly were generated using the automated genome annotation pipeline *MAKER* [53] which aligns and filters EST and protein homology evidence, identifies repeats, produces *ab initio* gene predictions, infers 5' and 3' UTR, and integrates these data to produce final downstream gene models along with quality control statistics. Inputs for maker included the *P. marinus* genome assembly, *P. marinus* ESTs, a species specific repeat library, and a protein databases containing all annotated proteins for human, mouse, chicken, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Xenopus tropicalis*, *Strongylocentrotus purpuratus* (sea urchin), *Branchiostoma floridae* (lancelet), *Lottia gigantea* (limpet), *Ciona intestinalis* (sea squirt), *Trichoplax adhaerens*, *Nematostella vectensis* (sea anemone), *Danio rerio* (zebrafish), and *Takifugu rubripes* (pufferfish) combined with the *Uniprot/Swiss-Prot* [54] protein database and all sequences for Chondrichthyes (cartilaginous fishes) and Myxinidae (hagfishes) in the *NCBI* protein database [55,56]. *Ab initio* gene predictions were produced inside of MAKER by the programs *SNAP* [57] and *Augustus* [58]. MAKER was also passed *P. marinus* mRNA-seq data processed by the programs *tophat* and *cufflinks* [59]. MAKER was run in a bootstrap fashion with the output gene models of one run acting as inputs for retraining *ab initio* gene predictors and better informing mRNA-seq alignment junctions for *tophat* and *cufflinks*. A total of three iterative runs of *MAKER* were used to produce the final gene set.

Following genome annotation, final gene models were then analyzed using the program *InterProScan* [60] to identify putative protein domains and *GO* [61] functions for each gene. All data was loaded into a *Chado* [62] database to allow for easy annotation distribution, online viewing, and remote annotation curation via a modified version of the program *Apollo* [63]. The final annotation set contained a total of 26,046 genes encoding 26,204 transcripts (an incomplete set of transcript variants), comparable with other large vertebrate genomes. The number of *MAKER* gene models compared to the true number of genes is likely to be somewhat inflated due to the splitting of genes that results from the fragmented nature of the genome assembly.

Identification of conserved noncoding elements (CNEs) in the lamprey genome

Previous studies have identified significant numbers of non-coding elements that have regulatory potential and remain highly conserved across the jawed vertebrate lineage. However, it has been suggested that only a small fraction of these are retained in lamprey [64]. The lamprey assembly was searched for sequences homologous to conserved non-coding sequences previously identified in comparisons between human and Fugu [65] and human and *Callorhinchus milii* [66]. BLASTn (2.2.25+) was used with word size set to 5, and gap existence and extension penalties of 1. A complete list of CNEs aligning to the lamprey genome with an e-value ≤ 0.005 are presented in Supplementary Table 6 (included as a separate excel file).

GC-content, codon usage bias and amino acid composition

Global GC heterogeneity in the lamprey genome

The entire genome assembly showed GC-content of 46%. This value resembled that for the repeat-masked genome assembly (46%) and whole genome shotgun reads (45%). To explore the GC heterogeneity within the lamprey genome, in comparison with other species, the entire genome assembly was analyzed with 10 kb non-overlapping windows (Supplementary Figure 5). The lamprey showed an intermediate degree of intragenomic GC heterogeneity whose peak was higher than mammals and birds, but much less sharp than invertebrates and teleost fishes. At the same time, the lamprey genome has been shown to be one of the most GC-rich among vertebrates whose genome has been already sequenced (Supplementary Figure 5).

Lamprey genes generally show high GC-content [67,68]. In mammals and birds with high intragenomic GC heterogeneity, protein-coding genes with high GC-content are often embedded in GC-rich genomic regions [69,70]. In those species, the GC-content for protein-coding regions, usually represented by GC-content of third codon positions (GC3), exhibits bimodal distribution [71]. Our analysis showed that GC3 of lamprey genes are generally high and sometimes close to 100% (Supplementary Figure 6). The distribution of lamprey GC3 was unimodal unlike in mammals and birds. To analyze the influence of local genomic GC-content to that of protein-coding regions harbored, we examined possible correlation between these two values (Supplementary Figure 7). This analysis did not reveal significant correlation (correlation coefficient r = 0.034), which indicates that in the lamprey genome the GC-content of protein-coding regions is not influenced by background GC-content of the genomic regions harboring the genes.

GC-content, codon usage bias and amino acid composition in protein-coding sequences

GC-content in protein-coding sequences significantly modulates frequencies of particular codons and amino acids. Major explanations for codon usage bias include mutational bias, translational selection and genetic drift. Mutational bias and translational selection which depends on tRNA abundance play important roles in the codon usage in prokaryotes [72-74], protists [75,76] and multicellular eukaryotes such as worms [77,78], flies [79] and plants [80,81]. In some vertebrates, there is a strong relationship between GC heterogeneity and codon usage bias [82]. Among vertebrates, teleost fishes [83] and amphibians [84] have less intragenomic GC heterogeneity, and translational efficiency influences codon usage bias in addition to mutational bias in their genomes [70,85]. A recent small-scale study showed that lamprey genes exhibit peculiar codon usage and amino acid composition [86]. Using the genome-wide dataset, below we examined the properties of protein-coding sequences in the lamprey genome.

To perform a genome-wide assessment of codon usage bias and amino acid composition in lamprey genes, we used the coding sequences predicted with MAKER [53] after discarding alternative splicing variants except the largest one per gene. To avoid any bias imparted by small sequences, we excluded those shorter than 300 bp. The remaining 18,444 coding sequences were analyzed as follows. In this dataset we identified 50 highly expressed and 50 lowly expressed genes based on RNA-seq reads to investigate possible influence of gene expression levels to codon usage bias and amino acid composition. To analyze the codon usage bias and amino acid composition, we performed correspondence analysis (COA) on the relative synonymous codon usage (RSCU) [87] and on the amino acid composition values using the software CodonW [88] (http://codonw.sourceforge.net). We also calculated overall GC-content and GC-content at third codon positions (GC3) for each protein-coding gene.

The principal axis (axis 1) generated by COA on RSCU values of lamprey genes represented 17% of the total variation while the second principal axis represented only 7% (grey dots in Supplementary Figure 8). This indicates that axis 1 explains a substantial proportion of variation of codon usage bias among the genes in the genome. The position of genes along axis1 was strongly correlated to their overall GC-content (r = 0.73), and especially GC3 (r = 0.96). However, when the coordinates of individual sequences along axis 1 and axis 2 were plotted for the highly and lowly expressed genes, they were completely intermingled (blue and red squares in Supplementary Figure 8). Furthermore, the comparison of RSCU values between the highly and lowly expressed genes did not yield any correlation with the tRNA abundance based on tRNA gene copy numbers included in Supplementary Table 7. Overall, the primary factor influencing codon usage bias in the lamprey is mutational bias (i.e. GC-content), and no obvious impact of translational efficiency on codon usage bias was detected.

To assess possible deviation of sequence properties of lamprey protein-coding regions in comparison to other species, we downloaded genome-wide protein-coding sequences for diverse vertebrates and invertebrates from *Ensembl*[89] and archives for individual genome projects. Using species-by-species concatenated protein-coding sequences, we calculated the RSCU and performed a correspondence analysis (Figure 2). The contribution of the first and second axis of the correspondence analysis is 74% and 13%, respectively. It is remarkable that the lamprey is located distantly from other vertebrates as well as from invertebrates. The principal axis (axis 1) was strongly correlated with GC3 (r = 0.99) (Supplementary Figure 9).

We also examined amino acid composition in deduced peptide sequences encoded by the predicted lamprey genes. Distribution of individual genes along the first and second principal axes of the COA on amino acid composition is shown in Supplementary Figure 10. The two principle axes generated by COA of the amino acid frequencies for lamprey genes explained 17% and 11.6% of the total variability. The first principal axis showed strong positive correlation with overall GC-content of the genes (r = 0.72), but not with GC3. However, neither the first axis nor the second axis generated by COA on amino acid composition demonstrated any discrimination of sequences based on expression level (Supplementary Figure 10).

A comparative analysis, similar to the one for the codon usage bias, was performed for amino acid composition after concatenating all the peptide sequences for each species (Figure 2). The contribution of the first and second axis of the amino acid correspondence analysis was 57% and 26%, respectively. Lamprey seems again to be an outlier in comparison with the other species, especially vertebrates. The primary axis is strongly influenced by the overall coding GC-content, as the correlation between the two is high (r = -0.89) (Supplementary Figure 11).

Taken together, these analyses suggest that the major factor influencing codon usage bias and amino acid composition in the lamprey genome coding GC-content. Although codon usage bias is often explained by translational efficiency in general, this does not seem to hold in the lamprey. The unique patterns of the lamprey in protein-coding GC-content, codon usage bias and amino acid composition as well as the relatively high GC-content in the genome assembly expand our knowledge of how far a vertebrate genome can deviate from the standard which we have learned from model vertebrate species.

Phylogenetic analysis of lamprey genes

A genome-wide phylogenetic analysis including 50 vertebrate genomes, 2 additional chordates and 3 outgroups was performed using the Ensembl tree reconstruction pipeline and Ensembl compara database, Build 64 [90]. All genes are clustered with hcluster_sg [91] according to their sequence similarity [92]. A multiple alignment is built for each cluster using MCoffee [93], then TreeBeST [91] is used to reconstruct a consensus tree for each family using 2 Maximum Likelihood (ML) and 3 neighbor-joining (NJ) trees. The final tree consists of parts from all 5-member trees, such that the number of duplications and gene losses are minimized and nodes with a higher bootstrap support are favored. Further details on the pipeline as well as the full set of trees are available in the release 64 of Ensembl (http://e64.ensembl.org)[94].

