Copy number and sequence variation of leucine-rich repeat modules suggests distinct functional constraints operating on variable lymphocyte receptors expressed by agnathan T-cell-like and B-cell-like lymphocytes

Yoichi Sutoh • Masanori Kasahara

Yoichi Sutoh • Masanori Kasahara (✉)

Department of Pathology, Hokkaido University Graduate School of Medicine, North 15 West 7, Sapporo 060-8638, Japan

e-mail: mkasaha@med.hokudai.ac.jp

*Present address: Emory Vaccine Center and Department of Pathology and Laboratory Medicine, Emory University, Atlanta, Georgia 30322, USA

Abstract Unlike jawed vertebrates that use T-cell and B-cell receptors for antigen recognition, jawless vertebrates represented by lampreys and hagfish use variable lymphocyte receptors (VLR) as antigen receptors. VLRs generate high levels of diversity by assembling variable leucine-rich repeat (LRR) modules. Of the three VLRs thus far identified, VLRB is expressed on B-cell-like lymphocytes and functions as antibodies, whereas VLRA and VLRC are expressed on T-cell-like lymphocytes and function as membrane-bound receptors. In the present study, we show that the copy number of LRRV modules in lamprey and hagfish VLRB transcripts follows a binominal distribution with the success rates of 15.5% and 22.4%, respectively. By contrast, the copy number distribution of LRRV modules in VLRA and VLRC transcripts deviates from the binominal distribution mainly because transcripts with two or less LRRV modules occur infrequently. Notably, the second LRRV module shows distinctive sequence signatures in VLRA and VLRC, but not in VLRB transcripts. These observations suggest that distinct functional constraints operate on VLRs expressed by agnathan T-cell-like and B-cell-like lymphocytes.

Keywords Binominal distribution • Jawless vertebrate • Leucine-rich repeat • Thymoid • Thymic selection

Introduction

Jawless vertebrates represented by lampreys and hagfish use variable lymphocyte receptors (VLR) as antigen receptors (Boehm et al. 2012; Cooper and Alder 2006; Kasahara and Sutoh 2014). VLRs generate diversity comparable to that of gnathostome T-cell and B-cell receptors by assembling variable leucine-rich repeat (LRR) modules. Thus far, three types of VLRs, designated VLRA, VLRB, and VLRC, have been identified in both lampreys (Kasamatsu et al. 2010; Pancer et al. 2004; Rogozin et al. 2007) and hagfish (Li et al. 2013; Pancer et al. 2005). VLRB is expressed on B-cell-like lymphocytes and secreted as antibodies in response to antigen stimulation (Alder et al. 2008; Alder et al. 2005; Guo et al. 2009). By contrast, VLRA and VLRC are expressed on T-cell-like lymphocytes and function as membrane-bound receptors (Guo et al. 2009; Hirano et al. 2013). Based on gene expression and tissue distribution profiles, it has been suggested that VLRA+ cells and VLRC+ cells are phylogenetically and functionally related to gnathostome αβ T cells and γδ T cells, respectively (Hirano et al. 2013).

Structurally, all VLR proteins are composed of an N-terminal cap (LRRNT), an 18-residue N-terminal LRR module (LRR1), multiple 24-residue variable LRR modules (LRRV), a 13-residue LRR known as the connecting peptide (CP), a C-terminal cap (LRRCT), and an invariant domain containing a stalk region (Boehm et al. 2012). The LRRV module has the consensus sequence XLXXLXXLXLXXNXLXXLPXXXFX (where X stands for any amino acid), with the most C-terminal LRRV module, known as LRRVe, displaying a distinct sequence signature (Alder et al. 2005). Sequence diversity is located primarily in the 3´-part of LRRNT (3´-LRRNT), LRR1, LRRV, LRRVe, CP, and the 5´-part of LRRCT (5´-LRRCT).

In lampreys, VLRA+ cells and VLRC+ cells assemble their antigen receptors in the thymoid, an organ located at the tips of the gill filaments with functions presumably equivalent to those of the thymus in jawed vertebrates (Bajoghli et al. 2011; Das et al. 2013; Hirano et al. 2013). Cytidine deaminase 1, an AID/APOBEC homolog expressed specifically in developing T-cell-like lymphocytes, is thought to trigger a gene conversion-like process leading to the assembly of *VLRA* and *VLRC* genes (Guo et al. 2009; Rogozin et al. 2007). Although the biologic events occurring in the thymoid remain poorly understood, the observation that misassembled *VLRA* and *VLRC* genes are hardly found in the peripheral blood but abundant in thymoids suggested that this organ performs quality control functions in the development of agnathan T-lineage cells (Bajoghli et al. 2011; Hirano et al. 2013).

Previously, it was suggested that the copy number distribution of LRRV modules follows the Poisson distribution in lamprey VLRB transcripts (Alder et al. 2005). In present study, we extended this work and analyzed the copy number and sequence variation of LRRV modules using currently available lamprey and hagfish *VLRA*, *VLRB*, and *VLRC* sequences.

