Supporting Information

Holland et al. 10.1073/pnas.1415655111

SI Materials and Methods

Determination of Variable Lymphocyte Receptor C Repertoires. *Petromyzon marinus* blood lymphocytes were sorted by flow cytomery on a FACSAria II cell sorter instrument after staining with monoclonal antibodies directed against variable lymphocyte receptor (VLR) A (R110) and VLRC (3A5) directly into single wells of a microtiter plate containing lysis buffer (10 mM Tris·HCl, pH 8.0; 250 ng/mL proteinase K) for subsequent PCR using the following primers: VLRC-F: 5′-AGTGTTGGGTCCCGTGCG-3′ and VLRC-R: 5′-TGCAACGGGGATGTCTCTACTTTA-3 as described (1). Under the sorting conditions used, about 1.5% of wells contained more than one cell. Sequences from thymoid and peripheral blood of *Lampetra planeri* larvae were obtained by PCR of genomic DNA as described (1–3). The VLRC clone selected for protein expression was derived as follows. Whole blood lymphocytes of *L. planeri* (specimen no. LP113, collected from small tributaries of the Rhine river near Freiburg, Germany) were sorted using forward and side light scatter parameters as previously described (4). Total RNA was extracted using Tri-Reagent (Sigma) and converted to cDNA using SuperScript II and random hexamer primers, following the manufacturer's instructions (Invitrogen). The sample was PCR amplified using polymerase with proofreading activity (Phusion; New England Biolabs) with primers specific for the VLRC 5′ signal peptide (VLRC_TB2: 5′-AGCCGAGCCGCGATGGGGTTTGTCGTG-3′ and 3′ terminal (VLRC_TB3: 5′-TACCACTCAATAACGGTGCAGC-3′) under the following cycling conditions: (98 °C for 30 s){(98 °C for 10 s)(56 °C for 20 s)(72 °C for 20 s) × 35}(72 °C for 5 min). Amplification products were gel purified, cloned in pGemT-easy (Promega), and sequenced.

Sequence Deposition. The *P. marinus* and *L. planeri* VLR sequences reported in this paper were deposited with GenBank (accession nos. KJ734027–KJ734078, KJ751404–KJ751454, KJ649525–KJ649607, KJ670495, KC732806–KC733164).

Protein Production and Purification. The diversity region of lamprey VLRC.1MP (GenBank accession no. KJ670495), from N-terminal LRR capping module (LRRNT) to LRR C-terminal capping module (LRRCT) (residues 1–221), was cloned into the expression vector pET26b (Novagen) and expressed as inclusion bodies in BL21(DE3) *Escherichia coli* cells (Agilent). Bacteria were grown at 37 °C in LB medium to an absorbance of 0.6 at 600 nm and induced with 1 mM isopropyl-b-D-thiogalactoside. After incubation for 3 h, the cells were centrifuged and resuspended in 50 mM Tris·HCl (pH 8.0), 0.1 M NaCl, and 2 mM EDTA. Following sonication for cell disruption, inclusion bodies were washed with 50 mM Tris·HCl (pH 8.0), 0.1 M NaCl, and 0.5% (vol/vol) Triton X-100, and then solubilized in 8 M urea and 100 mM Tris·HCl (pH 9.0). In vitro folding was carried out by drop dilution of inclusion bodies to a final concentration of 10 mg/L into 1.0 M arginine, 100 mM Tris·HCl (pH 9.0), 2 mM EDTA, 3 mM cysteamine, and 0.3 mM cystamine. After 3 d at 4 °C, the folding mixture was concentrated, dialyzed against 25 mM Tris·HCl (pH 9.0), and applied to a Sephadex 75 HR column (GE Healthcare). Further purification was carried out using a MonoQ column.

