Evolutionary conservation and diversification of Rh family genes and proteins

Cheng-Han Huang* and Jianbin Peng

Laboratory of Biochemistry and Molecular Genetics, Lindsley F. Kimball Research Institute, New York Blood Center, 310 East 67th Street,
New York, NY 10021

Communicated by Sydney Kustu, University of California, Berkeley, CA, September 9, 2005 (received for review May 19, 2005)

Rhesus (Rh) proteins were first identified in human erythroid cells and recently in other tissues. Like ammonia transporter (Amt) proteins, their only homologues, Rh proteins have the 12 transmembrane-spanning segments characteristic of transporters. Many think Rh and Amt proteins transport the same substrate, NH_3/NH_4^+, whereas others think that Rh proteins transport CO_2 and Amt proteins NH_3. In the latter view, Rh and Amt are different biological gas channels. To reconstruct the phylogeny of the Rh family and study its coexistence with and relationship to Amt in depth, we analyzed 111 Rh genes and 260 Amt genes. Although Rh and Amt are found together in organisms as diverse as unicellular eukaryotes and sea squirts, Rh genes apparently arose later, because they are rare in prokaryotes. However, Rh genes are prominent in vertebrates, in which Amt genes disappear. In organisms with both types of genes, Rh had apparently diverged away from Amt rapidly and then evolved slowly over a long period. Functionally divergent amino acid sites are clustered in transmembrane segments and around the gas-conducting lumen recently identified in *Escherichia coli* AmtB, in agreement with Rh proteins having new substrate specificity. Despite gene duplications and mutations, the Rh paralogous groups all have apparently been subject to strong purifying selection indicating functional conservation. Genes encoding the classical Rh proteins in mammalian red cells show higher nucleotide substitution rates at nonsynonymous codon positions than other Rh genes, a finding that suggests a possible role for these proteins in red cell morphogenetic evolution.

CO_2 channel | membrane proteins

Although the first Rhesus (Rh) protein was detected in human erythroid cells in 1939 (1), it has only recently been established that there are at least four Rh proteins in mammals, Rh30 and RhAG in red cells and RhBG and RhCG in other tissues (2–7). Rh homologues have also been found in simpler organisms, but relatively few have been identified and hence the origin and evolutionary history of Rh proteins remains elusive.

Rh proteins have 12 transmembrane (TM)-passing segments indicative of a transport function (2–7) with limited homology to microbial ammonium transporter (Amt) proteins first noticed by Marini *et al.* (8). Many research groups think that Amt proteins concentrate the $NH_4{}^+$ ion against a gradient, i.e., that they are NH_4^+ active transporters (9). Likewise, several groups think that human and mouse Rh proteins also transport ammonium and are Amt functional equivalents in mammals (10–16). Both findings have been challenged. Soupene *et al.* (17–20) think that Amt proteins are gas channels for NH_3, a view that has been substantiated by the high-resolution protein structures of *Escherichia coli* AmtB (EcAmtB) (21, 22). Moreover, Soupene *et al.* find that the substrate for the Rh1 protein of the green alga *Chlamydomonas reinhardtii*, is apparently CO_2 (23–25). They focused on this organism because it was one of the few microbes previously known to have an Rh protein (7, 23).

To probe the evolutionary history of Rh and Amt genes in depth we assembled the sequences of 111 Rh and 260 Amt and analyzed them phylogenetically and bioinformatically. Using this large data set, we explored particularly (*i*) the organismal distribution of Rh genes as to how often and widespread they coexisted with Amt in the same species (paralogous occurrence); (*ii*) whether there were distinct differences between Rh and Amt proteins, supporting physiological and genetic evidence that they have different substrate specificities (24, 25); and (*iii*) proliferation of Rh genes over evolutionary time and the degree of their conservation. Our data are consistent with functional conservation within the Rh family and functional diversification of Rh proteins from the distantly related Amt proteins.

Materials and Methods

Data Sets. Accession numbers and identifiers of Rh and Amt can be found in the supporting information, which is published on the PNAS web site. The Rh data set contains 111 nonredundant genes mostly of full-length cDNAs (see Table 1, which is published as supporting information on the PNAS web site). Mammalian Rh30 and RhAG were from GenBank via BLAST search (26); other Rh genes were mainly cloned in our laboratory. The Amt data set contains 260 nonredundant genes mostly retrieved from annotated GenBank entries (see the Amt data set, which is published as supporting information on the PNAS web site).

Sequence Alignment. Rh and/or Amt protein sequence alignments were obtained by using MUSCLE (Version 3.52; ref. 27) and were used to derive codon-based nucleotide sequence alignments. Homogeneities of amino acid or codon composition were measured by disparity index (28) as described in MEGA 3.0 (29). A biased codon usage in the first and third positions was noticed.

Phylogenetic Analysis. The Rh/Amt joint tree was reconstructed by using the maximum likelihood (ML) method as implemented in PHYML (Version 2004; ref. 30) under the Jones–Taylor–Thornton (JTT) + 4G (four categories of Gamma substitution rates) + I (invariable sites) model (31). The gene tree for coexisting Rh and Amt was reconstructed by using PHYML (29) and the Bayesian inference (BI) method MRBAYES (Version 3.0; ref. 32), and was rooted with EcAmt as an arbitrary outgroup. The BI gene tree for the Rh family was built as above and rooted with NeRh from *Nitrosomonas europaea*, which is the lowest in species order. To curtail codon bias in reconstructing this tree, first and third codon positions were converted to purines or pyrimidines (R/Y-coded), whereas nucleotides at second positions were retained. The model of two-state substitution + 4G + I was applied to the first and third

Abbreviations: 4G, four categories of Gamma substitution rates; Amt, ammonium transporter; BI, Bayesian inference; I, invariable sites; ML, maximum likelihood; PP, posterior probability; Rh, Rhesus; TM, transmembrane.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AF398238, AF447925, AF510715, AF529360, AF531094–AF531097, AY013262, AY116072–AY116077, AY129071–AY129073, AY139091, AY198126–AY198128, AY207445, AY227357, AY271818, AY332758, AY340237, AY353246, AY353247, AY363116, AY363117, AY377923, AY455819, AY613958, AY613959, AY618933, AY618934, AY619986, AY622224, AY622225, AY831675–AY831678, AY865609–AY865618, DQ011226, and DQ013062).