Materials and methods

Sequence analysis

Essentially all publicly available, non-germline VLR sequences from lampreys and hagfish were included in analysis (Table 1). Sequences excluded from the analysis were those that obviously originated from the same clones or had atypical domain structures or non-functional sequences. The sequences were aligned using the Clustal Omega program (https://www.ebi.ac.uk/Tools/msa/clustalo/) and analyzed on MEGA 5.2 (Tamura et al. 2011).

Estimation based on the binominal distribution (BD) model

The copy number distribution of LRRV modules predicted by the binominal probability b(*x*, *n*, *p*) was obtained using the "BINOM.DIST" function in Microsoft Excel 2010 based on the following formula:

$$\mathrm{b}(x, n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

where $\binom{n}{x}$ is the binomial coefficient defined as follows:

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

where *x*, *n*, and *p* represent the number of successful LRRV insertion, the number of LRRV insertion trials, and the probability of success (or the success rate) for each insertion event, respectively. The model distribution was estimated in the following range: *x* = 1, 2, 3, 4, 5, 6, 7, and 8; *n* = 8; and *p* = 10-30%. The parameter *n* was fixed to 8 because no VLR transcripts with more than eight LRRV modules have been identified.

To optimize the success rate in the BD model, we calculated the correlation coefficients between the model and the actual distribution of LRRV modules in VLRB transcripts at the following success rates: *p* = 10, 12.5, 15, 16, 16.7, 17, 17.5, 18, 19, 20, 22.5, 25, 27.5, and 30%. The correlation coefficients thus obtained were plotted in a two-dimensional system where the success rate and correlation coefficient were represented by *x* and *y*, respectively. Because this plot yielded a symmetric curve with a single peak, it was approximated with a quadratic equation using a least squares algorithm implemented in the graph drawing function of Microsoft Excel 2010. The optimized success rate was calculated from the quadratic equation by the following formula:

$$y = ax^2 + bx + c$$

where the success rate at the maximum correlation coefficient (*X*) is represented as

$$X = -\frac{b}{2a}$$

Results

The copy number distribution of LRRV modules in VLRB transcripts follows a BD

During lymphocyte development, the intervening sequence of the germline *VLR* gene is replaced by a gene conversion-like mechanism in a stepwise manner, beginning either from its 5'- or 3'-end, by inserting flanking modules, eventually forming a completely assembled *VLR* gene (Alder et al. 2005; Nagawa et al. 2007). Previous work has shown that the copy number distribution of LRRV modules shows the Poisson distribution in lamprey VLRB transcripts (Alder et al. 2005). Because the insertion of each LRRV module is likely an independent event with either success or failure, and because the success rate likely remains constant in each insertion trial, we reasoned that the copy number of LRRV modules should follow a BD. As expected, the copy number of LRRV modules in VLRB transcripts showed a distribution that closely matched a BD (Fig. 1). When the success rate at each LRRV insertion event was changed from 10% to 30%, the BD models assuming the success rates of 15.5% and 22.4% yielded the best correlation with the actual distribution of LRRV modules in lamprey and hagfish VLRB transcripts, respectively (Fig. 1A). Interestingly, transcripts with three LRRV modules were overrepresented in hagfish VLRB transcripts, producing a small hump in the distribution curve and resulting in an increased success rate (Fig. 1B).

The copy number distribution of LRRV modules in VLRA and VLRC transcripts deviates from a BD

We next examined whether the copy number of LRRV modules in VLRA and VLRC transcripts follows a BD (Fig. 2). Because the copy number distribution of LRRV modules is essentially the same in VLRA and VLRC transcripts, we combined these transcripts in the following analysis and called them T-lineage VLR transcripts (Fig. 2A). We found that the copy number distribution of LRRV modules deviates from the BD in T-lineage VLR transcripts (Fig. 2B), as indicated by a large positive value for kurtosis and skewness (Fig. 2C). This deviation occurred mainly because T-lineage VLR transcripts with two or less LRRVs were relatively rare (26%, 23%, 12%, and 12% in lamprey VLRA, hagfish VLRA, and lamprey VLRC, and hagfish VLRC transcripts, respectively), whereas such transcripts occupied 91% and 70% in lamprey and hagfish VLRB transcripts, respectively (Fig. 2C). As a result, the average copy number of LRRV modules was significantly larger in T-lineage VLR transcripts than in VLRB transcripts (2.932, 3.114, 3.023, and 1.537 in VLRA, VLRC, VLRA/VLRC combined, and VLRB transcripts, respectively) as pointed out previously (Li et al., 2013; Pancer et al., 2005).

Because the paucity of transcripts with two or less LRRV modules was an essential feature that distinguished T-lineage VLR transcripts from VLRB transcripts, we constructed the BD models for lamprey and hagfish *VLRA/VLRC* genes on the assumption that the insertion of an LRRV module occurs with the success rates identical to those in the lamprey and hagfish *VLRB* genes, and that the T-lineage transcripts with two or less LRRV modules are not allowed to occur. These adjusted BD models showed good correlations with the actual distribution of LRRV

modules in T-lineage VLR transcripts (Fig. 2D). The correlation coefficients were 0.9662 and 0.9625 for lamprey and hagfish T-lineage VLR transcripts, respectively. These results indicate that the copy number distribution of LRRV modules in T-lineage VLR transcripts basically follows the BD, but is deviated from it, because transcripts with two or less LRRVs are infrequent.