Among 10,402 Ensembl predicted lamprey protein coding genes initially fed into the pipeline, 9,888 are included in trees encompassing several species. Excluding the lamprey genome, the most recent common ancestor of all the other vertebrate genomes in Ensembl is the Euteleostomi (synonymous with Osteichthyan) ancestor. Prior to the addition of the lamprey genome, out of ca. 2,150 gene families could be tracked up to the Euteleostomi ancestor only. The lamprey genome brings evidence that at least a third of these genes also existed in the Vertebrata ancestor. From the full set of trees, we inferred 7,670 Vertebrata speciation nodes, 4,496 Vertebrata duplication nodes, and 1,796 lamprey-specific duplications.

Duplication events were inferred on the basis of tree structure and assigned to the most-inclusive phylogenetic node. Lamprey is deeply diverged from all other living animal groups and any phylogenetic analysis involving the species can only partially account for several confounding factors (i.e. substitution bias, mutational saturation, paralog loss over evolutionary time or via programmed deletion[95] and incomplete representation of genes in the current assembly). However, combined information from 8,693 trees with a node in Euteleostomi, Vertebrata, and Chordata (87.0% of the protein coding genes of 55 Ensembl species) reveals signatures consistent with 2R WGD and provides new insight into the timing of these duplications (Supplementary Figure 12). The size distribution of Euteleostomi (bony fish and all limbed vertebrates, but not lamprey) orthology groups is unimodal, indicating no large-scale signature of duplication in Euteleostomi lineages (except teleost fish). In contrast, the size distribution of Chordata and Vertebrata orthology groups are multimodal, indicating that several orthology groups retain a signature of one or two duplication events in which one or more duplicated paralogs have not been lost. Thus, inclusion of lamprey paralogs in gene tree reconstructions reveals numerous gene duplications that map to the vertebrate stem lineage.

Sizes of gene families

Any speciation node represents an ancestral gene found in several species, the extant copies of which define a group of homologous genes (homology groups). Under normal circumstances, we expect the average size of the homology groups of an ancestral node to be roughly equal to the number of species contained within the node (given either equal or low birth and death gene rates). A duplication event will result in a situation wherein nodes predating the duplication event contain approximately twice as many genes as there are taxa contained within the node. Following a WGD event, purifying selection tends to eliminate additional gene copies. However, one expects a small fraction of the duplicated copies to be retained. Under this model, a genome-wide tree reconstruction analysis should infer an excess of duplication events on branches with retained duplicates, leading to two sub-families. For genes with no retained duplicates, the distribution of the homology group sizes for different taxonomical nodes is expected to peak at a value corresponding to the number of extant species for nodes. Where duplicated have been retained, we expect the distribution of homology group sizes for nodes followed by one WGD to contain a second peak, centered at roughly twice the number of extant species under that node.

In the chordate lineage, the 1R/2R WGD events affected at least 49 gnathostome species present in Ensembl, and a third lineage-specific (3R) WGD event affected 5 teleost fish species [96]. The distribution of the sizes of the homology groups of deep chordate nodes is expected to be multi-modal, with modes that are roughly a linear combination of 49 and 5, depending of the position of those nodes with respect to the WGD events. The number of fish species being relatively small, we can ignore their effect as it is obscured by other peaks. In summary, we expect to find one peak only at ca. 50 genes for nodes that do not immediately postdate WGD events, a bi-modal distribution (peaks at 50 and 100 genes) for nodes that immediately postdate a single WGD event and a quadri-modal distribution (peaks at 50, 100, 150 and 200 genes) for nodes that immediately postdate 2 WGD events (Supplementary Figures 12 and 13).

This analysis can be biased by several factors and the results must be interpreted with caution. First, as lamprey gene annotation is incomplete we expect fewer Vertebrata speciation nodes, which could result in a failure to identify duplication events that occurred before the lamprey/gnathostome split (Supplementary Figure 14). Second, the tree reconstruction program (TreeBeST) is designed to optimize gene tree topology in order to minimize the number of duplications and gene losses. If a duplication event produced two Vertebrata families, but only one of two lamprey genes is annotated, TreeBeST will favor a Euteleostomi-specific duplication with the single lamprey gene as an outgroup (Supplementary Figure 15). Other analyses show that differential gene losses are a dominant effect in the comparison between the lamprey and the gnathostomes. Supplementary Figure 16 illustrates expected patterns of family size under such a scenario. In this case, the curves are expected to be lower, but with similar shape to Supplementary Figure 15. Finally, long-branch attraction between pairs of orthologous lamprey genes (explained by an atypic GC content and lack of sister taxa) can give the appearance that paralogous lamprey genes derive from a lamprey-specific duplication even if they are derived from a duplication event in the Vertebrata ancestor (Supplementary Figure 17). These factors would tend to prevent the identification of duplication events in the vertebrate ancestral lineage and affect the distribution of homology group sizes.

The distribution of homology group sizes mapping to the Euteleostomi node (Supplementary Figure 13) contains 16,637 homology groups and appears to be uni-modal. The 3R WGD should appear as an additional peak of families with ca. 55 genes (44 Tetrapoda + 2 x 5 Teleost). As mentioned earlier, this peak is not visible because the Teleost genomes included in this analysis represent only a small fraction of the total number of species: the peak is obscured by the large number of tetrapod genomes included in this analysis. The Chordata and Vertebrata distributions are quadri-modal (Chordata: N = 7,249 homology groups; Vertebrata: N = 7,670 homology groups). This result is consistent with two rounds of WGD after the Chordata node and before the Euteleostomi node. Because of the possible artifacts listed before, such analyses cannot conclusively determine whether the two rounds of WGD occurred before or after the Vertebrata node, but seem to provide evidence for the presence of two duplication events before diversification of the euteleostome lineage. Other criteria, such as conservation of synteny, are necessary to further test this hypothesis.

Evolution of gene families

We used CAFE [97] to study the evolution of gene families in the lamprey and the gnathostomes. Given a species tree and the counts of genes per species for each family, CAFE computes the most likely rate of gene birth / death ("lambda"), either globally or on specific branches. The output contains, for each family, its inferred size at each ancestral node and the p-value of expansion / contraction on each branch.

CAFE has several constraints regarding the input dataset. First, the list of species in the counts must exactly match the species tree. Second, the species tree should be ultrametric (ideally with branches reflecting time). We used the TimeTree database [98] to build such a species tree. Finally, each family must be present at the root of the species tree (the inferred gene count at this node is at least 1), which would favor one analysis per tree-root. On the other hand, computations of lambda and probabilities are more accurate with more families. As mentioned above, extreme caution should be used when interpreting the perceived absence of a lamprey gene (e.g. gene losses in lamprey and Euteleostomi-specific expansions)

Among 18,809 gene families contained in the Ensembl 64 database, we selected 10,801 families with genes in at least 2 species out of the 11 models (yeast, nematode, fly, *Ciona intestinalis, Ciona savignyi,* lamprey, zebrafish, medaka, chicken, human and mouse). We defined 3 datasets.

DATASET 1: 7,721 families were either present in one of the *Ciona*, or absent from *Ciona* but in 1 copy in one of the outgroups (fly, nematode, yeast).

DATASET 2: 663 families are absent from all outgroups, but present in the lamprey and at least 1 gnathostome.

DATASET 3: 130 families are absent from *Ciona*, and in more than one copy in outgroups. This dataset was not sufficiently large for subsequent analyses.

Dataset 1 was analyzed with the species tree (*Ciona*,(lamprey,gnathostomes)), as we can be positive that the selected families existed in at least 1 copy at the Vertebrata node. CAFE optimizes the initial family size if the family is found in *Ciona*, or use 1 as default, which is reasonable in these cases since there is 1 copy of the gene in at least one outspecies. We asked CAFE to infer a lambda for tetrapods, fish, *Ciona*, and each of the remaining branches (3 internal, and the lamprey terminal branch). In total, 30 families show a significant change in lamprey (p-value < 10% for lamprey expansions, p-value < 2.5% for lamprey contractions; Supplementary Table 8; included in a separate excel file). Specifically, 4 show an expansion in lamprey+gnathostomes, 2 in lamprey+fish, 3 expanded only in lamprey (but present in gnathostomes), 7 expanded only in lamprey (but almost absent from gnathostomes), 1 present only in lamprey (but lost in fish). On the other hand, CAFE detected 11 contractions / loss in the lamprey, and 2 in the gnathostomes.

Similarly, Dataset 2 was analysed with a tree (lamprey,gnathostomes), again with specific lambdas for tetrapods, fish, lamprey, and the remaining internal branches. In total, 34 families show a significant change in lamprey (p-value < 10%; Supplementary Table 8). 9 correspond to an expansion in lamprey, 9 in the lamprey+fish, 1 in lamprey+tetrapods, whereas 15 show contraction in lamprey.

Conserved synteny and genome duplication

Analyses of genome structure were performed as to avoid several potential pitfalls of phylogenetic analysis, which can be misled by several factors intrinsic to the lamprey genome, including:1) the deep divergence between lamprey and its nearest common ancestor, 2) broad variation in nucleotide content among taxa, and 3) pervasive G/C substitution bias in protein coding regions of the lamprey genome. These comparisons rely on two simple assumptions to identify putative orthologs and duplications. First, we assume that, within a genome, duplicated genes will diverge at relatively similar rates. As such, regions were considered putative orthologs if they yielded the highest-scoring alignment between the two genomes or an alignment score (bitscore) within 90% of the top-scoring alignment (tblastn of lamprey gene models to human or chicken genomes). This convention permits some variation in divergence rate and can be applied uniformly to the genome, but may fail to identify some duplicates that have undergone exceedingly rapid diversification following duplication. Second, we limit our analysis to duplicates that are broadly distributed through the genome and are present at a relatively low copy number. Comparative maps were consequently pruned to remove redundant copies of tandemly duplicated genes (i.e. lineage-specific gene amplifications) and homology groups that contained more than 6 homologs in either of the two species being compared in any pairwise analysis. These comparative maps are available as separate supplementary files (Supplementary Tables 9 and 10, included as separate excel files).

