Supporting Information

Material and Methods

Sequence Data. We retrieved the full complement of globin genes from the genome sequences of nine vertebrate taxa, including three teleost fish (medaka, *Oryzias latipes*; pufferfish, *Tetraodon nigroviridis*; and zebrafish, *Danio rerio*), one amphibian (western clawed frog, *Xenopus tropicalis*), one squamate reptile (green anole lizard, *Anolis carolinensis*), two birds (chicken, *Gallus gallus*; and zebra finch, *Taeniopygia guttata*), and two mammals (human, *Homo sapiens*; and platypus, *Ornithorhynchus anatinus*). The complete globin gene repertoire of each species was obtained by means of bioinformatic searches in the Genbank or Ensembl (release 55) databases. We broadened our phylogenetic coverage by adding globin sequences derived from mRNA or protein records from representative cartilaginous fish (class Chondrichthyes), the most basal lineage of extant gnathostomes, and from cyclostomes, the sister group to gnathostome vertebrates. In the case of cartilaginous fish, we obtained α- and β-Hb sequences from the red stingray (*Dasyatis akajei*) and gummy houndshark (*Mustelus antarcticus*), as well as Mb sequences from the latter species and the Port Jackson shark (*Heterodontus portusjacksoni*). In the case of cyclostomes, we included 12 sequences of functional Hbs from three representatives of subclasses Myxini and Hyperoartia: Five paralogous sequences from the sea lamprey (*Petromyzon marinus*, Hyperoartia), three from the Arctic lamprey (*Lethenteron japonicum*, Hyperoartia), and four from the hagfish (*Myxine glutinosa,* Myxini). We did not identify any previously undescribed globin genes in the Ensembl pre-release assembly of the sea lamprey genome. In addition, we included the previously reported globins from the sea squirt, *Ciona intestinalis* (1).

SI Table 1.
Sequences used in this study with the corresponding accession numbers.

Sequence Name	Accesion number	Source
Anole lizard Cygb	ENSACAG00000008394*	Ensembl
Anole lizard GbY	AAWZ01045931*	Ensembl
Anole lizard Hba	ENSACAG00000016421	Ensembl
Anole lizard Hbb	ENSACAG00000012173	Ensembl
Anole lizard Mb	ENSACAG00000016595	Ensembl
Chicken Cygb	NM_001008789	GenBank
Chicken GbE	NM_001008786	GenBank
Chicken Hba	NM_001004376	GenBank
Chicken Hbb	NM_001081704	GenBank
Chicken Mb	XM_416292	GenBank
Frog Cygb	NM_001006869	GenBank
Frog GbX	NP_001011196	GenBank
Frog GbY	BC158411	GenBank
Frog Hba	NM_203529	GenBank
Frog Hbb	NM_203528	GenBank
Frog Ngb	ENSXETG00000027106	Ensembl
Hagfish Hb1	AF156936	GenBank
Hagfish Hb2	AF157494	GenBank
Hagfish Hb3	AF184047	GenBank
Hagfish Hb4	AF184239	GenBank
Houndshark Hba	BAA75399	GenBank
Houndshark Hbb	BAA75400	GenBank
Houndshark Mb	P14399	GenBank
Human Cygb	NM_134268	GenBank
Human Hba	NM_000558	GenBank
Human Hbb	NM_000518	GenBank
Human Mb	NG_007075	GenBank
Human Ngb	NP_067080	GenBank
Lethenteron Hb 1	AB294235	GenBank
Lethenteron Hb 2	AB294236	GenBank
Lethenteron Hb 4	AB294237	GenBank
Medaka Cygb-1	NM_001104767	GenBank
Medaka Cygb-2	NM_001104768	GenBank
Medaka GbX	ENSORLG00000017054	Ensembl
Medaka Hba	ENSORLG00000005267	Ensembl
Medaka Hbb	ENSORLG00000003020	Ensembl
Medaka Mb	ENSORLG00000004130	Ensembl
Medaka Ngb	ENSORLT00000020359	Ensembl
Platypus GbY	AC203513	GenBank
Platypus Hba	AC203513	GenBank
Platypus Hbb	AC190020	GenBank
Platypus Hbw	AC203513	GenBank
Platypus Mb	XM_001513063	GenBank
Platypus Ngb	XP_001508417	GenBank
Port Jackson shark Mb	P02206	GenBank
Pufferfish Cygb-1	AJ635230	GenBank
Pufferfish Cygb-2	AJ635231	GenBank
Pufferfish GbX	CAG25725	GenBank
Pufferfish Hba	ENSTNIG00000018576	Ensembl
Pufferfish Hbb	ENSTNIG00000012913	Ensembl
Pufferfish Mb	ENSTNIG00000005518	Ensembl
Pufferfish Ngb	CAC59974	GenBank
Sea lamprey Hb PMII	AF248645	GenBank
Sea lamprey Hb1a	P09967	GenBank
Sea lamprey Hb1b	P21197	GenBank
Sea lamprey Hb2	Q9I9I3	GenBank
Sea lamprey Hb3	P09968	GenBank
Sea lamprey Hb5	P02208	GenBank
Sea squirt Glb1	CAD68145	GenBank
Sea squirt Glb2	CAD68146	GenBank
Sea squirt Glb3	CAD68147	GenBank
Sea squirt Glb4	CAD89600	GenBank
Stingray Hba	BAA75249	GenBank
Stingray Hbb	BAA75250	GenBank
Zebra finch Cygb	XM_002195407	GenBank
Zebra finch GbE	XM_002196350	GenBank
Zebra finch Hba	XM_002196096	GenBank
Zebra finch Hbb	XM_002190485	GenBank
Zebra finch Mb	XM_002199380	GenBank
Zebrafish Cygb1	NM_152952	GenBank
Zebrafish Cygb2	NM_001024224	GenBank
Zebrafish GbX	CAG25723	GenBank
Zebrafish Hba	NM_131257	GenBank
Zebrafish Hbb	NM_131020	GenBank
Zebrafish Mb	NM_200586	GenBank
Zebrafish Ngb	NP_571928	GenBank