Crystallization and Structure Determination. The hanging drop vapor diffusion method was used for crystallization of VLRC.1MP. Crystals were obtained at room temperature by mixing equal volumes of protein solution (10 mg/mL) and reservoir solution containing 0.1 M succinic acid:sodium dihydrogen phosphate: glycine buffer (SPG) (pH 9.0) and 20% (wt/vol) PEG 1500. For data collection, crystals of VLRC were cryoprotected with 40% (wt/vol) PEG 400 before flash cooling in liquid nitrogen. X-ray diffraction data were recorded in-house at 100 K with a Rigaku R-axis IV^{++} image plate detector. The data were indexed, integrated, and scaled using the program CrystalClear (5). Data collection statistics are summarized in Dataset S2. The structure of VLRC.1MP was solved by molecular replacement with the Phaser program (6). The search model was truncated hagfish VLRB.59 (Protein Data Bank, PDB accession code 2O6S) (7). Structure refinement was performed using Phenix (8). Rebuilding and modeling were accomplished manually with COOT (9) according to $2F_o - F_c$ and $F_o - F_c$ maps. Stereochemical parameters were evaluated by PROCHECK (10). Final refinement statistics are presented in Dataset S2. Figures were prepared using PyMol (http://www.pymol.org). Atomic coordinates and structure factors for VLRC.1MP have been deposited in the PDB under accession code 4PO4.

Sequence Analysis. All VLR protein sequences were searched using the BLASTP and PSI-BLAST programs (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Multiple sequence alignments for analysis of VLRs were performed using Kalign followed by manual adjustment using known structures as guides (11). Shannon entropy analysis was carried out using the sequences listed in Dataset S3. Note that the *VLRC* sequences from *Eptatretus burgeri* and *Eptatretus stoutii* (GenBank accession nos. AY964719–AY964931) are designated as *VLRA* in the database entries; a recent study (12), however, found that they in fact represent the *VLRC* genes of hagfish and reported the sequences of several bona fide hagfish *VLRA* assemblies (see GenBank accession nos. KF314046–KF314110).

Statistical Analysis. The coefficient of determination (R^2) was calculated using the linear regression analysis implemented in Prism 5 for Mac OS X (v5.0a). R values were derived from the aforementioned R^2 calculations. Shannon entropy analysis was used to determine and interpret the diversity of each amino acid position in a protein alignment according to Litwin and Jores as implemented at imed.med.ucm.es/PVS/pvs-help.html. Entropy scores were calculated using a custom PERL script that uses sequence alignments as input. The Shannon entropy formula, $H = -\sum_{i=1}^{M} P_i \log_z P_i$, where P_i is the fraction of a given amino acid residue i, and M, the total number of different amino acids, was used to calculate the entropy scores. Fisher's r to z transformation was used to compare correlation coefficients (R) to determine statistically significant differences between correlations from linear regression analyses; z scores generated using this transformation are compared using the Cohen and Cohen procedure implemented at www.quantpsy.org/corrtest/corrtest.htm.

1. Hirano M, et al. (2013) Evolutionary implications of a third lymphocyte lineage in lampreys. *Nature* 501(7467):435–438.
2. Bajoghli B, et al. (2011) A thymus candidate in lampreys. *Nature* 470(7332):90–94.
3. Das S, et al. (2013) Organization of lamprey *variable lymphocyte receptor C* locus and repertoire development. *Proc Natl Acad Sci USA* 110(15):6043–6048.
4. Mayer WE, et al. (2002) Isolation and characterization of lymphocyte-like cells from a lamprey. *Proc Natl Acad Sci USA* 99(22):14350–14355.
5. Pflugrath JW (1999) The finer things in X-ray diffraction data collection. *Acta Crystallogr D Biol Crystallogr* 55(Pt 10):1718–1725.
6. Storoni LC, McCoy AJ, Read RJ (2004) Likelihood-enhanced fast rotation functions. *Acta Crystallogr D Biol Crystallogr* 60(Pt 3):432–438.

7. Kim HM, et al. (2007) Structural diversity of the hagfish variable lymphocyte receptors. *J Biol Chem* 282(9):6726–6732.

