J Mol Evol (1999) 48:151–159

JOURNAL OF MOLECULAR EVOLUTION

The Members of the *RH* Gene Family (*RH50* and *RH30*) Followed Different Evolutionary Pathways

Giorgio Matassi,[1] Baya Chérif-Zahar,[1] Graziano Pesole,[2] Virginie Raynal,[1] Jean-Pierre Cartron[1]

[1] Unité INSERM U76, Institut National de la Transfusion Sanguine 6, rue Alexandre Cabanel, 75015 Paris, France
[2] Dipartimento di Biologia, DBAF, Università di Basilicata, via Anzio 10, 85100 Potenza, Italy

Received: 17 June 1998 / Accepted: 10 July 1998

Abstract. The evolution of the *RH* gene family is characterized by two major duplication events, the first one originating the *RH50* and *RH30* genes and the second one giving rise to *RHCE* and *RHD*, the two paralogous *RH30* genes which encode the Rh blood group antigens in human. The new sequence data obtained here for mouse *RH50* and *RH30* and for macaque *RH50* allowed us to compare the evolutionary rates of the two genes and to show that *RH50* evolved about 2.6 times more slowly than *RH30* at nonsynonymous positions. This result implies that Rh50 proteins were evolutionarily more conserved compared to Rh30 polypeptides, thus being indicative of the functional significance of the former protein in species as distantly related as sponge and human. The duplication event leading to *RH50* and *RH30* genes was estimated to have occurred between 250 and 346 million years ago. Moreover, we could also estimate that the duplication event producing the *RHCE* and *RHD* genes occurred some 8.5 ± 3.4 million years ago, in the common ancestor of human, chimpanzee, and gorilla. Interestingly, this event seems to coincide with the appearance in these species of a G-to-T mutation in the *RH50* gene which created a stop codon in the corresponding transcript. This led to an Rh50 C-terminal cytoplasmic domain shorter than that found in orangutan and early primates.

Key words: Nucleotide substitution rate — Divergence time — Rh antigenic complex — Gene duplication — Coevolution

Introduction

The Rh protein family in human comprises the RhD/RhCE (Rh30) polypeptides, which carry the DCcEe Rh blood group antigens, and the Rh50 glycoprotein (Matassi et al. 1998; Huang 1998). The *RH30* and *RH50* genes, which map on chromosomes 1p34.3–p36.1 and 6p11–p21.1, respectively (Chérif-Zahar et al. 1991, 1996), both consist of 10 exons and show strikingly similar exon–intron organizations (Matassi et al. 1998; Huang 1998). The corresponding proteins, of 417 and 409 residues, respectively, share about 36% amino acid identity and similar positioning of their 12 predicted transmembrane (TM) domains. These proteins are believed to interact with each other to form the core of a membrane complex thought to be a tetramer composed of two Rh30 and two Rh50 glycoprotein subunits, to which the accessory chains CD47, LW, and GPB, which are encoded by independent genes, are associated by noncovalent linkages. It is assumed that when one chain is missing, the Rh complex is not assembled or transported to the cell surface (for reviews see Agre and Cartron 1991; Cartron and Agre 1993; Anstee and Tanner 1993; Cartron 1994). Furthermore, the analyses carried out on Rh_{null} phenotypes strongly suggest that the Rh30 and Rh50 proteins are likely to be crucial for the transport of the Rh complex to the RBC membrane and/or for the stability of the Rh membrane complex (for review

Correspondence to: G. Matassi; *e-mail:* gmatassi@infobiogen.fr

see Cartron et al. 1998). Several structural features differentiate the members of the *RH* gene family in human. The *RH30* and *RH50* genes extend over about 69 and 32 kb, respectively. Interestingly, the GC content of the chromosomal environment differs between *RH30* and *RH50*. Indeed, the RH50 locus is embedded in a GC-poor L isochore, whereas RH30 is located in the GC-rich H1 family (Matassi et al. 1998). Accordingly, the G+C content at the weakly constrained sites (third codon positions—intronic regions) is lower in RH50 (49.8–35.0%) compared to RH30 (66.7–49.0%) genes, nonsynonymous sites showing similar GC values (Matassi et al. 1998).

A major role in shaping the evolution of the human RH30 locus is played by recombination. In fact, many Rh30 variant phenotypes have been described in which recombination events are probably ascribable to the high sequence similarity shared by the *D/CE* introns, as exemplified by two recombination hot spots recently identified in the intronic regions of the D^{VI}, Dc−, DFR, R^N, and D−variants (Kemp et al. 1996; Matassi et al. 1997; Wagner et al. 1998). Moreover, evidence has been provided which demonstrates that the ''*Ce*'' allele originated from a nonreciprocal intergenic exchange of *D* sequences into the ''*ce*'' allele of *CE,* a relatively recent event which seems to have taken place in the human lineage (Carritt et al. 1997).

The importance of the Rh50 protein from both functional and evolutionary standpoints has emerged only very recently, when it was demonstrated that in the majority of Rh_{null} individuals (so-called ''regulator type''), the lack of expression of Rh antigens and the severe reduction of the expression of the other proteins of the complex are ascribable to the absence of expression of the Rh50 protein, which is brought about by a heterogeneous mutation pattern in the *RH50* gene (Chérif-Zahar et al. 1996, 1998a; Hyland et al. 1998; Huang 1998).

