

An evolutionary ancient mechanism for regulation of hemoglobin expression in vertebrate red cells

Masato Miyata^{1#&}, Nynke Gillemans^{1&}, Dorit Hockman^{2,3&}, Jeroen Demmers⁴, Jan-Fang Cheng⁵, Jun Hou^{6#}, Matti Salminen⁷, Christopher A. Fisher⁸, Stephen Taylor⁹, Richard J. Gibbons⁸, Jared J. Ganis¹⁰, Leonard I. Zon¹⁰, Frank Grosveld¹, Eskeatnaf Mulugeta¹, Tatjana Sauka-Spengler², Douglas R. Higgs^{8*} and Sjaak Philipsen^{1*}

¹ Erasmus MC, Department of Cell Biology, P.O. Box 2040, 3000 CA Rotterdam, NL

² Nuffield Department of Clinical Laboratory Sciences, Weatherall Institute of Molecular Medicine, Headington, Oxford OX3 9DS, UK

³ Division of Cell Biology, Faculty of Health Sciences, University of Cape Town, Cape Town, RSA

⁴ Erasmus MC, Department of Biochemistry, P.O. Box 2040, 3000 CA Rotterdam, NL

⁵ Genomics Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, MS 84-171, Berkeley, CA 94720, USA

⁶ Erasmus MC, Department of Gastroenterology, P.O. Box 2040, 3000 CA Rotterdam, NL

⁷ Natural Resources Institute Finland, Latokartanonkaari 9, P.O. Box 2, 00791 Helsinki, FI

⁸ Molecular Haematology Unit, Weatherall Institute of Molecular Medicine, Headington, Oxford OX3 9DS, UK

⁹ MRC WIMM Centre for Computational Biology, Weatherall Institute of Molecular Medicine, Headington, Oxford OX3 9DS, UK

¹⁰ Children's Hospital Boston, 1 Blackfan Cir., Karp 7, Boston, MA 02115, USA

& Equal contribution

Current affiliation:

Masato Miyata: Temasek Polytechnic, School of Applied Science, 21 Tampines Avenue 1, Singapore 529757

Jun Hou: Guangdong General Hospital, School of Medicine, South China University of Technology, Guangzhou, Guangdong 510006, China

* Correspondence: Sjaak Philipsen, Department of Cell Biology Ee1071b, Erasmus MC, P.O. Box 2040, 3000 CA Rotterdam, NL, tel: +31-10-7044282, email: j.philipsen@erasmusmc.nl, and Doug Higgs, The Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Headington, Oxford OX3 9DS, UK, tel. +44-1865-222393, email: doug.higgs@imm.ox.ac.uk.

Counts

| | |
|------------|------|
| Abstract | 216 |
| Main text | 4000 |
| Figures | 5 |
| Tables | 0 |
| References | 45 |

Key points

Many properties are shared between the lamprey and human *NPRL3*-linked *HB* locus, including a remote erythroid enhancer in intron 7 of *NPRL3*

Linkage of multiple *globin* genes to the same adjacent gene explains how hemoglobins could undergo convergent evolution in different species

Abstract

The oxygen-transport function of hemoglobin (HB) is thought to have arisen ~500 million years ago, roughly coinciding with the divergence between jawless (*Agnatha*) and jawed (*Gnathostomata*) vertebrates. Intriguingly, extant HBs of jawless and jawed vertebrates were shown to have evolved twice, and independently, from different ancestral globin proteins. This raises the question whether erythroid-specific expression of HB also evolved twice independently. In all jawed vertebrates studied to date, one of the *Hb* gene clusters is linked to the widely expressed *Nprl3* gene. Here we show that the *nprl3*-linked *hb* locus of a jawless vertebrate, the river lamprey (*Lampetra fluviatilis*), shares a range of structural and functional properties with the equivalent jawed vertebrate *Hb* locus. Functional analysis demonstrates that an erythroid-specific enhancer is located in intron 7 of lamprey *nprl3*, which corresponds to the *NPRL3* intron 7 *MCS-R1* enhancer of jawed vertebrates. Collectively, our findings signify the presence of an *nprl3*-linked multi-globin gene locus, which contained a remote enhancer driving globin expression in erythroid cells, prior to the divergence of jawless and jawed vertebrates. Different globin genes from this ancestral cluster evolved in the current *nprl3*-linked *hb* genes in jawless and jawed vertebrates. This provides a solution for the enigma of how, in different species, globin genes linked to the same adjacent gene could undergo convergent evolution.

216 words

Introduction

Hemoglobin (Hb) is responsible for the oxygen transport function of erythrocytes, and comprises more than 90% of the soluble protein in these cells. Each erythrocyte contains approximately 250×10^6 hemoglobin molecules. To attain these high numbers, expression of the *HB* genes is activated by powerful distal erythroid-specific enhancers¹. Given the importance of Hb in human physiology and its role in hemoglobinopathies such as α -thalassemia, β -thalassemia and sickle cell disease, the structure and evolutionary origin of HB loci and proteins have been intensively studied^{2,3}. Globin-related proteins are present in all phyla of life⁴. Since the first single-celled organisms were anaerobic, the original function of globins was most likely in detoxification, acting as oxygen scavengers and peroxidases/deoxygenases^{5,6}. As aerobic multicellular organisms evolved and increased in size, they became dependent on globins to provide an oxygen transport and storage system. In extant mammals, the monomeric myoglobin (MB) is still used as an oxygen storage protein in muscle⁷. The oxygen-transport function of HB is thought to have arisen ~500 million years ago⁸, roughly coinciding with the divergence between jawless (*Agnatha*) and jawed (*Gnathostomata*) vertebrates⁹. *Cyclostomata* (lampreys and hagfish) are extant representatives of jawless vertebrates and therefore characterization of these species may provide important insights into the evolutionary origins of vertebrate genomes, loci and proteins¹⁰. It has been widely accepted that there is a common origin of HBs from a proto-HB protein that evolved to become an oxygen transporter in the common ancestor of all vertebrates^{4,11}. However, by contrast, recent phylogenetic analyses concluded that HBs arose twice, and independently, from different ancestral globin proteins in jawless and jawed vertebrates^{8,12,13}. This raises the question whether erythroid-specific expression of the *HB* genes also evolved twice. In jawed vertebrates, one of the *HB* gene clusters is linked to the widely expressed *NPRL3* gene. When investigated, distal erythroid regulatory elements which activate the linked *HB* genes are invariably present in the introns of the *NPRL3* gene^{14,15}. The two strongest enhancers, called regulatory multispecies conserved sequences or *MCS-Rs*, are *MCS-R1* and *MCS-R2*¹⁶. To investigate the evolutionary origin of vertebrate *HB* loci further we isolated and sequenced two cosmids covering the *hb* cluster linked to the *nprl3* gene in the river lamprey (*L. fluviatilis*), a jawless vertebrate. Our analysis demonstrates that this *L. fluviatilis* *hb* locus shares an uncanny range of structural and functional properties with the jawed vertebrate *NPRL3*-linked *HB* locus. Chromatin accessibility mapping and functional analysis demonstrates that an erythroid-specific enhancer is located in intron 7 of lamprey *nprl3*, which corresponds to the *NPRL3* intron 7 *MCS-R1* enhancer of jawed vertebrates. We infer that **multiple globin** genes may have been linked to *NPRL3* before the divergence of jawless and jawed vertebrates, **explaining how, in**

different species, *globin* genes linked to the same adjacent gene could undergo convergent evolution.

Materials and Methods

Fish

Adult European river lampreys (*Lampetra fluviatilis* taxonomy ID: 7748) were freshly caught from the Kymijoki river in south-east Finland. **Tissues** were collected and immediately frozen. Lamprey larvae (ammocoetes) were preserved in 70% ethanol.

Migratory-phase adult sea lampreys (*Petromyzon marinus* taxonomy ID: 7757) were trapped in rivers in Michigan (USA) by the Great Lakes Fisheries Commission and shipped overnight from USGS Hammond Bay Biological Station. Sea lamprey ammocoetes were captured and shipped by Lamprey Services (Ludington, MI). All sea lamprey were imported in accordance with a detrimental species permit approved by the CA Dept. of Fish and Wildlife, and maintained in tanks in accordance with the Caltech Institutional Animal Care and Use Committee (IACUC) protocol #1436.

L. fluviatilis cosmid library

Arrayed filters of a *L. fluviatilis* cosmid library were obtained from the German Science Centre for Genome Research (RZPD, Berlin, DE). Isolation and characterization of cosmids containing *hb* genes is described in Supplemental Materials and Methods. Two partially overlapping cosmids (99E08 and 109N16) were selected for sequence analysis.

Sequencing and contig assembly

Cosmid DNA was sheared into ~2kb fragments which were used for shotgun cloning and Sanger sequencing. Paired-end sequences were used to construct contigs (see Supplemental Materials and Methods). Potential exons and genes were identified by GenScan¹⁷, these initial gene structures were further refined by manual curation. Predicted protein sequences were used to search the genome databases at NCBI and Ensembl with the BLAST family of search algorithms^{18,19}.

Long-range amplification of genomic DNA

Long-range PCR using genomic *L. fluviatilis* genomic DNA as a template was used to generate 5 overlapping amplicons. **Primers** and annealing temperatures are listed in Supplemental Table 2. **Amplifications** were performed using Prime STAR GXL DNA polymerase (TaKaRa Bio Inc, Shiga, JP). Between 1.0-1.5 µg of amplicons were pooled to sequence with the Nanopore 1D Amplicon Sequencing strategy for the MinION using kit SQK-LSK108 (Oxford Nanopore Technologies, Oxford, UK). Additional details **are described** in Supplemental Materials and Methods.

Nanopore sequencing

MinION sequencing was performed as per the manufacturer's guidelines using an R9 flow cell (FLO-MIN106). Details of alignment and contig assembly **are described** in Supplemental Materials and Methods

Comparative analysis

Assemblies of the *L. fluviatilis* *hb* locus based on Sanger and Nanopore sequencing data were aligned and visualized using PipMaker²⁰ with default settings.

Globin and NPRL3 protein sequences were retrieved from the NCBI and Ensembl databases. Protein sequences are listed in Supplemental Information. Multiple sequence alignments and molecular phylogenetic analyses are described in Supplemental Materials and Methods.

Genomic sequences of the pufferfish and human *NPRL3* genes were retrieved from Ensembl. LAGAN²¹ was used for multiple alignment of the river lamprey, pufferfish and human *NPRL3* genes, and VISTA²² for visualization of the results.

Analysis of HB and NPRL3 expression

RNA was isolated from adult *L. fluviatilis* brain and blood, and used for oligo-dT-primed cDNA synthesis. Reverse primers specific for each of the six *L. fluviatilis* *hb* genes were located in the 3'UTR and combined with a common forward primer in exon 2. Each reverse/forward primer combination yielded a cDNA amplicon of unique size. For NPRL3, primers were designed spanning exons 2-5 and 13-14 of the *L. fluviatilis* *npnl3* gene. Primer sequences, PCR conditions and sizes of PCR products are listed in Supplemental Information.

Proteomics of L. fluviatilis HB proteins

Whole blood and gills of adult river lampreys, and gills of ethanol-fixed larvae, were lysed in SDS-PAGE loading buffer and ~10 µg of protein was separated on 12.5% SDS-PAGE. The area in the 15-20 kD range was cut into 2-mm slices and used for analysis on an LTQ-Orbitrap mass spectrometer (ThermoFischer Scientific, Waltham, MA). Details are described in Supplemental Materials and Methods.

DNaseI hypersensitive site mapping

Nuclei were isolated from freshly frozen blood and liver samples obtained from adult river lampreys. Aliquots were treated with increasing amounts of DNaseI^{23,24}. DNA of the DNaseI treatment series was digested with NcoI or PvuII, size-fractionated on 0.7% agarose gels and subjected to Southern blot analysis. Other details are described in Supplemental Materials and Methods.

ATAC-seq

Blood was collected from adult *P. marinus* individuals and a 70 mm *P. marinus* ammocoete. Adult and ammocoete blood was pelleted by centrifugation and washed several times with PBS. Erythrocyte concentration was determined using a hemocytometer. ~50,000 cells were then processed for ATAC-seq²⁵. Details of ATAC-seq including bioinformatics are described in Supplemental Materials and Methods.

Vector construction for enhancer reporter assays

Either the zebrafish *MCS-R2 hb* enhancer, intron 5 of the *L. fluviatilis nprl3* gene or intron 7 of the *L. fluviatilis nprl3* gene were inserted into the HLC GFP reporter vector ²⁶, along with the *L. fluviatilis hb2* promoter. Constructs were sequenced to confirm they carried the correct inserts. Other details are described in Supplemental Materials and Methods.

Enhancer reporter assays in sea lamprey embryos

ISec-I meganuclease-mediated transgenesis was performed in *P. marinus* embryos as described previously ²⁶⁻²⁸ in Prof. Marianne Bronner's laboratory (California Institute of Technology, USA). Embryos were imaged with a Zeiss AxioCam MRm camera and AxioVision Rel 4.6 software (Zeiss, Oberkochen, DE). Movies were compiled using ImageJ. Other details are described in Supplemental Materials and Methods.

Enhancer reporter assays in zebrafish embryos

Conventional zebrafish transgenesis was conducted using the Tol2 transposase system ²⁹. Embryos were imaged with a Zeiss AxioCam MRm camera and AxioVision Rel 4.6 software. Movies were compiled using ImageJ.

Data sharing statement

The *nprl3*-linked *L. fluviatilis hb* locus **sequence** has been submitted to the NCBI nucleotide database (accession number MK495953). Oxford Nanopore sequencing data is available from RJG (richard.gibbons@imm.ox.ac.uk); proteomics data is available from JD (j.demmers@erasmusmc.nl). ATAC-seq data have been deposited in the European Nucleotide Archive (<https://www.ebi.ac.uk/ena>), accession number PRJEB31091.

Results

Isolation and ab-initio sequencing of a river lamprey hb locus

Arrayed filters of a *L. fluviatilis* cosmid library were screened with a 1.2 kb PCR fragment of *L. fluviatilis* genomic DNA. The PCR primers were designed by aligning larval/adult hemoglobin cDNA sequences of *Lampetra zanandreae*³⁰. Positive cosmids were subjected to restriction mapping and Southern blot analysis according to standard procedures³¹. Two partially overlapping cosmids (99E08 and 109N16) were selected for analysis by Sanger sequencing of ~2 kb fragments. Since the genome of *L. fluviatilis* is rich in repetitive sequences, it was challenging to assemble a consensus sequence. We therefore confirmed the assembly using long-read single molecule sequencing (Nanopore sequencing; Suppl. Fig. 1a). GenScan analysis¹⁷ of the final assembly identified ten potential protein-coding genes including six hemoglobin genes (*hb1* to *hb6*). Manual curation confirmed their typical three exon/two intron structures, with splice donor/acceptor sites at positions conserved in all vertebrate *HB* genes studied to date³, and consensus polyadenylation sites (5'-AATAAA-3') following the third exon. These six hemoglobin genes were flanked by an apparent orthologue of gnathostome *NPRL3* (Fig. 1a; analyzed in more detail below).

Genomic structure of the river lamprey nprl3-linked hb locus and analysis of gene expression

Our analysis of the genomic structure of the *L. fluviatilis* *nprl3*-linked *hb* locus revealed that, remarkably, all six hemoglobin genes were oriented in the same transcriptional direction (from left to right in Fig. 1a). Such an arrangement is similar to that found in the human *NPRL3*-linked *HBA* locus³. We found evidence for recent duplications leading to the *L. fluviatilis* *hb1*, *hb2* and *hb3* genes (Suppl. Fig. 1a,b; Suppl. Table 1). Using RT-PCR we observed erythroid-specific expression of the *hb5* and *hb6* genes in adult *L. fluviatilis* blood (Fig. 1b and Suppl. Fig. 1c); expression of *hb5* appeared to be the most abundant. In addition, expression of *hb1,2,3* and *4* was also detected at low levels. To further substantiate these observations, we isolated and analyzed proteins from fresh adult blood and gills, and from the gills of ethanol-fixed larvae (Suppl. Fig. 2). In the larval samples, we found peptides mapping to HB1 to HB4. Most peptides mapped to all four proteins, but some mapped specifically to HB1/HB4 or HB2/HB3 (Fig. 1c). In the adult blood samples, peptides uniquely mapping to HB5 were abundant; two peptides uniquely mapping to HB6 were also detected (Fig. 1c). Differential expression of *HB* genes during development has been universally observed in vertebrates³, including lampreys^{30,32,33}. We conclude that HB1-4 are larval and HB5-6 adult hemoglobins. Furthermore, the results show that all six *hb* genes in this locus are active. These data are largely consistent with a previous survey of hemoglobin mRNA expression at different developmental stages of the sea lamprey *Petromyzon marinus*³³; HB6 (aHB11 in sea lamprey, see Supplemental Information) was described as an embryonic hemoglobin in this previous study. Of note, expression of all six *hb* genes is detectable in

adult blood by sensitive RT-PCR assays (Fig. 1b), such assays may therefore detect expression of the *hb6* gene at embryonic stages as observed by Rohlfing et al.³³. Multiple sequence alignments of the six lamprey HB proteins and gnathostome globin proteins selected from bony fish (pufferfish and zebrafish), an amphibian (African clawed frog), birds (chicken and zebra finch) and mammals (human and mouse) were used to analyze their phylogenetic relationships. Consistent with previous analyses^{8,12,13} we observed that the six lamprey HB proteins **form a sister group** with gnathostome cytoglobin proteins (CYGB), which do not have an oxygen transport function³⁴, **separate from the clades with gnathostome HB and MB proteins** (Suppl. Fig. 3).