Comparative mapping studies focused on two vertebrate genomes, chicken and human. These two genomes were selected because 1) they are relatively well assembled and curated, 2) they represent two evolutionary divergent vertebrate lineages, 3) rates of intrachromosomal rearrangement are relatively lower than other well curated mammalian genomes [99-102] and 4) they have not experienced recent whole genome duplications (i.e. 3R in teleost fishes) [103]. Overall, chicken is expected to show stronger conservation of synteny, due to the fact that chicken has experienced a lower rate of intrachromosomal rearrangement than human [99-102]. However, strong conservation of synteny (accounting for duplication) can be seen in both lamprey/chicken and lamprey/human comparisons (Supplementary Tables 9 and 10).

Consistent with the idea that a whole genome duplication event (or events) occurred in the vertebrate stem lineage, we found that lamprey and gnathostomes: 1) share several low-copy duplications/paralogy groups and 2) possess similar overall frequencies of gene duplication (Supplementary Tables 11 and 12). However, it is possible that a similar duplication structure might arise from several independent gene duplications or segmental duplications. Under scenarios of independent/segmental duplication, shared duplications might reflect selective forces acting against duplication, such that very few duplication events are evolutionarily viable. To further explore this possibility, we repeated our analysis, focusing only on those regions that provided orthogonal evidence for segmental or chromosomal duplication (interdigitated conserved syntenies). This subset shows significant bias toward shared duplication (observed = 0.184, expected = 0.067, $\chi 2$ = 147.8, P($\chi 2$) = 5e-34, P(Fisher's Exact) = 3e-12; Supplementary Table 13), and is consistent with patterns seen across the genome at large. Thus, regions with sufficient contiguity and evolutionary conservation as to permit the identification of large-scale genome duplication do not differ in duplication frequency from the rest of the genome. It therefore appears that genome-wide patterns of shared duplication cannot be described sufficiently by a model invoking recurrent selection against duplications across a majority of the genome. We therefore propose that patterns of shared duplication are indicative of a shared history of genome-wide duplication prior to the lamprey/gnathostome divergence.

In addition to primary analyses presented in the manuscript, we performed several additional analyses to assess the distribution of gene duplicates across the lamprey genome. Because each lamprey scaffold represents a small fraction of the genome, analyses of the genome-wide distribution of duplications must rely on long-range linkage information from other vertebrate species or otherwise account for the influence of scaffold length/quality in the identification of duplicates.

We used positional information from human and chicken to gain (albeit imperfect) insight into the genome-wide distribution of lamprey gene duplications. Supplementary Figures 18 and 19 show the proportion of loci with orthologous duplications in the lamprey genome over sliding windows of 50 homology-informative loci. These plots indicate that the incidence of lamprey gene duplication is relatively uniform across the genome, roughly 25% of loci being present as duplicates over any given genomic interval. These analyses make the assumption that the distribution of loci in chicken and human genomes approximates distribution of orthologous loci in the lamprey genome. Given the observation of conserved synteny across lamprey scaffolds, we infer that this assumption is not unreasonable. Although duplication frequencies fluctuate around ~25% in both plots, it may be worth noting that there are a few peaks that show a conspicuous excess of retained duplicates in the lamprey genome. Curiously, several of these peaks are adjacent to, although not necessarily including, Hox genes. This may be indicative of greater conservation of duplicates within a few specific intervals of the lamprey genome, however the evolutionary relevance of such intervals remains an open question.

On the basis of the lamprey genome assembly, it appears that paralogous duplicates are not rare within the lamprey genome and are broadly distributed across scaffolds (and likely chromosomes). However, we expect that variation in contig length and “quality” may prevent the identification of some paralogous duplicates. In order to gain some perspective on the influence of inter-scaffold variation in the identification of lamprey paralogs, we examined the distribution of paralogous duplicates, relative to the quality of scaffolds containing these genes. For these analyses we used a quality metric that is related to a scaffold’s information content, specifically: the number of unique protein coding domains within that scaffold. This measure effectively down-weights scaffolds that contain tandem/local duplications or few protein coding genes. Below, we compare the relationship between scaffold quality and the identification of gene duplications in lamprey and gnathostome genomes.

Our ability to detect gnathostome duplications varied only slightly with scaffold information content (Supplementary Figure 20), therefore fragmentation of the lamprey genome does not appear, in itself, to limit our ability to detect duplicated gnathostome orthologs of lamprey genes. The detection of lamprey duplications is also only weakly correlated with scaffold information content, but showed a more consistent pattern of variation (Supplementary Figure 21). Specifically, detection of lamprey duplications increased with increasing information content. An important caveat to such analyses is that the ability to identify lamprey duplicates is not simply a function of the quality of a given scaffold, but rather the quality of all scaffolds that contain paralogous duplicates of a given gene. Therefore quality of individual scaffolds may not, individually, be highly predictive of the ability to identify lamprey paralogs on that scaffold. Nonetheless, general trends in lamprey/human and lamprey/chicken comparative maps seem to indicate that fragmentation of the lamprey genome may have limited our ability to detect loci that are duplicated within the lamprey genome. As such, the true number of loci that are duplicated in the lamprey genome is likely to be higher than our estimates.

Lamprey Hox clusters

Among all genes in the genome, the Hox genes arguably have played the most seminal role in studies of vertebrate genome duplication: the four Hox clusters found in most vertebrate genomes are thought to have resulted from two rounds of whole genome duplication [104]. Indeed, the elephant shark possesses four Hox clusters [105], whereas the catshark and skate possess only three [106,107]. Targeted resequencing of lamprey scaffolds identified two distinct Hox clusters, which extend into syntenic flanking sequence containing additional non-Hox protein coding genes and show patterns of independent retention of paralogs in lamprey vs. gnathostomes (Figure 4B). This is the same number of Hox clusters identifiable in the chicken genome assembly (2), although chicken possesses four clusters. Lamprey has eight additional Hox genes that could not be assigned to the above clusters and at least two are linked to each other suggesting the existence of a third cluster (Figure 4B, Supplementary Figure 22). As such, the lamprey Hox content is not inconsistent with a pre-vertebrate 2R, given the nuances of genome assembly and overarching patterns of paralog loss following 2R.

To supplement the assembly of Hox-containing regions, we selected a series of BACs via hybridization to a Hox2 probe designed from known lamprey transcripts (Genbank: AY497314). Mapping and sequence analysis identified these as a set of overlapping BACs from a single cluster (Cluster 1 in Supplementary Figure 22). Another series of BACs were selected by hybridization to Hox4 or Hox9 homeodomain probes, pooled and sequenced by 454. In addition, a series of BACs were selected based on BAC end sequence data, which permitted the identification of HOX linked BACs. Manual curation of these BAC data together with the lamprey genome assembly, support the existence of two extended Hox clusters and linkages to adjacent conserved syntenic regions (Supplementary Figure 22).

Cluster 1 extends Scaffold_430, encompassing HOX 2, 3, 4, 5, 6, 7, 8, 9 and 11, and conserved syntenic genes downstream of the cluster (including homologs of BOLL, CYC and MTX2). Cluster 2 merges Scaffold_686, Scaffold_1553 and Scaffold_1243, encompassing HOX 1, 4, 5, 7, 8, 9, 10 and 11 and conserved syntenic genes upstream (homologs of TAX1BP1 and EVX1) and downstream (including homologs of CHORDC, SNX10, CBX1, THRA, RARB, FAM126B, and MRPL10) of the cluster. The scaffolds for these loci contain gaps (indicated by dotted regions between Hox genes), which could contain additional Hox members (indicated by boxes with question marks).
Seven additional unassigned Hox containing scaffolds are identified in the lamprey genome assembly (Supplementary Figure 22). These scaffolds are generally small, and do not contain other predicted Hox genes. An eighth (the C4/B4 homolog) was found in the Illumina sequenced BAC clone (66A06), which end-mapped to Sc_10557 (which contains a Hox1 homolog) and overlaps a number of other non-Hox containing scaffolds. An Illumina sequenced BAC (648A10), end-mapping to Sc _6993, was found to overlap with Sc_73, containing a cluster of seven Receptor-type tyrosine-protein phosphatases and four other genes, none of which resemble Hox or known Hox-syntenic loci.

The annotated Hox hexapeptide and homeodomain predicted amino-acid regions are shown aligned in Supplementary Figure 23. These data support the gene assignments in the assembles, confirm that these are distinct Hox genes in the lamprey genome.

Lamprey neuroendocrinology

Genome wide analyses of gene evolution in the ancestral vertebrate lineage verify that the development of the hypothalamus and pituitary axis was a seminal event in the evolution of vertebrates. Using the lamprey genome, we further examined the presence or absence of genes encoding hormones and related receptors, with particular attention to the gonadotropin-releasing hormones (GnRHs) [108], the master neurohormone regulators of vertebrate reproduction. Analyses of conserved synteny reveal several key features of GnRH evolution, and suggest an evolutionary scenario that differs substantially from existing paradigms (Figure 4A, Supplementary Table 14)[10,109-114]. Overall, lamprey synteny data suggest that all duplication events that generated the different fish and tetrapod GnRH groups likely took place before the divergence of the ancestral lamprey and gnathostome lineages. A GnRH1 paralog was lost from the lamprey genome, reminiscent of a parallel loss in zebrafish [115] and GnRH3 was lost in tetrapods rather than arising in the teleost lineage as a result of a teleost-specific whole genome duplication event (3R)[114,115]. The functional group IV GnRHs in lamprey (GnRH-I and -III) share a more recent common ancestry with GnRH2 and 3 paralogs (Supplementary Figure 24). Given a single amino acid difference between the mature (10 amino-acids) lamprey GnRH-II and GnRH2, we propose that an ancestral GnRH2-like gene existed before the lamprey/gnathostome split and that paralogous genes (GnRH-I/III and GnRH 3) independently evolved divergent structure/function in lamprey and gnathostome lineages. Intriguingly, previous data suggest that hagfish express two GnRH-2/3-like peptides[116,117] consistent with the idea that the tandem duplication that gave rise to lamprey GnRH-I and -III occurred after the lamprey/gnathostome split.