The second LRRV module in T-lineage VLR transcripts has distinctive sequence signatures

The deviated copy number distribution of LRRV modules in T-lineage transcripts prompted us to examine whether their LRRV modules have distinctive features not seen in VLRB transcripts. When the consensus sequences were analyzed by WebLogo 3 (http://weblogo.threeplusone.com/create.cgi) for each LRRV module, we noticed that the second LRRV module in VLRA and VLRC molecules have distinctive sequences (Fig. 3 and supplementary Fig. 1). In the LRRV modules of lamprey VLRA molecules, dominant amino acids at residues 4, 6, and 16 are lysine, threonine, and glutamine, respectively. However, the corresponding residues in the second LRRV module are glutamic acid or asparagine for residue 2, lysine or aspartic acid for residue 4, and lysine for residue 16 (Fig. 3A). These distinctive amino acid preferences are also observed in the second LRRV module of lamprey and hagfish VLC molecules. By contrast, neither lamprey nor hagfish VLRB molecules show such distinctive sequence preferences in their second LRRV modules (supplementary Fig. 1). When mapped onto the crystal structure of lamprey VLRA, residues 4 and 6 of the second LRRV module are on the antigen-binding surface (Fig. 3B), suggesting that amino acid substitutions at these positions may affect antigen recognition.

Discussion

We showed here that the copy number distribution of LRRV modules in VLRB transcripts can be accounted for by a simple mathematical model, namely a BD (Fig. 1). This indicates that jawless vertebrates have a simple, but sophisticated recombination mechanism that enables to insert LRRV modules at a constant success rate. The BD model yielded best matches to the actual data when the success rates in each insertion event were assumed to be 15.5% and 22.4% in lamprey and hagfish *VLRB* genes, respectively.

The most striking feature that characterizes VLRs expressed on T-cell-like lymphocytes was the paucity of transcripts with two or less LRRV modules (Fig. 2). When BD models were constructed on the assumption that the T-lineage transcripts with two or less LRRV modules are not allowed to occur and that the insertion of LRRV modules in lamprey and hagfish *VLRA/VLRC* genes occurs with the success rates identical to those in lamprey and hagfish *VLRB* genes, respectively, they accurately reproduced the actual distribution of LRRV modules in T-lineage VLR transcripts (Fig. 2). This indicates that the assembly process does not fundamentally differ between B-lineage and T-lineage *VLR* genes and that the same Bernoulli process applies to the assembly of all three *VLR* genes. It is therefore likely that VLRA and VLRC transcripts with two or less LRRV modules are selected against presumably in the thymoid (Bajoghli et al., 2011), resulting in the preferential, but not absolute elimination of lymphocyte clones expressing such receptors before reaching the periphery.

Another indication that some selection, which could be akin to positive thymic selection of gnathostome T cells, might operate on agnathan T-cell-like lymphocytes is that the second

LRRV module shows distinctive amino acid preferences in VLRA and VLRC, but not in VLRB molecules (Fig. 3 and supplementary Fig. 1). Remarkably, this feature is conserved not only across different species of lampreys or hagfish, but also between lampreys and hagfish. It is possible that lymphocyte clones, which express VLRA or VLRC molecules with an appropriate second LRRV module, have selective advantages. Indeed, all known VLRA sequences (249 sequences) have at least two LRRV modules. Also, of 528 VLRC sequences in the database, only eight have less than two LRRV modules and hence are devoid of the second LRRV module. Despite the presence of the second LRRV module, VLRA and VLRC transcripts with only two LRRV modules are relatively rare. Therefore, it appears that the second LRRV module with an appropriate sequence is required, but not sufficient for conferring full selective advantages upon VLRA+ or VLRC+ cells. Presumably, VLRA+ or VLRC+ lymphocyte clones, which have three or more LRRV modules and an appropriate second LRRV module, are preferentially selected.

Accumulated evidence indicates that jawless vertebrates do not have major histocompatibility complex (MHC) class I or class II molecules (Flajnik and Kasahara 2010; Smith et al. 2013). Nevertheless, their immune system is equipped with the essential features of adaptive immunity such as self-tolerance (Takaba et al. 2013) and immunologic memory (Finstad and Good 1964; Hildemann and Thoenes 1969). Therefore, jawless vertebrates might have molecules with functions equivalent to those of gnathostome MHC molecules. It is tempting to speculate that the distinctive sequence preferences seen in the second LRRV module of T-lineage VLR transcripts reflect the needs of VLRA and VLRC molecules to interact with such antigen-presenting molecules.

This work does not provide insights into the molecular mechanisms underlying the unsuccessful assembly of *VLR* genes. Previous work has shown that one mechanism for

unsuccessful assembly is the insertion of an LRRV module containing a stop codon (Bajoghli et al. 2011; Das et al. 2013). We suggest the premature termination of a gene conversion-like process as a potentially important mechanism; this may occur if a region of nucleotide identity required for priming is non-existent between the LRRV modules to be joined, or if such a region of homology is too short to initiate the gene conversion-like process.