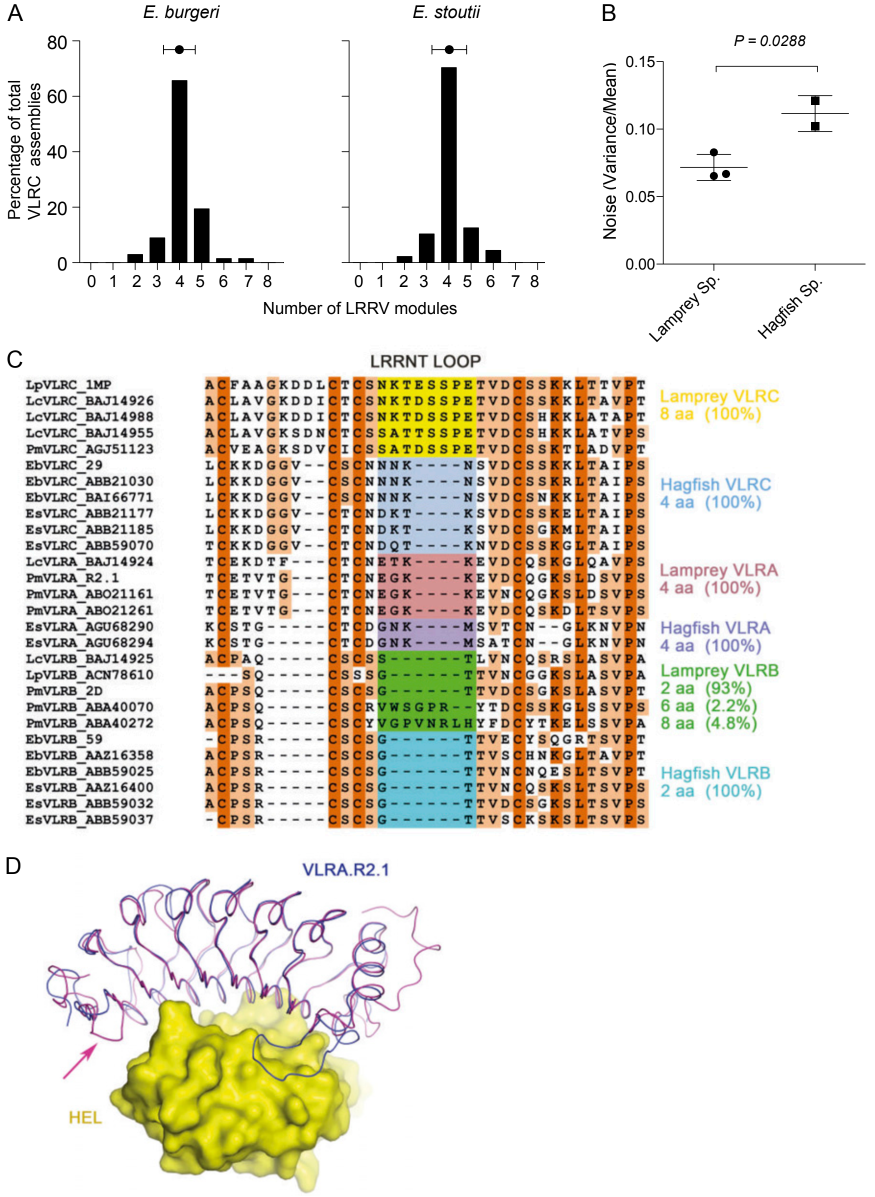
8. Adams PD, et al. (2010) *PHENIX*: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 66(Pt 2): 213–221.

9. Emsley P, Cowtan K (2004) *Coot*: Model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* 60(Pt 12 Pt 1):2126–2132.

10. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) *PROCHECK*: A program to check the stereo chemical quality of protein structures. *J Appl Cryst* 26:283–291.

11. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22(22):4673–4680.