Identification of vertebrate-specific gene families

First, using all *P. marinus* predicted peptides we performed BLASTP searches [47] against *Ensembl* peptides of all gnathostome species (version 58 [118]). Second, every gnathostome peptide sequence that exhibited the highest bit score of no less than 50 in each *BLASTP* search in the first step was used as query in a *BLASTP* search performed against invertebrate peptide sequences. This invertebrate database included all sequences available in *GenBank* and *Ensembl* for invertebrates as well as all peptides predicted in the genomes of *Schistosoma japonicum* [119], *S. mansoni* [120] and *Lottia gigantea* [55]. The gnathostome query sequences with the highest bit score of no more than 50 were selected as candidates of genes that have homologs in lamprey but not in any invertebrate. In both searches, the cases with bit scores between 50 and 60 were further examined, using an approach based on reciprocal best hit, where the homology was regarded as true if a reciprocal *BLASTP* search resulted in the best hit with the starting query sequence itself or its homolog with the bit score of no less than 50. After multiple members of gene families were removed to have only one representative per family, we identified 224 gene families that are found in vertebrates, but not in invertebrate lineages (Supplementary Table 15; included in a separate excel file). These gene families are considered to be unique to the vertebrate lineage and to have emerged before the radiation of all extant vertebrates, including lampreys.

Immunity in the lamprey

Adaptive immunity

An immune receptor system capable of generating enormous diversity via somatic rearrangement has recently been described in lamprey and hagfish, taxa representing the most primitive vertebrates but which also lack the hallmark components necessary for adaptive immunity in higher vertebrates, i.e., the rearranging genes of the immunoglobulin superfamily (IgSF) [121,122]. These variable lymphocyte receptors (VLRs) are produced through an entirely novel genomic mechanism in which large banks of leucine rich repeat (LRR) cassettes are used to build the diversity region of the receptor molecules in a process involving *de novo* genome rearrangements. Functional experiments have shown that the VLR molecules are capable of directly binding to antigens [123-125], and crystal structures have been produced of VLR molecules complexed with their cognate antigens [126,127]. Existence of this novel receptor system suggests that there were at least two independent solutions (Ig-mediated and VLR-mediated) to evolving an adaptive immune system in vertebrates. Interestingly, these two types of adaptive immune system use similar lymphocyte differentiation strategies involving two lymphocyte lineages that somatically assemble highly diverse antigen receptor repertoires [128,129]. Lampreys have three VLR types, VLRA, VLRB, and VLRC. Diverse repertoires of VLRA and VLRB are expressed by separate populations of lymphocytes that resemble mammalian T and B cells, respectively [124,128,130]. The germline VLRA and VLRB genes are both incomplete in that they encode only portions of the amino- and carboxy-terminal LRRs plus the complete stalk region. Hundreds (or thousands) of different LRR sequences flank the germline VLR-A and VLR-B genes, and these are randomly selected as templates to be copied in a piece-by-piece manner to complete a VLR-A or VLR-B gene during lymphocyte development [121,124,131,132].

Ontology analysis identified 572 and 294 putative genes for the terms “immune” and “inflammation”. These data are given in Supplementary Table 16 (included as a separate excel file) and provide a coarse glimpse as to presumptive immune molecules in the sea lamprey genome based on comparisons with the known (mammalian) immune proteome. Roughly 80% of these genes are involved in the immune response and in regulating the immune system. Notable genes have been identified from these analyses that are known to play a role in adaptive immunity of higher vertebrates. These include genes encoding: 1) numerous Fox transcription factors known to be involved in central tolerance, T-cell cytokine production and V(D)J recombination; 2) THEMIS1, a protein involved in selection in the thymus; 3) numerous proteins thought to be involved in rearrangements of immunoglobulin-type genes; 4) a single homolog of terminal deoxynucleotidyl transferase and DNA polymerase mu, known to be involved in end-joining of immunoglobulin-type genes and somatic hypermutation; 5) PMS2, known to be involved in somatic hypermutation of immunoglobulin genes; 6) numerous immune signaling molecules and cytokines; and 7) numerous proteins involved in development of hematolymphoid structures. Consistent with a previous EST study on lamprey lymphocytes [133,134], there is clearly a lower number of immunoglobulin-type molecules encoded in the lamprey genome (Supplementary Table 17).

Genomic loci encoding VLR

Blast searches of the genome assembly identified scaffolds encoding all three VLR loci (Supplementary Table 18). The gene organizations of these loci are consistent with previous descriptions [135,136]. From the original description of the VLRB locus based on P1 artificial chromosome (PAC) library screening/sequencing and pulsed field gel Southern genomic analysis [121], it was inferred that VLRB was a single locus. However, two separate PACs (PAC4 and PAC16) were described that encode highly similar, but not identical regions of the VLRB locus. PAC4 is encompassed in scaffold_256 whereas PAC16 is partly encompassed in scaffolds_3467 and 6374. To link the three scaffolds, we screened a lamprey genomic BAC library with VLRB-specific probes [4] and used 454 titanium sequencing to sequence two BACs as shown in Supplementary Figure 25. The two BAC clones enabled gap closure of the VLRB locus and resulted in a single scaffold of 717 kb. To confirm that the two presumptive transcriptional start sites (PAC4-type and PAC16-type) were indeed separate regions, chromosomal fluorescence *in situ* hybridization was carried out [3] using labeled PAC4 and PAC16 as probes (Supplementary Figure 26). The data confirm that the regions encompassed by PAC4 and PAC16 are indeed in close proximity since the hybridization respective spots were always next to one another and clearly discernible. The utilization of PAC4 versus PAC16 transcription start sites as well as the respective LRR components within the locus await further analyses. Based on searches of the lamprey genome assembly using various VLRB components as query sequences [132], it is likely that there are other scaffolds which likely contain VLRB-encoding modules such as N-terminal and C-terminal LRR cassettes. The relationship of these scaffolds to the 717 kb core VLRB scaffold is as yet unclear and requires more in-depth physical mapping and BAC characterization.

Transcriptomes of VLR-expressing cells

Transcriptome analysis of the VLRA and VLRB lymphocyte populations indicates that they have very different gene expression profiles [128]. The preferentially expressed genes for VLRB$^+$ lymphocytes have clear orthologs for several genes that are preferentially expressed in B cells of jawed vertebrates (Supplementary Figure 27). These include transcripts for the haematopoietic progenitor homing receptor CXCR4, the herpes virus entry mediator/tumor necrosis factor receptor superfamily member 14 (TNFRSF14), two components of the BCR-mediated signaling cascades, Syk and the B cell adaptor protein (BCAP), the chemotactic inflammatory cytokine IL-8, the IL-17 receptor, and the Toll-like receptors TLR2abc, TLR7 and TLR10, the ligation of which may induce B cell activation. Conversely, the VLRA$^+$ lymphocytes express genes orthologous to those typically expressed by T cells in the jawed vertebrates; these preferentially expressed genes include ones that encode the GATA2/3, c-Rel, aryl hydrocarbon receptor (AHR) and BCL11b transcriptional factors used for T cell differentiation, the CCR9 chemokine receptor that is involved in thymic homing of thymocyte progenitors, the Notch1 T cell fate-determining molecule, the CD45 tyrosine phosphatase receptor protein that is essential for T cell differentiation, the IL-17 and Mif pro-inflammatory cytokines and the CXCR2 IL-8 receptor (Figure 5, main text). The sequences of these genes are extended and confirmed by the current assembly (see below for more details). Activated VLRA$^+$ cells upregulate their expression of *IL-17* and *MIF*, whereas activated VLRB$^+$ cells upregulate their expression of *IL-8* [128,130]. Coupled with the reciprocal expression of *IL-17R* by

VLRB$^+$ cells and *IL-8R* by VLRA$^+$ lymphocytes, these findings suggest the potential for functional interactions between the lamprey T-like and B-like lymphocyte populations.

Innate Immunity

Identifying lamprey innate immune genes

Genes that encode elements of immunity often present unique challenges to annotation and orthology assignment. For many of these rapidly evolving genes, primary sequence similarity is quickly lost even in closely related taxa. Furthermore, many of the immune signaling and effector molecules, such as cytokines and antimicrobial peptides, are small and are difficult to capture in gene models. However, despite the divergence in sequence similarity, many immune mediators are characterized by the presence of conserved domains or combinations of domains. For the analysis presented here, we have largely followed previously described methods in which immune genes were identified in low stringency searches based on domain architecture [137]. Searches were performed on the *MAKER* gene model set, in addition to the translated genome to identify sequences not included in the models. The results of this search are shown in Supplementary Table 19. For comparison, the same analysis was run on the genomes from *Homo sapiens, Callorhincus milii, Ciona intestinalis, Branchiostoma floridae, and Strongylocentrotus purpuratus* (Supplementary Table 17). Consistent with previously discussed findings, this analysis suggests that the lamprey genome encodes fewer immunoglobulin domains than jawed vertebrates and some invertebrate deuterostomes, and that the lamprey innate immune system is more similar to that of jawed vertebrates in terms of the size of the gene families that encode innate immune receptors. In a number of cases, we were able to uncover divergent, unmodeled immune genes that have only previously been described in jawed vertebrates.

Pattern recognition receptors in the lamprey genome

In addition to the complex, immunoglobulin-based, adaptive immune system, jawed vertebrates maintain an innate immune system that has been conserved across bilaterian evolution. At the core of this system is a set of small multigene families of pattern recognition receptors (PRRs) that recognize broadly conserved microbial signatures and initiate an immune response. The Toll-like receptors (TLR), Nod-like receptors (NLR), scavenger-receptor cysteine-rich (SRCR), and Rig-I like receptors (RLR) are organized in small gene families (5-30 genes in most jawed vertebrates; see Supplementary Table 17). In contrast, the genomes of the lower deuterostomes are often characterized by significantly expanded families of homologs of these innate immune receptors. The TLR, NLR, and SRCR gene families in the sea urchin and amphioxus genomes are five to ten-times larger than those in vertebrates (Supplementary Table 17; [138-140]). Given its phylogenetic position, the lamprey offers a unique perspective on the evolution of these extensive innate immune gene families. The data presented here indicate that the gene families encoding lamprey PRRs are more similar in size to those of jawed vertebrates than the complex invertebrate deuterostomes (Supplementary Table 17) and suggest that the reduction in innate immune complexity, within the resolution of this phylogeny, was coincident with the introduction of lymphocyte-based adaptive immunity.