*Analysis of the *nprl3* gene in the river lamprey *hb* locus*

To obtain further insight into the evolutionary origin of extant jawed vertebrate *HB* loci, we turned our attention to the other genes identified by GenScan in the *nprl3*-linked *L. fluviatilis* *hb* locus (Fig. 1a). A large predicted peptide of 821 amino acids was derived from a retrotransposon-like element (*tr*), and was not further considered. Two small predicted peptides of 84 and 74 amino acids (*u1* and *u2*) did not bear resemblance to any currently known proteins; these were most likely false positives of the GenScan analysis. A predicted peptide of 632 amino acids was highly homologous to the *NPRL3* gene linked to the human *HBA* locus¹⁴. Of note, linkage of an *nprl3* homologue to *hb* genes has been reported for the sea lamprey and the arctic lamprey *Lethenteron camtschaticum*¹³, indicating that this a common feature of cyclostomes (see Supplemental Information). The river lamprey *nprl3* homologue is actively expressed, as revealed by RT-PCR using primer pairs spanning exons 2-5 and 13-14 (Fig. 2a,b). After manual curation, using human *NPRL3* as a reference, the river lamprey *nprl3* gene is predicted to encode a polypeptide of 581 amino acids which displays a remarkable homology (69% identity/80% similarity) to the human *NPRL3* protein. A phylogenetic tree of mammalian (human, mouse), bird (chicken, zebra finch), bony fish (pufferfish, zebrafish), cyclostome (river lamprey, sea lamprey) and insect (fruit fly) *NPRL3* further illustrates the orthologous relationships between the proteins (Fig. 2c). Alignment of pufferfish and human *NPRL3* genes to the *L. fluviatilis* *hb* locus showed that, with the exception of exons 12 and 13 which have merged into a single exon 12 in the *L. fluviatilis* gene, the exon-intron structures of all three *NPRL3* genes are identical (Fig. 2d). Finally, the river lamprey *nprl3* gene is located upstream of the larval *hb1* gene, and transcribed in the opposite direction to the *hb* genes (from right to left in Fig. 1a). This structural arrangement is common to the *NPRL3*-linked *HB* clusters in most jawed vertebrates studied to date^{1,3,14,15,23}.

*Chromatin accessibility of the lamprey *nprl3*-linked *hb* locus*

In jawed vertebrates, distal erythroid regulatory elements of the *HB* genes are invariably present in the introns of the linked *NPRL3* gene^{14,15}. Where tested, the two strongest enhancers are *MCS-R1* and *MCS-R2* which, in the mouse, contribute to ~40% and ~50% of

Hba expression respectively ¹⁶. *MCS-R1* is located in intron 7, and *MCS-R2* in intron 5, of the mouse *NPRL3* gene. Such elements display strong sensitivity to DNaseI digestion in erythroid cells ³⁵. We therefore initially used Southern blotting to map DNaseI hypersensitive sites (HSs) in the *L. fluviatilis nprl3* gene in liver and erythroid cells. Nuclei isolated from adult *L. fluviatilis* erythrocytes and liver were digested with increasing amounts of DNaseI. After purification the DNA samples were digested with NcoI or PvuII and subjected to Southern blot analysis. The locations of restriction sites and the probes used are shown in Suppl. Fig. 4a; owing to the virtually ubiquitous presence of simple and complex repeats in cyclostome genomes ¹⁰, it was not possible to design probes abutting the ends of the restriction fragments.

We found a DNaseI HS in intron 7, which was erythroid-specific since it was absent in liver nuclei (eHS, Suppl. Fig. 4). These data suggest that the eHS corresponds to jawed vertebrate *MCS-R1*. To investigate this further, we applied ATAC-seq ²⁵ to assess chromatin accessibility throughout the entire *hb* locus. These experiments were performed using erythrocytes isolated from sea lamprey larvae and adults. Consistent with the Southern blot analysis, we observed an ATAC site in intron 7 of *nprl3* which was present in samples of larval and adult origin (Fig. 3). In addition, the promoter regions of the *hb1-4* genes displayed an extensive open chromatin conformation in larval erythrocytes. In contrast, in adult erythrocytes the promoter area of the *hb5* gene was most highly accessible (Fig. 3c). In agreement with the RT-PCR and proteomics data (Fig. 1b,c), the accessibility of the *hb6* promoter area was unremarkable in both larval and adult erythrocytes, supporting the notion that *hb6* encodes a minor hemoglobin. Collectively, we conclude that the chromatin landscapes of the lamprey *hb* locus at these two developmental time points are remarkably similar to those observed for mammalian *HBA* loci ¹⁶.

Functional analysis of putative erythroid enhancers in transgenic lamprey embryos

To test the ability of these regions to act as erythroid-specific enhancers, intron 5 or 7 of the *L. fluviatilis nprl3* gene were linked to the *hb2* promoter and cloned into an HLC GFP reporter vector ²⁶. These constructs were used in I-SceI meganuclease-mediated transient transgenesis in the sea lamprey (Fig. 4a). Using the *L. fluviatilis nprl3* intron 7 element, we first noted erythroid-specific GFP expression in the circulation at 12 days post-fertilisation (dpf), which intensified by 17 dpf (Fig. 4b,c; Movie 1). We observed a similar level of GFP reporter expression in circulating blood cells using the reporter vector containing the zebrafish (*Danio rerio*) *MCS-R2 hb* enhancer ³⁶ (Fig. 4a,d; Movie 2). In contrast, we did not detect enhancer activity when intron 5 of the *L. fluviatilis nprl3* gene was used. Finally, a reporter vector using the standard *fos* promoter downstream of the *nprl3* intron 7 region yielded greatly reduced GFP expression in erythroid cells, **indicating that activation by the intron 7 enhancer is promoter-specific**. Thus, the DNaseI HS in intron 7 of the *L. fluviatilis*

np13 gene marks the location of an erythroid-specific enhancer that drives gene expression from the *hb2* promoter in a manner similar to the *MCS-R2 hb* enhancer of zebrafish. Inspection of the sequence of *L. fluviatilis np13* intron 7 revealed clustering of a number of potential binding sites for the well-known erythroid transcription factors GATA1 (GATA box), KLF1 (GC/GT box), NF-E2 (MARE) and TAL1 (E-box) (Suppl. Fig. 5). This is a hallmark of distal enhancers of *HB* gene activation in mammals^{1,3,14,15,23}. We conclude that this erythroid enhancer in river lamprey corresponds to the *MCS-R1* element present in jawed vertebrates.

Functional analysis of putative erythroid enhancers in transgenic zebrafish embryos

As the zebrafish *MCS-R2 hb* enhancer showed activity in the lamprey, we sought to determine if the reciprocal experiment would result in reporter activity in zebrafish erythroid cells. To investigate this, we used Tol2-mediated transient transgenesis of the HLC GFP reporter vectors. The zebrafish *MCS-R2 hb* enhancer drove GFP reporter expression from the *L. fluviatilis hb2* promoter in the blood islands at 30 hours post-fertilisation (hpf) and in circulating erythroid cells at 50 hpf (Movie 3). In contrast, we did not observe any reporter activity above background expression when intron 5 or intron 7 of the *L. fluviatilis np13* gene was used to drive GFP expression. These observations show that, while the zebrafish *MCS-R2* enhancer has retained properties that allow it to be recognized by the lamprey transcriptional machinery in a tissue-specific manner, the *L. fluviatilis MCS-R1* enhancer is not functional in the zebrafish.

Discussion

HB gene clusters have been studied intensively as models for tissue-specific, high level, developmentally regulated gene expression. The discovery and detailed molecular characterization of distal regulatory elements activating *HB* expression in erythroid cells have been instrumental for the development of current gene therapy vectors for the treatment of β -thalassemia and sickle cell disease patients³⁷⁻³⁹. While the evolutionary origin of the *HB* genes and proteins can be traced by multi-species sequence alignments followed by phylogenetic analyses^{3,4,8,13}, the distal regulatory elements often display poor sequence conservation even between relatively closely related species such as mouse and human^{1,14,15}. Identification of these elements therefore requires a different experimental approach which may include localization of clustered binding sites for specific transcription factors such as GATA1, KLF1, NF-E2 and TAL1, mapping of local chromatin properties such as histone modifications and DNaseI/ATAC hypersensitive sites, and functional analysis by linking putative distal regulatory elements to reporter genes in stable transgenesis assays¹. This approach has revealed the general molecular principles underlying developmentally regulated *HB* expression, addressing, for instance, the silencing mechanism of the fetal *HBG1/2* genes⁴⁰⁻⁴² and the interplay of multiple distal regulatory elements in *HBA1/2* gene activation¹⁶. Mammalian *HBA* genes are invariably linked to the ubiquitously expressed *NPRL3* gene. Major erythroid-specific distal regulatory elements are located in intron 5 and intron 7 of *NPRL3*. Bony fish also contain an *nprl3*-linked *hb* locus^{14,36} and in zebrafish, the presence of a distal regulatory element in *nprl3* intron 5 has been demonstrated by biochemical and transgenic analysis³⁶.

An ancient common evolutionary origin of NPRL3-linked HB loci

Based on comparative analysis of pufferfish and human globin loci, some of us previously proposed that the linkage of *HB* genes to *NPRL3* occurred after the diversification of the monomeric hemoglobins of jawless vertebrates into the tetrameric hemoglobins of jawed vertebrates²³. However, the striking structural and functional similarities between the *L. fluviatilis hb* and mammalian *HBA* loci reported here provide unambiguous support for an ancient, common evolutionary origin of *NPRL3*-linked *HB* loci. These similarities can be summarized as follows. Firstly, the lamprey *hb* locus contains multiple *hb* genes all oriented in the same transcriptional direction and arranged in order of developmental expression. Secondly, the lamprey *nprl3* gene is located upstream, and in the opposite transcriptional direction, to the larval *hb* genes. We note that, as in the human genome, the lamprey genome contains only one *nprl3* gene. Thirdly, the chromatin accessibility landscape of the lamprey *hb* locus displays an ATAC site at *nprl3* intron 7 in larval and adult erythroid cells. In contrast, the *hb* gene promoters display ATAC sites only when the genes are active, i.e. the larval genes in larval cells and the adult *hb5* gene in adult cells. Finally, functional analysis

shows that the lamprey *nprl3* intron 7 ATAC site corresponds to the mammalian MCS-R1 erythroid distal enhancer element, located in *NPRL3* intron 7. Thus, our data strongly support an ancient common evolutionary origin of *NPRL3*-linked *HB* loci in jawless and jawed vertebrates. Consequently, our previous model in which linkage of *HB* genes to *NPRL3* was proposed to occur after the diversification of jawless and jawed vertebrates²³ needs to be redrawn.

The evolutionary origin of the human HBA locus

An updated model for the evolutionary origin of the human *HBA* locus, taking these and other recent observations¹³ into account, is presented in Fig. 5. Of note, in *Tunicata* and *Cephalochordata*, which represent *Chordata* more primitive than *Vertebrata*, no linkage between *nprl3* and *globin* genes is observed^{43,44}. The globins do not have respiratory functions in these organisms. In the sea squirt *Ciona intestinalis*, a tunicate, *globin* genes are linked to *mpg* (on chromosome 3) and *rhbdf1* (on chromosome 1) (Fig. 5; ⁴⁴). *MPG* and *RHBDF1* are hallmark genes of jawed vertebrate *NPRL3*-linked *HB* loci¹. This suggests that the coupling between *MPG*, *RHBDF1*, *NPRL3* and multiple *globin* genes was first established in the vertebrate lineage (Fig. 5), providing a platform to develop erythroid-specific expression via the enhancers located in *NPRL3* introns, and oxygen transport via adaptations of the *globin* proteins⁴⁴. Despite this common evolutionary origin, comprehensive phylogenetic analysis of *globin* proteins supports an independent origin of the oxygen transport function of hemoglobins in extant jawless and jawed vertebrates^{8,13}. Our model provides a solution for this enigma. We propose that the different *globin* genes present in the ancestral *NPRL3*-linked multi-*globin* gene locus acquired oxygen transport functionality and were recruited by the *NPRL3*-linked enhancer to drive erythroid-specific expression (Fig. 5). Jawless and jawed vertebrates kept different *HB* genes from this ancestral locus, resulting in the *NPRL3*-linked *HB* loci of extant representatives of these two vertebrate classes.

Probing the origin of multi-gene loci

In conclusion, we show that comparison of the genomic environment in jawless and jawed vertebrates, including determination of tissue-specific chromatin accessibility, functional characterization of distal *cis*-acting elements, and analysis of developmentally regulated gene expression, provides critical information about the evolutionary origin of multi-gene loci.

Acknowledgments

This work was supported by the Netherlands Genomics Initiative (NGI), the Landsteiner Foundation for Blood Transfusion Research (1040 and 1627), and the Netherlands Scientific Organization ZonMw (DN 82-301, 912-07-019 and 40-00812-98-12128). DH was supported by an EMBO Short Term Fellowship (ASTF 337-2014). The nanopore sequencing was undertaken as part of the MinION Early Access Program; the MinION sequencing apparatus and flow cells were supplied by Oxford Nanopore Technologies without charge. Research in the laboratory of DRH was supported by the Medical Research Council (UK). We thank Jim Hughes and Maria Suciu for help with the initial chromatin accessibility experiments. **We would like to thank the reviewers for their thoughtful comments.**

Author contributions

Designed experiments: MM, DH, RJG, LIZ, FG, TS-S, DRH, SP

Performed experiments: MM, NG, DH, J-FC, EM, MS, AAF, JJG

Analyzed data: MM, NG, DH, JD, J-FC, JH, ST, EM, DRH, SP

Interpreted results: MM, NG, DH, RJG, LIZ, FG, EM, TS-S, DRH, SP

Wrote the manuscript: MM, DH, EM, DRH, SP

Reviewed the manuscript: all authors

Competing interests

None to declare

References

- Philipsen S, Hardison RC. Evolution of hemoglobin loci and their regulatory elements. *Blood Cells Mol Dis*. 2018;70:2-12.
- Gell DA. Structure and function of haemoglobins. *Blood Cells Mol Dis*. 2018;70:13-42.
- Hardison RC. Evolution of hemoglobin and its genes. *Cold Spring Harb Perspect Med*. 2012;2(12):a011627.
- Hardison RC. A brief history of hemoglobins: plant, animal, protist, and bacteria. *Proc Natl Acad Sci U S A*. 1996;93(12):5675-5679.
- Gardner PR, Gardner AM, Martin LA, Salzman AL. Nitric oxide dioxygenase: an enzymic function for flavohemoglobin. *Proc Natl Acad Sci U S A*. 1998;95(18):10378-10383.
- Minning DM, Gow AJ, Bonaventura J, et al. Ascaris haemoglobin is a nitric oxide-activated 'deoxygenase'. *Nature*. 1999;401(6752):497-502.
- Gros G, Wittenberg BA, Jue T. Myoglobin's old and new clothes: from molecular structure to function in living cells. *J Exp Biol*. 2010;213(Pt 16):2713-2725.
- Hoffmann FG, Opazo JC, Storz JF. Gene cooption and convergent evolution of oxygen transport hemoglobins in jawed and jawless vertebrates. *Proc Natl Acad Sci U S A*. 2010;107(32):14274-14279.
- Blair JE, Hedges SB. Molecular phylogeny and divergence times of deuterostome animals. *Mol Biol Evol*. 2005;22(11):2275-2284.
- Smith JJ, Kuraku S, Holt C, et al. Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. *Nat Genet*. 2013;45(4):415-421.
- Goodman M, Pedwaydon J, Czelusniak J, et al. An evolutionary tree for invertebrate globin sequences. *J Mol Evol*. 1988;27(3):236-249.
- Katoh K, Miyata T. Cyclostome hemoglobins are possibly paralogous to gnathostome hemoglobins. *J Mol Evol*. 2002;55(2):246-249.
- Schwarze K, Campbell KL, Hankeln T, Storz JF, Hoffmann FG, Burmester T. The globin gene repertoire of lampreys: convergent evolution of hemoglobin and myoglobin in jawed and jawless vertebrates. *Mol Biol Evol*. 2014;31(10):2708-2721.
- Flint J, Tufarelli C, Peden J, et al. Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the alpha globin cluster. *Hum Mol Genet*. 2001;10(4):371-382.
- Hughes JR, Cheng JF, Ventress N, et al. Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. *Proc Natl Acad Sci U S A*. 2005;102(28):9830-9835.
- Hay D, Hughes JR, Babbs C, et al. Genetic dissection of the alpha-globin super-enhancer in vivo. *Nat Genet*. 2016;48(8):895-903.
- Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*. 1997;268(1):78-94.
- Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25(17):3389-3402.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403-410.
- Elnitski L, Riemer C, Schwartz S, Hardison R, Miller W. PipMaker: a World Wide Web server for genomic sequence alignments. *Curr Protoc Bioinformatics*. 2003;Chapter 10:Unit 10 12.
- Brudno M, Do CB, Cooper GM, et al. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*. 2003;13(4):721-731.
- Mayor C, Brudno M, Schwartz JR, et al. VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*. 2000;16(11):1046-1047.
- Gillemans N, McMorrow T, Tewari R, et al. Functional and comparative analysis of globin loci in pufferfish and humans. *Blood*. 2003;101(7):2842-2849.
- Ellis J, Tan-Un KC, Harper A, et al. A dominant chromatin-opening activity in 5' hypersensitive site 3 of the human beta-globin locus control region. *EMBO J*. 1996;15(3):562-568.
- Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol*. 2015;109:21 29 21-29.
- Parker HJ, Bronner ME, Krumlauf R. A Hox regulatory network of hindbrain segmentation is conserved to the base of vertebrates. *Nature*. 2014;514(7523):490-493.