12. Li J, Das S, Herrin BR, Hirano M, Cooper MD (2013) Definition of a third *VLR* gene in hagfish. *Proc Natl Acad Sci USA* 110(37):15013–15018.



```
LpVLRC_1MP         ACFAAGKDDLCTCSNKTESSPETVDCSSKKLTTVPT
LcVLRC_BAJ14926    ACLAVGKDDICTCSNKTDSSPETVDCSSKKLTAVPT
LcVLRC_BAJ14988    ACLAVGKDDICTCSNKTDSSPETVDCSHKKLATAPT
LcVLRC_BAJ14955    ACLAVGKSDNCTCSSATTSSPETVDCSHKKLATVPS
PmVLRC_AGJ51123    ACVEAGKSDVCICSSATDSSPETVDCSSKTLADVPT
EbVLRC_29          LCKKDGGV--CSCNNNK----NSVDCSSKKLTAIPS
EbVLRC_ABB21030    LCKKDGGV--CSCNNNK----NSVDCSSKRLTAIPS
EbVLRC_BAI66771    LCKKDGGV--CSCNNNK----NSVDCSNKKLTAIPS
EsVLRC_ABB21177    LCKKDGGV--CTCNDKT----KSVDCSSKELTAIPS
EsVLRC_ABB21185    LCKKDGGV--CTCNDKT----KSVDCSGKMLTAIPS
EsVLRC_ABB59070    TCKKDGGV--CTCNDQT----KNVDCSSKGLTAIPS
LcVLRA_BAJ14924    TCEKDTF---CTCNETK----KEVDCQSKGLQAVPS
PmVLRA_R2.1        TCETVTG---CTCNEGK----KEVDCQGKSLDSVPS
PmVLRA_ABO21161    TCETVTG---CTCNEGK----KEVNCQGKSLDSVPS
PmVLRA_ABO21261    TCETVTG---CTCNEGK----KEVDCQSKDLTSVPS
EsVLRA_AGU68290    KCSTG-----CTCDGNK----MSVTCN--GLKNVPN
EsVLRA_AGU68294    KCSTG-----CTCDGNK----MSATCN--GLKNVPN
LcVLRB_BAJ14925    ACPAQ-----CSCSS------TLVNCQSRSLASVPA
LpVLRB_ACN78610    ---SQ-----CSSSG------TTVNCGGKSLASVPA
PmVLRB_2D          ACPSQ-----CSCSG------TTVDCSGKSLASVPT
PmVLRB_ABA40070    ACPSQ-----CSCRVWSGPR--YTDCSSKGLSSVPS
PmVLRB_ABA40272    ACPSQ-----CSCYVGPVNRLHYFDCYTKELSSVPA
EbVLRB_59          -CPSR-----CSCSG------TTVECYSQGRTSVPT
EbVLRB_AAZ16358    ACPSR-----CSCSG------TTVSCHNKGLTAVPT
EbVLRB_ABB59025    ACPSR-----CSCSG------TTVNCNQESLTSVPT
EsVLRB_AAZ16400    ACPSR-----CSCSG------TTVNCQSKSLTSVPS
EsVLRB_ABB59032    ACPSR-----CSCSG------TTVDCSGKSLTSVPS
EsVLRB_ABB59037    -CPSR-----CSCSG------TTVSCKSKSLTSVPS
```

Fig. S1. Analysis of the distribution of the number of LRRV modules in the VLRC repertoires. (*A*) Analysis of the distribution of LRRV module numbers in the VLRC repertoires of the two hagfish species *E. burgeri* and *E. stoutii* (the GenBank accession nos. for sequences used in this analysis are listed in *Materials and Methods* and Dataset S3). The dots correspond to overall mean LRRV number with error bars representing the SDs of the means. (*B*) Comparison of VLRC LRRV number "noise" (variance divided by the mean) as a way to measure the shape of the distribution in lampreys and hagfishes. Noise values were calculated for the three lamprey species analyzed in Fig. 1 *B*–*D* and the two hagfish species from *A* of this figure. The average noise between clades was analyzed using Student's unpaired *t* test. Horizontal lines represent the mean noise with error bars showing the SDs of the means. (*C*) Structure-based sequence alignment of VLR LRRNT modules. Residues between the β1 and β2 strands of LRRNT in representative lamprey and hagfish VLR sequences are highlighted in yellow (lamprey VLRC), cyan (hagfish VLRC), salmon (lamprey VLRA), purple (hagfish VLRA), green (lamprey VLRB), and teal (hagfish VLRB). (*D*) Superposition of VLRC.1MP (magenta) onto VLRA.R2.1 (blue) in the VLRA.R2.1–HEL complex (3M18). HEL (yellow) is shown as a surface representation. The protruding LRRNT loop of VLRC.1MP (red arrow) could potentially contact antigen.