Toll-like receptors

TLRs are transmembrane proteins with an extracellular ligand binding domain consisting of a series of leucine-rich repeats (LRRs) and an intracellular Toll/Interleukin-1 Receptor (TIR) domain that mediates signaling through TIR-containing adaptor molecules. In jawed vertebrates, five adaptors have been characterized: MyD88, Mal/TIRAP, TRIF/TICAM-1, TRAM/TICAM-2, and SARM (reviewed in [141]). The IL-1 receptor also contains a TIR domain, but has extracellular immunoglobulin domains. An initial analysis of the lamprey TLR genes identified 16 TLRs, and four homologs of the adaptor molecules (1 MyD88, 2 TICAM, and 1 SARM) [142]. We identified TIR domains (PFAM profile PF01582.12) in the *P. marinus* gene models and translated genome using HMMER 3.0, and candidates were further characterized using blast searches against the non-redundant database at NCBI and on the basis of domain architecture. In total, 24 TIR domains were identified: 19 TLRs, the four previously identified adaptors, and a single IL-1 receptor (Supplementary Table 19).

Phylogenetic analysis of the lamprey TLRs suggests that some of the lamprey receptors closely related to TLRs found in jawed vertebrates, and may reflect the TLR repertoire present in the common vertebrate ancestor (Supplementary Figure 28). There is a single homolog of TLR-3 and two sequences that are most closely related to the TLR-7/8/9 cluster. Eight of the lamprey sequences clustered with TLR-1, 6, and 10, one of which appears to be homologous to the zebrafish TLR-18. Three of the lamprey TLRs (here named Pm-TLR23-25) appear to be unique to the lamprey and did not exhibit similarity with any known gnathostome TLR. Notably, no orthologs of gnathostome TLR-4 or TLR-5 were identified. The lamprey genome is also lacking orthologs of the protostome-like TLR sequences that are present in *Ciona,* amphioxus and sea urchin (data not shown; [138-140]), which indicates that these sequences may have been lost prior to the emergence of vertebrates.

Nod-like receptors

NLRs are ancient intracellular sensors of microbial molecules [143,144]. These proteins are primarily expressed in epithelial cells, particularly in the gut, but are also expressed by some immune cells. NLRs are multidomain proteins that consist of an N-terminal effector binding domain, a central NACHT domain, and a series of C-terminal LRRs. In vertebrates, the NLRs are organized into five families based on the presence of different effector domains: caspase recruitment domains (CARD), pyrin domains, acidic domains, baculovirus inhibitor repeats (BIR), or domains of unknown homology [143]. In humans, 23

NLRs have been characterized, whereas over 200 NLRs have been identified in the sea urchin genome, and at least 92 in amphioxus (Supplementary Table 17; [145]). CARD, pyrin, Death, and DED domains are all related members of the death domain-fold superfamily that primarily function to regulate apoptotic and inflammatory processes. Teleost genomes also possess an expanded NLR subfamily in which several hundred genes encode NLRs with a C-terminal B30.2 domain [146]. To identify homologs of NLRs in the lamprey, both the gene models and the translated genome were searched for NACHT domains (PFAM profile PF05729.4). A total of 34 putative homologs of vertebrate NLRs were identified that contain NACHT domains, the majority of which also encode effector CARD domains (Supplementary Table 20).

NLR genes are particularly difficult to identify using standard gene modeling methods. In mammalian NLRs, the C-terminal LRRs are individually encoded in single exons that are commonly missed in the modeling. Notably, only 19 of the 34 NACHT domains identified here were captured by the gene models, which underscores the importance of searching the genome directly for homologs of immune molecules that may be expressed at very low levels. Without reliable gene models, it is difficult to characterizes the complete domain structures beyond the conserved NACHT domains. Of the death domain superfamily subtypes that typically characterize vertebrate NLRs, only CARD domains were located near the lamprey NACHT domains. No pyrin domains were identified in the genome. Phylogenetic analysis of the lamprey NACHT domains indicates that, with the exception of three genes, the lamprey NLRs are the result of an independent expansion in this lineage, and do not share clear orthology with NLR subfamilies from either the jawed vertebrates or the lower deuterostomes (Supplementary Figure 29). Of the remaining three sequences, there is a homolog of human NLRX1 (Pm-NLRX1), a homolog of mammalian Nod1/2 (Pm-Nod1/2), and a homolog of NLRC4 (Pm-NLRC4) (Supplementary Table 20). Twelve of the lamprey NLR genes are clustered in the genome on scaffold_357 (Pm-NLR1-12; Supplementary Figure 29).

Homologs of several mediators of mammalian NLR signaling are also present in the lamprey genome. Mammalian Nod proteins recruit RIPK2 through CARD:CARD interactions [143]. The lamprey has a homolog of RIPK2 (PMZ_0017431-RA). Furthermore, an isolated CARD domain in the middle of the NLR cluster on scaffold_357 has some similarity with the mammalian NLR adaptor Card6 (Supplementary Figures 29 and 30). There are several caspases in the lamprey genome, including a homolog of caspase-1 (PMZ_0005524-RA), which cleaves pro-IL1β into its secreted form. Other NLR mediators, however, including Asc (apoptosis-associated speck-protein containing a CARD) and Aim2 (also known as Pyhin) were not identified. A number of CARD domain containing proteins, however, were identified within gene models that lack domain architecture or sequence similarity with other known mediators of NLR signaling, and may be involved in these pathways in the lamprey.

Scavenger receptors and Rig-I like receptors

Two remaining well-characterized vertebrate PRR gene families are the SRCRs and the RLRs. The sea urchin and amphioxus genomes both contain an expanded repertoire of SRCR domains (Supplementary Table 17). However, only 100 SRCR domains were identified in the lamprey genome, comparable in number to those in higher vertebrates. The lamprey genome also contains a homolog of MDA5, which is a member of the RLR family (PMZ_0010575-RA). This appears to be the only RIG-I like receptor represented in the lamprey. RLR proteins are critical for detecting viral RNA in the cytoplasm (reviewed in [147]).

Intercellular signaling: cytokines, chemokines, and receptors

Cytokines are among the most challenging components of the immune system to identify in divergent species. In teleosts, notable progress has been recently made in discovering both homologs of known mammalian cytokines as well as novel signaling molecules [148]. We searched the lamprey gene models and translated genome for divergent homologs of mammalian cytokines. The following cytokines were not identified: IL-2, IL-3, IL-4, IL-5, IL-7, IL-9, IL-11, IL-13, IL-15, IL-21, GM-CSF, OSM, LIF, erythropoietin, IFN-α, IFN-β, or IFN-γ. Two cytokines have been previously described in the lamprey: IL-8 [128] and Mif [149]. Our analysis confirms the presence of a single gene encoding IL-8 and a single gene encoding Mif (Supplementary Table 21).

IL-17.

IL-17 is a key cytokine that defines the T helper subset 17 (Th17) cells, and is also critical for mucosal immunity, particularly in the gut [150]. A single IL-17 ligand has been previously characterized in *Lethenteron japonicum* [151]. A lamprey IL-17 receptor has also been identified [128]. Our analysis of the genome sequence identified four IL-17 ligand genes, four genes that encode IL-17 receptors, and a homolog of the adaptor molecule Act1 (also known as CIKS/TRAF3IP2), which mediates downstream signaling (Supplementary Table 21).

IL-1.

We also identified a homolog of IL-1 by searching the translated lamprey genome for the IL-1 domain (PF00340.11). This is the first description of IL-1 outside of jawed vertebrates. Transcriptome sequencing provided further validation of this locus.

IL-6.

There are two predicted gene models that contain IL-6/G-CSF domains (PF00489.10; Supplementary Table 22). These cytokines form a family based on structural similarity [152], and, for these divergent molecules, homology assignations are difficult to determine. The two molecules share only 16% amino acid identity with each other, and, despite the prediction of the IL-6 domain with high probability, neither sequence shares sufficient similarity with mammalian IL-6 to be identified via *BLAST* searches.

Complement

Complement components have been identified in both lampreys and hagfish [153,154], and the analysis of the genome sequence offers more details about this innate effecter system. In chordates, collectins initiate the lectin cascade through members of the mannose-binding protein (MBP)–associated protease (MASP)/C1r/C1s family. Several genes encoding collectins, C1q and MBP members of the MASP/C1r/C1s family were present in lampreys (Supplementary Table 16, included in a separate excel file). VLRB was found to be physically associated with one of MASP/C1r/C1s family, MASPb, in lamprey blood [155], suggesting the existence of the complement classical pathway that mediates innate and adaptive immunity.

Conclusions

Overall, these findings from analysis of the genome suggest that the innate immune system of the lamprey is much more similar in character to those of the jawed vertebrates, particularly with respect to the multiplicity of the gene families that encode the pattern recognition receptors. Furthermore, although we identified many divergent molecules in this animal, we expect that additional homologs of cytokines and signaling molecules exist in the lamprey that are beyond our ability to identify using solely computational methods that rely on sequence similarity. Additional transcriptome data will contribute to this by uncovering genes that respond transcriptionally to immune challenge.