* These sequences were re-annotated manually.

Results

Sensitivity analysis. Because changes in sequence alignment and substitution model are known to influence the results of phylogenetic analyses (2, 3),we explored sensitivity of our results to variation in alignment method, substitution model, and choice of outgroup. We aligned sequences with 10 alternative methods: Dialign (4), Kalign2 (5), the E-INS-i, G-INS-i, and L-INS-i strategies from Mafft v6.17 (6), Muscle v3.5 (7), Prank (8), Probalign (9), Probcons (10), and PROMALS3d (11). For each alignment we performed maximum likelihood searches under the JTT (12), LG (13), and mixed models of amino acid substitution and Bayesian analyses under the JTT (12), and mixed models of amino acid substitution. Maximum likelihood searches were implemented in Treefinder version October 2008 (14), and support for the nodes was evaluated with 1,000 bootstrap pseudoreplicates. Bayesian analyses were conducted using MrBayes version 3.1.2 (15), setting two independent runs of four simultaneous chains for 10,000,000 generations, sampling every 2,500 generations, and using default priors. Once convergence was verified, support for the nodes and parameter estimates were derived from a majority rule consensus of the last 2,500 trees. In maximum likelihood we used constrained searches to compare the likelihood scores of the 'single co-option' hypothesis (Fig. 2A, SI Fig 1A), the 'parallel co-option or single co-option/secondary loss' hypothesis (Fig. 2B, SI Fig 1B), and the 'convergent co-option' hypothesis (Fig. 2C, SI Fig 1C). Finally, we added vertebrate Globin X sequences and *Ciona* globin sequences to the alignment as additional outgroup sequences. A data file containing the complete set of sequence alignments is provided in the Supporting Information online.

SI Table 2. Results of the sensitivity analysis. Maximum likelihood scores of the best unconstrained tree, and the three competing hypotheses of globin gene family evolution, plus support for the node joining cyclostome hemoglobins with gnathostome cytoglobin. This first set of analyses included all the vertebrate-specific globins, plus six vertebrate neuroglobin sequences as outgroup sequences.