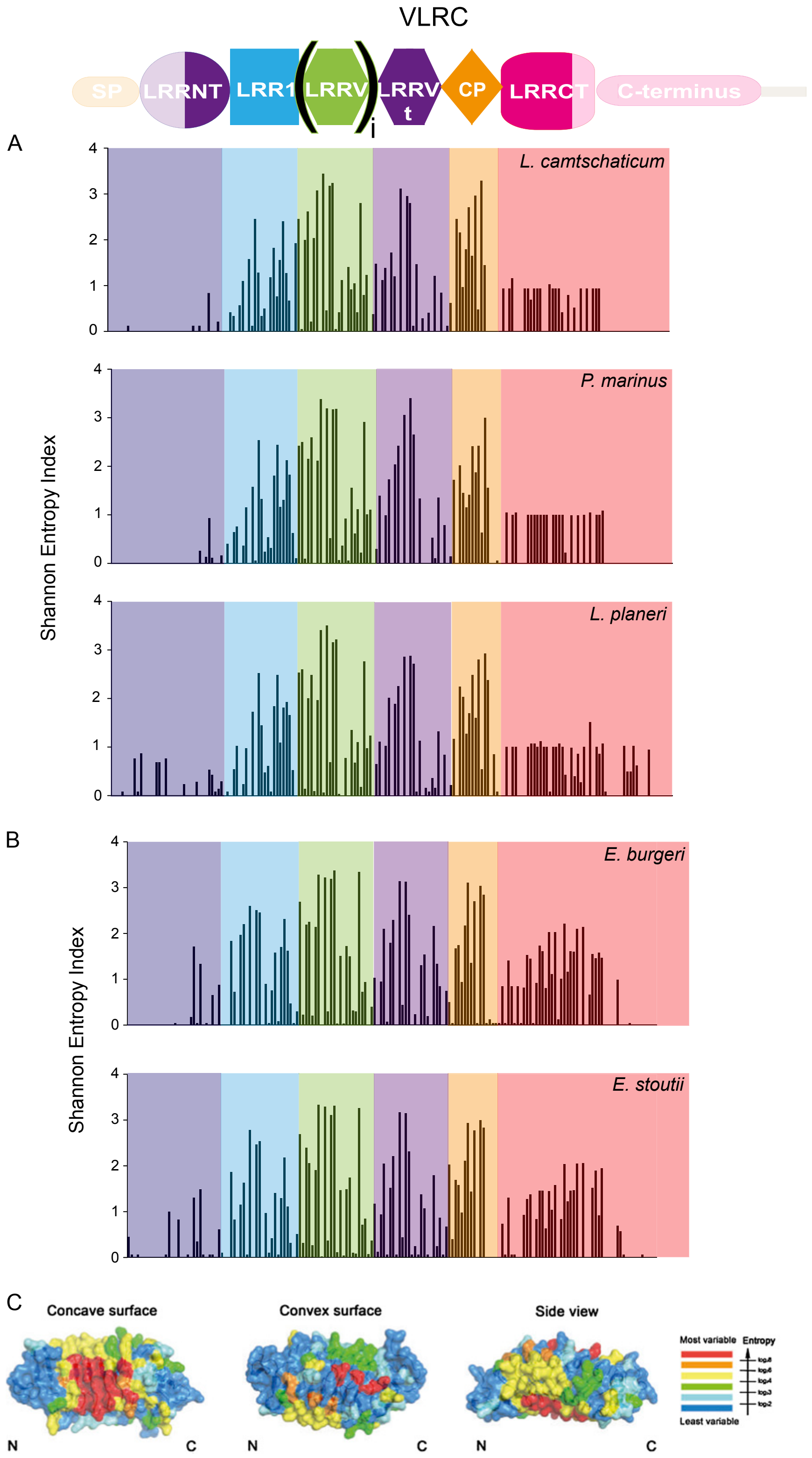


Fig. S2. Shannon entropy indices for VLRC molecules of jawless vertebrates. The individual molecules (schematic above the figure) are color coded for orientation. (*A*) Distribution of sequence diversity of lamprey VLRCs. (*B*) Distribution of sequence diversity of hagfish VLRCs. (*C*) Distribution of sequence diversity of hagfish VLRCs mapped onto the structure of hagfish VLRC.29 (2O6S) (7). Shannon entropy scores are color coded.