In contrast, because of their importance in transfusion medicine, Rh30 proteins which carry the Rh allotypes were the first studied to delineate their evolutionary pathway. In human, the *RHD* and *RHCE* genes are highly homologous (~3.5% divergence over the coding region; intronic regions are also highly conserved). These genes are believed to be the result of a duplication, most likely from a ''*c-like*'' allele (Cartron 1994; Salvignol et al. 1995; Carritt et al. 1997). Indeed, serological studies showed that ''*c*'' antigens could be detected on erythrocytes in all anthropoid apes as well as in Old and New World monkeys, whereas anti-D antibodies reacted only with gorilla and chimpanzee (for review see Socha and Ruffié 1983; Blancher and Socha 1997). The number of *RH30* genes varies among nonhuman primates. Only one *RH30*-like gene has been found in orangutan, gibbon, and Old and New World monkeys, whereas at least two genes were detected in gorilla and chimpanzee (Cartron 1994; Salvignol et al. 1995; Blancher and Socha 1997). The functional importance of the Rh30 proteins was also demonstrated by showing that in some Rh_{null} individuals (so-called ''amorph type''), the lack of expression of Rh antigens is caused by splice site and frameshift mutations in *RH30* genes (Chérif-Zahar et al. 1998b).

Here we investigate the changes that affected *RH50* and *RH30* genes during evolution and show that these genes followed quite distinct evolutionary pathways.

Materials and Methods

PCR Amplification and Cloning of the Macaque Rh50-like cDNA. Macaque *(Macaca mulatta)* RNAs were extracted from whole blood (300 μl) using the acid–phenol–guanidinium method (Lazano et al. 1993) and were reverse transcribed in a total volume of 33 μl using the First Strand cDNA synthesis kit (Pharmacia, Uppsala, Sweden). Five microliters of cDNA products was amplified by PCR using primers deduced from the human Rh50 cDNA sequence (Ridgwell et al. 1992): 5′-GTGGCCTCTGTCCTTTGCCACAA-3′ (sense, nt −26 to −3; + 1 refers to the position of the first nucleotide of the initiation ATG codon);5′-AAGGACATTTTTTACACTGGCCATTTGGG-3′ (antisense, nt +1397 to +1424). Amplification reactions were carried out in a total volume of 50 μl using the mix of thermostable DNA polymerases provided in the Advantage cDNA PCR kit (Clontech, Palo Alto, CA) with 200 μ*M* dNTP and 20 pmol of each primer. PCR conditions were denaturation for 1 min at 94°C (first cycle) and 30 s at 94°C, 3 min at 68°C, for 30 cycles. PCR products were separated on a 1% agarose gel, purified using the Advantage PCR-Pure Kit (Clontech), and cloned into a PCR II vector using the TA cloning kit (Invitrogen, The Netherlands).

PCR Amplification and Cloning of the Mouse Rh50-like cDNA. The clone 940c01 (GenBank W81811), containing the 3′ region of the mouse Rh50 cDNA, was obtained from the IMAGE LLNL consortium via the UK HGMP Resource Centre. The 5′ region of the cDNA was amplified using a 5′ Marathon Race reaction (Clontech) carried out on a mouse 15-day embryo Marathon-Ready cDNA template with the Ap1 sense primer provided in the kit and the mouse Rh50-specific oligonucleotide 5′-GAACCCCACATGTATCATGGATCATCAG-3′ as antisense primer (complementary to nt 994 to 1021 of the cDNA) derived from the sequencing of the 940c01 clone. The amplification reaction was performed as described above.

Sequencing of Rh50 and Rh30 cDNAs. The clone 1490b20 (GenBank AA168171), containing the complete mouse Rh30 cDNA, was obtained from the IMAGE LLNL consortium via the UK HGMP Resource Centre. The sequencing of macaque and mouse Rh50 and mouse Rh30 cDNAs was performed using the Thermo Sequenase sequencing kit (Amersham International) following the manufacturer's instruction. A primer/template molar ratio of 40 was used. Sequences were analyzed on an automated fluorescence-based ALF Express sequencing system (Pharmacia).

PCR Amplification of the Nucleotide Region Surrounding the Stop Codon of the RH50 *Gene in Primates.* The following primates were analyzed: lemur (*Lemur catta;* prosimian), spider monkey (*Ateles paniscus;* New World monkey), rhesus macaque (*Macaca mulatta;* Old World monkey), gibbon (*Hylobates lar*), orangutan (*Pongo pigmaeus*), gorilla (*Gorilla gorilla*), and chimpanzee (*Pan troglodytes*). Primate genomic DNAs were a gift from Professor Damian Labuda (University of Montreal, Canada). The nonhuman primate genomic DNA fragment, corresponding to the region surrounding the human *RH50* exon 10, was PCR amplified using the following primers: 5′-ACTGCTATGAT-

GATTCTGTTTATTGG-3′ (sense primer, nt 1184 to 1209 of the human cDNA) and 5′-AAGGACATTTTTTACACTGGCCATTTGGG-3′ (antisense primer, nt 1397 to 1424). The nucleotide sequences of sense and antisense primers, lying in exon 9 and exon 10, respectively, are identical between human and macaque. Amplification reactions were carried out in a total volume of 50 μl containing 200 μ*M* dNTP, 20 pmol of each primer, *Taq* buffer (20 m*M* Tris–HCl, pH 8.4, 50 m*M* KCl, 1.5 m*M* $MgCl_2$), and 2.5 U of *Taq* DNA polymerase (GIBCO BRL). PCR conditions were denaturation for 3 min at 94°C (first cycle) and 30 s at 94°C, annealing for 30 s at 58°C, and extension of 1 min at 72°C, for 30 cycles. PCR products were purified through a Microcon ultrafiltration membrane (Amicon) and directly sequenced as described above.

Estimation of Nucleotide Substitution Rates and Divergence Times. The multiple alignment of the nucleotide sequences under investigation was derived from the corresponding protein sequence alignment obtained using the program CLUSTALW (Thompson et al. 1994). The general time reversible (GTR) model (Saccone et al. 1990; Yang 1994), also known as Stationary Markov Model, was used to calculate the number of synonymous (K_S) and nonsynonymous (K_A) substitutions per site.