*To whom correspondence should be addressed. E-mail: chuang@nybloodcenter.org.

codon positions, and the general time reversible (GTR) + 4G + I model was applied to the second positions. The ML tree of the Rh gene family was reconstructed by using PHYML. The GTR + 4G + I model was applied to the original (non-R/Y-coded) first and second codon positions, and a nonparametric bootstrap test was conducted at 500 replicates.

Modeling the 3D Structure of Rh on Amt. The 3D structure of Rh was simulated on the template of EcAmtB (PDB entry 1U7G; ref. 21) by using MODELLER (Version 8.1; ref. 33). The resultant structures were prepared by using PYMOL (Version 0.98; http://pymol.sourceforge.net).

Functional Divergence and Substitution Patterns. Type I divergence between Rh clusters was measured by the coefficient of functional divergence, θ_λ, as implemented in DIVERGE 1.04 (34). Given that θ_λ tests are biased by alignment error, care was taken to have the alignment as reliable as possible. Significant evolutionary rate shift sites between Rh clusters were identified by using the more rigorous likelihood-ratio tests (35). These sites were visualized by bar chart and sequence logo. Ratios of nonsynonymous versus synonymous substitution rates (d_N/d_S) were calculated to assess substitution patterns, using a model-based ML approach as implemented in PAML 3.12 (36).

Results

Rh Distribution and Coexistence with Amt. Although Rh is extremely rare in bacteria and does not appear to be found in archaea or vascular plants, the data set shows a wide distribution of Rh genes in 41 species from the chemolithoautotroph *N. europaea* to humans. Rh and Amt genes are found together in organisms as diverse as unicellular eukaryotic microbes (e.g., green alga, slime mold, and water molds) and invertebrate animals (e.g., nematodes, arthropods, echinoderms, and ascidians) (see Table 2, which is published as supporting information on the PNAS web site) and thus appear to have coexisted for a long period of evolutionary time. All of the vertebrate animals examined have multiple Rh genes but lack Amt genes.

Rh and Amt Are Distantly Related. The joint tree of Rh and Amt proteins from all domains of life revealed two far-separated clusters with no interim or mixed branches (Fig. 1*A*). The tree built from 17 Rh and 25 Amt from organisms with both showed again that the two groups are distantly related (Fig. 1*B*) [mean identity = 14%, a value similar to that of a small nonparalogous data set (8)]. The ancestral Rh gene, which could be inferred from Rhp1 group, apparently went through a rapid evolution and then long periods of purifying selection from unicellular eukaryotes to ascidians.

Analysis of Rh and Amt proteins from organisms with both revealed 96 significant sites of evolutionary rate shift, providing evidence for functional diversification (Fig. 1*C*; see the alignment of coexistent Rh and Amt, which is published as supporting information on the PNAS web site). A larger sampling from organisms lacking one or the other showed a nearly identical pattern with a few more sites (data not shown). Twenty-two of the 96 sites are conserved in each group with different amino acid residues. For mutually exclusive conserved vs. varied sites, Rh has 42 conserved and 32 varied sites, and vice versa for Amt. Some involve different charges in the two groups (Fig. 1*C*, logo), but a total of 70 sites (70/96 = 73%) reside in TM spans (62 sites) and TM turns (8 sites) (e.g., Fig. 1*D Right*). To test whether such changes in the primary sequence affect higher-order structure, a model of Rh was built on the EcAmtB template (21) with rate shift sites mapped (shown is CiRh from *Ciona intestinalis*). The CiRh model differs from EcAmtB by 17% rate shift sites (16/96) in secondary structures (Fig. 1*D*; see the supplementary data for CiRh modeling, which are published as supporting information on the PNAS web site). Furthermore, Rh differs from Amt in three striking ways (Fig. 1*D*). (*i*) Rh has rate shift sites clustered in TM segments equivalent to TM2, -4, -7, and -8 of Amt. (*ii*) Some of the 20 residues lining the channel lumen are missing or lie in close proximity to rate shift sites. (*iii*) All Rh proteins share a unique TM1 and absolutely conserved signature motifs absent from Amt (7). The coexistence, distant relationship, evolutionary rate shift, and structural differences suggest that Rh proteins could have evolved a new function from their ancestor(s) related to ancient Amt genes.