Ontology database

All ontology analyses were performed using a custom ontology database, using *Blast2GO* [156]. To generate an ontology database from the complete set of lamprey gene models, predicted amino acid sequences were first aligned to the *SwissProt* database (1E-6 BLAST threshold), and best *BLAST* hits were pulled from all available organisms. Resulting homologies were searched against the *B2G-GO-Database* to collect available gene homologs. The distribution of sequence identity statistics for these best blast hits was centered ~61% amino acid identity (max 100%, min 31%, mean 64%, mode 61%; see Supplementary Figure 31), which is within the range of expectation for homologous relationships given the deep evolutionary divergence between lamprey and other organisms. Ontology information from homologous sequences was processed by *Blast2GO* to compile a set of likely ontologies for each lamprey gene model. The resulting ontology database provides two key functionalities that are of particular relevance to our reported analyses of the lamprey genome: 1) it permits rapid identification of genes that are likely to be associated with a particular biological process, molecular function or cellular compartment and 2) it provides a framework for identifying ontologies that are over- (or under-) represented in a subset of loci, relative to the lamprey genome as a whole.

Identification of genes associated with myelin and neurodegenerative diseases in the lamprey genome

The assembly was manually searched using *TBLASTN* in order to identify lamprey sequences that are homologous to genes found in the mouse genome. Specifically, we searched for known genes that are related to myelin (Supplementary Table 23) or have associations to diseases or disorders of the human central nervous system (Supplementary Table 24). To be considered homologous, the lamprey genes identified had to cover at least 50% of the total mammalian protein length (RefSeq database), and they had to share at least 25% identity. Alignments for MBP and MPZ were further examined to evaluate homology among copies of the gene from diverse vertebrate taxa (Supplementary Figure 32). Results returned from the lamprey genome browser were independently searched in the NCBI Blast server to verify that the same protein was returned.

Notably, the lamprey genome contains homologs of human genes that are linked to neurodegenerative diseases, such as Alzheimer's (APP and presenilin), Parkinson's (synuclein), Huntington's (huntingtin), and autism (neurexin) (Supplementary Table 24). Of these, synucleins appear to comprise a vertebrate-specific gene family. In addition, lampreys possess homologs of a diverse array of axon guidance molecules and extracellular matrix molecules that are known to change in response to injury in the human nervous system (Supplementary Table 24). As several lamprey nervous system cell types are large and identifiable, making them experimentally tractable, presence of these disease- and injury-associated factors in lamprey provides an excellent opportunity to study the basic, ancestral functions of proteins relevant to these pathobiologies.

Signaling pathways in appendage evolution

Tbx 4 and Tbx 5 are the earliest expressed transcription factors known to be required to initiate forelimb and hind limb outgrowth (Tbx5 for forelimb and Tbx4 for the hindlimb). The Tbx gene family codes for transcription factors playing major roles in embryogenesis. During limb development six Tbx-genes are expressed: Tbx2, -3, -4, -5, -15, and -18 with Tbx2 and -3, Tbx4 and -5, and Tbx15 and -18 as paralogous pairs. Interestingly the fore and hind limb specific roles of Tbx5 and Tbx4, respectively and the fact that they are still present as a single gene in the limbless amphioxus, suggests a correlation of the duplication status with the evolution of their roles in limb development. In amphioxus and lamprey the Tbx4/5 gene is thought to play an ancestral role in heart development, and its modification in *cis*-regulatory regions must have provided the limb forming function in the lineage leading to limbed vertebrates [157,158]. This observation raises the question whether the Tbx genes are present in cognate pairs in the lamprey. We identified nine Tbx family members in the lamprey genome, which correspond to the number found in amphioxus [159]. None of the genes expressed in limb development in gnathostomes were found in multiple 2R duplicates, further strengthening the argument that the status of the duplication is correlated to the occurrence of paired appendages.

Downstream of the Tbx genes, the limb bud forms as a result of an interplay between fibroblast growth factors (in particular Fgf8 and 10*)* and Wnt signaling. These signals operate in the context of a transcriptional pre-pattern in the posterior limb buds (Hand2, Gli3, Hox genes etc).

In the lamprey genome we identified Fgf10, and Fgf8/17 was previously identified in another lamprey species [160], however other definitive homologs of Fgf family members that are expressed in the apical ectodermal ridge were not identified (Fgf4, Fgf9). Some Wnt genes known to play a role in limb development were identified (Wnt3, Wnt5a, Wnt7a, and Wnt10a), while other important members such as Wnt3a, Wnt2b or Wnt8c were not identified. Similar to the case of the Tbx5 and 4 genes it might be interesting to note that genes thought to specifically play a role in the forelimb (Wnt2b) or hind limb (Wnt8c) development are not identified in the lamprey genome assembly. The presence of some limb patterning genes such as Hand2 and Gli3 is consistent with their pleiotropic roles during different stages of development. This supports the notion that many of the signaling systems, which are used during development, plausibly including median fin development, are reused for paired fin development. On the other hand that many family members of genes could not be identified in the lamprey genome suggests that retention of some 2R duplicates may have facilitated the development and evolution of modern paired appendages. Whether these were absolutely required for the evolution and development of the first paired appendages needs further investigation.

References

1. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695-716 (2004).
2. Waterston, R.H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562 (2002).
3. Smith, J.J., Baker, C., Eichler, E.E. & Amemiya, C.T. Genetic consequences of programmed genome rearrangement. *Current biology : CB* **22**, 1524-9 (2012).
4. Smith, J.J., Stuart, A.B., Sauka-Spengler, T., Clifton, S.W. & Amemiya, C.T. Development and analysis of a germline BAC resource for the sea lamprey, a vertebrate that undergoes substantial chromatin diminution. *Chromosoma* **119**, 381-389 (2010).
5. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**, 1596-9 (2007).
6. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic acids research* **33**, 511-8 (2005).
7. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25**, 955-964 (1997).
8. Hofacker, I.L. *et al.* Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie* **125**, 167-188 (1994).
9. Gorbman, A. & Sower, S.A. Evolution of the role of GnRH in animal (Metazoan) biology. *General and comparative endocrinology* **134**, 207-13 (2003).
10. Silver, M.R., Kawauchi, H., Nozaki, M. & Sower, S.A. Cloning and analysis of the lamprey GnRH-III cDNA from eight species of lamprey representing the three families of Petromyzoniformes. *General and comparative endocrinology* **139**, 85-94 (2004).
11. Forey, P. & Janvier, P. Evolution of the early vertebrates. *American Scientist* **82**, 554-565 (1994).
12. Nixon, M. & Whiteley, D. *The Oxford Book of Vertebrates: Cyclostomes, Fish, Amphibians, Reptiles, and Mammals*, (Oxford University Press, 1972).
13. Gans, C. & Northcutt, R.G. Neural crest and the origin of vertebrates: a new head. *Science* **220**, 268 (1983).
14. Shimeld, S.M. & Holland, P.W.H. Vertebrate innovations. *Proceedings of the National Academy of Sciences* **97**, 4449 (2000).
15. Forey, P. Agnathans and the origin of jawed vertebrates. *Nature* **361**, 129-134 (1993).
16. Hubbs, C. Distribution, phylogeny and taxonomy. In ‘The Biology of Lampreys’. Vol. 1.(Eds MW Hardisty, IC Potter.) pp. 1–65. (Academic Press: London., 1971).

17. Kuratani, S., Kuraku, S. & Murakami, Y. Lamprey as an evo−devo model: Lessons from comparative embryology and molecular phylogenetics. *genesis* **34**, 175-183 (2002).

18. Murakami, Y., Uchida, K., Rijli, F.M. & Kuratani, S. Evolution of the brain developmental plan: Insights from agnathans. *Developmental biology* **280**, 249-259 (2005).

19. Bodznick, D. & Northcutt, R.G. Electroreception in lampreys: evidence that the earliest vertebrates were electroreceptive. *Science* **212**, 465 (1981).

20. Bodznick, D. & Preston, D.G. Physiological characterization of electroreceptors in the lampreysIchthyomyzon unicuspis andPetromyzon marinus. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology* **152**, 209-217 (1983).

21. Ronan, M. & Bodznick, D. End buds: non-ampullary electroreceptors in adult lampreys. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology* **158**, 9-15 (1986).

22. Fritzsch, B., Crapon, C.M.D., Wächtler, K. & Körtje, K. Neuroanatomical evidence for electroreception in lampreys. *Zeitschrift fur Naturforschung. Section C: Biosciences* **39**, 856 (1984).

23. Gill, H.S., Renaud, C.B., Chapleau, F., Mayden, R.L. & Potter, I.C. Phylogeny of living parasitic lampreys (Petromyzontiformes) based on morphological data. *Journal Information* **2003**(2003).

24. Potter, I.C. & Gill, H.S. Adaptive radiation of lampreys. *Journal of Great Lakes Research* **29**, 95-112 (2003).

25. Potter, I., Hilliard, R. & Bird, D. Stages in metamorphosis. *The biology of lampreys* **4**, 137-164 (1982).

26. Li, W. *et al.* Bile acid secreted by male sea lamprey that acts as a sex pheromone. *Science* **296**, 138 (2002).

27. Siefkes, M.J., Scott, A.P., Zielinski, B., Yun, S.S. & Li, W. Male sea lampreys, Petromyzon marinus L., excrete a sex pheromone from gill epithelia. *Biol Reprod* **69**, 125-32 (2003).

28. Johnson, N.S., Luehring, M.A., Siefkes, M.J. & Li, W. Mating pheromone reception and induced behavior in ovulating female sea lampreys. *North American journal of fisheries management* **26**, 88-96 (2006).

29. Johnson, N.S., Siefkes, M.J. & Li, W. Capture of ovulating female sea lampreys in traps baited with spermiating male sea lampreys. *North American journal of fisheries management* **25**, 67-72 (2005).

30. Johnson, N.S., Yun, S.S., Thompson, H.T., Brant, C.O. & Li, W. A synthesized pheromone induces upstream movement in female sea lamprey and summons them into traps. *Proceedings of the National Academy of Sciences* **106**, 1021 (2009).

31. Siefkes, M.J., Winterstein, S.R. & Li, W. Evidence that 3-keto petromyzonol sulphate specifically attracts ovulating female sea lamprey, Petromyzon marinus. *Animal behaviour* **70**, 1037-1045 (2005).

32. Li, W., Scott, A.P., Siefkes, M.J., Yun, S.S. & Zielinski, B. A male pheromone in the sea lamprey (Petromyzon marinus): an overview. *Fish Physiology and Biochemistry* **28**, 259-262 (2003).