Data: vertebrate-specific globins + Neuroglobin

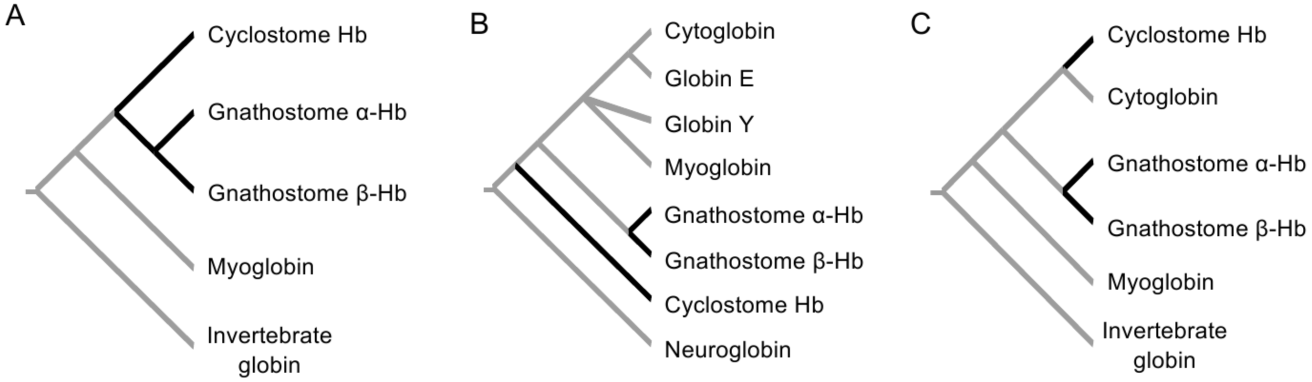
model = JTT		Likelihood scores			Support for the node joining cyclostome Hbs and cytoglobin	
Alignment	best tree	'single co-option'	'parallel co-option or single co-option/secondary loss'	'convergent co-option'	ML bs	MrBayes pp
dialign	-13480.4	-13491.6	-13491.3	-13481.1	63%	1.00
kalign	-13354.1	-13366.1	-13361.7	-13354.1	58%	1.00
mafft_einsi	-13351.2	-13361.6	-13361.0	-13351.2	65%	1.00
mafft_ginsi	-13325.5	-13332.5	-13331.0	-13322.7	66%	1.00
mafft_linsi	-13351.2	-13361.6	-13361.0	-13351.2	64%	1.00
muscle	-13417.7	-13430.3	-13426.8	-13416.7	76%	1.00
prank	-13255.0	-13265.8	-13262.3	-13255.0	73%	1.00
probalign	-13476.1	-13487.4	-13486.1	-13475.6	65%	1.00
probcons	-13433.3	-13445.6	-13442.7	-13433.3	59%	1.00
promal_wPDB	-13512.2	-13522.4	-13520.9	-13512.8	65%	1.00
model = LG						
dialign	-13434.8	-13443.9	-13444.6	-13429.3	54%	--
kalign	-13293.4	-13300.7	-13300.0	-13296.8	< 50%	--
mafft_einsi	-13279.8	-13291.0	-13289.1	-13279.8	57%	--
mafft_ginsi	-13253.7	-13261.1	-13260.3	-13252.3	58%	--
mafft_linsi	-13279.8	-13291.0	-13289.1	-13279.8	54%	--
muscle	-13355.7	-13367.2	-13364.4	-13356.2	67%	--
prank	-13197.1	-13205.3	-13201.3	-13197.1	68%	--
probalign	-13415.2	-13425.5	-13423.6	-13417.1	59%	--
probcons	-13362.0	-13370.5	-13369.1	-13362.0	< 50%	--
promal_wPDB	-13457.0	-13465.8	-13464.4	-13457.0	58%	--
model = mixed						
dialign	-13406.7	-13417.1	-13415.2	-13406.9	57%	1.00
kalign	-13279.1	-13286.0	-13285.2	-13279.1	54%	0.99
mafft_einsi	-13272.0	-13281.6	-13280.7	-13272.4	59%	1.00
mafft_ginsi	-13225.4	-13235.0	-13234.0	-13225.2	59%	1.00
mafft_linsi	-13272.0	-13281.6	-13280.7	-13272.4	55%	0.99
muscle	-13339.6	-13350.9	-13347.8	-13340.1	70%	1.00
prank	-13151.6	-13162.7	-13159.1	-13151.6	70%	0.99
probalign	-13388.9	-13399.6	-13397.1	-13388.9	63%	1.00
probcons	-13351.5	-13357.4	-13357.5	-13349.7	50%	0.99
promal_wPDB	-13428.4	-13437.9	-13436.0	-13428.5	60%	1.00

SI Table 3. Results of the sensitivity analysis. Maximum likelihood scores of the best unconstrained tree, and the three competing hypotheses of globin gene family evolution. This second set of analyses included all the vertebrate-specific globins, plus four vertebrate Globin X sequences, four *Ciona* globin sequences, and the six vertebrate neuroglobin sequences as outgroup sequences.