Acknowledgements This work was supported by Grants-in-Aid for Scientific Research from The Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) and a grant from The Uehara Memorial Foundation, Tokyo, Japan. Y. S. was the recipient of a Research Fellowship for Young Scientists from the Japan Society for the Promotion of Science.

Conflict of interest The authors declare they have no conflict of interest.

References

Alder MN, Herrin BR, Sadlonova A, Stockard CR, Grizzle WE, Gartland LA, Gartland GL, Boydston JA, Turnbough CL, Jr., Cooper MD (2008) Antibody responses of variable lymphocyte receptors in the lamprey. Nat Immunol 9:319-327

Alder MN, Rogozin IB, Iyer LM, Glazko GV, Cooper MD, Pancer Z (2005) Diversity and function of adaptive immune receptors in a jawless vertebrate. Science 310:1970-1973

Bajoghli B, Guo P, Aghaallaei N, Hirano M, Strohmeier C, McCurley N, Bockman DE, Schorpp M, Cooper MD, Boehm T (2011) A thymus candidate in lampreys. Nature 470:90-94

Boehm T, McCurley N, Sutoh Y, Schorpp M, Kasahara M, Cooper MD (2012) VLR-based adaptive immunity. Annu Rev Immunol 30:203-220

Cooper MD, Alder MN (2006) The evolution of adaptive immune systems. Cell 124:815-822

Das S, Hirano M, Aghaallaei N, Bajoghli B, Boehm T, Cooper MD (2013) Organization of lamprey variable lymphocyte receptor C locus and repertoire development. Proc Natl Acad Sci USA 110:6043-6048

Deng L, Velikovsky CA, Xu G, Iyer LM, Tasumi S, Kerzic MC, Flajnik MF, Aravind L, Pancer Z, Mariuzza RA (2010) A structural basis for antigen recognition by the T cell-like lymphocytes of sea lamprey. Proc Natl Acad Sci USA 107:13408-13413

Finstad J, Good RA (1964) The evolution of the immune response. III. Immunologic responses in the lamprey. J Exp Med 120:1151-1168

Flajnik MF, Kasahara M (2010) Origin and evolution of the adaptive immune system: genetic events and selective pressures. Nat Rev Genet 11:47-59

Guo P, Hirano M, Herrin BR, Li J, Yu C, Sadlonova A, Cooper MD (2009) Dual nature of the adaptive immune system in lampreys. Nature 459:796-802

Hildemann WH, Thoenes GH (1969) Immunological responses of Pacific hagfish. I. Skin transplantation immunity. Transplantation 7:506-521

Hirano M, Guo P, McCurley N, Schorpp M, Das S, Boehm T, Cooper MD (2013) Evolutionary implications of a third lymphocyte lineage in lampreys. Nature 501:435-438

Kanda R, Sutoh Y, Kasamatsu J, Maenaka K, Kasahara M, Ose T (2014) Crystal structure of the lamprey variable lymphocyte receptor C reveals an unusual feature in its N-terminal capping module. PLoS ONE 9:e85875

Kasahara M, Sutoh Y (2014) Two forms of adaptive immunity in vertebrates: similarities and differences. Adv Immunol 122:59-90

Kasamatsu J, Sutoh Y, Fugo K, Otsuka N, Iwabuchi K, Kasahara M (2010) Identification of a third variable lymphocyte receptor in the lamprey. Proc Natl Acad Sci USA 107:14304-14308

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

Nagawa F, Kishishita N, Shimizu K, Hirose S, Miyoshi M, Nezu J, Nishimura T, Nishizumi H, Takahashi Y, Hashimoto S, Takeuchi M, Miyajima A, Takemori T, Otsuka AJ, Sakano H (2007) Antigen-receptor genes of the agnathan lamprey are assembled by a process involving copy choice. Nat Immunol 8:206-213

Pancer Z, Amemiya CT, Ehrhardt GR, Ceitlin J, Gartland GL, Cooper MD (2004) Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. Nature 430:174-180

Pancer Z, Saha NR, Kasamatsu J, Suzuki T, Amemiya CT, Kasahara M, Cooper MD (2005) Variable lymphocyte receptors in hagfish. Proc Natl Acad Sci USA 102:9224-9229

Rogozin IB, Iyer LM, Liang L, Glazko GV, Liston VG, Pavlov YI, Aravind L, Pancer Z (2007) Evolution and diversification of lamprey antigen receptors: evidence for involvement of an AID-APOBEC family cytosine deaminase. Nat Immunol 8:647-656

Smith JJ, Kuraku S, Holt C, Sauka-Spengler T, Jiang N, Campbell MS, Yandell MD, Manousaki T, Meyer A, Bloom OE, Morgan JR, Buxbaum JD, Sachidanandam R, Sims C, Garruss AS, Cook M, Krumlauf R, Wiedemann LM, Sower SA, Decatur WA, Hall JA, Amemiya CT, Saha NR, Buckley KM, Rast JP, Das S, Hirano M, McCurley N, Guo P, Rohner N,