33. Wagner, C.M., Jones, M.L., Twohey, M.B. & Sorensen, P.W. A field test verifies that pheromones can be useful for sea lamprey (Petromyzon marinus) control in the Great Lakes. *Canadian Journal of Fisheries and Aquatic Sciences* **63**, 475-479 (2006).
34. Osoegawa, K. *et al.* An improved approach for construction of bacterial artificial chromosome libraries. *Genomics* **52**, 1-8 (1998).
35. Waterston, R.H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-62 (2002).
36. Warren, W.C. *et al.* The genome of a songbird. *Nature* **464**, 757-62 (2010).
37. Warren, W.C. *et al.* Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**, 175-83 (2008).
38. Jaffe, D.B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91-96 (2003).
39. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS biology* **5**, e254 (2007).
40. Green, P. Phrap http://www.phrap.org/phredphrapconsed.html. *University of Washington, Department of Genome Sciences* (1994).
41. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research* **8**, 175-185 (1998).
42. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* **8**, 186-194 (1998).
43. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *Journal of computational biology : a journal of computational molecular cell biology* **7**, 203-14 (2000).
44. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-1067 (2007).
45. Parra, G., Bradnam, K., Ning, Z., Keane, T. & Korf, I. Assessing the gene space in draft genomes. *Nucleic acids research* **37**, 289-97 (2009).
46. Bao, Z. & Eddy, S.R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome research* **12**, 1269-76 (2002).
47. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-3402 (1997).
48. Morgenstern, B. DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. *Nucleic acids research* **32**, W33-6 (2004).
49. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nature reviews. Genetics* **8**, 973-82 (2007).
50. Smit, A.F.A., Hubley, R. & Green, P. RepeatMasker Open-3.0. in *httpwwwrepeatmaskerorg* (http://www.repeatmasker.org/, 1996).
51. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* **110**, 462-467 (2005).
52. Tahara, Y. Normal stages of development in the lamprey Lampetra reissneri (Dybowski). *Zoolog.Sci.* **5**, 109-118 (1988).

53. Cantarel, B.L. *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research* **18**, 188-96 (2008).
54. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Research* **38**, D142-D148 (2010).
55. Simakov, O. *et al.* Insights into bilaterian evolution from three spiralian genomes. *Nature* (2012).
56. Pruitt, K.D., Tatusova, T., Klimke, W. & Maglott, D.R. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Research* **37**, D32-D36 (2009).
57. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
58. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435-W439 (2006).
59. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).
60. Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Research* **37**, D211-D215 (2009).
61. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **25**, 25-29 (2000).
62. Mungall, C.J. & Emmert, D.B. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics* **23**, i337-i346 (2007).
63. Ed, L., Nomi, H., Mark, G., Raymond, C. & Suzanna, L. Apollo: a community resource for genome annotation editing. *Bioinformatics* **25**, 1836-7 (2009).
64. McEwen, G.K. *et al.* Early evolution of conserved regulatory sequences associated with development in vertebrates. *PLoS genetics* **5**, e1000762 (2009).
65. Kenyon, E.J., McEwen, G.K., Callaway, H. & Elgar, G. Functional analysis of conserved non-coding regions around the short stature hox gene (shox) in whole zebrafish embryos. *PLoS One* **6**, e21498 (2011).
66. Woolfe, A. *et al.* CONDOR: a database resource of developmentally associated conserved non-coding elements. *BMC Dev Biol* **7**, 100 (2007).
67. Kuraku, S. & Kuratani, S. Time scale for cyclostome evolution inferred with a phylogenetic diagnosis of hagfish and lamprey cDNA sequences. *Zoolog Sci* **23**, 1053-64 (2006).
68. Kuraku, S. Insights into cyclostome phylogenomics: pre-2R or post-2R. *Zoolog Sci* **25**, 960-8 (2008).
69. Clay, O., Cacciò, S., Zoubak, S., Mouchiroud, D. & Bernardi, G. Human coding and noncoding DNA: compositional correlations. *Molecular Phylogenetics and Evolution* **5**, 2-12 (1996).
70. Musto, H., Romero, H., Zavala, A. & Bernardi, G. Compositional correlations in the chicken genome. *Journal of Molecular Evolution* **49**, 325-329 (1999).
71. Kuraku, S. *et al.* cDNA-based gene mapping and GC3 profiling in the soft-shelled turtle suggest a chromosomal size-dependent GC bias shared by

sauropsids. *Chromosome research an international journal on the molecular supramolecular and evolutionary aspects of chromosome biology* **14**, 187-202 (2006).

72. Ikemura, T. Correlation between the abundance of Escherichia-coli transfer-RNAs and the occurrence of the respective codons in its protein genes - a proposal for a synonymous codon choice that is optimal for the Escherichia-coli translational system. *Journal of Molecular Biology* **151**, 389-409 (1981).
73. Shields, D.C., Sharp, P.M., Higgins, D.G. & Wright, F. "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. *Molecular Biology and Evolution* **5**, 704-716 (1988).
74. Das, S., Pan, A., Paul, S. & Dutta, C. Comparative analyses of codon and amino acid usage in symbiotic island and core genome in nitrogen-fixing symbiotic bacterium Bradyrhizobium japonicum. *Journal of biomolecular structure dynamics* **23**, 221-232 (2005).
75. Ikemura, T. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer R. *Journal of Molecular Biology* **158**, 573-597 (1982).
76. Sharp, P.M., Tuohy, T.M. & Mosurski, K.R. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research* **14**, 5125-5143 (1986).
77. Stenico, M., Lloyd, A.T. & Sharp, P.M. Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases. *Nucleic Acids Research* **22**, 2437-2446 (1994).
78. Duret, L. tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. *Trends in Genetics* **16**, 287-289 (2000).
79. Vicario, S., Moriyama, E.N. & Powell, J.R. Codon usage in twelve species of Drosophila. *BMC Evolutionary Biology* **7**, 226 (2007).
80. Mukhopadhyay, P., Basak, S. & Ghosh, T.C. Differential selective constraints shaping codon usage pattern of housekeeping and tissue-specific homologous genes of rice and arabidopsis. *DNA Res* **15**, 347-56 (2008).
81. Liu, H. *et al.* Analysis of synonymous codon usage in Zea mays. *Molecular Biology Reports* **37**, 677-684 (2010).
82. Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y. & Ikemura, T. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *Journal of Molecular Evolution* **53**, 290-298 (2001).
83. Costantini, M., Auletta, F. & Bernardi, G. Isochore patterns and gene distributions in fish genomes. *Genomics* **90**, 364-371 (2007).
84. Fortes, G.G., Bouza, C., Martínez, P. & Sánchez, L. Diversity in isochore structure among cold-blooded vertebrates based on GC content of coding and non-coding sequences. *Genetica* **129**, 281-289 (2007).

85. Romero, H., Zavala, A., Musto, H. & Bernardi, G. The influence of translational selection on codon usage in fishes from the family Cyprinidae. *Gene* **317**, 141-147 (2003).

86. Qiu, H., Hildebrand, F., Kuraku, S. & Meyer, A. Unresolved orthology and peculiar coding sequence properties of lamprey genes: the KCNA gene family as test case. *BMC Genomics* **12**, 325 (2011).

87. Sharp, P.M. & Li, W.H. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research* **15**, 1281-1295 (1987).

88. Peden, J.F. Analysis of codon usage. in *DNA Repair* (University of Nottingham, 2000).

89. Flicek, P. *et al.* Ensembl 2011. *Nucleic acids research* **39**, D800-6 (2011).

90. Vilella, A.J. *et al.* EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19**, 327-35 (2009).

91. Ruan, J. *et al.* TreeFam: 2008 Update. *Nucleic Acids Res* **36**, D735-40 (2008).

92. Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. *J Mol Biol* **147**, 195-7 (1981).

93. Wallace, I.M., O'Sullivan, O., Higgins, D.G. & Notredame, C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res* **34**, 1692-9 (2006).

94. Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Res* (2011).

95. Smith, J.J., Antonacci, F., Eichler, E.E. & Amemiya, C.T. Programmed loss of millions of base pairs from a vertebrate genome. *Proc.Natl.Acad.Sci.U.S.A* **106**, 11212-11217 (2009).

96. Van de Peer, Y., Maere, S. & Meyer, A. The evolutionary significance of ancient genome duplications. *Nat Rev Genet* **10**, 725-32 (2009).

97. De Bie, T., Cristianini, N., Demuth, J.P. & Hahn, M.W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269-71 (2006).

98. Kumar, S. & Hedges, S.B. TimeTree2: species divergence times on the iPhone. *Bioinformatics* **27**, 2023-4 (2011).

99. Smith, J.J. & Voss, S.R. Gene order data from a model amphibian (Ambystoma): new perspectives on vertebrate genome structure and evolution. *BMC.Genomics* **7**, 219 (2006).

100. Bourque, G., Pevzner, P.A. & Tesler, G. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome research* **14**, 507-16 (2004).

101. Bourque, G., Zdobnov, E.M., Bork, P., Pevzner, P.A. & Tesler, G. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome research* **15**, 98-110 (2005).