Data: vertebrate-specific globins + Neuroglobin + Globin X + *Ciona* globins

model = JTT			Likelihood scores	
Alignment	best tree	'single co-option'	'parallel co-option or single co-option/secondary loss'	'convergent co-option'
dialign	-16406.3	-16412.3	-16405.0	-16402.3
kalign	-16279.1	-16284.0	-16281.2	-16278.6
mafft_einsi	-16253.2	-16260.3	-16254.8	-16254.0
mafft_ginsi	-16221.8	-16227.5	-16225.9	-16221.0
mafft_linsi	-16264.2	-16268.7	-16267.4	-16262.4
muscle	-16521.8	-16529.8	-16523.5	-16521.6
prank	-16143.5	-16150.7	-16143.0	-16142.1
probalign	-16513.6	-16521.5	-16516.1	-16514.5
probcons	-16449.4	-16453.5	-16450.6	-16449.4
promal_wPDB	-16440.8	-16448.5	-16442.3	-16440.8
model = LG				
dialign	-16349.3	-16349.9	-16350.5	-16342.6
kalign	-16212.4	-16218.3	-16208.2	-16206.9
mafft_einsi	-16178.6	-16185.1	-16179.3	-16178.6
mafft_ginsi	-16146.3	-16153.1	-16147.0	-16146.3
mafft_linsi	-16187.7	-16194.6	-16188.3	-16187.7
muscle	-16462.1	-16471.0	-16463.0	-16462.1
prank	-16069.9	-16078.5	-16070.3	-16069.5
probalign	-16447.7	-16454.6	-16448.0	-16446.7
probcons	-16373.4	-16377.8	-16374.1	-16374.0
promal_wPDB	-16382.0	-16382.0	-16378.5	-16377.6
model = mixed				
dialign	-16304.7	-16313.8	-16312.6	-16305.4
kalign	-16192.1	-16199.0	-16193.6	-16192.1
mafft_einsi	-16151.7	-16160.6	-16152.4	-16151.7
mafft_ginsi	-16113.0	-16120.3	-16113.6	-16113.0
mafft_linsi	-16154.2	-16160.5	-16154.9	-16154.2
muscle	-16425.4	-16433.6	-16427.0	-16425.1
prank	-16028.3	-16037.2	-16028.1	-16028.0
probalign	-16412.6	-16417.3	-16412.8	-16410.9
probcons	-16341.8	-16348.9	-16343.4	-16341.8
promal_wPDB	-16349.0	-16352.5	-16347.8	-16346.9

SI Fig 1. Previous hypotheses regarding phylogenetic relationships between gnathostome and cyclostome Hbs. According to the traditional view (A), the Hbs of cyclostomes and gnathostomes are orthologous proteins that derive from a proto-Hb precursor protein that had evolved an O_2-transport function in the vertebrate common ancestor (16). Under the 'parallel co-option or single co-option/secondary loss' hypothesis (B), the O_2-transport function evolved independently in both Hb lineages, or alternatively, an ancestral O_2-transport function was secondarily lost in the remaining gnathostome-specific globins (17). Finally, under the 'convergent co-option' hypothesis (C), O_2-transport functions evolved independently in the Hbs of cyclostomes and gnathostomes, or alternatively, an ancestral O_2-transport function was secondarily lost in gnathostome cytoglobin (18).



References

1 Ebner B, Burmester, T, Hankeln, T (2003) Globin genes are present in *Ciona intestinalis*. *Mol Biol Evol* 20:1521-1525.

2 Sullivan J, Joyce P (2005) Model Selection in Phylogenetics. *Annu Rev Ecol Evol Syst* 36:445-466.

3 Wong KM, Suchard MA, Huelsenbeck JP (2008) Alignment uncertainty and genomic analysis. *Science* 319:473-476.

4 Morgenstern B (2004) DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. *Nucleic Acids Res* 32:33-36.

5 Lassmann T, Frings O, Sonnhammer ELL (2009) Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic Acids Res* 37:858-865.

6 Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* 9:286-298.

7 Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792-1797.

8 Löytynoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci USA* 102:10557-10562.

9 Chikkagoudar S, Roshan U, Livesay D (2007) eProbalign: generation and manipulation of multiple sequence alignments using partition function posterior probabilities. *Nucleic Acids Res* 35:675-677.

10 Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res* 15:330-340.

11 Pei J, Tang M, Grishin NV (2008) PROMALS3D web server for accurate multiple protein sequence and structure alignments. *Nucleic Acids Res* 36:30-34.

12 Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8:275-282.

13 Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. *Mol Biol Evol* 25:1307-1320.

14 Jobb G, von Haeseler A, Strimmer K (2004) TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol* 4:18.

15 Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572-1574.

16 Goodman M, Pedwaydon J, Czelusniak J, Suzuki T, Gotoh T, Moens L, Shishikura F, Walz D, Vinogradov S (1988) An evolutionary tree for invertebrate globin sequences. *J Mol Evol* 27:236-249.

17 Burmester T, Hankeln T (2009) What is the function of neuroglobin? *J Exp Biol* 212:1423-1428.

18 Katoh K, Miyata T (2002) Cyclostome hemoglobins are possibly paralogous to gnathostome hemoglobins. *J Mol Evol* 55:246-249.