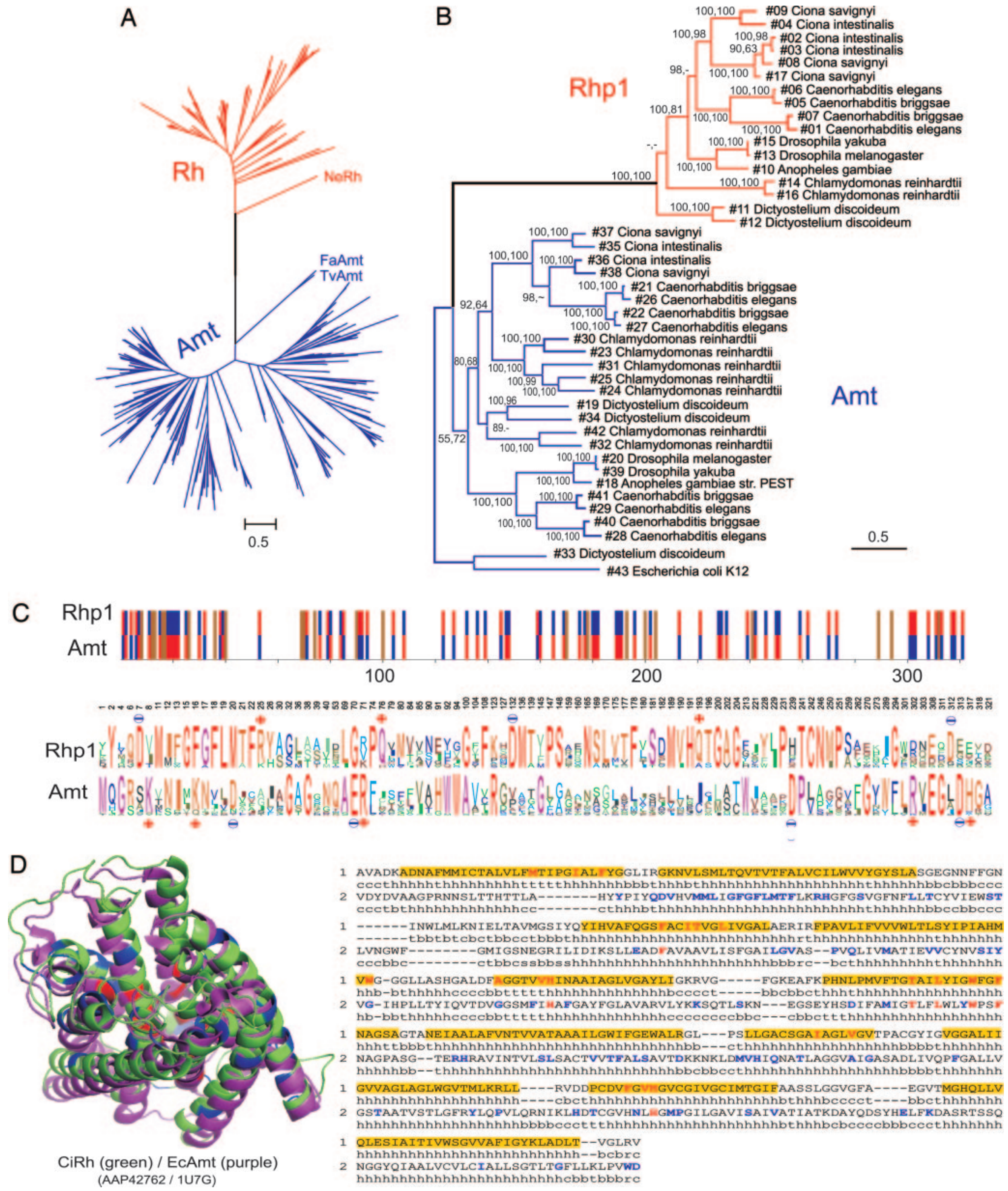
Rh Gene Clusters. Having shown the divergence of Rh from Amt, we examined the evolutionary pathway of the Rh family in detail (Fig. 2; see the alignment of the 111 Rh protein sequences, which is published as supporting information on the PNAS web site). Four well separated clusters, Rh30, RhAG, RhBG, and RhCG were found as the main paralogous groups in vertebrates from fish to mammals. In addition, two primitive clusters, Rhp1 and Rhp2 containing members from microbes to invertebrates and non-mammalian vertebrates, respectively, extend the lower part of the tree. With few exceptions, the tree topology for the Rh family is congruent with known species orders.

The red cell-specific clusters RhAG and Rh30 share a common ancestor, but in all organisms from zebrafish to humans the Rh30 cluster has longer branches (Fig. 2). Hence, the Rh30 cluster underwent fast evolution after the separation of its ancestor from an ancestral RhAG early in fish speciation. Likewise, the nonerythroid clusters RhBG and RhCG emerged from another common ancestor, and they were subject to slow evolution, with RhBG in chicken (the only bird examined) as an apparent exception (Fig. 2). The erythroid and nonerythroid sister groups appear to have evolved in parallel and probably diverged from an ancestral gene similar to one in the Rhp group.

Phylogenetic analysis and data mining yielded additional insight into the origin and duplications of Rh genes (Fig. 2). (*i*) Some vertebrates have extra paralogues: like human or chimp, chicken has two Rh30 genes. Fish have two RhCG genes. (*ii*) A novel gene occurs in fish, frog, and chicken (not shown on the tree), and its members define Rhp2 as a special non-mammalian cluster. This cluster is probably ancient because it is expressed specifically in the gut, the oldest organ, and, apart from one intron in *Danio rerio*, it lacks introns. In addition, the Rhp2 cluster is more closely linked than other Rh genes to NeRh and the Rhp1 genes. (*iii*) As a most diverse cluster, Rhp1 consists of members from organisms that have both Rh and Amt genes. Species of the group have from one to three Rh genes, the latter having arisen from gene duplication events (Fig. 2). In *Ciona* the three genes arose by two steps of duplication: one early interspecific event and one late intraspecific event. The above findings together pinpoint the complexity in the birth and death of Rh genes and establish that the four-gene framework did not develop until the initiation of vertebrate speciation.

Divergence and d_N/d_S Ratios Between Rh Clusters. Estimation of θ_λ, the coefficient of functional divergence, reveals the varying degrees of divergence at the cluster level (Fig. 3*A*; see the alignment of the 111 Rh protein sequences, which is published as supporting information on the PNAS web site). The lowest divergence is between RhBG and RhCG (0.36), and the highest is between Rh30 and RhBG (0.71)/RhCG (0.69). RhAG shows much higher divergence with Rh30 (0.59) than with RhBG or RhCG (0.38). The divergence of RhAG with RhBG or RhCG is at the same level; likewise, the divergence of Rh30 with RhBG or RhCG is also comparable. These data indicated that RhBG and RhCG are evolutionally more conserved than red cell paralogues and that Rh30 may have diverged for a red cell-specific functional modification.

The d_N/d_S ratios were computed to detect positive selection over codon sites. Despite a few sites in Rh30, the averaged ratio for each of the four cluster is <1, thus signifying purifying selection (function constraint) on the Rh family as a whole. The plot of nonsynonymous