27. Hockman D, Chong-Morrison V, Green SA, et al. A genome-wide assessment of the ancestral neural crest gene regulatory network. *Nat Commun.* 2019;10(1):4689.
28. Parker HJ, Sauka-Spengler T, Bronner M, Elgar G. A reporter assay in lamprey embryos reveals both functional conservation and elaboration of vertebrate enhancers. *PLoS One.* 2014;9(1):e85492.
29. Kawakami K. Transgenesis and gene trap methods in zebrafish by using the Tol2 transposable element. *Methods Cell Biol.* 2004;77:201-222.
30. Lanfranchi G, Pallavicini A, Laveder P, Valle G. Ancestral hemoglobin switching in lampreys. *Dev Biol.* 1994;164(2):402-408.
31. Sambrook J, MacCallum P, Russell D. Molecular Cloning: A Laboratory Manual. 2001(third edition).
32. Lanfranchi G, Odorizzi S, Laveder P, Valle G. Different globin messenger RNAs are present before and after the metamorphosis in *Lampetra zanandreaei*. *Dev Biol.* 1991;145(2):367-373.
33. Rohlfing K, Stuhlmann F, Docker MF, Burmester T. Convergent evolution of hemoglobin switching in jawed and jawless vertebrates. *BMC Evol Biol.* 2016;16:30.
34. Liu X, El-Mahdy MA, Boslett J, et al. Cytoglobin regulates blood pressure and vascular tone through nitric oxide metabolism in the vascular wall. *Nat Commun.* 2017;8:14807.
35. Higgs DR, Wood WG. Long-range regulation of alpha globin gene expression during erythropoiesis. *Curr Opin Hematol.* 2008;15(3):176-183.
36. Ganis JJ, Hsia N, Trompouki E, et al. Zebrafish globin switching occurs in two developmental stages and is controlled by the LCR. *Dev Biol.* 2012;366(2):185-194.
37. Ribeil JA, Hacein-Bey-Abina S, Payen E, et al. Gene Therapy in a Patient with Sick Cell Disease. *N Engl J Med.* 2017;376(9):848-855.
38. Thompson AA, Walters MC, Kwiatkowski J, et al. Gene Therapy in Patients with Transfusion-Dependent beta-Thalassemia. *N Engl J Med.* 2018;378(16):1479-1493.
39. Marktel S, Scaramuzza S, Cicalese MP, et al. Intrabone hematopoietic stem cell gene therapy for adult and pediatric patients affected by transfusion-dependent ss-thalassemia. *Nat Med.* 2019;25(2):234-241.
40. Masuda T, Wang X, Maeda M, et al. Transcription factors LRF and BCL11A independently repress expression of fetal hemoglobin. *Science.* 2016;351(6270):285-289.
41. Liu N, Hargreaves VV, Zhu Q, et al. Direct Promoter Repression by BCL11A Controls the Fetal to Adult Hemoglobin Switch. *Cell.* 2018;173(2):430-442 e417.
42. Martyn GE, Wienert B, Yang L, et al. Natural regulatory mutations elevate the fetal globin gene via disruption of BCL11A or ZBTB7A binding. *Nat Genet.* 2018;50(4):498-503.
43. Ebner B, Panopoulou G, Vinogradov SN, et al. The globin gene family of the cephalochordate amphioxus: implications for chordate globin evolution. *BMC Evol Biol.* 2010;10:370.
44. Wetten OF, Nederbragt AJ, Wilson RC, Jakobsen KS, Edvardsen RB, Andersen O. Genomic organization and gene expression of the multiple globins in Atlantic cod: conservation of globin-flanking genes in chordates infers the origin of the vertebrate globin clusters. *BMC Evol Biol.* 2010;10:315.
45. Burmester T, Ebner B, Weich B, Hankeln T. Cytoglobin: a novel globin type ubiquitously expressed in vertebrate tissues. *Mol Biol Evol.* 2002;19(4):416-421.

Figure legends

Figure 1. Analysis of the *hb* genes in the *nprl3*-linked *L. fluviatilis* *hb* locus

a) Schematic drawing (to scale) of the *L. fluviatilis* *hb* locus. Exons of *hb* genes are shown as red boxes. To indicate the direction of transcription, gene names are positioned next to the first exon. *tr* = transposon; *u1* and *u2* are predicted genes. **b)** Expression of the *L. fluviatilis* *hb* genes assessed by RT-PCR. Red asterisks: fragments of expected size for cDNA amplicons; blue asterisks: fragments of expected size for genomic amplicons. **c)** Proteomic analysis of *L. fluviatilis* HB proteins in larvae and adults. Peptides identified by mass spectrometry are indicated by colored bars. Light blue: peptides unique to HB1-4; purple: peptides unique to HB1 and HB4; pink: peptides unique to HB2 and HB3; red: peptides unique to HB5; orange: peptides unique to HB6.

Figure 2. Analysis of the *L. fluviatilis* *nprl3* gene

a) Structure of the *L. fluviatilis* *nprl3* gene. The predicted exons (green: untranslated regions; purple: coding regions) are numbered. Primers used for RT-PCR are shown, with expected fragment sizes in base pairs (genomic/cDNA). **b)** Expression of *L. fluviatilis* NPRL3 mRNA assessed by RT-PCR, using primers shown in **a)**, amplifying exons 2-5 or 13-14. Purple asterisks: fragments of expected size for cDNA amplicons; blue asterisk: fragment of expected size for genomic amplicon. **c)** Phylogenetic relationship of fruit fly (*Drosophila melanogaster* (*Dm*)), river lamprey (*Lampetra fluviatilis* (*Lf*)), sea lamprey (*Petromyzon marinus* (*Pm*)), pufferfish (*Tetraodon nigroviridis* (*Tn*)), zebrafish (*Danio rerio* (*Dn*)), chicken (*Gallus gallus* (*Gg*)), zebra finch (*Taeniopygia guttata* (*Tg*)), human (*Homo sapiens* (*Hs*)) and mouse (*Mus musculus* (*Mm*)) NPRL3 proteins inferred by using the Maximum Likelihood method. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Size of the proteins is indicated (aa). **d)** VISTA plot displaying alignments of the *L. fluviatilis*, *T. nigroviridis* and *H. sapiens* NPRL3 genes. Note that the first (non-coding) exon of *L. fluviatilis* *nprl3* is not included in the drawing.

Figure 3. Chromatin accessibility of the *nprl3*-linked *hb* locus mapped by ATAC-seq

a) Schematic drawing of the *nprl3*-linked *hb* locus. Intron 7 of the *nprl3* gene is marked by a red arrow. Other details are as in Fig. 1a. **b)** ATAC-seq analysis of larval (orange) and adult (red) lamprey blood. Light blue shading indicates areas enlarged in **c)**.

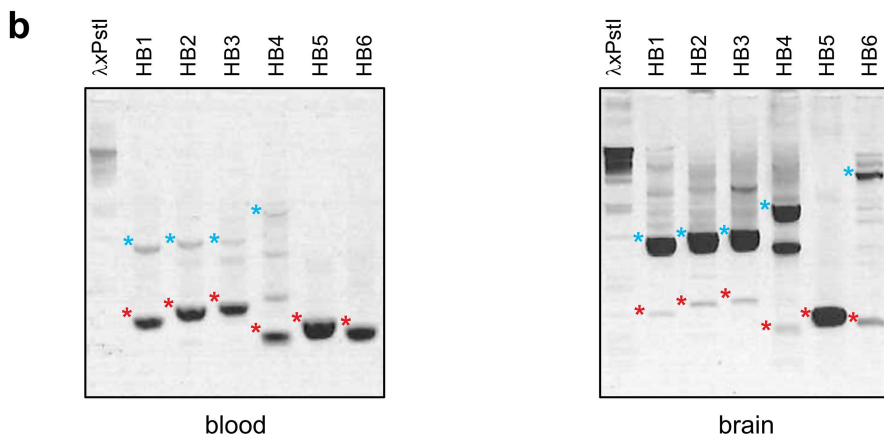
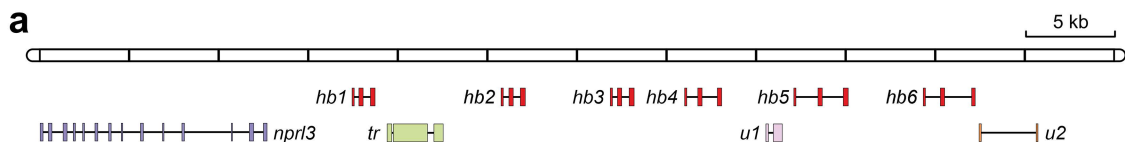
Figure 4. Erythroid-specific enhancer activity of intron7 of *L. fluviatilis* *npnl3*

a) Diagram of the *L. fluviatilis* *npnl3* gene, and GFP reporter vectors used in transgenic enhancer reporter assays in *P. marinus*. **b)** Diagram of a 17 dpf *P. marinus* embryo showing the circulatory system (red-dashed lines) of the head and branchial arches. Dashed box indicates the region shown in **c** and **d**. **c)** Still from Movie 1 showing erythroid-specific GFP reporter expression in circulation when intron7 of *L. fluviatilis* *npnl3* is used to drive GFP expression from the *L. fluviatilis* *hb2* promoter. **d)** Still from Movie 2 showing erythroid-specific GFP reporter expression in circulation when the zebrafish *MCS-R2* *hb* enhancer is used to drive GFP expression from the *L. fluviatilis* *hb2* promoter. Dotted white line in **c** and **d** outlines the embryo. ba, branchial arches; h, heart; m, mouth.

Figure 5. Model for the evolutionary origin of the human *HBA* locus

Based on the results reported in this paper, our previous model for the evolutionary origin of the human *HBA* locus²³ has been revised. Genes are color-coded and indicated by gene symbols or names at first appearance. Shaded grey bars indicate when invertebrate chordates, Agnathans, Gnathostomes (fish, mammals) first appeared in geologic time. Extant representative species are indicated on the right (Sea squirt, Lamprey, Pufferfish, Platypus and Human). Loci from these species are shown on a cyan background; **inferred loci are shown on a grey background**. Green lines indicate the trajectory of evolution of the human *HBA* locus. Distal erythroid enhancers in introns of the *NPRL3* gene are indicated by a red circle. Gene symbols: ***HB* hemoglobin (encoding a globin with oxygen transport function)**; MPG N-methylpurine DNA glycosylase; *NPRL3* nitrogen permease regulator-like 3, GATOR1 complex subunit; *RHBDF1* rhomboid 5 homolog 1. The evolutionary time scale is based on

45.