Tabin CJ, Piccinelli P, Elgar G, Ruffier M, Aken BL, Searle SM, Muffato M, Pignatelli M, Herrero J, Jones M, Brown CT, Chung-Davidson YW, Nanlohy KG, Libants SV, Yeh CY, McCauley DW, Langeland JA, Pancer Z, Fritzsch B, de Jong PJ, Zhu B, Fulton LL, Theising B, Flicek P, Bronner ME, Warren WC, Clifton SW, Wilson RK, Li W (2013) Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. Nat Genet 45*:*415-421

Takaba H, Imai T, Miki S, Morishita Y, Miyashita A, Ishikawa N, Nishizumi H, Sakano H (2013) A major allogenic leukocyte antigen in the agnathan hagfish. Sci Rep 3*:*1716

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28*:*2731-2739

Figure legends

Fig. 1 Copy number distribution of LRRV modules in VLRB transcripts. **A.** The success rates yielding best approximations between the BD models and actual distribution were determined using a least squares algorithm as described in Materials and Methods. Success rates (horizontal axis) were plotted against correlation coefficients between the BD models and the actual distribution of LRRV modules (vertical axis). Blue squares and triangles indicate the actual distribution of LRRV modules in lamprey and hagfish VLRB transcripts, respectively. Regression curves (yellow and green dotted lines for lamprey and hagfish VLRB transcripts, respectively) were drawn to maximize the correlation coefficients between the BD model and actual distribution. For lamprey VLRB transcripts, the optimized success rate *x* was given by the following formula where *y* is a correlation coefficient: $y = -0.0022x^2 + 0.0683x + 0.4458$. The maximum coefficient of determination (R^2 value) for the regression formula was 0.9941. The success rate of 15.5% yielded the maximum correlation coefficient of 0.9759. For hagfish VLRB transcripts, the optimized success rate *x* was given by the following formula where *y* is a correlation coefficient: $y = -0.0022x^2 + 0.0985x - 0.1358$. The R^2 value for the regression formula was 0.9977. The success rate of 22.4% yielded the maximum correlation coefficient of 0.9667. **B.** BD models constructed using the success rates of 15.5% and 22.4% were compared with the actual distribution of LRRV modules in lamprey (blue squares) and hagfish (blue triangles) VLRB transcripts, respectively. Yellow and green dotted lines represent the copy number distribution of LRRV modules predicted by the BD model using the success rates of 15.5% and 22.4%, respectively.

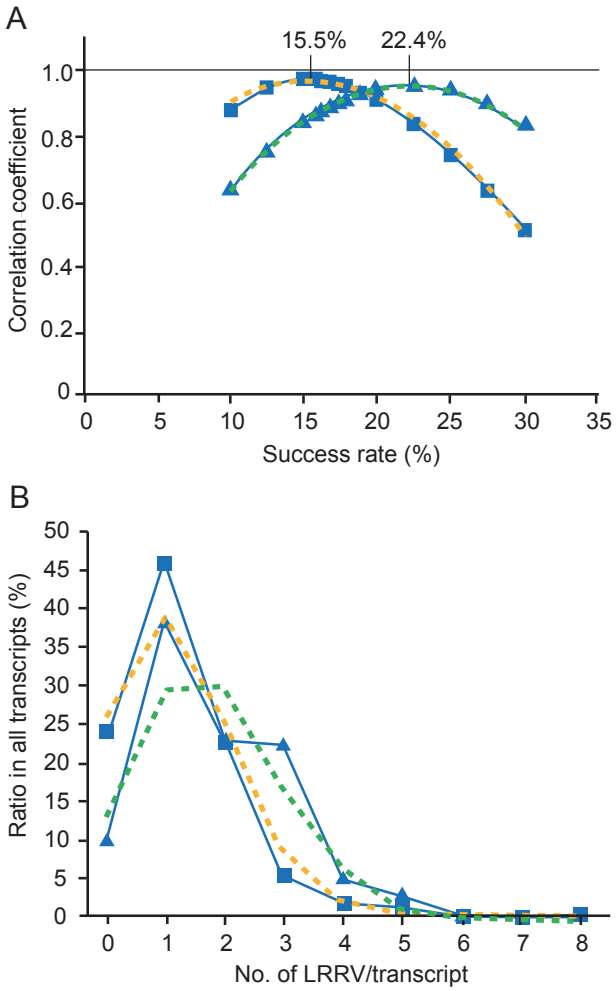
Fig. 2 Copy number distribution of LRRV modules in VLRA and VLRC transcripts. **A.** Copy number distribution of LRRV modules in VLRA (green), VLRB (blue), and VLRC (yellow) transcripts in lampreys (square) and hagfish (triangle). **B.** The actual distribution of LRRV modules in lamprey (square) and hagfish (triangle) T-lineage VLR transcripts is indicated in red. Yellow and green dotted lines represent the distribution of LRRV modules predicted by the BD models with the success rates of 15.5% and 22.4%, respectively. Squares and triangles indicate data for lampreys and hagfish, respectively. **C.** T-lineage VLR transcripts have a larger number of LRRV modules than VLRB transcripts. The copy number distribution of LRRV modules has higher kurtosis and skewness values in T-lineage transcripts than in VLRB transcripts. **D.** BD models were constructed on the assumption that transcripts with two or less LRRV modules are not allowed to occur in T-lineage VLR transcripts. The models were constructed using the success rates of 15.5% and 22.4% for lamprey and hagfish T-lineage transcripts, respectively. Symbols are as in Figure 2B.