A
Rh
NeRh
FaAmt
TvAmt
Amt
0.5
B
Rhp1
100,100 #09 Ciona savignyi
#04 Ciona intestinalis
100,98
100,98 #02 Ciona intestinalis
90,63 #03 Ciona intestinalis
#08 Ciona savignyi
98,-
100,100 #17 Ciona savignyi
100,100 #06 Caenorhabditis elegans
#05 Caenorhabditis briggsae
100,81
100,100 #07 Caenorhabditis briggsae
100,100 #01 Caenorhabditis elegans
-,-
100,100 #15 Drosophila yakuba
#13 Drosophila melanogaster
100,100 #10 Anopheles gambiae
100,100
100,100 #14 Chlamydomonas reinhardtii
#16 Chlamydomonas reinhardtii
100,100 #11 Dictyostelium discoideum
#12 Dictyostelium discoideum
100,100 #37 Ciona savignyi
#35 Ciona intestinalis
100,100
100,100 #36 Ciona intestinalis
#38 Ciona savignyi
100,100 #21 Caenorhabditis briggsae
98,~ #26 Caenorhabditis elegans
92,64
100,100 #22 Caenorhabditis briggsae
100,100 #27 Caenorhabditis elegans
100,100 #30 Chlamydomonas reinhardtii
#23 Chlamydomonas reinhardtii
80,68
100,100 #31 Chlamydomonas reinhardtii
100,99 #25 Chlamydomonas reinhardtii
100,100 #24 Chlamydomonas reinhardtii
100,96 #19 Dictyostelium discoideum
#34 Dictyostelium discoideum
89.- #42 Chlamydomonas reinhardtii
55,72
100,100 #32 Chlamydomonas reinhardtii
100,100 #20 Drosophila melanogaster
#39 Drosophila yakuba
100,100 #18 Anopheles gambiae str. PEST
100,100 #41 Caenorhabditis briggsae
100,100 #29 Caenorhabditis elegans
100,100 #40 Caenorhabditis briggsae
100,100 #28 Caenorhabditis elegans
#33 Dictyostelium discoideum
#43 Escherichia coli K12
Amt
0.5
C
Rhp1
Amt
100
200
300
Rhp1
Amt
D
CiRh (green) / EcAmt (purple)
(AAP42762 / 1U7G)

```
1 AVADKADNAFMMICTALVLFMTIPGIALFYGGLIRGKNVLSMLTQVTVTFALVCILWVVYGYSLASGEGNNFFGN
  cccthhhhhthhhhhhhhhhttttthhhhhhhbbbthhhbthhhhhhhhhhhhhhhhhhhhthhhhhhbbcbbbccc
2 VDYDVAAGPRNNSLTTHTTLA-------HYYPIYQDVHVMMLIGFGFLMTFLKRHGFGSVGFNFLLTCYVIEWST
  ccctbthhhhhhhhhhhhhcc-------cthhhbthhhhhhhhhhhhhhhhhhhhhhhhhhthhhhhhbbccbbccc

1 ------INWLMLKNIELTAVMGSIYQYIHVAFQGSFACITVGLIVGALAERIRFPAVLIFVVVWLTLSYIPIAHM
  ------tbbtbtcbctbbcctbbbthhhhhhhhhhhhhhhhhhhhhhhhhhrthhhhhhhhhhhhhhthhhhhhh
2 LVNGWF--------GMIGSNEGRILIDIKSLLEADFAVAAVLISFGAILGVAS--PVQLIVMATIEVVCYNVSIY
  cccbbc--------ctbbcssbbsshhhhhhhhhhhhhhhhhhhhhhhbbbrc--bcthhhhhhhhhthhhhhhh

1 VWG-GGLLASHGALDFAGGTVVHINAAIAGLVGAYLIGKRVG-----FGKEAFKPHNLPMVFTGTAILYIGWFGF
  hhb-bthhhhhccccbbtttttthhhhhhhhhhhhhhbcccct-----bbcbbcthhhthhhhhhhhhhhhhhhh
2 VG-IHPLLTYIQVTDVGGSMFIHAFGAYFGLAVARVLYKKSQTLSKN---EGSEYHSDIFAMIGTLFLWLYWPSF
  hb-bbthhhhhccccbbtttttthhhhhhhhhhhhhhcccccccccbbc---cbbctttthhhhhhhhhhhhhhhh

1 NAGSAGTANEIAALAFVNTVVATAAAILGWIFGEWALRGL---PSLLGACSGAIAGLVGVTPACGYIGVGGALII
  hhhttbbbthhhhhhhhhhhhhhhhhhhhhhhhhhhhhbb---cthhhhhhhhhhhhhhhhhhhbbbbthhhhhhh
2 NAGPASG--TERHRAVINTVLSLSACTVVTFALSAVTDKKNKLDMVHIQNATLAGGVAIGASADLIVQPFGALLV
  hhhhhbb--thhhhhhhhhhhhhhhhhhhhhhhhhhhbbcbbccthhhhhhhhhhhhhhhhhhbbbbthhhhhhh

1 GVVAGLAGLWGVTMLKRLL----RVDDPCDVFGVMGVCGIVGCIMTGIFAASSLGGVGFA----EGVTMGHQLLV
  hhhhhhhhhhhhhhhhhhh----cbcthhhhhhhhhhhhhhhhhhhhhhhbthhhbcccct----bbcthhhhhhh
2 GSTAATVSTLGFRYLQPVLQRNIKLHDTCGVHNLHGMPGILGAVISAIVATIATKDAYQDSYHELFKDASRTSSQ
  hhhhhhhhhhhhhhhhhhhbbbbcccttthhhhhhhhhhhhhhhhhhhhhbthhhbcbccccbbbcccthhhhhhh

1 QLESIAITIVWSGVVAFIGYKLADLT--VGLRV
  hhhhhhhhhhhhhhhhhhhhhhhhhhh--bcbrc
2 NGGYQIAALVCVLCIALLSGTLTGFLLKLPVWD
  hhhhhhhhhhhhhhhhhhhhhhhhhcbbtbbbrc
```

Fig. 1. Rh and Amt are distant relatives. (*A*) The ML optimal tree of 111 Rh (red) and 260 Amt (blue) proteins. NeRh (*N. europaea*) and FaAmt/TvAmt (archaea *Ferroplasma acidarmanus* and *Thermoplasma volcanium*) are at the base of Rh and Amt clusters, separated by a large distance. (*B*) The partitioned ML-like BI tree for Rh and Amt genes from species that have both. The values >50% at nodes are PP (posterior probability) from BI (left) or bootstrap proportion from ML (right). (*C*) Bar chart and logo of significant rate shift sites of Rhp1 vs. Amt. Blue, conserved; red, varied; brown, conserved with different residues. In the logo the height of a residue corresponds to its level of conservation. ⊕ and ⊖, charge differences. Positions refer to the gap-free alignment (see the alignment of coexistent Rh and Amt in supporting information). (*D*) The 3D fold of CiRh superimposed with EcAmt (*Left*; see the supplementary data for CiRh modeling in the supporting information). The rate shift sites of CiRh are in blue, and the 20 residues of the Amt channel lumen are in red. EcAmtB (1) and CiRh (2) are aligned to show secondary structures and rate shift sites (*Right*). In EcAmtB TM segments are shaded in yellow, and lumen residues are shaded in bold red. In CiRh rate shift sites are in bold blue. c, coil; h, helix; t, turn; b, bend; r, bridge; u, undecided.

substitution rates d_N as a function of species orders (37) (Fig. 3*B*; see the inferred ancestor sequences of vertebrate Rh, which are published as supporting information on the PNAS web site) reveals that the Rh30 trend line is high above that of the other clusters, and the rate for Rh30 is lifted 26% from fish to human. By comparison, the mean rate of Rh30 is 2.08 times higher than RhAG in fish (0.52/0.25), but that rises to 2.64 in chimp or human (0.66/0.25). Overall, Rh30 exhibited a trend of increasing d_N rates from fish to human, whereas RhAG showed a fluctuated pattern (Fig. 3*B*). For RhBG and RhCG, the mean rates are also low (<0.3035), but their trends are opposite in direction. As an outlier, chicken RhBG is high in both d_N and d_S (Fig. 3 *B* and *C*), maintaining a low d_N/d_S ratio (0.1192) and thus low functional divergence. Further studies are needed to determine whether this unique pattern is discernable in other birds.