c

larval

>HB1|153 aa

MPIVDSGSVGAISAAEKSLIVSAWAPVYAKYEEAGVDILVKFFAANPEAQAFFPKFKGLDSADQLKKSPAVRWHAER

IINAVNDAVVALDDPAKQSLQLKALSQKHAHELNVDPNYFKVLAGVISDAVVKSGDAKAAYDKFLSQVVILLKSAY

>HB2|153 aa

MPIVDSGSVGAISAAEKSLIVSAWAPVYAKYEEAGVDILVKFFAANPEAQAFFPKFKGLDSADQLKKSPAVRWHAER

IINAVNDAVVALDDPAKQSLQLKALSQKHAHELNVDPKYFKVLAGVISDAVVKSGDAKAAYDKFLSQVVILLKSAY

>HB3|153 aa

MPIVDSGSVGAISAAEKSLIVSAWAPVYAKYEEAGVDILVKFFAANPEAQAFFPKFKGLDSADQLKKSPAVRWHAER

IINAVNDAVVALDDPAKQSLQLKALSQKHAHELNVDPKYFKVLAGVISDAVVKSGDAKAAYDKFLSQVVILLKSAY

>HB4|153 aa

MPIVDSGSVGALSAAEKSLVSAWAPVYAKYEEAGVDILVKFFSDNPGVQDFFPKFKGLDSADQLKKSPAVRWHAER

IINAVNDAVVALDEPAKLSLKLKGLSKKHAQELNVDPQYFKVLAGVISDAVVKSGDAKAAYDKFLSQVVILLKFAY

adult

>HB5|150 aa

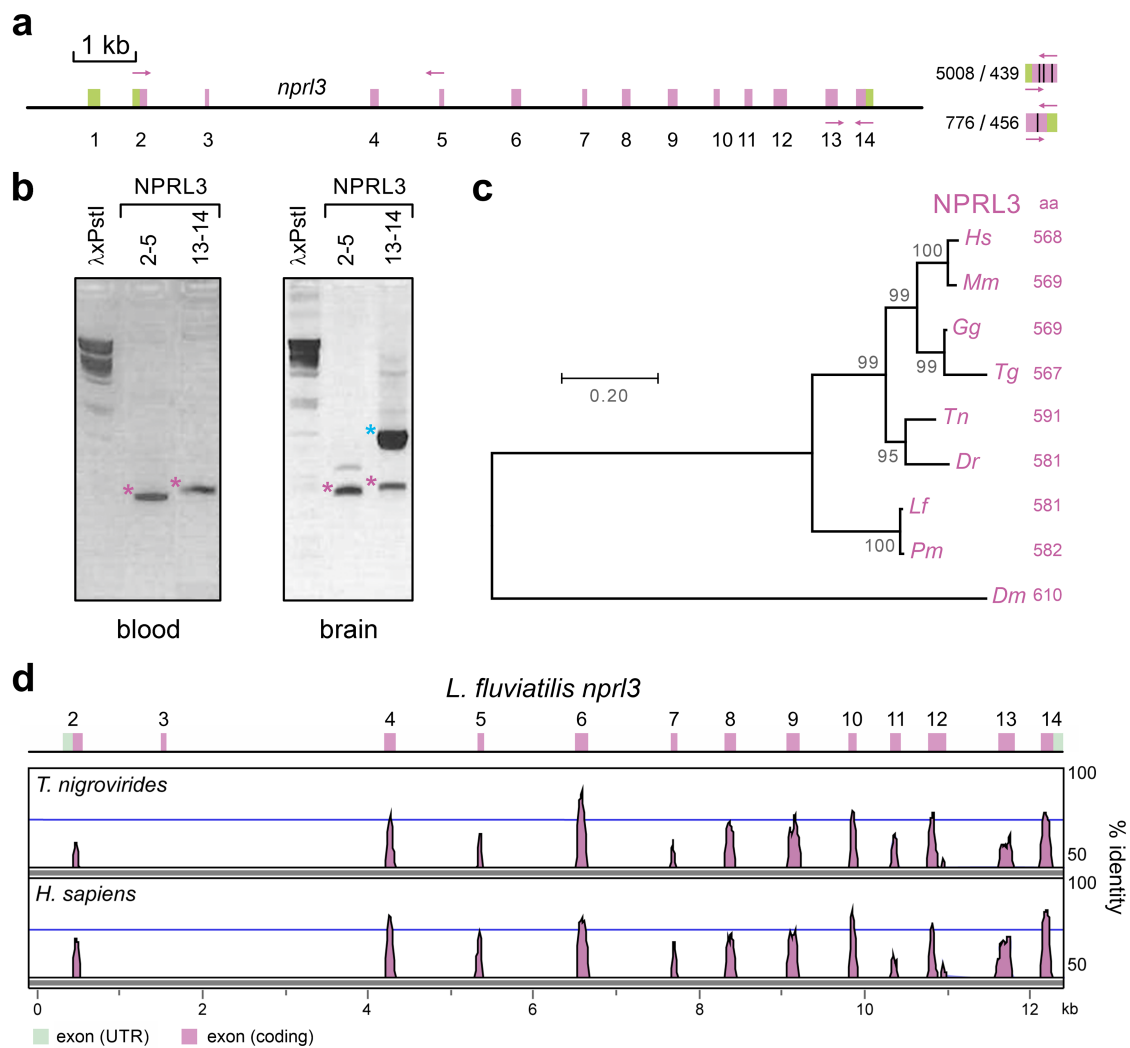
MPIVDSGSVPALTAEEKATIRTAWAPVYAKYQSTGVDILIKFFTSNPAAQEFFPKFKGLTSADQLKKSMQVVRWHAER

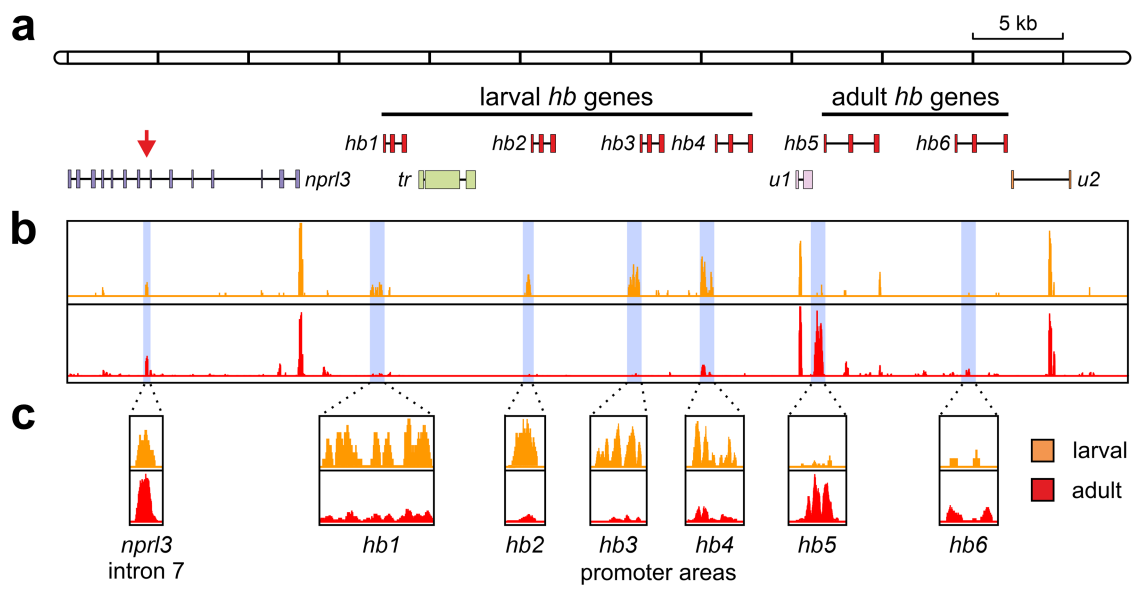
IINAVNDAVVAAMDTERMSLKLNELSSKHAQSFQVDPQYFKVLAIVVDTVLPDAGLEKLSMICILLRSSY

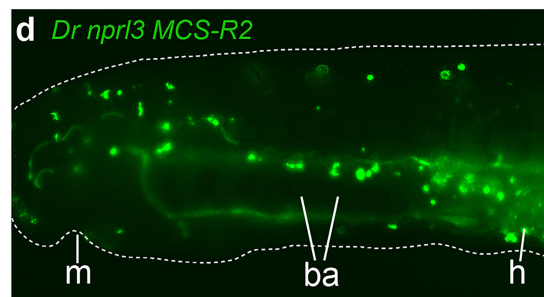
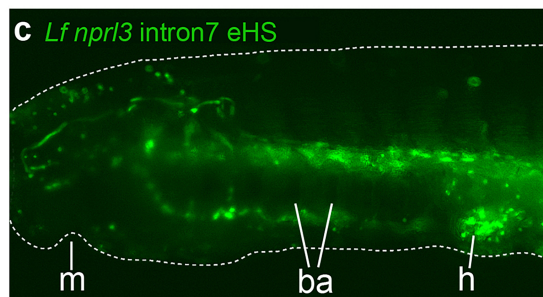
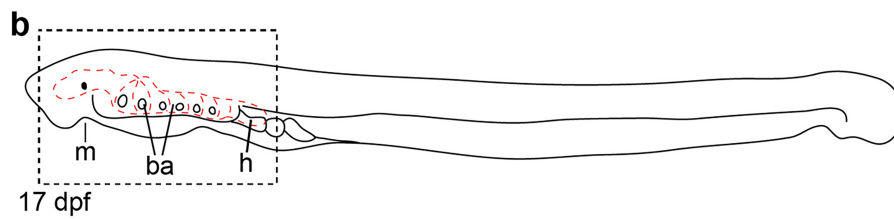
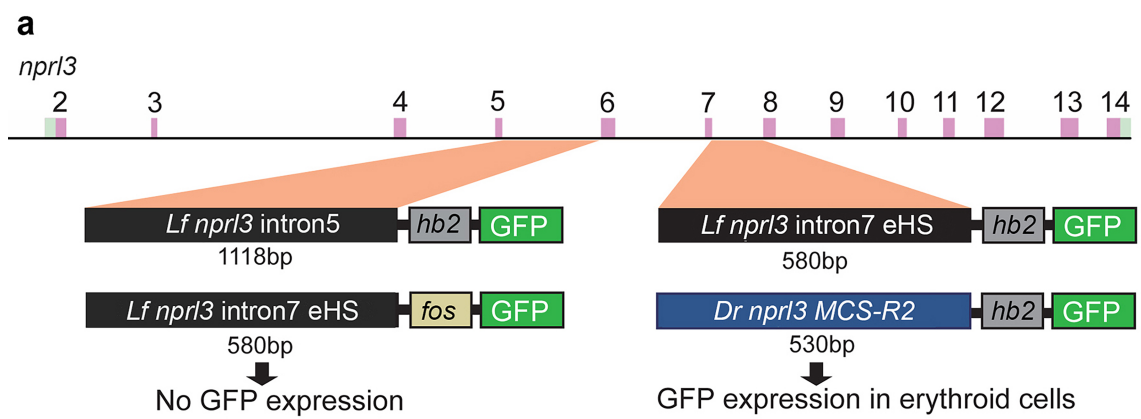
>HB6|158 aa

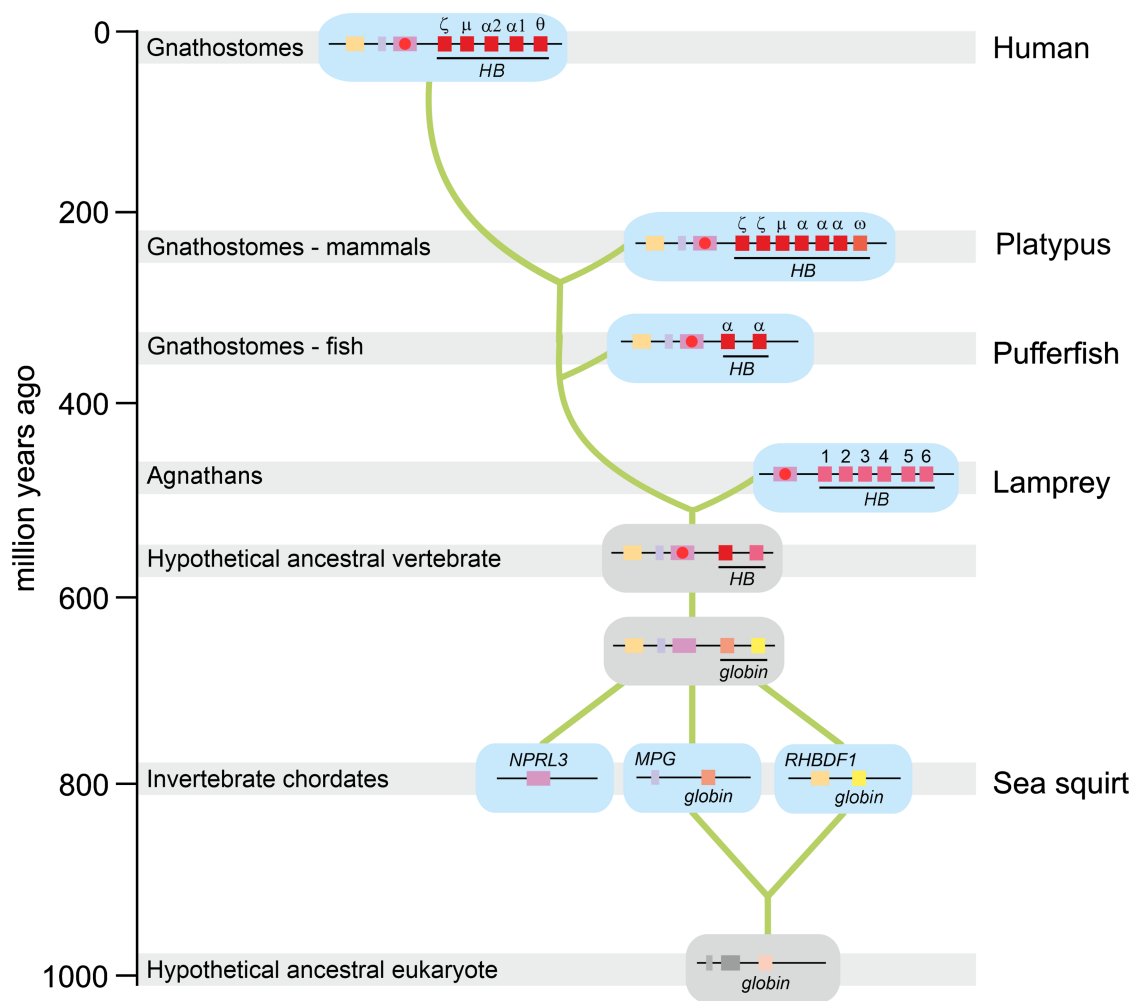
MPIVDSGSVGALSAAEKAIIDSWKVYVYADYEAAGKAILIKFFTSNAGVQDFFPKFKGLDSADQLSKSAAVRWHAER

IINAVNDAVVALDDPEKLSLKLKALSQKHAQEFNVDPQYFKVLSANVLEQVAAANGGLSAAEQGAWEKLLSIISILLKSQY









An evolutionary ancient mechanism for regulation of hemoglobin expression in vertebrate red cells

Masato Miyata^{1#&}, Nynke Gillemans^{1&}, Dorit Hockman^{2,3&}, Jeroen Demmers⁴, Jan-Fang Cheng⁵, Jun Hou^{6#}, Matti Salminen⁷, Christopher A. Fisher⁸, Stephen Taylor⁹, Richard J. Gibbons⁸, Jared J. Ganis¹⁰, Leonard I. Zon¹⁰, Frank Grosveld¹, Eskeatnaf Mulugeta¹, Tatjana Sauka-Spengler², Douglas R. Higgs^{8*} and Sjaak Philipsen^{1*}

Supplemental Information

Supplemental Material and Methods

Fish

Adult European river lampreys (*Lampetra fluviatilis* taxonomy ID: 7748) were freshly caught from the Kymijoki river in south-east Finland. Blood, gills, and livers were collected and immediately frozen. Lamprey larvae (ammocoetes) were preserved in 70% ethanol.

Migratory-phase adult sea lampreys (*Petromyzon marinus* taxonomy ID: 7757) were trapped in rivers in Michigan (USA) by the Great Lakes Fisheries Commission and shipped overnight from USGS Hammond Bay Biological Station. Sea lamprey ammocoetes were captured and shipped by Lamprey Services (Ludington, MI). All sea lamprey were imported in accordance with a detrimental species permit approved by the CA Dept. of Fish and Wildlife, and maintained in tanks in accordance with the Caltech Institutional Animal Care and Use Committee (IACUC) protocol #1436.

L. fluviatilis cosmid library

Arrayed filters of a *L. fluviatilis* cosmid library were obtained from the German Science Centre for Genome Research (RZPD), Berlin, Germany. To isolate cosmids containing *hb* genes, PCR primers (forward: 5'-ccatgcctatcgctcgactctg-3', reverse: 5'-ggttgcgctgcggtttaat-3') within highly conserved regions in lamprey globins were designed by aligning larva/adult hemoglobin cDNA sequences of *Lampetra zanandreae* ¹. These primers were used for amplification of *L. fluviatilis* genomic DNA (a kind gift of Dr. Akie Sato, Max-Planck-Institut für Biologie, Tübingen DE). A PCR product of 1.2 kb in length was obtained and used as a probe for screening the cosmid library. Cosmids were subjected to restriction mapping and Southern blot analysis according to standard procedures ². Two partially overlapping cosmids (99E08 and 109N16) were selected for sequence analysis.

Sequencing and contig assembly

Cosmid DNA was sheared into ~2kb fragments which were used for shotgun cloning and Sanger sequencing. Paired-end sequences were used to construct contigs. Assembly of the *L. fluviatilis* *hb* locus sequence was hampered by the abundant presence of repetitive elements and simple repeats typical of *Cyclostome* genomes ³, and in addition some fragments were refractory to sequencing. Using genomic *L. fluviatilis* DNA as template, PCR products bridging the gaps in the initial assembly were generated and directly sequenced. Long-range single molecule Nanopore sequencing (see below) was used to confirm the final assembly. Potential exons and genes were identified by GenScan ⁴, these initial gene structures were further refined by manual curation. **The location of the promoter of the *npr3* gene was predicted using the Promoter 2.0 Prediction Server (<http://www.cbs.dtu.dk/services/Promoter/>).** Predicted protein sequences were used to search the genome databases at NCBI and Ensembl with the BLAST family of search algorithms ^{5,6}.

Long-range amplification of genomic DNA

Long-range PCR using genomic *L. fluviatilis* genomic DNA as a template was used to generate 5 overlapping amplicons (A-E). The primers and their annealing temperatures are listed in Supplemental Table 2. All amplifications were performed using Prime STAR GXL DNA polymerase (TaKaRa Bio Inc, Shiga, JP) with two-step cycling conditions. An initial denaturation of 98°C for 30 seconds then a denaturation of 98°C for 10 seconds followed by an annealing extension step at the annealing temperature (Supplemental Table 2) for 10 minutes for 32 cycles. In a 25 µl reaction volume 2.5 ng of template was added to 5 µl of 5X PrimeSTAR GXL Buffer, 2 µl of dNTP mixture (2.5 mM each), 0.5 µl of each primer (10 µM), 0.5 µl PrimeSTAR GXL DNA Polymerase (1.25 U/µl) and made up to 25 µl with water. PCR product C had 2 bands and the larger band was excised from an agarose gel and cleaned using Zymoclean Gel DNA Recovery Kit (Zymo Research, Irvine, CA). The other amplicons were all cleaned up using Agencourt AMPure XP (Beckman Coulter, Brea, CA). Between 1.0 and 1.5 µg of amplicons were pooled to sequence with the Nanopore 1D Amplicon Sequencing strategy for the MinION using kit SQK-LSK108 (Oxford Nanopore Technologies, Oxford, UK). The amplicons were end repaired using Ultra II End-Prep enzyme (NEB, Ipswich, MA) and then cleaned up with AMPure XP beads. A blunt ended ligation reaction then added the adaptors. The mixture was again cleaned and the library eluted in ABB ready for loading on to a MinION flow cell.

Nanopore sequencing

MinION sequencing was performed as per the manufacturer's guidelines using an R9 flow cell (FLO-MIN106). MinION sequencing was controlled using Oxford Nanopore Technologies MinKNOW software.

Nanopore sequencing bioinformatics

The base calling program Albacore/1.2.4 (available from the Oxford Nanopore Technologies user community) was used to process .fast5 read files generated by the MinION. Fasta files with sequence reads in the range 5K to 18K were extracted from the processed .fast5 files using Poretools⁷. The sequence was assembled using canu/1.5 with the following parameters: -useGrid=0 -stopOnReadQuality=false -contigFilter="2 1000 1.0 1.0 2" -correctedErrorRate=0.075⁸. The draft assembly was then iteratively processed by the nano-polish consensus-calling algorithm using the signal-level data measured by the MinION sequencer to produce an improved consensus sequence⁹. The reads were mapped to the draft assembly using BWA-MEM with the "-x ont2d" option. Assemblies of the *L. fluviatilis* *hb* locus based on Sanger and Nanopore sequencing data were aligned and visualized using PipMaker¹⁰ with default settings.

Comparative analysis

Globin protein sequences of a cyclostome (sea lamprey, *Petromyzon marinus*), bony fish (pufferfish, *Tetraodon nigroviridis*, and zebrafish, *Danio rerio*), an amphibian (African clawed frog, *Xenopus tropicalis*), birds (chicken, *Gallus gallus*, and zebra finch, *Taeniopygia guttata*) and mammals (human, *Homo sapiens*, and mouse, *Mus musculus*) were retrieved from the NCBI and Ensembl databases. NPRL3 protein sequences were retrieved for an insect (fruit fly, *Drosophila melanogaster*), a cyclostome (sea lamprey, *Petromyzon marinus*), bony fish (pufferfish, *Tetraodon nigroviridis*, and zebrafish, *Danio rerio*), birds (chicken, *Gallus gallus*, and zebra finch, *Taeniopygia guttata*) and mammals (human, *Homo sapiens*, and mouse, *Mus musculus*). All protein sequences are listed in Supplemental Information. Multiple sequence alignments of proteins were performed using the progressive alignment algorithm implemented in CLC Genomics Workbench 12.0.3 (Qiagen, Hilden, DE). Molecular phylogenetic analysis using Maximum Likelihood approach was performed using Molecular Evolutionary Genetics Analysis software (MEGA version 7.0). Best protein substitution model for Maximum Likelihood approach was inferred using MEGA version 7.0¹¹ and ProtTest (prottest-3.4.2)¹². The reliability of the inferred tree was tested using bootstrapping (1000 replications).

Genomic sequences of the pufferfish and human *NPRL3* genes were retrieved from Ensembl. LAGAN¹³ was used for multiple alignment of the river lamprey, pufferfish and human *NPRL3* genes, and VISTA¹⁴ for visualization of the results.

Analysis of HB and NPRL3 expression

RNA was isolated from adult *L. fluviatilis* brain and blood, and used for oligo-dT-primed cDNA synthesis. Reverse primers specific for each of the six *L. fluviatilis hb* genes were located in the 3'UTR and combined with a common forward primer in exon 2. Each reverse/forward primer combination yielded a cDNA amplicon of unique size. For NPRL3, primers were designed spanning exons 2-5 and 13-14 of the *L. fluviatilis nprl3* gene. RT-PCR was performed as described¹⁵, using 30 cycles of 30 s 58°C, 30 s 72°C and 30 s 92°C. PCR products were analyzed on 2% agarose gels². Primer sequences and sizes of PCR products are listed in Supplemental Information.

Proteomics of L. fluviatilis HB proteins

For proteomics analysis of *L. fluviatilis* HB proteins, whole blood and gills of adult river lampreys, and the gills of ethanol-fixed larvae, were lysed in SDS-PAGE loading buffer and ~10 µg of protein was separated on 12.5% SDS-PAGE. The area in the 15-20 kD range was cut into 2-mm slices using an automatic gel slicer and subjected to in-gel reduction with dithiothreitol, alkylation with iodoacetamide and digestion with trypsin (sequencing grade, Promega, Madison, WI). Subsequently, nanoflow LC-MS/MS was performed on an 1100 series capillary LC system (Agilent Technologies, Santa Clara, CA) coupled to an LTQ-Orbitrap mass spectrometer (Thermo Fischer Scientific, Waltham, MA) operating in positive mode and equipped with a nanospray source.