The divergence time for the *RH50/RH30* gene duplication was estimated according to the method of Rambaut and Bromham (1998), which, given as an input the divergence times between two pairs of taxa (see Results), calculates the divergence date of their common ancestor. The final estimate was calculated by averaging results obtained using all possible pairs of *RH50* and *RH30* genes. Moreover, nucleotide sequences were assumed to evolve following the GTR model, with rate heterogeneity shaped by a discrete gamma distribution using a maximum-likelihood estimate of the gamma shape parameter obtained by the PAUP package as implemented in the GCG package [Wisconsin Package Version 9.1, Genetics Computer Group (GCG), Madison, WI].

All other divergence times (*T*) were calculated from GTR nonsynonymous distances (*K*) using the human–macaque divergence (25 Mya) as a calibrating date, according to the formula $T_{pair} = (K_{pair}/K_{human–macaque}) \times T_{human–macaque}$.

The molecular clock hypothesis was checked by carrying out relative rate tests according to the method of Muse and Gaut (1994) on all possible triplets including two ingroups and one outgroup sequence (e.g., for *RH30*, human and macaque sequences as ingroups and bovine or mouse sequences as outgroups).

Results

The available sequence data were clearly inadequate to analyze the overall evolutionary pathway of the *RH* gene family. Indeed, aside from the human Rh50 cDNA, Rh50-like sequences were described only in the sponge *Geodia cydonium* (Seack et al. 1997) and in the nematode *Caenorhabditis elegans.* Several Rh30 cDNAs were sequenced in primates (for a review see Blancher and Socha 1997), whereas only the bovine sequence (Méténier-Delisse et al. 1997) was known for nonprimate mammals. In this study, we have obtained new cDNA sequence data for macaque and mouse Rh50 and for mouse Rh30.

GC Content Variation in the RH *Gene Family*

Table 1 shows the GC levels of the *RH50* and *RH30* coding sequences presently available. The GC content at nonsynonymous positions (mainly first and second codon positions, GC1+GC2) is similar between *RH50* and *RH30.* In contrast, GC content at synonymous sites (mainly third codon positions, GC3) is clearly lower in *RH50* compared to *RH30.* An even lower GC3 value is exhibited by one of the nematode *RH50*-like genes (B0240.1). However, the sponge *RH50*-like gene shows a GC content clearly higher than that shared by the majority of the *RH50* sequences.

Table 1. GC levels in the *RH* gene family

	GC1 + GC2[a]	GC3[a]	GenBank
RH50			
Human	45.3	49.9	X64594
Macaque	46.3	49.6	AF058917[b]
Mouse	45.7	44.5	AF065395[b]
Nematode	45.3	44.4	U64847 (F08F3.3)
Nematode	44.7	21.0	Z74026 (B0240.1)
Sponge	47.9	60.7	Y12397
RH30			
Human *RHcE*	47.2	66.6	X54534
Human *RHD*	47.2	66.6	X63097
Macaque	45.8	66.6	S70343
Bovine	45.2	68.5	U59270
Mouse	47.8	71.2	AF047827[b]

[a] GC1, GC2, and GC3 are the GC levels at the first, second, and third codon positions, respectively.
[b] This work.

Estimation of the Evolutionary Rates of RH50 *and* RH30 *Genes*

The number of synonymous (K_S) and nonsynonymous (K_A) substitutions per site were estimated according to the general time reversible (GTR) model (Saccone et al. 1990; Yang 1994) on *RH50* and *RH30* gene pairs. Absolute nucleotide substitution rates (*R*) were calculated from GTR distances (*K*) by fixing the times of divergence (*T*) between human and macaque, bovine, and mouse at 25, 70, and 100 Mya, respectively, according to $R = K/2T$ (see also Materials and Methods).

The results are shown in Table 2. Most interestingly, *RH50* and *RH30* genes were found to evolve at remarkably different rates at nonsynonymous positions, *RH30* evolving about 2.6 times faster than *RH50,* the average absolute substitution rates (per site per year) being $1.48 \pm 0.49 \times 10^{-9}$ and $0.57 \pm 0.25 \times 10^{-9}$, respectively. Similar results were obtained independently by T. Kitano and N. Saitou (personal communication). Synonymous sites appeared to evolve at a rather uniform rate in both genes (2.01 ± 0.86 and $2.29 \pm 1.08 \times 10^{-9}$). Although there are other examples of duplicated genes that have evolved at different rates at nonsynonymous sites, e.g., α-fetoprotein was shown to evolve 1.7 times faster than its albumin paralogue (Minghetti et al. 1985), the majority of duplicated genes appears not to show striking changes in the nonsynonymous rate (our unpublished data).

Table 2. Nucleotide substitution rates (per site per year × 10^{-9}) calculated on synonymous and nonsynonymous codon positions in *RH50* and *RH30* genes

	RH50		*RH30*	
Pair	Syn.	Nonsyn.	Syn.	Nonsyn.
Human–macaque	1.08 ± 0.66	1.00 ± 0.54	1.50 ± 0.98	2.30 ± 0.92
Human–mouse	2.53 ± 1.05	0.49 ± 0.20	2.70 ± 1.20	1.32 ± 0.42
Macaque–mouse	2.52 ± 1.06	0.53 ± 0.23	2.46 ± 1.10	1.34 ± 0.42
Average	2.01 ± 0.86	0.57 ± 0.25	2.29 ± 1.08	1.48 ± 0.49

We noted that, in the human–macaque lineages, for both the *RH50* and the *RH30* genes there seems to be a slowdown of the synonymous rate (particularly for the *RH50* gene), while the nonsynonymous rate is about twofold than that found in the other comparisons. In addition, we also observed that, in the case of *RH30*, nonsynonymous substitutions outnumbered synonymous ones. This was also the case in all possible pairwise comparisons between the available *RH30* sequences from macaque, gorilla, chimpanzee, and human (data not shown). Therefore, these data tend to suggest the existence of some sort of positive selection acting on both *RH50* and *RH30* genes in primates and confirm similar conclusions made for *RH30* genes by T. Kitano and N. Saitou (personal communication).