Evolutionary Rate Shift in the Rh Family. To dissect the functional divergence of Rh clusters after gene duplication, we analyzed the distribution and number of evolutionary rate shift sites. We found 75 sites for Rh30 vs. RhAG and 72 sites for RhBG vs. RhCG (Fig. 4; see the alignments of vertebrate Rh, which are published as supporting information on the PNAS web site); their distribution showed similar ratios in TM domains vs. loop regions for Rh30:RhAG pair (1.27) and RhBG:RhCG pairs (1.25). This is in sharp contrast with the 2:1 ratio observed for the 96 rate shift sites identified in comparisons between Rh and Amt proteins (Fig. 1 *C* and *D*).

Of the 75 sites (75/354 = 21%) for Rh30/RhAG, 14 are conserved in both but differ in amino acids. Twenty-five are conserved in Rh30 and varied in RhAG, but 36 are conserved in RhAG and varied in Rh30 (Fig. 4*A*). In Rh30, 32 sites (32/376 = 8.5% with PP ≥ 95%) fall into class 3 (d_N/d_S = 1.1341), likely being subject to weak positive selection, whereas all of the others are under purifying selection (Fig. 4*B*). In RhAG, only 17 sites (17/380 = 4.5% with PP ≥ 95%) fall into class 3 (d_N/d_S = 1.038) (data not shown); these may be under neutral evolution or very weak positive selection. Of the 72 sites (72/404 = 18%) for RhBG/RhCG, 16 are conserved with different residues, but the 28 mutually exclusive conserved vs. varied sites are identical in location and number (Fig. 4*C*, blue/red or vice versa). These data support a strong purifying selection on both RhBG and RhCG, conforming to their extremely similar codon composition and highest protein sequence identity.

Discussion

We here carried out two lines of investigation on genes and proteins of the Rh family, whose function is in dispute (10–16, 23, 24). The first line, which focused on the occurrence of Rh and Amt in the same organism and their sequence similarity vs. divergence, revealed that the two families are distant relatives having independent evolutionary histories (Fig. 1). The second line of study, which dealt with a thorough analysis of the Rh phylogeny with 111 branches in 41 species from a bacterium to humans, uncovered the highly conserved evolution among the Rh clusters themselves (Figs. 2–4). The strong purifying selection on the Rh family, together with its discrete organismal distribution, evolutionary divergence, and structural features, suggests that Rh proteins could have acquired a new function different from that of Amt proteins. Because CO_2,

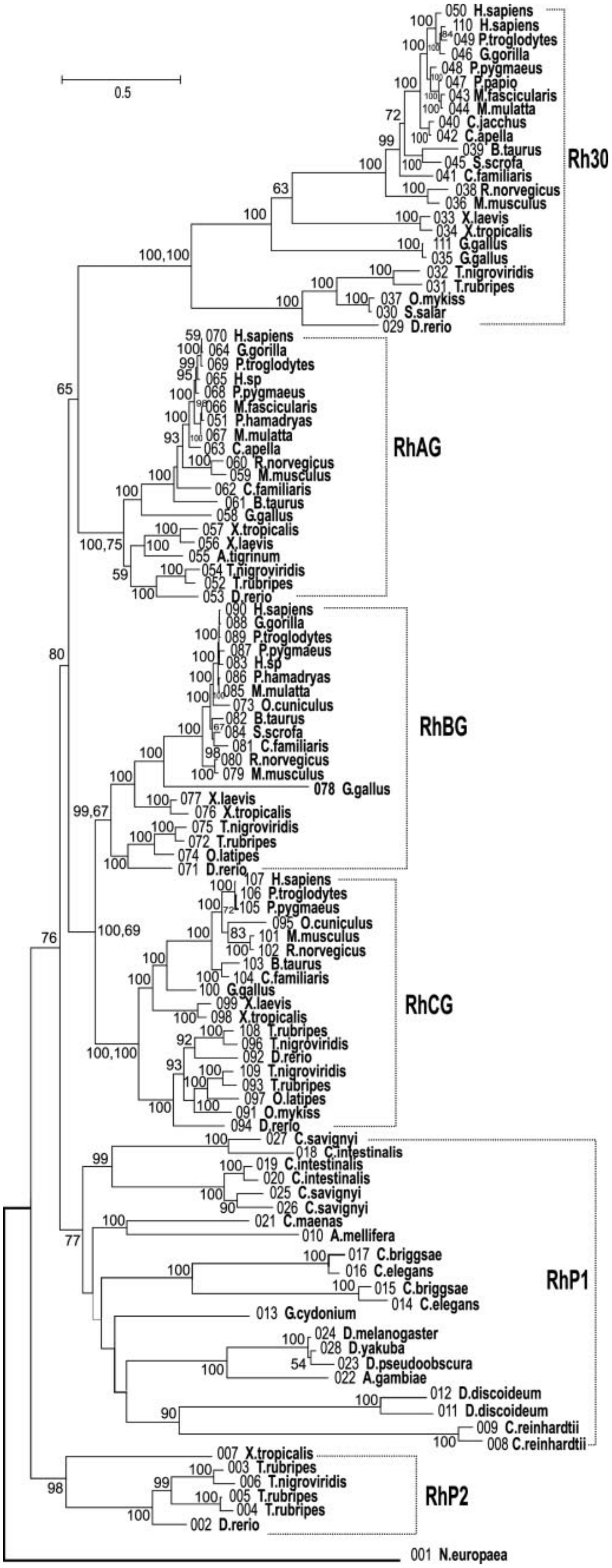


Fig. 2. Rh family gene tree. The BI tree of 111 Rh genes was built by using three independent models of substitution. Four increasingly heated Metropolis-coupled Markov chain Monte Carlo simulations were run 1.2 × 10^6 generations, with trees and parameters being sampled every 10 generations. The consensus tree was derived from the 65,000 trees sampled after the initial burn-in period. The values >50% at nodes are PP from BI (*Left*) and bootstrap proportion from ML [*Right*; using the general time reversible (GTR) + 4G + I model, 500 replicates, and labeled only on ancestor nodes of the four common clusters]. (Scale bar: 0.5 substitutions per nucleotide.) The outgroup is NeRh, because it is at the base of the Rh family (Fig. 1*A*), and *N. europaea* is the lowest in species order.