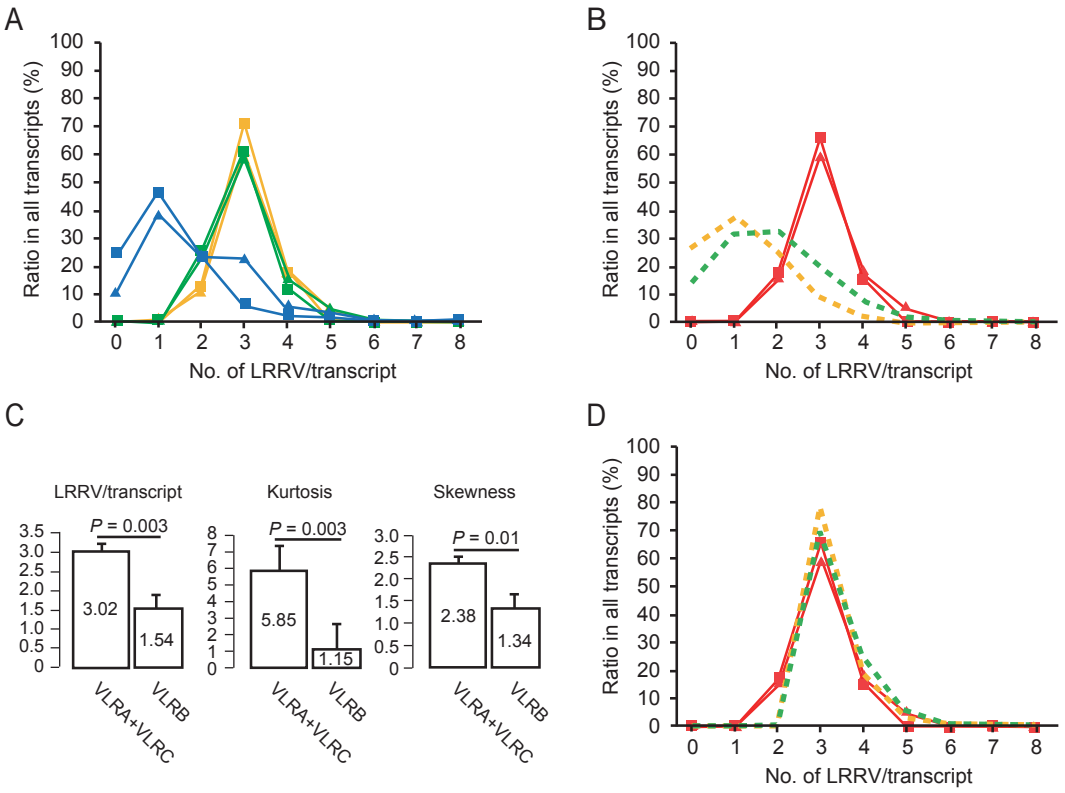
Fig. 3 The second LRRV module shows distinctive amino acid preferences in VLRA and VLRC molecules. **A.** Amino acid compositions at residues 4 (top), 6 (middle), and 16 (bottom) in LRRV modules. 1, 2, and 3 stand for the first, second, and third LRRV module, respectively. Amino acids are written in a single letter code and color-coded. Pm, *Petromyzon marinus* (sea lamprey); Es, *Eptatretus stoutii* (Pacific hagfish); Lj, *Lethenteron japonicum* (Japanese lamprey); and Eb, *Eptatretus burgeri* (inshore hagfish) **B.** Residues 4, 6, and 16 in the second LRRV module were mapped onto the VLRA crystal structure (PDB: 3M18) (Deng et al. 2010). The structure was displayed by Cuemol (http://www.cuemol.org/en/). Residues 4 and 6 were predicted to be on the antigen-binding surface, whereas residue 16 was predicted to be on the convex surface (red). β-