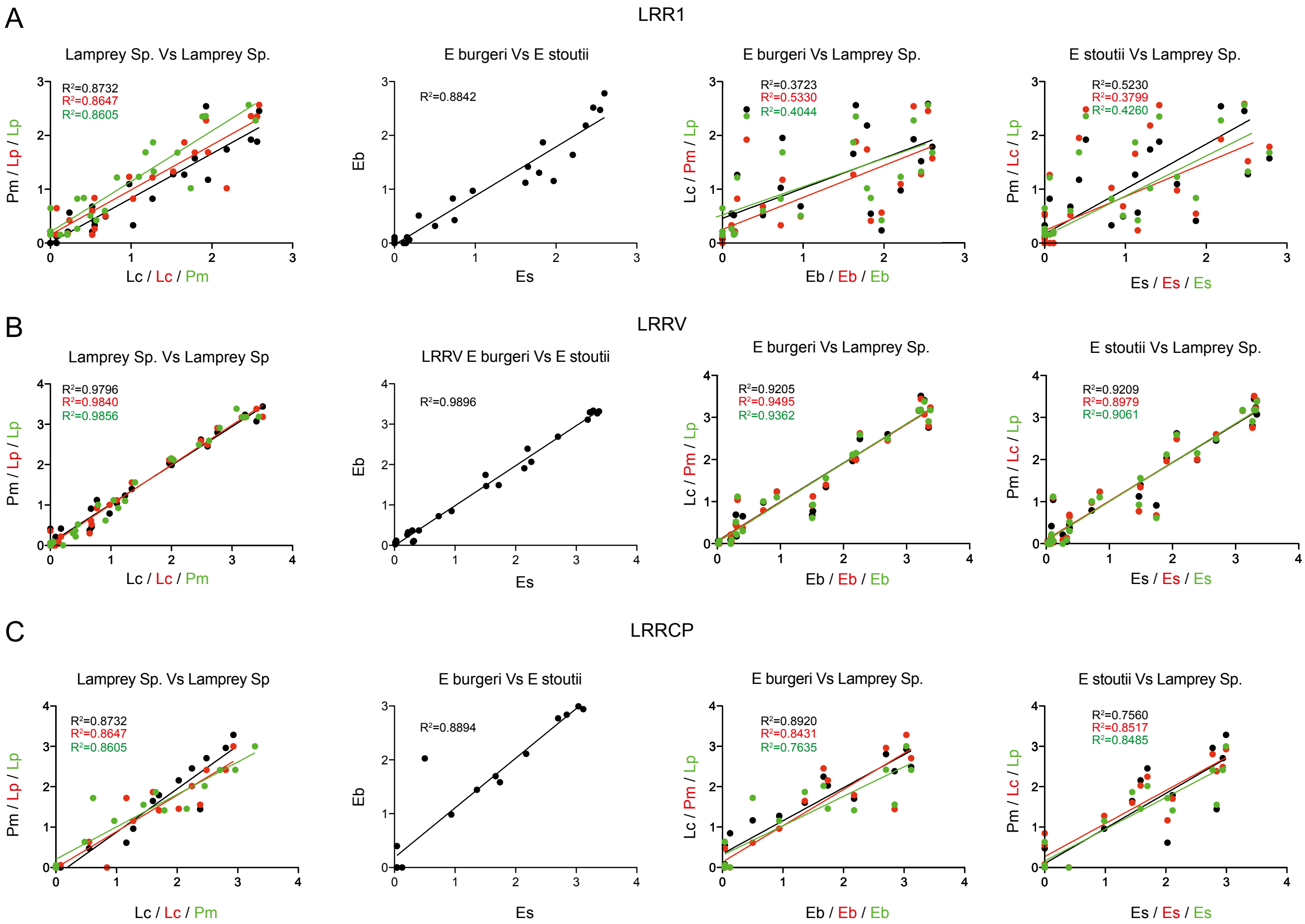


Fig. S3. LRRV and LRRCP but not LRR1 diversity is strongly correlated between hagfish and lamprey species. Scatterplots of module diversity were generated as in Fig. 4, but here isolating each individual species from the hagfish and lamprey clades. Intraclade comparisons (lamprey versus lamprey and hagfish versus hagfish; first and second panels from the *Left*, respectively) were compared with interclade comparisons (*E. burgeri* versus each lamprey species and *E. stoutii* versus each lamprey species; third and fourth panels from the *Left*, respectively) for the LRR1 (*Top*), LRRV (*Middle*), and LRRCP (*Bottom*) VLRC modules. Specific comparisons are color coded accordingly with their respective coefficient of determination (R^2). Eb, *E. burgeri*; Es, *E. stoutii*; Lc, *L. camtschaticum*; Lp, *L. planeri*; Pm, *P. marinus*.