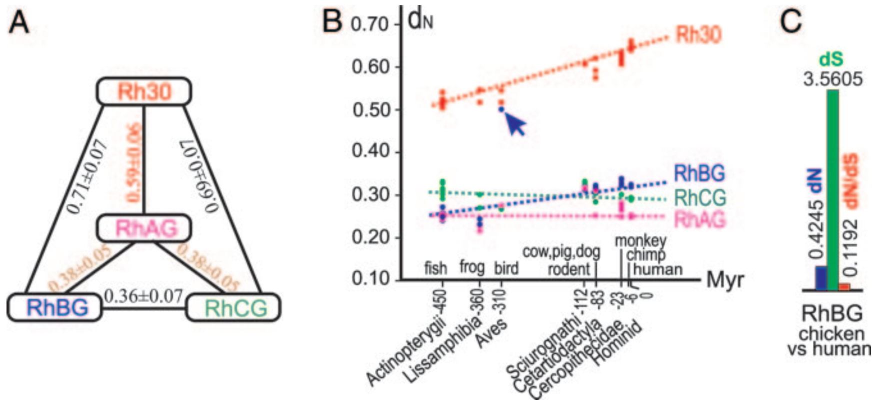


Fig. 3. Cluster divergence and substitution trend. (*A*) Cluster correlations. Lower θ_λ values (mean ± SE) denote lower functional divergence (see the alignment of the 111 Rh protein sequences in the supporting information). (*B*) Plot of d_N rates of Rh genes against species orders. The rate was computed by alignment of 83 Rh genes each with the ancestor inferred at the node where the four clusters merge (see the inferred ancestor sequences of vertebrate Rh in the supporting information). In each case the trend line directly links fish to human. [Scale: million years (Myr).] Red, Rh30; magenta, RhAG; blue, RhBG; green, RhCG. (*C*) Diagram for the high d_N and d_S rates yielding a low d_N/d_S ratio for chicken RhBG vs. other species (e.g., human RhBG).

and not NH_3/NH_4^+, is the physiologic substrate of Rh1 in green alga (23, 24), our results support the view that Rh proteins in all organisms mainly function as CO_2 channels. Transport of NH_3/NH_4^+ by Rh proteins under high, nonphysiologic concentrations (10–16) may reflect retention of the residual ancestral function related to ancient Amt proteins.

Rh was absent, whereas Amt was prominent in the 25 archaeal and over 350 bacterial genomes examined, except for the few bacteria including *N. europaea* (38) that have an Rh gene. Rh and Amt coexisted in organisms ranging from unicellular eukaryotes to sea squirts (Fig. 1*B*); thus, the period of their cooccurrence extended over vast stretches of evolutionary time. Before or during this overlapping period, the ancestral Rh gene(s) could already have diverged away from its related ancestral Amt, given the inferred ancestor of Rhp1 group remaining close to NeRh but far from Amt (data not shown). This divergence could be driven in response to a new selective pressure. In vertebrates the Rh family expanded and Amt genes disappeared, whereas in vascular plants Amt genes remained prominent and Rh genes were lost. Expansion of the Rh family to four major paralogous groups in vertebrates may have had to do with the usefulness of CO_2 channels in control of pH homeostasis and in waste disposal by vital organs. The absence of Rh in plants may reflect the fact that it worked well only at high CO_2 concentrations, as demonstrated in the green alga (23, 24). Loss of Amt genes in vertebrates may have occurred because the extremely toxic NH_3/NH_4^+ derived from amino acid catabolism is salvaged and reused by reversing the glutamate dehydrogenase reaction, which is integrated with biosynthesis and excretion (39). By contrast, retention of Amt genes in plants reflects the fact that ammonia remains an excellent source for nitrogen assimilation (40).

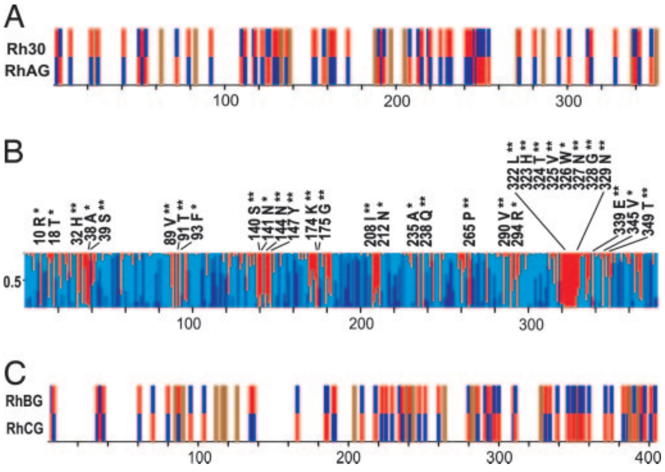


Fig. 4. Functional divergence related to sister groups (see the alignments of vertebrate Rh in the supporting information). (*A*) Schematic of significant rate shift sites for Rh30 vs. RhAG. Colored half-bar symbols are as in Fig. 1. (*B*) Diagram of d_N/d_S patterns of Rh30. The codon sequence alignment is gap-stripped. Three classes of d_N/d_S (blue, 0.05; cyan, 0.35; red, 1.13) denote most negative to most positive selection. The vertical coordinate is PP scale, and the height of each color bar indicates the site-specific PP value. The 32 sites under positive selection are denoted: one star, PP ≥ 95%; two stars, PP ≥ 99%. (*C*) Schematic of significant rate shift sites of RhBG vs. RhCG.