102. Burt, D.W. *et al.* The dynamics of chromosome evolution in birds and mammals. *Nature* **402**, 411-3 (1999).

103. Jaillon, O. *et al.* Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* **431**, 946-57 (2004).
104. Larhammar, D., Lundin, L.G. & Hallbook, F. The human Hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. *Genome research* **12**, 1910-20 (2002).
105. Ravi, V. *et al.* Elephant shark (Callorhinchus milii) provides insights into the evolution of Hox gene clusters in gnathostomes. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 16327-32 (2009).
106. Oulion, S. *et al.* Evolution of Hox gene clusters in gnathostomes: insights from a survey of a shark (Scyliorhinus canicula) transcriptome. *Molecular biology and evolution* **27**, 2829-38 (2010).
107. King, B.L., Gillis, J.A., Carlisle, H.R. & Dahn, R.D. A natural deletion of the HoxC cluster in elasmobranch fishes. *Science* **334**, 1517 (2011).
108. Sower, S.A., Freamat, M. & Kavanaugh, S.I. The origins of the vertebrate hypothalamic-pituitary-gonadal (HPG) and hypothalamic-pituitary-thyroid (HPT) endocrine systems: new insights from lampreys. *General and comparative endocrinology* **161**, 20-9 (2009).
109. Fernald, R.D. & White, R.B. Gonadotropin-releasing hormone genes: phylogeny, structure, and functions. *Frontiers in neuroendocrinology* **20**, 224-40 (1999).
110. Parhar, I.S. Cell migration and evolutionary significance of GnRH subtypes. *Progress in brain research* **141**, 3-17 (2002).
111. Morgan, K. & Millar, R.P. Evolution of GnRH ligand precursors and GnRH receptors in protochordate and vertebrate species. *General and comparative endocrinology* **139**, 191-7 (2004).
112. Zhang, L., Tello, J.A., Zhang, W. & Tsai, P.S. Molecular cloning, expression pattern, and immunocytochemical localization of a gonadotropin-releasing hormone-like molecule in the gastropod mollusk, Aplysia californica. *General and comparative endocrinology* **156**, 201-9 (2008).
113. Kavanaugh, S.I., Nozaki, M. & Sower, S.A. Origins of gonadotropin-releasing hormone (GnRH) in vertebrates: identification of a novel GnRH in a basal vertebrate, the sea lamprey. *Endocrinology* **149**, 3860-9 (2008).
114. Kim, D.K. *et al.* Revisiting the evolution of gonadotropin-releasing hormones and their receptors in vertebrates: secrets hidden in genomes. *General and comparative endocrinology* **170**, 68-78 (2011).
115. Kuo, M.W., Lou, S.W., Postlethwait, J. & Chung, B.C. Chromosomal organization, evolutionary relationship, and expression of zebrafish GnRH family members. *Journal of biomedical science* **12**, 629-39 (2005).
116. Sower, S.A., Nozaki, M., Knox, C.J. & Gorbman, A. The occurrence and distribution of GnRH in the brain of Atlantic hagfish, an agnatha, determined by chromatography and immunocytochemistry. *General and comparative endocrinology* **97**, 300-7 (1995).

117. Braun, C.B., Wicht, H. & Northcutt, R.G. Distribution of gonadotropin-releasing hormone immunoreactivity in the brain of the Pacific hagfish, Eptatretus stouti (Craniata: Myxinoidea). *The Journal of comparative neurology* **353**, 464-76 (1995).
118. Hubbard, T.J. *et al.* Ensembl 2009. *Nucleic Acids Res* **37**, D690-7 (2009).
119. The Schistosoma japonicum genome reveals features of host-parasite interplay. *Nature* **460**, 345-351 (2009).
120. Berriman, M. *et al.* The genome of the blood fluke Schistosoma mansoni. *Nature* **460**, 352-8 (2009).
121. Pancer, Z. *et al.* Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. *Nature* **430**, 174-80 (2004).
122. Pancer, Z. *et al.* Variable lymphocyte receptors in hagfish. *Proc Natl Acad Sci U S A* **102**, 9224-9 (2005).
123. Alder, M.N. *et al.* Antibody responses of variable lymphocyte receptors in the lamprey. *Nat Immunol* **e**(2008).
124. Alder, M.N. *et al.* Diversity and function of adaptive immune receptors in a jawless vertebrate. *Science* **310**, 1970-3 (2005).
125. Tasumi, S. *et al.* High-affinity lamprey VLRA and VLRB monoclonal antibodies. *Proc Natl Acad Sci U S A* **106**, 12891-6 (2009).
126. Han, B.W., Herrin, B.R., Cooper, M.D. & Wilson, I.A. Antigen recognition by variable lymphocyte receptors. *Science* **321**, 1834-7 (2008).
127. Velikovsky, C.A. *et al.* Structure of a lamprey variable lymphocyte receptor in complex with a protein antigen. *Nat Struct Mol Biol* **16**, 725-30 (2009).
128. Guo, P. *et al.* Dual nature of the adaptive immune system in lampreys. *Nature* **459**, 796-801 (2009).
129. Hsu, E. The invention of lymphocytes. *Curr Opin Immunol* **23**, 156-62 (2011).
130. Saha, N.R., Smith, J. & Amemiya, C.T. Evolution of adaptive immune recognition in jawless vertebrates. *Semin Immunol* **22**, 25-33 (2010).
131. Nagawa, F. *et al.* Antigen-receptor genes of the agnathan lamprey are assembled by a process involving copy choice. *Nat Immunol* **8**, 206-13 (2007).
132. Rogozin, I.B. *et al.* Evolution and diversification of lamprey antigen receptors: evidence for involvement of an AID-APOBEC family cytosine deaminase. *Nat Immunol* **8**, 647-56 (2007).
133. Mayer, W.E. *et al.* Isolation and characterization of lymphocyte-like cells from a lamprey. *Proc Natl Acad Sci U S A* **99**, 14350-5 (2002).
134. Uinuk-Ool, T. *et al.* Lamprey lymphocyte-like cells express homologs of genes involved in immunologically relevant activities of mammalian lymphocytes. *Proc Natl Acad Sci U S A* **99**, 14356-61 (2002).
135. Herrin, B.R. & Cooper, M.D. Alternative adaptive immunity in jawless vertebrates. *J Immunol* **185**, 1367-74 (2010).
136. Kasamatsu, J. *et al.* Identification of a third variable lymphocyte receptor in the lamprey. *Proc Natl Acad Sci U S A* **107**, 14304-8 (2010).
137. Buckley, K.M. & Rast, J.P. Characterizing immune receptors from new genome sequences. *Methods Mol Biol* **748**, 273-98 (2011).

138. Hibino, T. *et al.* The immune gene repertoire encoded in the purple sea urchin genome. *Dev Biol* **300**, 349-65 (2006).

139. Holland, L.Z. *et al.* The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res* **18**, 1100-11 (2008).

140. Huang, S. *et al.* Genomic analysis of the immune gene repertoire of amphioxus reveals extraordinary innate complexity and diversity. *Genome Res* **18**, 1112-26 (2008).

141. Kenny, E.F. & O'Neill, L.A. Signalling adaptors used by Toll-like receptors: an update. *Cytokine* **43**, 342-9 (2008).

142. Kasamatsu, J., Oshiumi, H., Matsumoto, M., Kasahara, M. & Seya, T. Phylogenetic and expression analysis of lamprey toll-like receptors. *Dev Comp Immunol* **34**, 855-65 (2010).

143. Franchi, L., Warner, N., Viani, K. & Nunez, G. Function of Nod-like receptors in microbial recognition and host defense. *Immunol Rev* **227**, 106-28 (2009).

144. Lange, C. *et al.* Defining the origins of the NOD-like receptor system at the base of animal evolution. *Mol Biol Evol* **28**, 1687-702 (2011).

145. Messier-Solek, C., Buckley, K.M. & Rast, J.P. Highly diversified innate receptor systems and new forms of animal immunity. *Semin Immunol* (2009).

146. Laing, K.J., Purcell, M.K., Winton, J.R. & Hansen, J.D. A genomic view of the NOD-like receptor family in teleost fish: identification of a novel NLR subfamily in zebrafish. *BMC Evol Biol* **8**, 42 (2008).

147. Yoneyama, M. & Fujita, T. RNA recognition and signal transduction by RIG-I-like receptors. *Immunol Rev* **227**, 54-65 (2009).

148. Wang, T. *et al.* Identification of a novel IL-1 cytokine family member in teleost fish. *J Immunol* **183**, 962-74 (2009).

149. Sato, A. *et al.* Macrophage migration inhibitory factor (MIF) of jawed and jawless fishes: implications for its evolutionary origin. *Dev Comp Immunol* **27**, 401-12 (2003).

150. Iwakura, Y., Ishigame, H., Saijo, S. & Nakae, S. Functional specialization of interleukin-17 family members. *Immunity* **34**, 149-62 (2011).

151. Tsutsui, S., Nakamura, O. & Watanabe, T. Lamprey (Lethenteron japonicum) IL-17 upregulated by LPS-stimulation in the skin cells. *Immunogenetics* **59**, 873-82 (2007).

152. Rose, T.M. & Bruce, A.G. Oncostatin M is a member of a cytokine family that includes leukemia-inhibitory factor, granulocyte colony-stimulating factor, and interleukin 6. *Proc Natl Acad Sci U S A* **88**, 8641-5 (1991).

153. Fujii, T., Nakamura, T. & Tomonaga, S. Component C3 of hagfish complement has a unique structure: identification of native C3 and its degradation products. *Mol Immunol* **32**, 633-42 (1995).

154. Matsushita, M. *et al.* Origin of the classical complement pathway: Lamprey orthologue of mammalian C1q acts as a lectin. *Proc Natl Acad Sci U S A* **101**, 10127-31 (2004).

155. Hirano, M., Das, S., Guo, P. & Cooper, M.D. The evolution of adaptive immunity in vertebrates. *Adv Immunol* **109**, 125-57 (2011).

156. Gotz, S. *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**, 3420-3435 (2008).
157. Minguillon, C., Gibson-Brown, J.J. & Logan, M.P. Tbx4/5 gene duplication and the origin of vertebrate paired appendages. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 21726-30 (2009).
158. Kokubo, N. *et al.* Mechanisms of heart development in the Japanese lamprey, Lethenteron japonicum. *Evolution & development* **12**, 34-44 (2010).
159. Ruvinsky, I., Silver, L.M. & Gibson-Brown, J.J. Phylogenetic analysis of T-Box genes demonstrates the importance of amphioxus for understanding evolution of the vertebrate genome. *Genetics* **156**, 1249-57 (2000).
160. Shigetani, Y. *et al.* Heterotopic shift of epithelial-mesenchymal interactions in vertebrate jaw evolution. *Science* **296**, 1316-9 (2002).