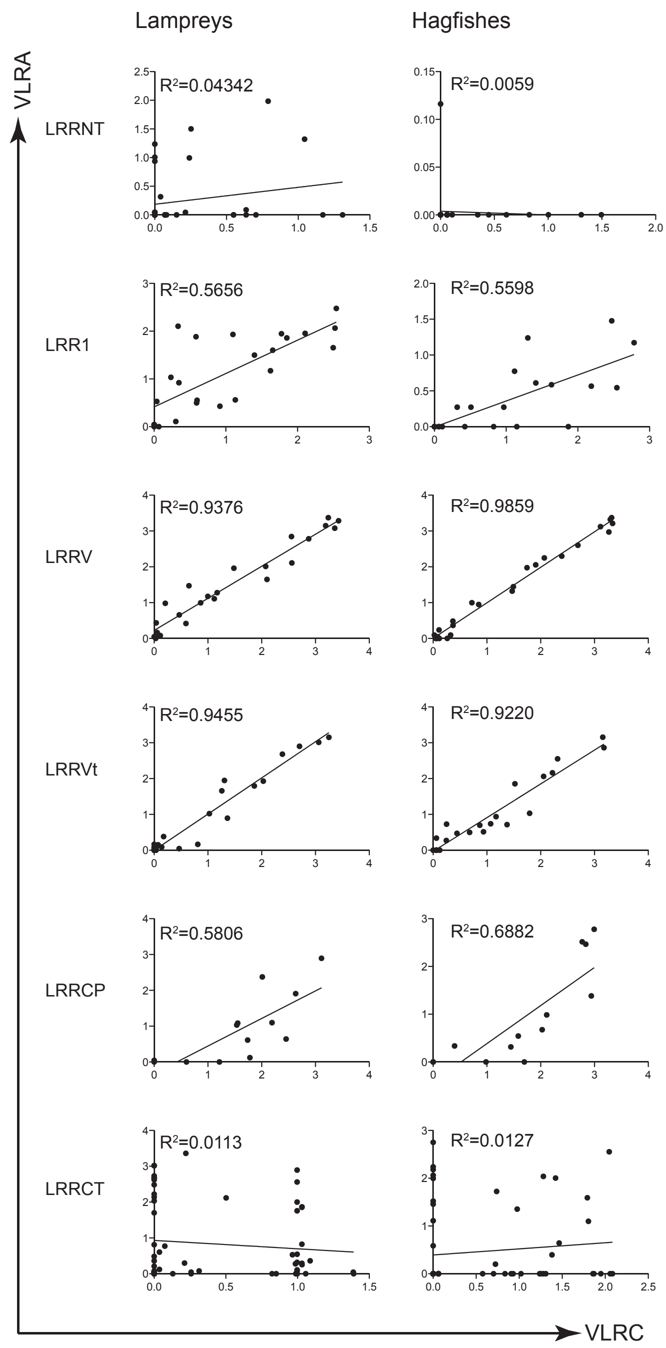


Fig. S4. LRRNT, LRR1, LRRCP, and LRRCT module diversity is poorly correlated between VLRA and VLRC of lamprey and hagfish. Scatterplots of module diversity (Dataset S3) were generated for VLRCs of three lamprey species and two hagfish species and compared with their corresponding VLRA modules from the same clade. Coefficient of determination (R^2) is displayed for each comparison. Accession nos. of sequences used for the analysis are listed in *Materials and Methods* and Dataset S3.

Other Supporting Information Files

Dataset S1 (PDF)
Dataset S2 (XLSX)
Dataset S3 (PDF)
Dataset S4 (XLSX)