Significantly, we observed that the evolutionary rate shift amino acid residues between Rh and Amt proteins differ from those between the Rh paralogous groups, in both their physical location and chemical nature. Such functionally divergent amino acid sites in Rh proteins are largely clustered in the TM domains and regions that correspond to the packing and formation of the lumen essentially conserved for Amt channel function (21, 22). This finding further supports the view that the two families of proteins differ in their transport function or substrate specificity. Taken together, our data lead to the hypothesis that construction of Rh as a primitive CO_2 channel could have been attained by recruiting an ancient Amt, which would have a preformed gas conductance fold like EcAmtB (21, 22).

Study of the Rh family as a whole gave insight into its origin and gene duplications. The birth of Rh and its split from Amt might have occurred in the bacteria (Fig. 1). Nonetheless, although duplications and sometimes triplications of Rh genes occurred in a vast array of species from unicellular eukaryotes to sea squirts, clearly discernable clusters were not established in these taxa (Fig. 2). We here show that the paralogous clusters have apparently arisen in vertebrates, including one uncommon cluster, Rhp2, and four common clusters, Rh30, RhAG, RhBG, and RhCG. Apart from erythroid-specific Rh30 and RhAG (2–4), RhBG and RhCG are found in epithelial tissues of a variety of important organs (5–7, 41–44), congruent with their having physiological roles in CO_2 waste disposal and/or buffering of body fluids (24). These proteins may also play roles in sensing CO_2 (45) and are essential for normal embryonic development in model organisms (C.-H.H., unpublished data). Intriguingly, there are more Rh genes in fish than in mammals, but this larger number does not directly reflect genome-wide duplication events (46), because only RhCG occurs in extra copies. The fact that the evolutionary divergence of all Rh clusters is limited may reflect a theme on which to build cell- or tissue-specific modulations of the conserved CO_2 channel function.

Rh is one of the most ancient proteins of red cell membranes, to which human homologues carrying the classical Rh antigens belong (4, 7). We show here that the Rh30 cluster as a whole is conserved throughout vertebrates but did exhibit relatively fast evolution, as has been observed in mammals (47–50). Significantly, a trend of increasing d_N rates with a higher d_N/d_S ratio at a few specific sites in Rh30 as well as fluctuated d_N rates in RhAG was observed. These distinct patterns suggest a functional modification specific for red cells during vertebrate evolution from nucleate elliptocyte in fish (51) to enucleate biconcave disk in mammals (52). Such a role for the two Rh proteins, particularly Rh30, may be secondary and related to enhancing their heteromeric interactions (49) and increasing the surface area-to-volume ratio of red cells for gas movement (24). This view is consistent with the shape change of Rh_{null} red cells, which lack the two proteins and manifest spherostomatocytes (3, 4). It will be interesting to explore whether the fast evolution of Rh30 and fluctuated changes in RhAG exerted a threshold effect or occurred in concert with positive selection of other membrane and cytoskeleton proteins to drive red cell morphogenetic evolution.

In the red cell membrane, Rh proteins and band 3, the anion exchanger for Cl^-/HCO_3^-, appear to form a macromolecular complex dubbed "gas exchange metabolon" (53), suggesting that Rh is part of the CO_2 transport machinery. Indeed, Forster *et al.* (54) first detected such a CO_2 transport activity in addition to band 3 across the membrane of intact human red cells. Notably, the Cl^-/HCO_3^- exchange mechanism operates in teleost fish but not jawless fish, because zebrafish has a genuine band 3 (55), whereas the red cell membrane of hagfish is impermeable to HCO_3^- (56). These data suggest that hagfish red cells lack a functional band 3 for Cl^-/HCO_3^- exchange and may rely on a gas channel to facilitate CO_2 movement. In light of the intimate relationship of green algal Rh1 to CO_2 (23, 24) and the apparent early origin of Rh genes (as compared with band 3), it is tempting to speculate that the CO_2 gas channel mechanism evolved before the Cl^-/HCO_3^- exchange mechanism. Given the existence of genuine Rh genes in hagfish (unpublished data), it will be of great interest to study the roles of these Rh proteins in CO_2 conductance across the red cell membrane of this organism.

We thank Dr. Sydney G. Kustu for her ongoing support of this research and for her insightful editorial instructions. We are also grateful to many colleagues who provided us with RNA or DNA samples that have made this multiyear project possible. The excellent technical assistance of Ying Chen in the cloning and sequencing of Rh genes is highly appreciated. This work was supported by National Institutes of Health Grants HD62704 and HL66274.