Peptide mixtures were trapped on a ReproSil C18 reversed phase column (column dimensions 1.5 cm × 100 µm, packed in-house; Dr. Maisch GmbH, Ammerbuch, DE) at a flow rate of 8 µl/min. Peptide separation was performed on ReproSil C18 reversed phase column (column dimensions 15 cm × 50 µm, packed in-house; Dr. Maisch GmbH) using a linear gradient from 0 to 80% B (A = 0.1 % formic acid; B = 80% (v/v) acetonitrile, 0.1 % formic acid) in 70 min and at a constant flow rate of 200 nl/min using a splitter. The column eluent was directly sprayed into the ESI source of the mass spectrometer. Mass spectra were acquired in continuum mode; fragmentation of the peptides was performed in data-dependent mode. Peak lists were automatically created from raw data files using the Mascot Distiller software (version 2.1; Matrix Science, Boston, MA). The Mascot search algorithm (version 2.2, Matrix Science) was used for searching against a custom database containing the predicted *L. fluviatilis* HBs. The peptide tolerance was set to 10 ppm and the fragment ion tolerance to 0.8 Da. A maximum number of 2 missed cleavages by trypsin were allowed. Carbamido-methylated cysteine was set as a fixed modification and oxidized methionine and phosphorylation on serine, threonine and tyrosine were set as variable modifications. The Mascot score cut-off value for a positive protein hit was set to 60. Individual peptide MS/MS spectra with Mascot scores below 40 were checked manually and either interpreted as valid identifications or discarded.

DNaseI hypersensitive site mapping

Nuclei were isolated from freshly frozen blood and liver samples obtained from adult river lampreys. Aliquots were treated with increasing amounts of DNaseI^{16,17}. DNA of the DNaseI treatment series was digested with NcoI or PvuII and size-fractionated on 0.7% agarose gels. After blotting, the membranes were hybridized to probes close to the ends of the restriction fragments. Probe fragments were generated by PCR amplification of *L. fluviatilis* genomic DNA, carefully avoiding repetitive elements including transposons and simple repeat DNA sequences. To this end, sequences were masked using RepeatMasker¹⁸ and the masked sequences were used for BLAST searches of the *P. marinus* preliminary genome assembly (21 Feb 2007) at Ensembl to detect lamprey-specific repeats. After hybridization and washing, the membranes were exposed to phosphor storage screens, which were scanned with a Typhoon Trio instrument (GE healthcare, Chicago, IL). Oligonucleotides and sequences of the probes are listed in Supplemental Information.

ATAC-seq

Adult *P. marinus* individuals were euthanised in MS222 (0.5 g/l). Blood was collected by puncturing the heart with a needle and syringe containing 1 mg/ml heparin in phosphate buffered saline (PBS). To obtain blood from a 70 mm *P. marinus* ammocoete, the individual was euthanised in MS222 (0.2 g/L), decapitated at the level of the gill arches and allowed to bleed out into a Petri dish containing 1 mg/ml heparin in PBS. Adult and ammocoete blood was pelleted by

centrifugation and washed several times with PBS. Erythrocyte concentration was determined using a hemocytometer. ~50,000 cells were then processed for ATAC as described previously ¹⁹, with the following modifications. The transposition reaction was stopped by adding EDTA to a final concentration of 50 nM and incubating the samples at 50°C for 30 min. Transposed fragments were amplified and indexed by an 11-14 cycle PCR and excess primers removed using AMPure XP beads (Beckman Coulter). Library quality was assessed using a TapeStation (Agilent Technologies, Santa Clara, CA) and QuBit Fluorometric Quantitation (ThermoFisher Scientific). Libraries were pooled, quantified using Kapa Library Quantification and sequenced on the NextSeq (75 base pairs; paired-end; Illumina, San Diego, CA). For bioinformatic analysis, the sequence of the *L. fluviatilis hb* locus was masked using RepeatMasker ¹⁸. Further repeat masking was necessary to block lamprey-specific repeats. This was done manually by aligning 5kb sections of the *L. fluviatilis hb* locus to the Ensembl *P. marinus* genome assembly (version 7.0) (<https://www.ensembl.org/index.html>) with the BLAST algorithm ⁶. ATAC-seq .fastq files were aligned using the masked *L. fluviatilis hb* locus as the reference genome with Bowtie2 ²⁰ with no or 3 mismatches allowed. The aligned reads were then sorted and indexed in SAMtools ²¹ using the default parameters. The alignments were visualized against the reference genome using the Integrated Genome Browser ²².

Vector construction for enhancer reporter assays

Either the zebrafish *MCS-R2 hb* enhancer, intron 5 of the *L. fluviatilis nprl3* gene or intron 7 of the *L. fluviatilis nprl3* gene were inserted into the HLC GFP reporter vector ²³, along with the *L. fluviatilis hb2* promoter, by In-Fusion cloning (Takara). PCR Primers for In-Fusion cloning were designed with SnapGene software (Takara). PCR to amplify the zebrafish or *L. fluviatilis* genomic fragments was performed using PfuUltra II Fusion HS DNA polymerase (Agilent Technologies) with 5 cycles of 30 s 95°C, 30 s 50-56°C and 30 s 72°C, followed by 25-35 cycles of 30 s 95°C, 30 s 56-62°C and 30 s 72°C. Gel-extracted PCR products were combined with 100 ng of linearised HLC GFP reporter vector at 1:3 molar ratio for the *hb2* promoter and 1:5 molar ratio for the intronic elements, and incubated for 15 min at 50°C with 1x In-Fusion HD Enzyme Premix (Takara). After transformation, constructs were sequenced to confirm they carried the correct inserts.

Enhancer reporter assays in sea lamprey embryos

ISec-I meganuclease-mediated transgenesis was performed in *P. marinus* embryos as described previously ²³⁻²⁵ in Prof. Marianne Bronner's laboratory (California Institute of Technology, USA). At 4-6 hours post fertilisation, single-cell embryos were injected with the ISec-I vector digestion mix at 20 ng/μl and maintained at 18 °C in Marc's Modified Ringers buffer (0.1x) for the remainder of their development. At 1 dpf embryos were transferred to 96 well plates until 6 dpf when they were returned to Petri dishes, and screened daily for reporter expression using an Olympus MVX10

microscope (Olympus, Tokyo, JP). Embryos were imaged with a Zeiss AxioCam MRm camera and AxioVision Rel 4.6 software (Zeiss, Oberkochen, DE). Movies were compiled using ImageJ.

Enhancer reporter assays in zebrafish embryos

Conventional zebrafish transgenesis was conducted using the Tol2 transposase system ²⁶. Single cell zebrafish embryos were injected with the HLC GFP reporter vector ²³ at 60 ng/μl together with Tol2 mRNA at 40 ng/μl. Embryos were maintained at 28.5°C in E3 medium and screened daily for reporter expression using an Olympus MVX10 microscope. Embryos were imaged with a Zeiss AxioCam MRm camera and AxioVision Rel 4.6 software. Movies were compiled using ImageJ.

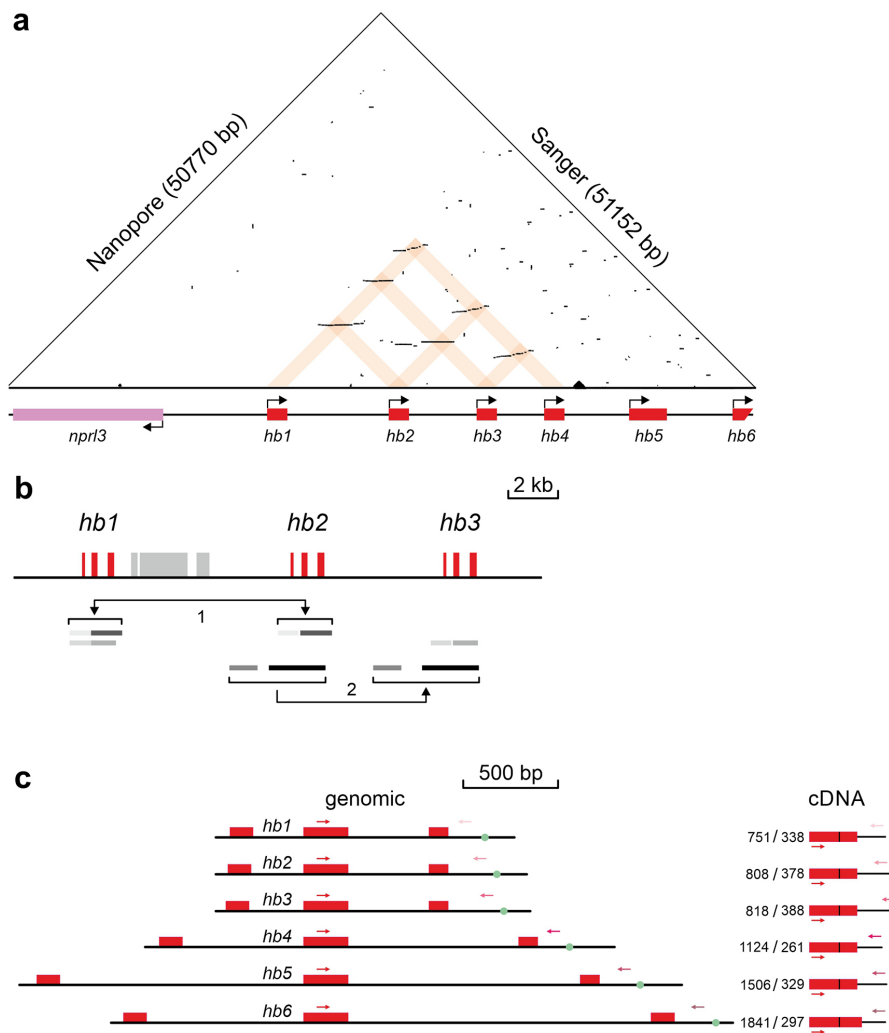
Movies

Movie 1. Erythroid-specific GFP reporter expression in circulation in the head region of a 17 dpf *P. marinus* embryo when intron7 of *L. fluviatilis nprl3* is used to drive GFP expression from the *L. fluviatilis hb2* promoter.

Movie 2. Erythroid-specific GFP reporter expression in circulation in the head region of a 17 dpf *P. marinus* embryo when the zebrafish *MCS-R2 hb* enhancer is used to drive GFP expression from the *L. fluviatilis hb2* promoter.

Movie 3. Erythroid-specific GFP reporter expression in circulation in the tail region of a 50 hpf zebrafish embryo when the zebrafish *MCS-R2 hb* enhancer is used to drive GFP expression from the *L. fluviatilis hb2* promoter.

Supplemental Figure 1



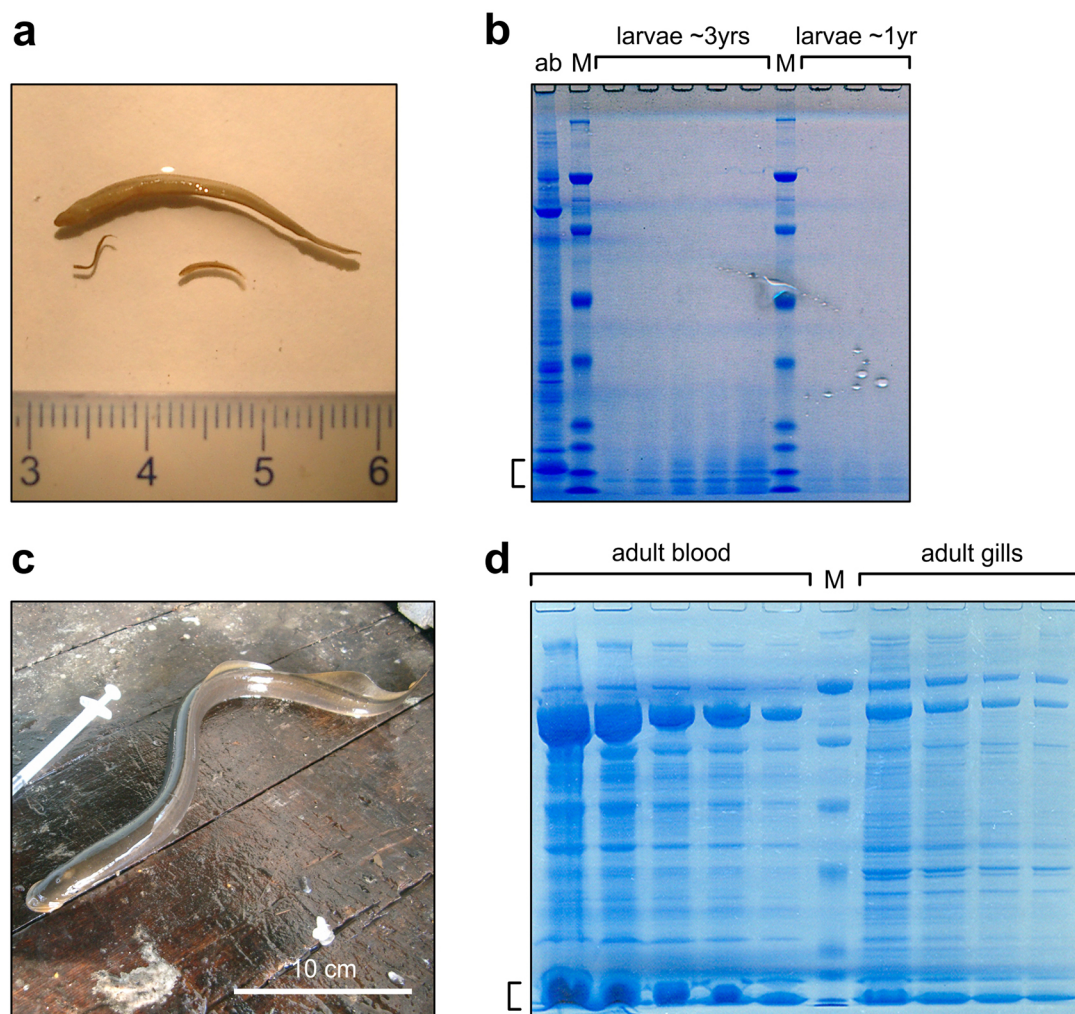
Supplemental Figure 1. Analysis of the *L. fluviatilis* *hb* locus

a) Percentage identity plots of alignments generated with PipMaker¹⁰. River lamprey *hb* locus assemblies based on Sanger and Nanopore sequencing were aligned as indicated. The horizontal line at the bottom of the plot indicates contiguous alignment of the two assemblies. Large repeats due to duplications of the larval globin genes (*hb1-4*) are highlighted.

b) Evidence for recent gene duplications in the *L. fluviatilis* *hb* locus. Blocks of direct repeats are shown; darker is higher homology. The first gene duplication event (1) resulted in the *hb1* and *hb2* genes. This was followed by duplication of the *hb2* gene yielding the *hb3* gene (2). The coordinates and percentage homology of the duplicated regions are shown in Suppl. Table 1. Note that *hb4* is not included here since it is more distantly related to *hb1-3* (see Supplemental Figure 3).

c) Strategy for expression analysis of *L. fluviatilis* HB mRNAs. A primer common to exon 2 of all six *L. fluviatilis* *hb* genes was used in combination with a gene-specific primer in the 3'UTR. This yields amplicons of unique size for each cDNA. Expected amplicon sizes (Genomic/cDNA) are shown in base pairs. Putative poly-adenylation sites are indicated by green dots.

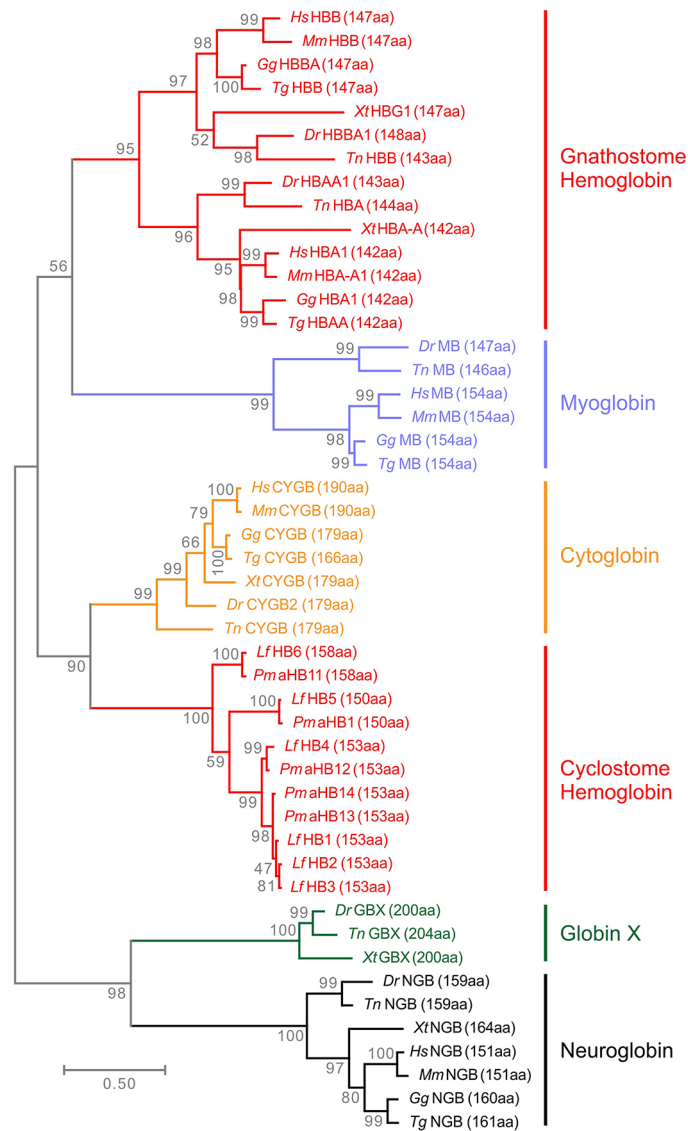
Supplemental Figure 2



Supplemental Figure 2. Proteomics of *L. fluviatilis* HBs

a) Larvae (ammocoetes) of *L. fluviatilis*, ~1- and ~3-years old. **b)** SDS-PAGE gel of proteins extracted from larval gills. **c)** Adult *L. fluviatilis*. **d)** SDS-PAGE gel of proteins extracted from adult blood and gills. The areas of the gels used for mass spectrometry analysis are indicated by vertical brackets.