Estimation of the Divergence Time Between RH50 *and* RH30 *Genes*

Due to the strong compositional heterogeneity observed at synonymous positions (see Table 1) and to avoid divergence underestimates because of saturation of nucleotide substitutions, all divergence times were inferred by analysing nonsynonymous sites. Moreover, the gene-specific evolution observed for *RH50* and *RH30* genes (with the latter evolving about 2.6 times faster than the former) does not permit the assumption of the molecular clock hypothesis to estimate divergence times. Consequently, we have used the method of Rambaut and Bromham (1998), which does take into account rate heterogeneity (see Materials and Methods).

This analysis allowed us to estimate the maximum-likelihood value of the divergence date between *RH50* and *RH30* genes at 292 Mya, with a 95% confidence interval between 250 and 346 Mya.

Dating the RHCE–RHD *Gene Duplication*

Relative rate tests, carried out using the method of Muse and Gaut (see Materials and Methods), showed that, taken separately, both RH50 and RH30 lineages did evolve in a clock-like manner at synonymous and nonsynonymous positions (data not shown). In addition, under the molecular clock assumption, we estimated the divergence dates between orthologous genes within the two groups of *RH50* and *RH30* genes. Fixing the divergence between human and macaque at 25 Mya, as a calibrating date, we estimated at 65.3 ± 20.2 Mya the divergence date between primates and bovine (based on *RH30* gene comparison) and at 70.6 ± 24.1 Mya the one between human and mouse (average value calculated from both *RH50* and *RH30* gene comparisons). These times, taking into account the statistical fluctuations, were in good agreement with paleontological estimates, thereby confirming a clock-like behavior of both *RH* genes.

Under these premises, GTR nonsynonymous distances were used to calculate the divergence time between *RHCE* and *RHD* genes at 8.5 ± 3.4 Mya (see Materials and Methods), suggesting that the duplication originating the two paralogous *RH30* genes predated the divergence between human, chimpanzee, and gorilla, whose common ancestor is dated at 6.7 ± 1.3 Mya (Kumar and Hedges 1998). This finding is in agreement with serological studies (see Introduction) and previous estimates (Cartron 1994; Carritt et al. 1997).

Gross Evolutionary Changes in the Rh50 Protein

The human Rh50 protein shares about 85% identity with the macaque homologue, 72% with the mouse, and 41 and 39% with the distantly related nematode and sponge sequences, respectively. Incidentally, the human Rh30 protein shows about 27% identity with the Rh50-like proteins of the latter two species. The multiple alignment of the Rh50 amino acids sequences revealed additional events which shaped this protein in the course of evolution (Fig. 1). It can be observed that the length of these proteins progressively shortened during evolution leading to the human lineage. In fact, Rh50 comprises 523 amino acid residues in sponge, 463 and 457 in nematode, 438 in mouse, 428 in macaque, and 409 residues in human. In contrast, the size of the Rh30 protein seems to have remained fairly constant during evolution: 417 residues in primates and bovine and 418 in mouse.