1. Levine, P. & Stetson, R. E. (1939) *J. Am. Med. Assoc.* **113,** 126–127.
2. Anstee, D. J. & Tanner, M. J. (1993) *Baillieres Clin. Haematol.* **6,** 401–422.
3. Cartron, J.-P. (1999) *Baillieres Best Pract. Res. Clin. Haematol.* **12,** 655–689.
4. Huang, C.-H., Liu, Z. & Cheng, G. (2000) *Semin. Hematol.* **34,** 150–165.
5. Liu, Z., Chen, Y., Mo, R., Hui, C.-c., Cheng, J.-F., Mohandas, N. & Huang, C.-H. (2000) *J. Biol. Chem.* **275,** 25641–25651.
6. Liu, Z., Peng, J., Mo, R., Hui, C. & Huang, C.-H. (2001) *J. Biol. Chem.* **276,** 1424–1433.
7. Huang, C.-H. & Liu, P. Z. (2001) *Blood Cells Mol. Dis.* **27,** 90–101.
8. Marini, A.-M., Urrestarazu, A., Beauwens, R. & Andre, B. (1997) *Trends Biochem. Sci.* **22,** 460–461.
9. Ludewig, U., von Wiren, N., Rentsch, D. & Frommer, W. B. (2001) *Genome Biol.* **2,** 1010.1–1010.5.
10. Marini, A.-M., Matassi, G., Raynal, V., Andre, B., Cartron, J.-P. & Cherrif-Zahar, B. (2000) *Nat. Genet.* **26,** 341–344.
11. Westhoff, C. M., Ferreri-Jacobia, M., Mak, D. O. & Foskett, J. K. (2002) *J. Biol. Chem.* **277,** 12499–12502.
12. Hemker, M. B., Cheroutre, G., van Zwieten, R., Maaskant-van Wijk, P. A., Roos, D., Loos, J. A., van der Schoot, C. E. & vondem Borne, A. E. (2003) *Br. J. Haematol.* **122,** 333–340.
13. Ludwig, U. (2004) *J. Physiol. (Paris)* **559,** 751–759.
14. Bakouh, N. L., Benjelloun, F., Hulin, P., Brouillard, F., Edelman, A., Cherif-Zahar, B. & Planelles, G. (2004) *J. Biol. Chem.* **279,** 15975–15983.
15. Ripoche, P., Bertrand, O., Gane, P., Birkenmeier, C., Colin, Y. & Cartron, J.-P. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 17222–17227.
16. Nakhoul, N. L., DeJong, H., Abdulnour-Nakhoul, S. M., Boulpaep, E. L., Hering-Smith, K. & Hamm, L. L. (2005) *Am. J. Physiol.* **288,** F170–F181.
17. Soupene, E., He, L., Yan, D. & Kustu, S. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 7030–7034.
18. Soupene, E., Ramirez, R. M. & Kustu, S. (2001) *Mol. Cell. Biol.* **21,** 5733–5741.
19. Soupene, E., Chu, T., Corbin, R. W., Hunt, D. F. & Kustu, S. (2002) *J. Bacteriol.* **184,** 3396–3400.
20. Soupene, E., Lee, H. & Kustu, S. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 3926–3931.
21. Khademi, S., O'Connell, J., III, Remis, J., Robels-Colmennares, Y., Miercke, L. J. W. & Stroud, R. M. (2004) *Science* **305,** 1587–1594.
22. Zheng, L., Kostrewa, D., Berneche, S., Winkler, F. K. & Li, X.-D. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 17090–17095.
23. Soupene, E., King, N., Feild, E., Liu, P., Niyogi, K. K., Huang, C.-H. & Kustu, S. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 7769–7773.
24. Soupene, E., Inwood, W. & Kustu, S. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 7787–7792.
25. Kim, K.-S., Feild, E., King, N., Yaoi, T., Kustu, S. & Inwood, W. (2005) *Genetics* **170,** 631–644.
26. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
27. Edgar, R. C. (2004) *Nucleic Acids Res.* **32,** 1792–1797.
28. Kumar, S. & Gadagkar, S. R. (2001) *Genetics* **158,** 1321–1327.
29. Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001) *Bioinformatics* **17,** 1244–1245.
30. Guindon, S. & Gascuel, O. (2003) *Syst. Biol.* **52,** 696–704.
31. Felsenstein, J. (1996) *Methods Enzymol.* **266,** 418–427.
32. Huelsenbeck, J. P. & Ronquist, F. (2001) *Bioinformatics* **17,** 754–755.
33. Fiser, A. & Sali, A. (2003) *Methods Enzymol.* **374,** 461–491.
34. Gu, X. (1999) *Mol. Biol. Evol.* **16,** 1664–1674.
35. Knudsen, B., Miyamoto, M. M., Laipis, P. J. & Silverman, D. N. (2003) *Genetics* **164,** 1261–1269.
36. Yang, Z. & Bielawski, J. P. (2000) *Trends Ecol. Evol.* **15,** 496–503.
37. Kumar, S. & Hedges, S. B. (1998) *Nature* **392,** 917–920.
38. Chain, P., Lamerdin, J., Larimer, F., Regala, W., Lao, V., Land, M., Hauser, L., Hooper, A., Klotz, M., Norton, J., *et al.* (2003) *J. Bacteriol.* **185,** 2759–2773.
39. Lehninger, A. (1975) *Biochemistry* (Worth, New York), 2nd Ed., pp. 579–584.
40. von Wiren, N., Gazzarrini, S., Gojon, A. & Frommer, W. B. (2000) *Curr. Opin. Plant Biol.* **3,** 254–261.
41. Eladari, D., Cheval, L., Quentin, F., Bertrand, O., Mouro, I., Cherrif-Zahar, B., Cartron, J.-P., Paillard, M., Doucet, A. & Chambrey, R. J. (2002) *J. Am. Soc. Nephrol.* **13,** 1999–2008.
42. Quentin, F., Eladari, D., Cheval, L., Lopez, C., Goossens, D., Colin, Y., Cartron, J.-P., Paillard, M. & Chambrey, R. J. (2003) *J. Am. Soc. Nephrol.* **14,** 545–554.
43. Verlander, J. W., Miller, R. Y., Frank, A. E., Royaux, I. E., Kim, Y.-H. & Weiner, I. D. (2003) *Am. J. Physiol.* **284,** F323–F337.
44. Weiner, I. D., Miller, R. T. & Verlander, J. W. (2003) *Gastroenterology* **124,** 1432–1440.
45. Mulkey, D. K., Stornetta, R. I., Weston, M. C., Simmons, J. R. Parker, A., Bayliss, D. A. & Guyenet, P. G. (2004) *Nat. Neurosci.* **7,** 1360–1368.
46. Volff, J.-N. (2005) *Heredity* **94,** 280–294.
47. Kitano, T., Sumiyama, K., Shiroishi, T. & Saitou, N. (1998) *Biochem. Biophys. Res. Commun.* **249,** 78–85.
48. Matassi, G., Cherif-Zahar, B., Pesole, G., Raynal, V. & Cartron, J.-P. (1999) *J. Mol. Evol.* **48,** 151–159.
49. Huang, C.-H., Liu, Z., Apoil, P.-A. & Blancher, A. (2000) *J. Mol. Evol.* **51,** 76–87.
50. Kitano, T. & Saitou, N. (2000) *Immunogenetics* **51,** 856–862.
51. Thisse, C. & Zon, L. I. (2002) *Science* **295,** 457–462.
52. Steck, T. L. (1989) in *Cell Shape: Determinants, Regulation, and Regulatory Role*, eds. Stein, W. & Brouner, F. (Academic, New York), pp. 205–246.
53. Bruce, L. J., Beckmann, R., Ribeiro, M. L., Peters, L. L., Chasis, J. A., Delaunay, J., Mohandas, N., Anstee, D. J. & Tanner, M. J. (2003) *Blood* **101,** 4180–4188.
54. Forster, R. E., Gros, G., Lin, L., Ono, Y. & Wunder, M. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 15815–15820.
55. Paw, B. H., Davidson, A. J., Zhou, Y., Li, R., Pratt, S. J., Lee, C., Trede, N. S., Brownlie, A., Donovan, A., Liao, E. C., *et al.* (2003) *Nat. Genet.* **34,** 59–64.
56. Peters, T., Forster, R. E. & Gros, G. (2000) *J. Exp. Biol.* **203,** 1551–1560.