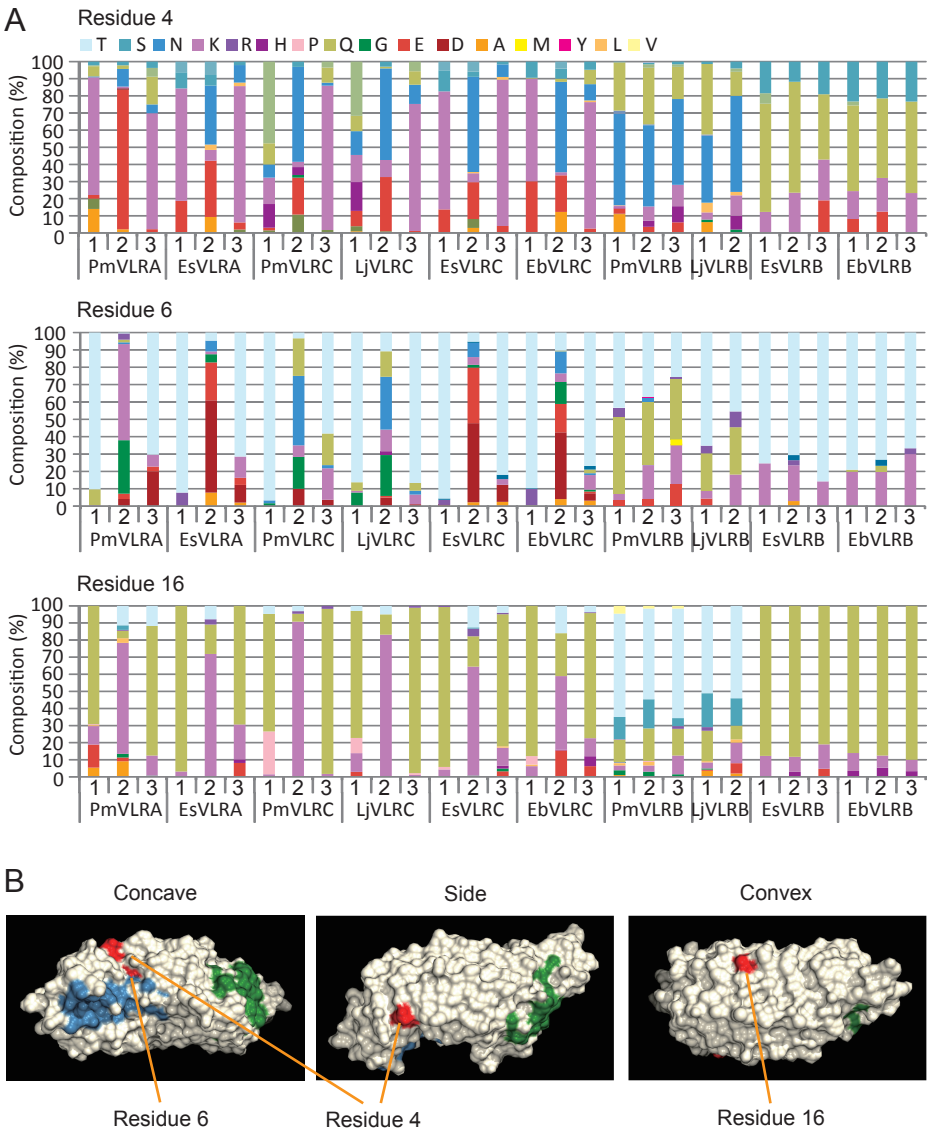
sheet and α-helix are indicated in blue and green, respectively. Mapping on the crystal structure of lamprey VLRC (Kanda et al. 2014) produced similar results (data not shown).

Supplementary Fig. 1

Consensus sequences were obtained for each LRRV module using the WebLogo 3 program (Crooks et al., 2004). The second LRRV module in VLRA and VLRC transcripts shows distinctive amino acid preferences (red arrowheads).


A
15.5%
22.4%
Correlation coefficient
1.0
0.8
0.6
0.4
0.2
0
0
5
10
15
20
25
30
35
Success rate (%)
B
Ratio in all transcripts (%)
50
45
40
35
30
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
No. of LRRV/transcript


A
Ratio in all transcripts (%)
No. of LRRV/transcript
B
Ratio in all transcripts (%)
No. of LRRV/transcript
C
LRRV/transcript
P = 0.003
3.02
1.54
VLRA+VLRC
VLRB
Kurtosis
P = 0.003
5.85
1.15
VLRA+VLRC
VLRB
Skewness
P = 0.01
2.38
1.34
VLRA+VLRC
VLRB
D
Ratio in all transcripts (%)
No. of LRRV/transcript


A
Residue 4
T S N K R H P Q G E D A M Y L V
Composition (%)
100 90 80 70 60 50 40 30 20 10 0
1 2 3 PmVLRA
1 2 3 EsVLRA
1 2 3 PmVLRC
1 2 3 LjVLRC
1 2 3 EsVLRC
1 2 3 EbVLRC
1 2 3 PmVLRB
1 2 LjVLRB
1 2 3 EsVLRB
1 2 3 EbVLRB
Residue 6
Residue 16
B
Concave
Side
Convex
Residue 6
Residue 4
Residue 16