If the most parsimonious evolutionary scheme is adopted, namely, considering the sponge sequence as the closest to the ancestral gene, the size changes in the Rh50 protein can be accounted for by several deletion events, and probably one insertion in the mouse sequence, within

```
          1                                                                  70
Sponge    ..MDWAKMLL PGFLLVFQVI FIILYGLLVR YDDTGDAIR. .....NDTTI SDVSNLDSYR STLKVYPFFQ
NemF      MRSPLHQNQL TLILGLFQVV FLVIFALYGS YDASALPSE  .........T SETKNVEEAA RMTNLYPLFQ
NemB      MWSVLHRRQF AIIAGLMQTV FIVLFAKYVK YIDPLDDSR. .......... ....RVYSGT ....DYPLFQ
Mouse     .....MRFKF PLMAISLEVA MIVLFGLFVE YETPQNASQK NASHQNASQQ GNTSSSAKKD QFFQLYPLFQ
Macaque   .....MRLKF PLMAIVLEIA MIVLFALFVE YEMDQTTPQ. .........Q LNITNSTDMG KFLELYPLFQ
Human     .....MRFTF PLMAIVLEIA MIVLFGLFVE YETDQTVLE. .........Q LNITKPTDMG IFFELYPLFQ

         71                                                                 140
Sponge    DVHVMIFVGF GFLMTFLRRY GFGSISFNLL LASFAIQWST LTSGVFQFID QSDAGDCCTI NVNLETLVGA
NemF      DTHVMIFIGF GFLMTFLKRY GFSAVSINML LAVFTIQWGI IVRGMASAHH GF......KF TISLEQLLTA
NemB      DVHLMIFVGF GFLMAFLKRY GFSAVSVNLL LSAFVIQFAM LLRGFMTVAF QETG....LF SIGIPEMISA
Mouse     DVHVMIFVGF GFLMTFLKKY GFSGVGFNLF LAALGLQWGT IMQGLLHSHG KE.......F HFGIYNMINA
Macaque   DVHVMIFVGF GFLMTFLKKY GFSSVGINLL IAALGLQWGT VVQGILHSQG QK.......I TIGIKNMINA
Human     DVHVMIFVGF GFLMTFLKKY GFSSVGINLL VAALGLQWGT IVQGILQSQG QK.......F NIGIKNMINA

        141                                                                 210
Sponge    DFAGAAVLIT MGAVLGKASP FQLVIIAFFE LIFYSCNEAL NVHVFMAADI GGSMLIHTFG AYFGLAVSLM
NemF      DFAAAVILIS MGAMLGKLSP SQYVIMAFFE TPVALIVEHI CVHNLQINDV GGSIIVHAFG AYFGLACAKG
NemB      ESSCAAVLIT MGVLLGRLTP VQFLLLAFFE TGINVLVEHY VFNYLHVNDS GRSLSVHTFG AYFGLAAACV
Mouse     DFSTATVLIS FGAVLGKTSP IQMLIMTILE IAVFAGNEYL VTELFEASDT GASMTIHAFG AYFGLAVAGV
Macaque   DFSTATVLIS FGAVLGKTSP TQMLIMTIIE IAVFAGNEYL VGEIFKASDI GASMTIHAFG AYFGLAVAGI
Human     DFSAATVLIS FGAVLGKTSP TQMLIMTILE IVFFAHNEYL VSEIFKASDI GASMTIHAFG AYFGLAVAGI

        211                                                                 280
Sponge    LYNKDARDNE KNS.TVYHSD LFSMIGTLFL WLFWPSFNGV LAS.GNAQTR AVINTYYAMT ASVLGTFIFS
NemF      FGKKEQR.GH TNEGSTYHTD IFAMIGAIFL WIYWPSFNAA VAATDDARQR AVANTFLSLC ACTMTTFLVS
NemB      GHKKNVM.EM DEHGGIHHSD LFSMIGTLLL WVFFPSFNAA IQEPEDARHR AIMNTYLAMA SGTVTTFMIS
Mouse     LYRPGLRCEH PNDESVYHSD LFAMIGTLFL WIFWPSFNSA IADPGDHQYR AIVNTYMSLA ACVITAYALS
Macaque   LYRSALRRGH KNEESTYYSD LFAMIGTLFL WMFWPSFNSA IAEPGDKQSR AIVNTYFSLV ACVVTAFAFS
Human     LYRSGLRKGH ENEESAYYSD LFAMIGTLFL WMFWPSFNSA IAEPGDKQCR AIVDTYFSLA ACVLTAFAFS

        281                                                                 350
Sponge    LLFSKKKGKL SMTHVQNATL AGGVAVGAMA DMVIQPWAAL VIGLLAGLIS VFGYKFLSPL LEKYLYIQDT
NemF      QAVDKH.KRF DMVHIANSTL AGGVAIGTTA NVVLEPYHAM IIGVIAGAVS VIGYKYITPF LSEKLGIHDT
NemB      SCVDTL.GRF NMIHIQSSTL AGGVAIGSSA NAVLHPYHAV IVGVIAALLS VIGHAWISPR LERTFHLFDT
Mouse     SLVERR.GRL DMVHIQNATL AGGVAVGTCA DMEIPLYAAM TIGSIAGIIS VLGYKFFSPL LANKLMIHDT
Macaque   SLVERR.GKL NMVHIQNATL AGGVAVGTCA DMAIHPFGSM TLGSIAGAVS VIGYKFLTPL FTTKLGIHDT
Human     SLVEHR.GKL NMVHIQNATL AGGVAVGTCA DMAIHPFGSM IIGSIAGMVS VLGYKFLTPL FTTKLRIHDT

        351                                                                 420
Sponge    CGVHNLHGMP GVFAGIGSFV AAVLASYSGG GNRIEYGDSL FVVFPARAPS SSSELTPSQM NLGVETGDGR
NemF      CGVNNLHGMP GLIAGFASIA FLFIYDET.. ....RYPAQY DKIYPGMAR. .......... ..GEDRTRMF
NemB      CGVHNLHGMP GILAGLLSIG FAYFYEPE.. ....SYGKTL YHIYPYWIG. .......... ..GELHGDR.
Mouse     CGVHNLHGLP GVFGGLASIV AISWGMST.. .......... .......... .......... ..........
Macaque   CGVHNLHGLP GVVGGLAGIA AVALGASN.. .......... .......... .......... ..........
Human     CGVHNLHGLP GVVGGLAGIV AVAMGASN.. .......... .......... .......... ..........

        421                                                                 490
Sponge    SAGVQAGFQW ACLATTLALA IIGGLTTGVI VRWLPKLKGE NEIDDDHLFD DQIYWELPDD ADKYLPIEEL
NemF      DEKTQALNQL MAIGLVFLAS TVSGYLTGLL L....KLKIW DQVRDDEYYA DGDYFETPGD YDFTSRIVTS
NemB      ENVSQAQYQA LGLLTILVTA VIGGLLTGCI L....KIKVW NQVDDPDFPH GEMNYYAQSD VNFLSKYKHA
Mouse     ...ASMAMQA AALGSSIGSA IVGGLLTGLI L....KLPIW NQPPDEYCYD DSVSWKVPKF RELDNRFFQH
Macaque   ...TSVAMQA AALCSSIGAA VVGGLITGLI L....KLPFW GQPSDQDCYD DSVYWEVPIL REPDHHFHGH
Human     ...TSMAMQA AALGSSIGTA VVGGLMTGLI L....KLPLW GQPSDQNCYD DSVYWKVPKT R.........

        491                                                 533
Sponge    SRSRERIEAI GLRHRGVPAA DSPPVSGETG QQTNEENKQE TSI
NemF      VKQIEVAEYN PLSQKEV... .......... .......... ...
NemB      QEQERLRERE QMQEIY.... .......... .......... ...
Mouse     ANHNHVEHEV .......... .......... .......... ...
Macaque   GDHSQLEPEV .......... .......... .......... ...
Human     .......... .......... .......... .......... ...
```

Fig. 1. Multiple sequence alignment of Rh50 proteins. The analysis was performed using the program CLUSTALW (Thompson et al. 1994). The 12th predicted TM domain of the human protein is *underlined* and the beginning of its potential C-terminal cytoplasmic domain is indicated by the *arrow* (as from SWISS-PROT entry Q02094). *Dots* indicate indels. GenBank accession numbers are given in Table 1. NemF and NemB refer to nematode F08F3.3 and B0240.1 products, respectively.

the first potential extracellular loop (see Fig. 1). All indels concerned mainly gain or loss of hydrophilic residues. However, the most notable modifications seemed to have affected the C-terminal region of the protein. Two major deletions were visible; the first one was located in the last extracellular loop of the sponge–nematode proteins (roughly between residues 319 and residue 423; see Fig. 1), whereas the second one involved the C-terminal cytoplasmic domain. In all sequences under investigation, hydropathy analysis was carried out on the residues corresponding to the last predicted TM domain and to the C-terminus of the Rh50 protein. The results, shown in Fig. 2, clearly visualize that the potential C-terminal cytoplasmic domain is the longest in sponge, its length decreasing in nematode, mouse, and macaque (identical in these two species), to be the shortest in human (see also Fig. 1).