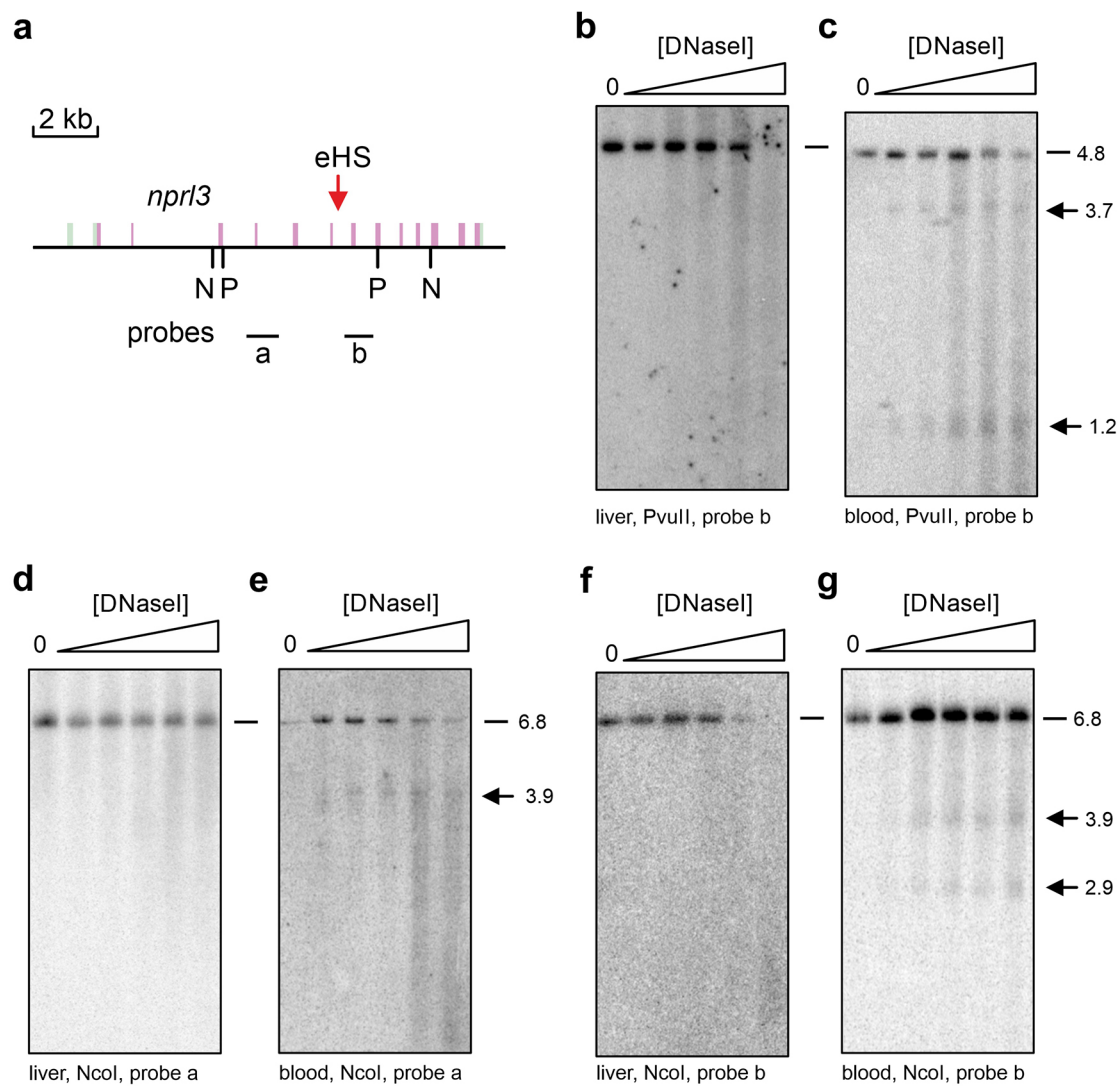
Supplemental Figure 3



Supplemental Figure 3. Molecular phylogenetic relationships of selected globins

Multiple sequence alignments of the *nprl3*-linked HBs from agnathans (river lamprey *Lampetra fluviatilis*, *Lf*, and sea lamprey *Petromyzon marinus*, *Pm*) with globins from bony fish (pufferfish *Tetraodon nigroviridis*, *Tn*, and zebrafish *Danio rerio*, *Dn*) an amphibian (African clawed frog *Xenopus tropicalis*, *Xt*), birds (chicken *Gallus gallus*, *Gg*, and zebra finch *Taeniopygia guttata*, *Tg*) and mammals (human (*Homo sapiens*, *Hs*, and mouse *Mus musculus*, *Mm*). Multiple sequence alignments were performed using the progressive alignment algorithm implemented in CLC Genomics Workbench 12.0.3. The evolutionary history was inferred using the Maximum Likelihood method. The best protein substitution model for Maximum Likelihood approach was inferred using MEGA version 7.0 and ProtTest 3.4.2. The reliability of the inferred tree was tested using bootstrapping (1000 replications). The tree is drawn to scale, with branch lengths measured by the number of substitutions per site. NGB = neuroglobin; GBX = globin X; MB = myoglobin; CYGB = cytoglobin; HB = hemoglobin; HBA = α -hemoglobin; HBB = β -hemoglobin HBG = γ -hemoglobin. Color scheme adopted from ²⁷.

Supplemental Figure 4



Supplemental Figure 4. An erythroid-specific DNaseI hypersensitive site in intron 7 of the *L. fluviatilis nprl3* gene

a) Drawing of *L. fluviatilis nprl3* with the positions of the erythroid-specific DNaseI hypersensitive site (eHS), probes and restriction sites (N = NcoI; P = PvuII) indicated. **b,c)** Liver and blood DNA digested with PvuII, hybridized with probe b. **d,e)** Liver and blood DNA digested with NcoI, hybridized with probe a. **f,g)** Liver and blood DNA digested with NcoI, hybridized with probe b. Fragment sizes are indicated in kilobases, hypersensitive sites by arrows. See Supplemental Information for additional details.

Supplemental Figure 5

L. fluviatilis *Nprl3* intron 7 eHS area – potential transcription factor binding sites

Exon7

GCTGGCTCGGGACCTAAAGGAGGCCTATGACAA GTAAGCTTGCTTTGCCGCGCAGGGCAC
 GACCAACTCAGAGTGCATGAGGAATGCGGT GTGGTGGG CTCTCTCG CAGATGTCTTTTA
 CAGATG
 ATCCCATTATTCTGGAACTCTTGTTACAAGAGAGAGAATGGATGGACTGAATAATTGCT
 TAAGTGCACAAGAGGCTA CATGACT GCCCCTTGTGGAATATGAACCATGTGCTTTCGTG
 TATTCAATA ACCGCCCC TATCAGAG GTGTTGAAGGCTTGTGATTGTAAACACATGTGGC
 CATGTG
 TCTGTGTGAATCGTCTTATAAC GTTATCAT TCTTGCGAGCAGAAGAATTCAAAATCTGAA
 AAATAGTGGAACGCTGGGG CAGATAGT GTTAAGTAGTTATTACAATACATTACAAATG
 GGGAAATCTTTAGTTGTTATACAGGGTGTGTCTTTTAAATTATAGCCAAAGTAACGTTAT
 TTCCTATACTTGGGTATATTTTTTAAAAGATGCACCCTGCTTAACCGGGTCTGGTGAGTT
 TTAATTAACGTTTGGATGATGGTAGAACAGAATGAGATGCTTCTCTAAGTGCGTGCTGTG
 TGTTCCTTCTCCAG CCTCTGTTCCACAGGTTTGGTGCAGCTATACATCAACAACTGGCTG

Exon8

GATA GC/GT box E box MARE

Supplemental Figure 5. Potential binding sites for erythroid regulators in *L. fluviatilis* *nprl3* intron 7

The sequence of *L. fluviatilis* *nprl3* intron 7 is displayed, with parts of the flanking exons 7 and 8 (purple). Potential binding sites for erythroid transcription factors are highlighted. Yellow: GATA factors; green: SP/KLF factors; grey: E2A/TAL1 factors; red: NF-E2 related factors.

Supplemental Table 1

| Duplications in <i>L. fluviatilis</i> <i>hb</i> locus | | | | | |
|---|-------|--------------|----------|----------------|-------|
| size (bp) | start | identity (%) | gaps (%) | genes involved | event |
| 855 | 17122 | 82 | 7 | <i>hb1</i> | (2) |
| 825 | 31649 | | | <i>hb3</i> | |
| | | | | | |
| 855 | 17122 | 81 | 7 | <i>hb1</i> | 1 |
| 819 | 25521 | | | <i>hb2</i> | |
| | | | | | |
| 1253 | 17995 | 88 | 3 | <i>hb1</i> | 1 |
| 1274 | 26404 | | | <i>hb2</i> | |
| | | | | | |
| 1005 | 18006 | 88 | 3 | <i>hb1</i> | (2) |
| 1028 | 32548 | | | <i>hb3</i> | |
| | | | | | |
| 1131 | 23554 | 85 | 4 | <i>hb2</i> | 2 |
| 1132 | 29333 | | | <i>hb3</i> | |
| | | | | | |
| 2277 | 25163 | 96 | 1 | <i>hb2</i> | 2 |
| 2283 | 31293 | | | <i>hb3</i> | |

* Refers to duplication events proposed in Suppl. Fig. 1b. Note that the apparent duplications involving the *hb1-hb3* genes (bracketed) are the consequence of a more recent duplication event involving the *hb2-hb3* genes.

Supplemental Table 2

| Product | Size (bp) | Primer name | Direction | Sequence (5' - 3') | Annealing temperature (°C) |
|---------|-----------|-------------|-----------|--------------------------------|----------------------------|
| A | 12199 | Lam-1 | Forward | GAAAGTGGGAAAAGGCTCTAGACGA | 69 |
| | | Lam-2 | Reverse | GCGTTACTCGTGAAAGCGTACTTGT | |
| B | 13549 | Lam-13 | Forward | CCAACGCCCTGTGCAAGATTACTAT | 69 |
| | | Lam-14 | Reverse | CCTCAAAATGTGCTGTACCTATGG | |
| C | 12977 | Lam-18 | Forward | TTTCGTGCAACAGGTTTTTCGCCGCCTATC | 75 |
| | | Lam-20 | Reverse | CTCACGCTGCTATTTACCCTTGCCCGTTCT | |
| D | 13786 | Lam-29 | Forward | CACAATCACATCCGCACAATTGCAT | 70.5 |
| | | Lam-21 | Reverse | TGTAGCCGATGCCAAAGCGTATGTC | |
| | 12452 | Lam-29 | Forward | CACAATCACATCCGCACAATTGCAT | 70.5 |
| | | Lam-28 | Reverse | TTTTGACCCACCCCCTAGTCACAC | |
| E | 11426 | Lam-9 | Forward | GTGTCCGTTGAGTGTAATGTGACG | 69 |
| | | Lam-10 | Reverse | GCAGGCTGTACACCACTTCTTGTTT | |

Primers and annealing temperature used for long-range amplification of *L. fluviatilis* genomic DNA.

Supplemental Information

Presence of *nprl3*-linked *hb* genes is a common feature of lamprey species

Comparison of *nprl3*-linked *hb* genes in three lamprey species.

| <i>Lampetra fluviatilis</i> [#] | <i>Lethenteron camtschaticum</i> ^{&} | <i>Petromyzon marinus</i> ^{&} |
|--|---|--|
| <i>hb1</i> | <i>aHb14a</i> | - |
| <i>hb2</i> | <i>aHb13</i> | <i>aHb13</i> |
| <i>hb3</i> | <i>aHb14b</i> | <i>aHb14</i> |
| <i>hb4</i> | <i>aHb12</i> | <i>aHb12</i> |
| <i>hb5</i> | <i>aHb1</i> | <i>aHb1</i> |
| <i>hb6</i> | <i>aHb11</i> | <i>aHb11</i> |

[#] This paper

[&] Schwarze *et al* (2014) ²⁸

Oligonucleotides used for RT-PCR

| Oligonucleotide | PCR (genomic) | PCR (cDNA) |
|--|---------------|------------|
| common exon2 forward GATCATCAACGCCGTCAAC | | |
| Hb1 3'UTR reverse TGCGTACAGCGATAGCAAAA | 751 bp | 338 bp |
| Hb2 3'UTR reverse ACAACACGGTTTCGTTTCCC | 808 bp | 378 bp |
| Hb3 3'UTR reverse TTTCATGCACAACACGGTTT | 818 bp | 388 bp |
| Hb4 3'UTR reverse GAGGTCGTAGGAAGGGAGGT | 1124 bp | 261 bp |
| Hb5 3'UTR reverse AACGACGCTTCATTCCTGAT | 1506 bp | 329 bp |
| Hb6 3'UTR reverse CGATCGGTGTGTCACATTTT | 1841 bp | 297 bp |
| Npr13 exon 2 forward GGGCGATGTCTCTTTGAGTC Npr13 exon 5 reverse CAAACACCACGTTGAACAGG | 5008 bp | 439 bp |
| Npr13 exon 13 forward GAGCAGCGATGATCTTACCC Npr13 exon 14 reverse TGGGAAAAGGCTCTAGACGA | 776 bp | 456 bp |

Globin protein sequences

| Species | Abbreviation |
|---------|--------------|
|---------|--------------|

Mammals

| | |
|---------------------|-----------|
| <i>Homo sapiens</i> | <i>Hs</i> |
| <i>Mus musculus</i> | <i>Mm</i> |

Birds

| | |
|----------------------------|-----------|
| <i>Gallus gallus</i> | <i>Gg</i> |
| <i>Taeniopygia guttata</i> | <i>Tg</i> |

Amphibians

| | |
|---------------------------|-----------|
| <i>Xenopus tropicalis</i> | <i>Xt</i> |
|---------------------------|-----------|

Bony fish

| | |
|-------------------------------|-----------|
| <i>Tetraodon nigroviridis</i> | <i>Tn</i> |
| <i>Danio rerio</i> | <i>Dr</i> |

Jawless fish

| | |
|-----------------------------|-----------|
| <i>Lampetra fluviatilis</i> | <i>Lf</i> |
| <i>Petromyzon marinus</i> | <i>Pm</i> |

>Hs HBA1 (142aa)

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR

>Hs HBB (147aa)

MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAMGNPK
VKAHGKKVLGAFSDGLAHLNKLKGTFTATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
KEFTTPVQAAAYQKVAVGVANALAHKYH

>Hs CYGB (190aa)

MEKVPGEIMEIERERSEELSEAERKAVQAMWARLYANCEDVGVAAILVRFFVNFPSAQYF
SQFKHMEDEPLEMERSPQLRKHACRVMGALNTVVENLHDPDKVSSVLALVGKAHALKHKVE
PVYFKILSGVILEVVAEEFASDFPETQRAWAKLRGLIYSHVTAAYKEVGWVQQVNPATT
PPATLPSSGP

>Hs MB (154aa)

MGLSDGEWQLVLNVWGKVEADIPGHGQEVILIRLFKGHPELTLEKFDKFKHLKSEDEMKASE
DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH
PGDFGADAQGAMNKALELFRKDMASNYKELGFQG

>Hs NGB (151aa)

MERPEPELIRQSWRAVSRSPLHGTVLFARLFALPDLLPLFQYNCRQFSSPEDCLSSPE
FLDHIRKVMLVIDAAVTNVEDLSSLEEYLASLGRKHRAVGVKLSSFSTVGESLLYMLEKC
LGPAFTPATRAAWSQLYGAVVQAMSRGWDGE

>Mm HBA-A1 (142aa)

MVLSGEDKSNIAAWGKIGGHGAIEYGAEALERMFAFPTTKTYFPHFDVSHGSAQVKGHG
KKVADALANAAGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPADFTP
AVHASLDKFLASVSTVLTSKYR

>Mm HBB -BS (147aa)

MVHLTDAEKA AVSGLWGKVN ADEVGGEALGRLLV VYPWTQRYFDSFGDLSSASAIMGNAK
VKAHGKKVITAFNDGLNHLDSLKGTFASLSELHCDKLHVDPENFRLLGNMIVIVLGHHLG
KDFTPAAQA AFQKV VAGVAAALAHKYH

>Mm CYGB (190aa)

MEKVPGDMEIERRERSEELSEAERKAVQATWARLYANCEDVGVAILVRFFVNFPSAKQYF
SQFRHMEDPLEMERSPQLRKHACRVMGALNTVVENLHDPDKVSSVLALVGKAHALKHKVE
PMYFKILSGVILEVIAEEFANDFPVETQKAWAKLRGLIYSHVTAAYKEVGWVQQVPNTTT
PPATLPSSGP

>Mm MB (154aa)

MGLSDGEWQLVLNVWGKVEADLAGHGQEVLI GLFKTHPETLDKFDKFKNLKSEEDMKGSE
DLKKHGCTVLTALGTILKKKGQHAAEIQLAQSHATKHKIPVKYLEFISEIIIEVLKRRH
SGDFGADAQ GAMSKALELFRNDIAAKYKELGFQG

>Mm NGB (151aa)

MERPESELIRQSWRVVSRSPLEHGTVL FARLFALEPSLLPLFQYNGRQFSSPEDCLSSPE
FLDHIRKVMLVIDAAVTNVEDLSSLEEYLTSLGRKHRAGVRLSSFSTVGESLLYMLEKC
LGPDFTPATRTAWSRLYGAVVQAMSRGWDGE

>Gg HBA1 (142aa)

MVLSAADKNNVKGIFTKIAGHAE EYGAETLERMFTTYPPTKTYFPHFDLSHGSAQIKGHG
KKVVAALIEAANHIDDIAGTLSKLSDLHAHKL RVDPVNFKLLGQCFLVVVAIHHPAALTP
EVHASLDKFLCAVGTVLTAKYR

>Gg HBBA (147aa)

MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPM
VRAHGKKVLT SFGDAVKNL DNKNTFSQLSELHCDKLHVDPENFRLLGDILIIIVLAAHFS
KDFTPECQA AWQKLVRVVAHALARKYH

>Gg CYGB (179aa)

MEKVQGEMEIERWERSEEISDAEKKVIQETWSRVYANCEDVGVSI LIRFFVNFPSAKQYF
SQFKHMDDTLEMERSLQLRKHAQRVMGAIN TVVENLDDPEKVSSVLALVGKAHALKHKVE
PVYFKKLTGVMLEVIAEAYGNDFTPEAHGAWTKMRTL IYTHVTAAYKEAGWVSYP SATL

>Gg MB (154aa)

MGLSDQEWQQVLTIWGKVEADIAGHGHEVLMRLFHDHPETLDRFDKFKGLKTPDQMGSE
DLKKHGATVLTQLGKILKQGNHESELKPLAQTHATKHKIPVKYLEFISEV IIKVIAEKH
AADFGADSQAAMKKALELFRNDMASKYKEFGFQG

>Gg NGB (160aa)

MESGMLSRTQQALIRESWRRVSGSPVQHGVVLF SRLFDLDPDLLPLFQYNCKRFAS PQEC
LAAPEFLDHIRKVMLVIDAAVSHLEDLPCLEEYLCNLGKKHQAVGVKVESFSTVGESLLY
MLEKCLGA AFSPDVREAWIELYS AVVKAMQRGWEVLPEGD

>Tg HBAA (142aa)

MVLSAGDKSNVKAVFGKIGGQADEYGADALERM FATYPQTKTYFPHFDLGKGS AQVKGHG
KKVAAALVEAANNVDDL AGALSKLSDLHAQKL RVDPVNFKLLGQCFLVVVATRNP SLLTP
EVHASLDKFLCAVGTVLTAKYR

>Tg HBB (147aa)

MVNWTAEEKQLVTTLWGRVNVDECGAEALARLLV VYPWTQRFFVSFGNMSSPTAVLGNPM
VRAHGKKVLT SFGEAVKNLDSIKNTFSQLSELHCDKLHVDPENFRLLGDILVVVLA AHFG

KDFTPDCQAAWQKLVRVVAHALARKYH

>Tg CYGB (166aa)

MEIERWERSEEISDAEKKVIQEIWSRVYANCEDVGVSILIRFFVNFPSAKQYFSQFKHME
DPLEMERSLQLRKHARRVMGAINTVVENLNDSEKVSSVLALVGKAHALKHKVEPIYFKKL
TGVMLEVIAEEYPNDFTPEAHGAWTKMKTLIYTHVTAAYKEVGWAQ

>Tg MB (154aa)

MGLSDQEWQQVLTWVGKVESDLAGHGHQILMRLFQDHPETLDRFEKFKGLKTPDAMKGSE
DLKKHGVTVLTQLGKILKAKGNHEAELKPLAQTHATKHKIPVKYLEFISEVLIKVLAEKH
AADFGADAQAAMKKALELFRNDMATKYKEFGFQG

>Tg NGB (161aa)

MESGMRLSGGQRALIRESWQRVSGSPVQHGLVLFTRLFDLDPDLLPLFQYNCKQFASPQE
CLSAPEFLDHIRKVMLVIDAAVSHLENLSCLEEYLCNLGKKHQAVGVKVESFSTVGESLL
YMLEKCLGAAFSPEVQEAWSKLYNAVVKAMQRGWETLPEGD

>Xt HBA-A (142aa)

MHLTADDDKKHIKAIWPSVAAHGDKYGGEALHRMFMCAPKTKTYFPDFDFSEHSKHILAHG
KKVSDALNEACNHLNDNIAGCLSKLSDLHAYDLRVDPGNFPLLAHQILVVVAIHFPKQFDP
ATHKALDKFLVSVSNVLTISKYR

>Xt HBG1 (147aa)

MVNLTAKERQLITGTWSKICAKTLGKQALGSMLYTYPWTQRYFSSFGNLSSIEAIFHNAA
VATHGEKVLTSIGEAIKHMDDIKGYAQLSKYHSETLHVDPYNFKRFCSTIISMAQTLO
EDFTPELQAAFEKLFAAIADALGKGYH

>Xt CYGB (179aa)

MEKVQGENDMERWERLEEITESERGVIKETWARVYANCEDVGVSILIRFFVNFPSAKQHF
SQFKHMEDPLEMEGSQLRKHARRVMGAVNSVVENLGDPEKITTVLSIVGKSHALKHKVD
PVYFKILTGVMLEVIAEEYAKDFTPDVQLAWNKLRSPLYSHVLSAYKEAGWTQYPSNSV

>Xt GBX (200aa)

MGCILSSLGWQWRDSDLHTTETSPLLPTNLSEQQQQLLVESWRLIQHDIKVGVLVFLVRL
FETHPECKDVFFLFRDVEDDLQALRANKDLRAHGLRVLSFVEKSVARIADCARLEELALEL
GRSHYRYNAPPRYYQYVGTEFISAVRPMLQDKWTAEEVEEAWKGLFAYICTVMERGYQEEE
RRHSDGRSLIDGLQGNKGLI

>Xt NGB (164aa)

MEKDQLSGPQKELIRESWQTVSQDQLHHGTVLFSRLFELEPELVFLFQYNSSHFQSKVQDC
LSSAEFTEHIRKVMTVIDAAVSSLDCLSSLDEYLTSLGRKHRAVGKLESFNTVGESLLF
ALESCLGDAFTSDTREAWSLLYANVVQSMSRGWHRDSQEOREGI

>Dr HBAA1 (143aa)

MSLSDTDKAVVKAIWAKISPKADEIGAEALARMLTVYPQTKTYFSHWADLSPGSGPVKKH
GKTIMGAVGEAISKIDDLVGGLAALSELHAFKLRVDPANFKILSHNVIVVIAMLFPAFT
PEVHVSVDKFFNNLALALSEKYR

>Dr HBBA1 (148aa)

MVEWTDARTAILGLWGKLNIDEIGPQALSRCCLIVYPWTQRYFATFGNLSSPAAIMGNPK
VAAHGRTVMGGLERAIKNMDNVKNTYAALSVMHSEKLVDPDNFRLLADCITVCAAMKFG
QAGFNADVQEAQKFLAVVVSALCRQYH

>Dr CYGB2 (179aa)
MEKEREDEETEGRERPEPLTDVERGI IKDTWARVYASCEDVGVTILIRFFVNFPSAKQYF
SQFQDMEDPEEMEKSSQLRKHARRVMNAINTVVENLHDPEKVSSVLVLVGKAHAFKYKVE
PIYFKILSGVILEILAEFEGECFTPEVQTSWSKLMAALYWHITGAYTEVGWVKLSSSAV

>Dr MB (147aa)
MADHDLVLKCGAVEADYAANGGEVLNRLFKEYPDTLKLFPKFSGISQGDLAGSPAVAAH
GATVLKKLGELLKAKGDHAALLKPLANTHANIHKVALNNFRLITEVLVKVMAEKAGLDAA
GQALRRVMDAVIDGIDIDGYEIGFAG

>Dr GBX (200aa)
MGCAISGSGLTARAPEIRAGEEETPAGLTANHIRLIKESWRLIQEDIAKVGIIMFVRLFE
THPECKDVFFLFRDVEDLERLRSTSRELRAHGLRVMSFIEKSVARLDQLERLETALALELKG
SHYRYNAPPKYGYVGAEFICAVRPILKDRWTPPELEEAWKTLFQYVTSIMREGFLEEERN
KRSNTQTSSRERPDKRSTAI

>Dr NGB (159aa)
MEKLSEKDKGLIRDSWESLGKNKVPBGIVLFTLRFELDPALLTLFSYSTNCGDAPECLSS
PEFLEHVTKVMLVIDAAVSHLDDLHTLEDFLNLGRKHQAVGVNTQSFALVGESLLYMLQ
SSLGPAYTTSLRQAWLTMYSIVVSAMTRGWAKNGEHKSN

>Tn HBA (144aa)
MTSLSTKDKETVRAFWAKVASNREEIGASALCRLLSVYPQTKTYFSHWKDQSPNSASAKK
HGITIMNAVGDVASKIDDLKTGLFNLSELHAFTLRVDPANFKLLAQCMVVIIMYPADF
TPEVHVAMDKFLASLALALSEKYR

>Tn HBB (143aa)
MVVWTDQERAIIDNIFSNLDYEDVGSKALIRCLIVYPWTQRYFSSFGNLYNAEAIARNPN
VAKHGVTVLHGLDRALKNMNIDKEEYKKHSEKLHVDPDNFKLLSDCLTVVIAGKLGSKFT
PEYQAALQKFLAVVVSALGRQYH

>Tn CYGB (179aa)
MERMQRDGEVDHVEQPGPLTEKEKVRIQDSWAKVFQSCDDAGVAILVRFFVNFPSKQFF
KDFKHMEPEEMQSVQLRKHAHRVMTALNTLVESLNNADRVASVLKSVGRAHALKHNV
PKYFKILSGVILEVLGEAFTEIITA EVASAWTKLLANMCCGIAAVYKEAGWTELSSSVE

>Tn MB (146aa)
MGDFDMVLKFWGPVEADYSAHGGMVLTRLFTENPETQQLFPKFVGIAQSELAGNAAVSAH
GATVLKKLGELLKAKGNHAAILQPLANSHTKHKIPIKNFKLIAEVIGKVMAEKAGLDTA
GQQALRNIMATIIADIDATYKELGFS

>Tn GBX (204aa)
MGCAISSLGAKAEFGDRSAEEEDAAAAA VVYPREDQIQMIKDSWKVIRDDIAKVGIIMF
VRLFETHPECKDVFFLFRDVEDLERLRSSRELRAHGLRVMSFIEKSVARLDQQDRLEALA
VELGKSHYHYNAPPKYYSYVGAEFICAVQPILKERFTSELEEAWKTLFQYVTGLMRKGHQ
EEGSRQRHLALPPKDGPEKRTSAL

>Tn NGB (159aa)
MEKLSSKDKELIRGSWDSLGNKVPBGIVLFSRLFELDPPELLNLFHYTTNCGSTQDCLSS
PEFLEHVTKVMLVIDAAVSHLDDLHSLLEDFLNLGRKHQAVGVKPKQSFAMVGESLLYMLQ
CSLGQAYTASLRQAWLNMYSVVVASMSRGWAKNGEDKAD

>Lf HB1 (153aa)

MPIVDSGSVGAISAAEKSLIVSAWAPVYAKYEEAGVDILVKFFAANPEAQAFFPKFKGLD
SADKLKSPAVRWHAERIINAVNDAVVALDDPAKQSLQLKALSQKHAHELNVDPNYFKVL
AGVISDAVVKSGDAKAAVDKFLSQVVILLKSAY

>Lf HB2 (153aa)

MPIVDSGSVGAISAAEKSLIVSAWAPVYAKYEEAGVDILVKFFAANPEAQAFFPKFKGLD
SADQLKKSPAVRWHAERIINAVNDAVVALDDPAKQSLQLKELSQKHAHELNVDPKYFKVL
AGVISDAVVKSGDAKAAVDKLLSQVVILLKSAY

>Lf HB3 (153aa)

MPIVDSGSVGAISAAEKSLIVSAWAPVYAKYEEAGVDILVKFFAANPEAQAFFPKFKGLD
SADQLKKSPAVRWHAERIINAVNDAVVALDDPAKQSLQLKALSQKHAHELNVDPKYFKVL
AGVISDAVVKSGDAKAAVDKLLSQVVILLKSAY

>Lf HB4 (153aa)

MPIVDSGSVGELSAAEKSLVVSAPVYAKYEEAGVDILVKFFSDNPGVQDFFPKFKGLD
SADQLKKSPAVRWHAERIINAVNDAVVALDEPAKLSLKLKGLSKKHAQELNVDPQYFKVL
AGVISDAVVKSGDAKAAVDKFLSQVVILLKFAY

>Lf HB5 (150aa)

MPIVDSGSVPALTAAEKATIRTAWAPVYAKYQSTGVDILIKFFTSNPAAQEFFPKFQGLT
SADQLKKSM DVRWHAERIINAVNDAVVAMDDTEKMSLKLNELSSKHAKSFQVDPQYFKVL
AAVIVDTVLPGDAGLEKLMSMICILLRSSY

>Lf HB6 (158aa)

MPIVDSGSVGALSAAEKAI IADSWKV VYADYEAAGKAILIKFFTSNAGVQDFFPKFKGLD
SADQLSKSAAVRWHAERIINAVNDAVVALDDPEKLSLKLKALS KKHAEFNVDPQYFKVL
SANVLEQVAAANGGLSAEAQGAW EKLLSIISILLKSQY

>Pm aHB1 (150aa)

MPIVDSGSVPALTAAEKATIRTAWAPVYAKYQSTGVDILIKFFTSNPAAQAFFPKFQGLT
SADQLKKSM DVRWHAERIINAVNDAVVAMDDTEKMSLKLRELSGKHAKSFQVDPQYFKVL
AAVIVDTVLPGDAGLEKLMSMICILLRSSY

>Pm aHB11 (158aa)

MPIVDSGSAGALSAAEKAIITDSWKV VYADYEAAGKAILIKFFTSNPGVQDFFPKFKGLD
SADQLSKSAAVRWHAERIINAVNDAVVALDDPEKQSLKLKALS KKHAEFNVDPQYFKVL
SANVLEQVAAANGGLSAEAQGAW EKLLSIISILLKSQY

>Pm aHB12 (153aa)

MPIVDSGSVGEFSAAEKSLIVSAWAPVYAKYEEAGVDILVKFFSDNPGVQDFFPKFKGLD
SADQLKKSPAVRWHAERIINAVNDAVVALDDPPKLSLKLKALS KKHAEFNVDPQYFKVL
AGVISDAVAKSGDEKAAVDKFLSQVVILLKFAY

>Pm aHB13 (153aa)

MPIVDSGSVGAISAAEKSLIVSAWAPVYAKYEEAGVDILVKFFAANPEAQAFFPKFKGLD
SADQLKKSPAVRWHAERIINAVNDAVVALDDPAKQSLQLKALSQKHAHELNVDP SYFKVL
AGVISDAVAKSGDAKAAVDKFLSQVVILLKSAY

>Pm aHB14 (153aa)

MPIVDSGSVGAISAAEKSLIVSAWAPVYAKYEEAGVDILVKFFAANPEAQAFFPKFKGLD
SADKLKSPAVRWHAERIINAVNDAVVALDDPAKQSLQLKALSQKHAHELNVDP SYFKVL
AGVISDAVAKSGDAKAAVDKFLSQVVILLKSAY

NPRL3 protein sequences

| Species | Abbreviation |
|---------|--------------|
|---------|--------------|

Mammals

| | |
|---------------------|-----------|
| <i>Homo sapiens</i> | <i>Hs</i> |
| <i>Mus musculus</i> | <i>Mm</i> |

Birds

| | |
|----------------------------|-----------|
| <i>Gallus gallus</i> | <i>Gg</i> |
| <i>Taeniopygia guttata</i> | <i>Tg</i> |

Bony fish

| | |
|-------------------------------|-----------|
| <i>Tetraodon nigroviridis</i> | <i>Tn</i> |
| <i>Danio rerio</i> | <i>Dr</i> |

Jawless fish

| | |
|-----------------------------|-----------|
| <i>Lampetra fluviatilis</i> | <i>Lf</i> |
| <i>Petromyzon marinus</i> | <i>Pm</i> |

Insects

| | |
|--------------------------------|-----------|
| <i>Drosophila melanogaster</i> | <i>Dm</i> |
|--------------------------------|-----------|

> Hs NPRL3 (568aa)

MRDNTSPIISVILVSSGSRGNKLLFRYPFQRSQEH PASQTSKPRSRYAASNTGDH ADEQDG
DSRFSDVILATILATKSEMCGQKFELKIDNVR FVGHP TLLQHALGQISKTDPSPKREAPT
MILFNVVFALRANADPSVINCLHNLSRRIATVLQHEERRCQYLTREAKLILALQDEV SAM
ADGNEGPQSPFHHILPKCKLARDLKEAYDSLCTSGVVRLHINSWLEV SFCLPHKIH YAAS
SLIPPEAIERSLKAIRPYHALLLLSDEKSLLGELPIDCSPALVRVIKTTSAVKNLQQLAQ
DADLALLQVFQLAAHLVYWGKAI I IYPLCENNVM LSPNASVCLYSPLAEQFSHQFPSHD
LPSVLAKFSLPVSLSEFRNPLAPAVQETQLIQMVVWMLQRRLLIQLHTYVCLMASPSEEE
PRPREDDVPFTARVGGRSLSTPNALSFGSPTSSDDMTLTSPSMDNSSAELLPSGDSPLNQ
RMTENLLALSEHERAAILSVPA AQNPEDLRMFARLLHYFRGRHHLEEIMYNENTRRS QLL
MLFDKFRSVLVVTTHEDPVI AVFQALLP

> Mm NPRL3 (569aa)

MGDNTSPIISVILVSSGSRGNKLLFRYPFQRSQEH PASQTNKPRSRYA VNNTGEHADDQDG
DSRFSDVILATILATKSEMCGQKFELKIDNVR FVGHP TLLQHALGQVSKTDPSPKREAPT
MILFNVVFALRANADPSVINCLHNLSRRIATVLQHEERRCQYLTREAKLILALQDEV SAM
ADANEGPQSPFQHILPKCKLARDLKEAYDSLCTSGVVRLHINSWLEV SFCLPHKIH YAAS
SLIPPEAIERSLKAIRPYHALLLLSDEKSLLSELPIDCSPALVRVIKTTSAVKNLQQLAQ
DADLALLQVFQLAAHLVYWGKAVI I IYPLCENNVM SPNASVCLYSPLAEQFSRQFPSHD
LPSVLAKFSLPVSLSEFRSPLAPPAQETQLIQMVVWMLQRRLLIQLHTYVCLMASPSEEE
PRLREDDVPFTARVGGRSLSTPNALSFGSPTSSDDMTLTSPSMDNSSAELLPSGDSPLNK
RMTENLLASLSEHERAAILNVPAAQNPEDLRMFARLLHYFRGRHHLEEIMYNENTRRS QL
LMLFDKFRSVLVVTTHEDPVI AVFQALLT