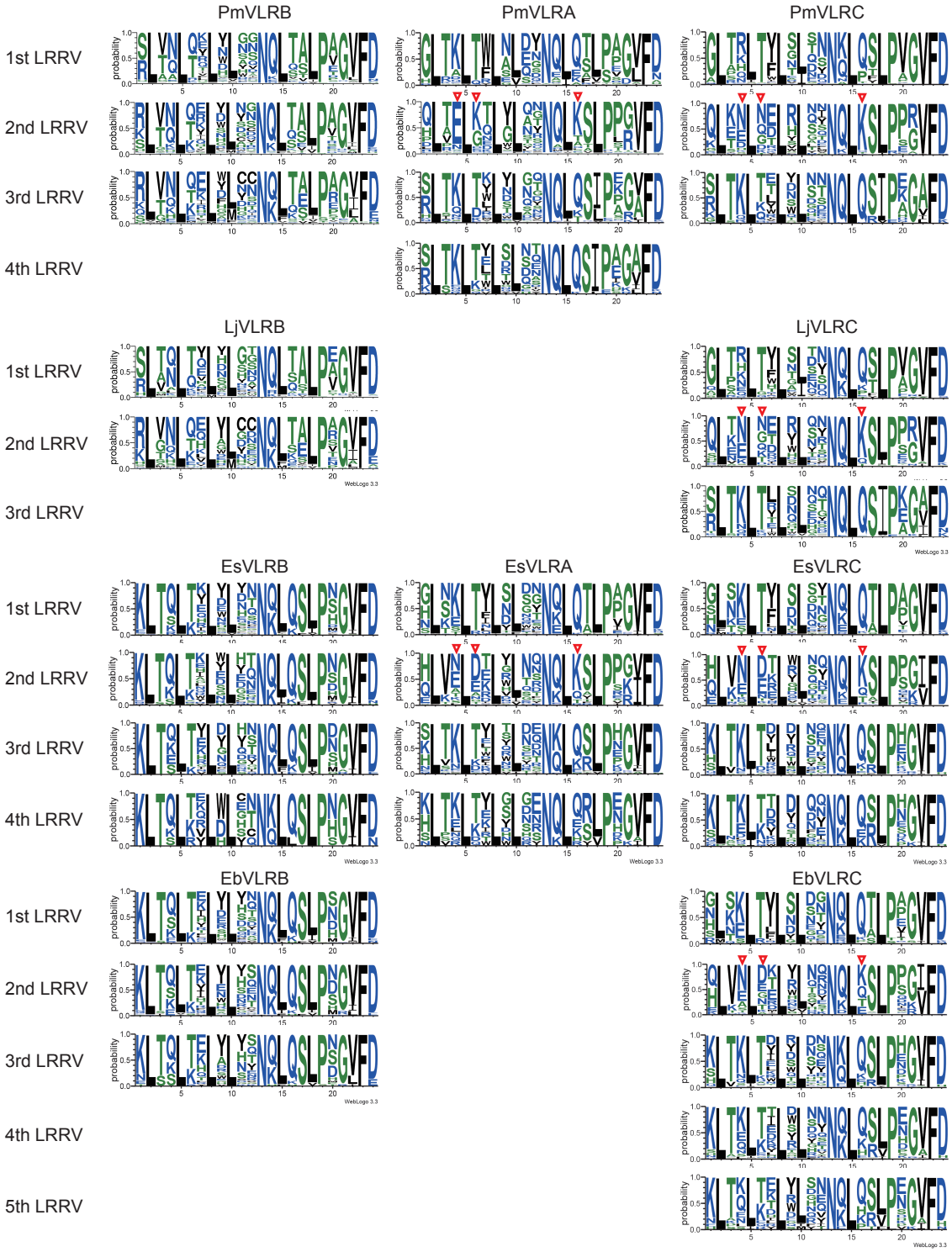

PmVLRB
PmVLRA
PmVLRC
1st LRRV
2nd LRRV
3rd LRRV
4th LRRV
LjVLRB
LjVLRC
1st LRRV
2nd LRRV
3rd LRRV
EsVLRB
EsVLRA
EsVLRC
1st LRRV
2nd LRRV
3rd LRRV
4th LRRV
EbVLRB
EbVLRC
1st LRRV
2nd LRRV
3rd LRRV
4th LRRV
5th LRRV
probability
WebLogo 3.3

Table 1. VLR sequences subjected to analysis

Species	Genes	Accession numbers of sequences retrieved from the database	Accession numbers of excluded sequences	Total number of sequences analyzed
Petromyzon marinus	*VLRA*	EF094629-EF094811, FJ794792-FJ794804	FJ794792-FJ794803, KF385955, KF385954	183
	VLRB	ABO15151-ABO15218, ABQ08670, ACT31435-ACT31445, AFU50752, AAT70240-AAT70356, ABA39891-ABA40278	AAT70311, AAT70317, ABO08646-ABO08669 ABO08671-ABO08695, ABA39234-ABA39237	591
	VLRC	KC244050-KC244109, KF385949-KF385954	KF385954	65
Lethenteron japonicum	*VLRB*	AB272387-AB272577	AB275668-AB275724, AB272526, AB272527	189
	VLRC	AB50727-AB507373		102
Eptatretus stoutii	*VLRA*	KF314046-KF314110	KF314110	64
	VLRB	AY965535-AY965612	AY965601, AY965590, AY965588, AY965572, AY965537	70
	VLRC	AY964783-AY964931	AY964824, AY964823, AY964919, AY964907, AY964882, AY964873, AY964810, AY964794	137
Eptatretus burgeri	*VLRB*	AB519991-AB520070, AY965520-AY965575		98
	VLRC	AB519822-AB519981, AY964719-AY964789	AY965678, AY964782, AY964721, AB519978, AB519884	224