Noteworthily, the C-terminal cytoplasmic domain in macaque is 19 amino acids longer than in human (see Fig. 1). This difference is brought about by a G-to-T transversion in the human DNA sequence that creates a stop codon in the corresponding transcript. The stop codon lies at identical positions in the macaque and mouse sequences.

Analysis of the Nucleotide Region Surrounding the Stop Codon of the RH50 *Gene in Primates*

From the above results, it can be predicted that the G-to-T mutation, which resulted in the reduction of the length of the C-terminal cytoplasmic domain of Rh50 during evolution, should have taken place in the ape lineage, after the divergence between macaque and human.

In order to test this hypothesis, the genomic sequence surrounding the *RH50* exon 10 region was amplified in primate lineages, from prosimians to chimpanzee, and directly sequenced (see Materials and Methods). The results demonstrate that the G nucleotide, present in early primates, is replaced by a T_{stop} nucleotide (arrow) only in human, chimpanzee, and gorilla (Fig. 3). Reverting this mutation back to G would allow translation to continue in these species until the (ancestral?) stop codon used in all other primates (and mouse), which lies about 55 nucleotides downstream. It should be noted that the nucleotide sequence after the human stop codon is highly conserved among all primates.

Discussion

In this study, we have shown that the members of the Rh protein family followed quite distinct pathways during evolution.

The first striking finding is that the Rh50 glycoprotein

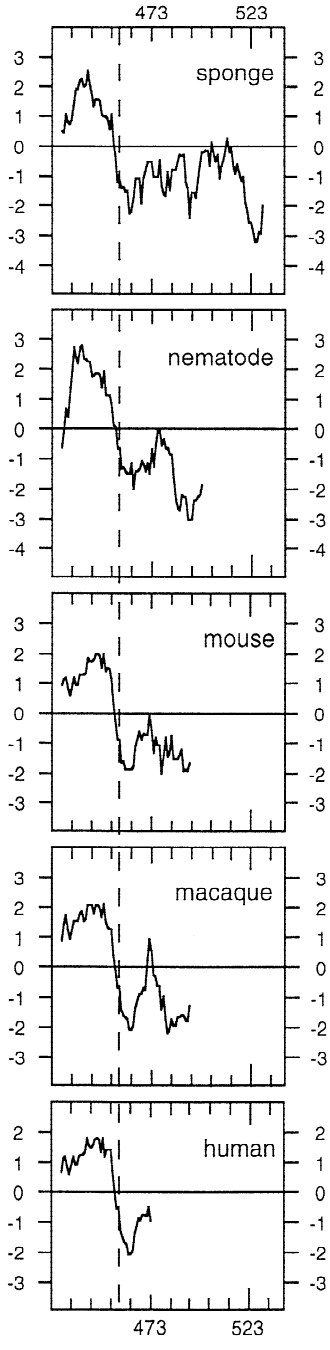

Fig. 2. Comparison of the hydropathy plots of the Rh50 C-terminal regions. The analysis was carried out using the method of Kyte and Doolittle (1982) with a window of nine amino acids. The results show that the length of the cytoplasmic domain decreases from sponge to human (see text). Hydrophobicity and hydrophilicity values are shown *above* and *below the horizontal line*, respectively. The *abscissa* indicates the residue number from Fig. 1. Species are ordered from top to bottom according to the date of divergence from the lineage leading to human. Residue 424, immediately following an indel (see Fig. 1), was arbitrarily chosen as a starting point for the analysis. The *vertical dashed line* identifies the beginning of the potential cytoplasmic domains based on the features of the human sequence (from the SWISS-PROT entry Q02094; this domain is believed to start at residue 456 in Fig. 1). Product B0240.1 was chosen to represent both nematode sequences.

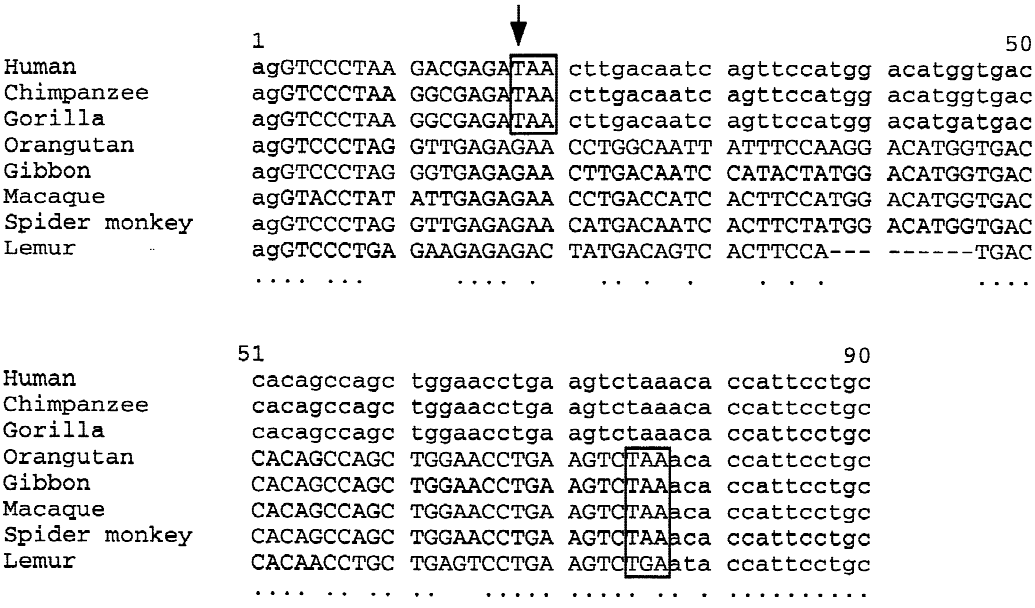

Fig. 3. The *RH50* exon 10 nucleotide region in primate lineages. Species are ordered from human to lemur, according to the currently accepted primate phylogeny (Goodman et al. 1994). The *arrow* indicates the G-to-T transversions in human, chimpanzee, and gorilla. Stop codons are *boxed.* Coding nucleotides are in *uppercase letters.* Positions 1 and 2 are the ''ag'' invariant nucleotides of the 3′ splice site of intron 9. *Dots* and *dashes* represent sequence identity and nucleotide indels, respectively.