> Gg NPRL3 (569aa)

MGESTSPIISVILVSSGSRGNKLLFRFPFQGAEH PAAQANKPRSRYA VNSSGDTSEDQDG
DSRFSDVILATILATKSDMCGKKFELKIDNVR FVGHP TLLQHALGQVSKTDPSPKREMP T
MILFNVVFALRANADPSVINCLHNLSRRIAIVLQHEERRCQYLTREAKLILAIQDEV SAM
SETTEGPQSPFHHILPKCKLARDLKETYS LCTTG VVRLHINN WLEV SFCLPHKIH YVAT
NFIPPEAIERSLKSI RPYHALLLLNDEKSLLNELPLDCSPALVRVIKTTSAVKNLQQLAQ
DADLALLQVFQLAAHLVYWGKAI I IYPLCENNVM LSPNASVCLYSPLADAFSCQFRGHN
LPSMLSKFSLPVSLSEFKNPLVPPVQETQLIQMVIWMLQHRLLIQLHTYVCLMVPPNEEE

FRAQDEDMPFTARVGGRLSTPNALSFGSPTSSDDMTLTSPSMDNSSAELIPGGDSPLNK
RMTENLLASLLEHEREAILNVPAAQNPEDLRMFARLLHYFRGRHHLEEIMYNENMRRSQL
LMLFDKFRSVLVVTSHEDPVISVFQSLLK

> Tg NPRL3 (567aa)

MGESTSPIISVILVSSGSRGNKLLFRFPFQRGAEHPAAQDNKPRSRVAVNGSGDTTEDQDG
DSRFSDVILATILATKSDMCGKKFELKIDNVRVFGHPTLLQHALGQVSKTDPSPKREMP
MILFNVVFALRANADPSVISCLHNLSRRIAIVLQHEERRCQYLTREAKLILAIQDEVSAM
SETTEGPQSPFHHILPKCKLARDLKETYSLSCTTGVRRLHINNWLVSFCLPHKIHVAT
NFIPPEAIERSLKSIRPYHALLLLNDKSLNELPLDCSPALVRVIKTTSAVKNLQQLAQ
DADLHCKKVFQLAAHLVYWGKAI I IYPLCENNVMYMLSPNASVCLYALKQGCSKCFTRADG
VLKFRVCTFPLLLTSLTSSLVSPLOTQLIQMVIWMLQHRLLIQLHTYVCLMVPNEEELR
APDEDMPFTARVGGRLSTPNALSFGSPTSSDDMTLTSPSMDNSSAELIPGGDSPLNKRM
TENLLASLLEHEREAILNVPAAQNPEDLRMFARLLHYFRGRHHLEEIMYNENMRRSQLLM
LFDKFRSVLVVTSHEDPVISVFQSLLK

> Tn NPRL3 (591aa)

MAFNPLAGFSDDHKSLYSHLQNTTYKSGEKTSPISVILVSSGSRGNRLLFRYPFQRVTEC
PASLAAKQRSRYALNTTGDVVEDQDGSDFSDIILATILATKSDICGKKFELKIDNVRV
GHPTLLQHPPVIQVSKTDPSPKREMPMILFNVVFALRANADPSVISCMHNLSRRLAIAL
QHEERRCQYLTREAKLMLAIQEEITTETDGNPQSPFRQILPKCKLARDLKEAYDSLCTTG
VVRRLHINNWLVSFCLPHKIHRIIGGSYIPPEALEQSLKAIRPYHTLLLESEKTLSQLPL
DCSPAMVRLIKTCSAVKNLQQLAQDADLALLQIFQIAAHLVYWGKAI I IYPLCENNVMYML
SPHANICLYSSLAQQFSQQFPGHDLPTMLAKFSLPVSLAEFRNPLEAPAQEAQLIQMVVW
MLQRLLIQLHTYVCLMVPSEDEPSARDEDPPIRVGGRLSTPSALSFGSPTSSDDMTL
TSPSMDNSSAELLPGGDSPLNKRITETLLASLSEHERQVILSIPAAQNPEDLRMFARLLH
YFRGHHHLEEIMYNENMRRSHLKTFLDKFRSVLVVTNHEDPVISIFQSPMD

> Dr NPRL3 (581aa)

MWKQSESSLQPNGEKTSPISVILVSSGSRGNKLLFRYPFQRASENTSSATSKQRSYPVLN
TSGDSTEDQDGSDFSDIILATILATKSDMCGKKFELKIDNVRVFGHPTLLQPPHTIQAS
KTDPSPKRELPTMILFSVVFALRANADASVISCMHNLSRRIAIALQHEERRCQYLTREAK
LMLVMQDEVTTITDSDGSPQSPFRQILPKCKLARDLKEAYDSLCTTGVRRLHINNWLVS
FCLPHKIHVRVGKHIPLAELERSLKAIRPYHALLLLENEKVLLAQLPLDCSPALVRLIKT
CSAVKNLQQLAQDADLALLQIFQIAAHLVYWGKAI I IYPLCENNVMYMLSPHANICIYSPL
AEHFAVQFPGHDLPSMLAKFSLPVSLSEFRNPLDAPVHEAHLIQMVVWMLQHRLLFQLHT
YVCLLVPPNEEEPGLRDEELPIVTRVTGRSLSTPSALSFGSPTSSDDMTLTSPSMDNSSA
ELQTGGDSPLNKRMETETLLASLTEHERQAILRVPAQNPEDLRLFARLLHYFRGHHHLEE
IMYNENLRRSQLKTFLDKFRSVLVITNHEDPIISLFQSPPE

> Lf NPRL3 (581aa)

MAGGDQTSPIISVILVSSGSRGNKLLFRYPFQQRKEQFQAATSKHRSPFALHTPSEPEDQD
GDVRFSDVILATILAPKSELGKRFELKIDNVLFVGHPTLLQHSQLPQVSKTDPSPKRET
PTMILFNVVFALRAHADPSVLGCMHNLSRRMAVALRHEERRCQYLTREAKLMLAVQDEIS
ALHDGSPPPSPFHHILPKSKLARDLKEAYDNLCTGLVQLYINNWLVSFCLPHKVHKIS
NNYISPEAIERSLKSIRPYHALLLLGEESLSQLPADSSPSLVRLIKVTSPMKTLLQQLA
QDADLALLQVFLAAHLVYWGKATVIYPLCETNVYMLSPSSNTYIHSPLYSEQFAQQFPGC
ELSTTLTEFSLPTPLAEHRNPLGTPVQQTQLVHTVMWMLQRRMLMQLHTYVCLMVPAGEE
SASSIGGGGGGGDGLQCAGSAARATGCAFSAPTALVFGSPSSDDLTLTSPSLEASSVE
LGACGEDSPVNRVTENLLASLPEHERSCINAVPAAHNPEDLRLFARLLPYFRGYHHLEE
IMYHENVRRSQLLTLDKFRSVLVVLTHEDPIISIFHSPT

>Pm NPRL3 (582aa)

MAGGDQTSPISVILVSSGSRGNKLLFRYPFQRQKEQFQAATSKHRSPFALHTPSEPEDQD
GDVRFSDVILATILAPKSELCKGRFELKIDNVLFVGHPTLLQHSQLPQVSKTDPSPKRET
PTMILFNVVFALRAHADPSVLGCMHNLSRRMAVALRHEERRCQYLTREAKMLLAVQDEIS
ALHDAGSPPSPFHHILPKSKLARDLKEAYDNLCSTGLVQLYINNWLEVSFCLPHKVHKI
SNNYISPEAIERSLKSIRPYHALLLLGEKSLLSQLPADSSPSLVRLIKVTSPMKTLQQL
AQDADLALLQVFQLAAHLVYWGKATVIYPLCETNVYMLSPSSNTYIHSPYSEQFAQQFPG
CELSTTLTEFSLPTPLAEHRNPLGTPVQQQTQLVHTVMWMLQRRMLMQLHTYVCLMVPAGE
ESASSSGSGGGGGDQELQCAGGVARAIACAFSAPTALVFGSPSSDDLTLTSPSLEASSV
ELGACGEDSPVNRRTVENLLASLPEHERSCINAVPAAHNPEDLRLFARLLPYFRGHHHLE
EIMYHENVRRSQLLTLIDKFRSVLVVLTHEDPIISIFHSPT

> Dm NPRL3 (610aa)

METNVNPLAVILVYFDSKGDRLLYRYPYQTLGQTEVANDEQRKSRKRNPYAVANTDDLQ
TPTHLGAAKSQGQLQGFADEVLSALFAVKPQLCNQKFELKLNDVRFVSHPTLI PQKEQRS
GPMAKQQMLINIVFALHAQASYSIVKCYHELKRLGLALKFEEQSRGYLTEQTAQMARTH
DEQQQQPLERTLELIAERCSLAQALRSIFHDLCTTGLLSTSLNHNLTLCFCLPAKAHQLH
KKGSMVDPETIDRCLRALKPYHGMLLLVDFAELLDCVPPTGARMLWQLVDVYDPLISLQS
MSSNADLSIEHVYKLVSHLVYWAKATIIYPLCETNVYVIAPDAPLHTKSHLVEKFSARFA
GMSLFEVISDFSLPTSIGHLTTPQQPARQGILAQMVIMWMLQHLLMQLHTYVQFMPSED
EFGDSASCSNHLRDAISDEEGDQEPDADELHGSMSSSHPLPVPAVLVGGHRREASEDH
SSLASDNIAVQPSSSHKSNFSITASMSTDNCDSLDSMEDEQKLKELLQVFSDADRAAIRR
IPASANVDDL SLLVKLYQMGYFKSEHHLEEIMYFENLRRSQLLQLLDKFRDVLIIYETED
PAIASMYNTK

Probes for Southern blots of *L. fluviatilis* genomic DNA (HS mapping)

PCR primers for generation of **probe a** from *L. fluviatilis* genomic DNA

First part of probe a:

ACAGGTTTGTTCAGCCAAGT (s) 386 bp product
ATTACTTGACTGGATATCGG (a)

>probe a, first part

ACAGGTTTGTTCAGCCAAGTAGAGGCAAAGCCTAGCCCGTACGAACATTAACGGTGATTG
TAAAAGCTCACCTTTGTTGAACCTGCAGTAATCCTATAGAAGTCGGAAGATTGGTCAATT
AAGTCAGTTCATCGACTACCCCTTTTGTGATCGATAACATTGGCCCTCTTAACAATGGCT
GCTGCCCCCACCATCCTACCTGTCTGAAGGTAGGAAACCCCGAGAGGCCACACCCAGCGTG
ACGTGACATGATGCCGATGGTGTCTACTAAGCCCTACCCATCACCAACTTTTAGCAGCAT
GAGTGAGAAGAGGGGTGAGGCTTCTTCAATATATTTACATGTCATCCAATCAGTCCACTA
AGCGGACCGATATCCAGTCAAGTAAT

Second part of probe a:

CAATTCCTTAAATCATCGGG (s) 209 bp product
AAACACCACGTTGAACAGGA (a)

>probe a, second part

CAATTCCTTAAATCATCGGGTAAATTGGCTATATAATCCTATCAATAACCCAGGTTATAC
AAATCTACACTGTTGGCGATACCTTGTGTCTGTATGAATTAGCGCACCGTTCTTTGTCAT
GTGAAGTTTGTCTTCTCAAAAGGTTTCCAAGACAGATCCATCTCCCAAACGAGAAACTC
CAACTATGATCCTGTTCAACGTGGTGTCT

PCR primers for generation of **probe b** from *L. fluviatilis* genomic DNA

GACAGTGGAACGACCAAACC (s) 919 bp product
ACAGAGGCTGGAGAAGGAAC (a)

>probe b

GACAGTGGAACGACCAAACCCCTGTCTCGGCAGCTAAAAGGTCAGACACCCTGTACAGGG
CGTGAAGGGGACAATGGAGTGGGCGGCATCGCTGTGCTTTCTGTTCTCTGCCACAAGCTG
TTAGTTTTGCGTGCTCTGCCGCCAATCAGAGAACGTGAATACGGTTGCTTTAGAGTCTCA
TGATGACGCTCCGCGGTGTTGAGGGGTGCAACCTTTTTTAATTCGCTCATTGTTGTGTTT
TTGCCCATGCAGGCAGTCCCTCCGCCTTCTCCCTTCCACCACATCCTGCCCCAAAGCAAGC
TGGCTCGGGACCTAAAGGAGGCCTATGACAAGTAAGCTTGCTTTGCCGCGCAGGGCACGA
CCAATCAGAGTGCATGAGGAATGCGGTGTGGTGGGCTCCTCTCGCAGATGTCTTTTAAT
CCCATTATTCTGGAACTCTTGTTACAAGAGAGAGAATGGATGGACTGAATAATTGCTTA
AGTGACAAGAGGCTACATGACTGCCCCCTTGTTGAATATGAACCATGTGCTTTCGTGTA
TTCAATAACCGCCCCCTATCAGAGGTGTTGAAGGCTTGTCGATTGTAAACACATGTGGCTC
TGTGTGAATCGTCTTATAACGTTATCATTCTTGCGAGCAGAAGAATTCAAATCTGAAAA
ATAGTGGGAACGCTGGGGCAGATAGTGTTAAGTAGTTATTACAATACATTACAAATGGG
GAAATCTTTAGTTGTTATACAGGGTGTGTCTTTTAAATTATAGCCAAAGTAACGTTATTT
CACTATACTTGGGTATATTTTTAAAGATGCACCCTGCTTAACCGGGTCTGGTGAGTTTT
AATTAACGTTTGGATGATGGTAGAACAGAATGAGATGCTTCTCTAAGTGCGTGCTGTGTG
TTCCTTCTCCAGCCTCTGT

Supplemental References

1. Lanfranchi G, Pallavicini A, Laveder P, Valle G. Ancestral hemoglobin switching in lampreys. *Dev Biol.* 1994;164(2):402-408.
2. Sambrook J, MacCallum P, Russell D. Molecular Cloning: A Laboratory Manual. 2001(third edition).
3. Smith JJ, Kuraku S, Holt C, et al. Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. *Nat Genet.* 2013;45(4):415-421.
4. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol.* 1997;268(1):78-94.
5. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25(17):3389-3402.
6. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403-410.
7. Loman NJ, Quinlan AR. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics.* 2014;30(23):3399-3401.
8. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 2017;27(5):722-736.
9. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods.* 2015;12(8):733-735.
10. Elnitski L, Riemer C, Schwartz S, Hardison R, Miller W. PipMaker: a World Wide Web server for genomic sequence alignments. *Curr Protoc Bioinformatics.* 2003;Chapter 10:Unit 10 12.
11. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol.* 2016;33(7):1870-1874.
12. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics.* 2011;27(8):1164-1165.
13. Brudno M, Do CB, Cooper GM, et al. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* 2003;13(4):721-731.
14. Mayor C, Brudno M, Schwartz JR, et al. VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics.* 2000;16(11):1046-1047.
15. van Dijk TB, Gillemans N, Pourfarzad F, et al. Fetal globin expression is regulated by Friend of Prmt1. *Blood.* 2010;116(20):4349-4352.
16. Gillemans N, McMorro T, Tewari R, et al. Functional and comparative analysis of globin loci in pufferfish and humans. *Blood.* 2003;101(7):2842-2849.
17. Ellis J, Tan-Un KC, Harper A, et al. A dominant chromatin-opening activity in 5' hypersensitive site 3 of the human beta-globin locus control region. *EMBO J.* 1996;15(3):562-568.
18. Smit AFA, Hubley R, Green P. Unpublished, Current Version: open-3.3.0 (RMLib: 20110419). 2011.
19. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol.* 2015;109:21 29 21-29.
20. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9(4):357-359.
21. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25(16):2078-2079.
22. Freese NH, Norris DC, Loraine AE. Integrated genome browser: visual analytics platform for genomics. *Bioinformatics.* 2016;32(14):2089-2095.
23. Parker HJ, Bronner ME, Krumlauf R. A Hox regulatory network of hindbrain segmentation is conserved to the base of vertebrates. *Nature.* 2014;514(7523):490-493.
24. Hockman D, Chong-Morrison V, Green SA, et al. A genome-wide assessment of the ancestral neural crest gene regulatory network. *Nat Commun.* 2019;10(1):4689.
25. Parker HJ, Sauka-Spengler T, Bronner M, Elgar G. A reporter assay in lamprey embryos reveals both functional conservation and elaboration of vertebrate enhancers. *PLoS One.* 2014;9(1):e85492.
26. Kawakami K. Transgenesis and gene trap methods in zebrafish by using the Tol2 transposable element. *Methods Cell Biol.* 2004;77:201-222.
27. Hoffmann FG, Opazo JC, Storz JF. Gene cooption and convergent evolution of oxygen transport hemoglobins in jawed and jawless vertebrates. *Proc Natl Acad Sci U S A.* 2010;107(32):14274-14279.
28. Schwarze K, Campbell KL, Hankeln T, Storz JF, Hoffmann FG, Burmester T. The globin gene repertoire of lampreys: convergent evolution of hemoglobin and myoglobin in jawed and jawless vertebrates. *Mol Biol Evol.* 2014;31(10):2708-2721.