evolved at a rate which is about 2.6 times slower than that of the Rh30 polypeptides. After gene duplication, it is often observed that one member of the duplicated gene family preserves its original function, under the same level of functional constraints, whereas the other is free to gain a new function, accumulating nonsynonymous substitutions more rapidly (Ohta 1989). Therefore, our data suggest that Rh30 may have experienced either a relaxation of selection due to a possible dispensability for the presence of the other protein (Rh50) or acquired a new function accompanied by loss of glycosylation and gain of palmitoylation sites (for review see Cartron 1994).

The evolutionary behavior of human/macaque *RH* genes is puzzling, as we observed some sort of coevolution between synonymous and nonsynonymous sites (i.e., the higher nonsynonymous rate is balanced by lower synonymous rate). The data suggest that nucleotide sequences in human and macaque may have a superimposed level of functional constraints in addition to those ascribed to protein coding (see also below). This is also confirmed by the high level of conservation of the primate *RH50* nucleotide sequences in the noncoding part immediately adjacent to the stop codon of the human, chimpanzee, and gorilla sequences (see also Fig. 3).

The two major events which shaped the evolution of the *RH* gene family were the split between the *RH50* and the *RH30* evolutionary pathways and the subsequent duplication event which gave rise to the *RHCE* and *RHD* genes. Here we have estimated at 250–346 Mya the evolutionary time at which the *RH50* and *RH30* genes separated from each other. From this result, it can be predicted that *RH30*-like sequences may also be present in birds, whose divergence time with mammals is estimated at about 310 Mya (Kumar and Hedges 1998). In agreement with this prediction, a positive hybridizing signal was detected in chicken DNA by low-stringency Southern blot analysis using the human Rh30 cDNA probe (Westhoff and Wylie 1994).

Altogether these results imply a more important and evolutionary conserved function of the Rh50 protein with respect to Rh30. This view is consistent with a possible role of Rh50 in the NH_4^+ transport system, as suggested by the finding that the human Rh50 protein shares a certain degree of similarity (~20–27% identity) with the Mep/Amt family of NH_4^+ permeases (Marini et al. 1997; Matassi et al. 1998).

However, as far as the functional role of Rh50 is concerned, the picture is probably far more complicated. Indeed, *C. elegans* is known to possess at least four ammonium transporter genes in addition to the two *RH50*-like sequences. In this respect, another exciting alternative is that Rh50 may (also) be used by the organism as an ammonium sensor. This suggestion is based on two observations. First, among the three members of the Mep protein family in yeast, the human Rh50 glycoprotein shares the highest similarity with MEP2 (24% identity, 40% similarity). Second, recent findings indicate that MEP2 may act as both transporter and ammonium sensor in yeast (Lorenz and Heitman 1998). However, it is also conceivable that the Rh50 protein may have modulated its function in a species-specific manner during evolution.

Despite following quite different evolutionary path-

ways, Rh50 and Rh30 proteins are functionally related, since they interact with each other in the human RBC membrane, to form the Rh complex together with other accessory chains, namely, CD47, LW, and GPB glycoproteins (see Introduction). Previous studies suggested that Rh50 and Rh30 proteins may interact via their N-terminal regions (Eyers et al. 1994). However, recent experimental evidence suggests that also the C-terminal region of the Rh50 protein may play a key role in protein–protein interactions in the assembly of a functional Rh complex (Chérif-Zahar et al. 1996, 1998a; Huang 1998).

In this respect, our results may shed new light on the formation of the Rh complex during evolution. Indeed, we have shown that a G-to-T_{stop} mutation in the *RH50* gene occurred in the common ancestor of anthropoid apes (human, chimpanzee, and gorilla) (see Fig. 3), resulting in a shorter C-terminal cytoplasmic domain of the protein. Most interestingly, this event appears to coincide with the second major change in the evolution of the *RH* gene family, namely, the duplication giving rise to the *RHCE* and *RHD* genes, which we have estimated here at 8.5 Mya. These results strongly suggest that an Rh complex, equivalent to that found in human, should also be present in chimpanzee and gorilla. Indeed, in these species the GPB protein is present (Xie et al. 1997), and most likely also CD47 and LW, whose homologues are found in mouse (Lindberg et al. 1993; P. Bailly, personal communication). Moreover, it can be speculated that in human (but probably in all apes), only the portion of the Rh50 C-terminal cytoplasmic domain, necessary for a functional Rh complex, has been conserved (i.e., the last 26 residues; see Fig. 1). In this respect, it is striking to note that the MEP2 C-terminal cytoplasmic domain comprises about 80 amino acids, the last 55 of which seem not to be required for both ammonium sensing and transport (Lorenz and Heitman 1998).

The structure of the Rh complex, if present in orangutan and in other early primates—compatible with the coevolution observed here between *RH50* and *RH30* genes in human and macaque—as well as in nonprimate mammals, may prove different. Further studies are currently under way in our laboratory aimed at elucidating whether protein–protein interactions among all the currently identified members of the Rh complex have been established earlier in evolution.

Acknowledgments. We thank Prof. Damian Labuda for the gift of primate DNAs. We also wish to thank Prof. Bruno Andre for helpful discussion. G.M. thanks ORTHO Clinical Diagnostics for financial support. G.P. benefited from partial funding by EC Grant BIO4-CT95-0130.

References

Agre P, Cartron JP (1991) Molecular biology of Rh antigens. Blood 78:551–563

Anstee DJ, Tanner MJ (1993) Biochemical aspects of the blood group Rh (rhesus) antigens. Baillieres Clin Haematol 6:401–422

Blancher A, Socha WW (1997) The Rhesus system. In: Blancher A, Klein J, Socha WW (eds) Molecular biology and evolution of blood group and MHC antigens in primates. Springer-Verlag, Berlin, pp 147–218

Carritt B, Kemp TJ, Poulter M (1997) Evolution of the human RH (rhesus) blood group genes: a 50 year old prediction (partially) fulfilled. Hum Mol Genet 6:843–850

Cartron JP (1994) Defining the Rh blood group antigens. Biochemistry and molecular genetics. Blood Rev 8:199–212

Cartron JP, Agre P (1993) Rh blood group antigens: protein and gene structure. Semin Hematol 30:193–208

Cartron JP, Bailly P, Le Van Kim C, Chérif-Zahar B, Matassi G, Bertrand O, Colin Y (1998) Insights into the structure and function of membrane polypeptides carrying blood group antigens. Vox Sang 74 (Suppl 2):29–64

Chérif-Zahar B, Mattei MG, Le Van Kim C, Bailly P, Cartron JP, Colin Y (1991) Localization of the human Rh blood group gene structure to chromosome region 1p34.3–1p36.1 by in situ hybridization. Hum Genet 86:398–400

Chérif-Zahar B, Raynal V, Gane P, Mattei MG, Bailly P, Gibbs B, Colin Y, Cartron JP (1996) Candidate gene acting as a suppressor of the RH locus in most cases of Rh-deficiency. Nature Genet 12:168–173

Chérif-Zahar B, Matassi G, Raynal V, Gane P, Delaunay J, Arrizabalaga B, Cartron JP (1998a) Rh-deficiency of the regulator type caused by splicing mutations in the human RH50 gene. Blood 92:2535–2540

Chérif-Zahar B, Matassi G, Raynal V, Gane P, Mempel W, Perez C, Cartron JP (1998b) Molecular defects of the *RHCE* gene in Rh-deficient individuals of the amorph type. Blood 92:639–646

Eyers SA, Ridgwell K, Mawby WJ, Tanner MJ (1994) Topology and organization of human Rh (rhesus) blood group-related polypeptides. J Biol Chem 269:6417–6423

Goodman M, Bailey WJ, Hayasaka K, Stanhope MJ, Slightom J, Czelusniak J (1994) Molecular evidence on primate phylogeny from DNA sequences. Am J Phys Anthropol 94:3–24

Huang CH (1998) The human Rh50 glycoprotein gene. Structural organization and associated splicing defect resulting in Rh(null) disease. J Biol Chem 273:2207–2213

Hyland CA, Chérif-Zahar B, Cowley N, Raynal V, Parkes J, Saul A, Cartron JP (1998) A novel single missense mutation identified along the RH50 gene in a composite heterozygote Rh_{null} blood donor of the regulator type. Blood 91:1458–1463

Kemp TJ, Poulter M, Carritt B (1996) A recombination hot spot in the Rh genes revealed by analysis of unrelated donors with the rare D—phenotype. Am J Hum Genet 59:1066–1073

Kumar S, Hedges SB (1998) A molecular timescale for vertebrate evolution. Nature 392:917–920

Kyte J, Doolittle RF (1982) A simple method for displaying the hydrophobic character of a protein. J Mol Biol 157:105–132

Lazano EM, Grau O, Romanowski V (1993) Isolation of RNA from whole blood for reliable use in RT-PCR amplification. Trends Genet 9:296

Lindberg FP, Gresham HD, Schwarz E, Brown EJ (1993) Molecular cloning of integrin-associated protein: An immunoglobulin family member with multiple membrane-spanning domains implicated in alpha v beta 3-dependent ligand binding. J Cell Biol 123:485–496

Lorenz MC, Heitman J (1998) The MEP2 ammonium permease regulates pseudohyphal differentiation in Saccharomyces cerevisiae. EMBO J 17:1236–1247

Marini AM, Urrestarazu A, Beauwens R, Andre B (1997) The Rh (rhesus) blood group polypeptides are related to NH4+ transporters. Trends Biochem Sci 22:460–461

Matassi G, Chérif-Zahar B, Mouro I, Cartron JP (1997) Characterization of the recombination hot spot involved in the genomic rear-

rangement leading to the hybrid D-CE-D gene in the D(VI) phenotype. Am J Hum Genet 60:808–817

Matassi G, Chérif-Zahar B, Raynal V, Rouger P, Cartron JP (1998) Organization of the human RH50A gene (RHAG) and evolution of base composition of the RH gene family. Genomics 47:286–293

Méténier-Delisse L, Hayes H, Leroux C, Giraud-Deville C, Levéziel H, Guérin G, Martin P, Grosclaude F (1997) Isolation and molecular characterization of bovine Rhesus-like transcript and chromosome mapping of the relevant locus. Anim Genet 28:202–209

Minghetti PP, Law SW, Dugaiczyk A (1985) The rate of molecular evolution of alpha-fetoprotein approaches that of pseudogenes. Mol Biol Evol 2:347–358

Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11:715–724

Ohta T (1989) Role of gene duplication in evolution. Genome 31:304–310

Rambaut A, Bromham L (1998) Estimating divergence dates from molecular sequences. Mol Biol Evol 15:442–448

Ridgwell K, Spurr NK, Laguda B, MacGeoch C, Avent ND, Tanner MJ (1992) Isolation of cDNA clones for a 50 kDa glycoprotein of the human erythrocyte membrane associated with Rh (rhesus) blood-group antigen expression. Biochem J 287:223–228

Saccone C, Lanave C, Pesole G, Preparata G (1990) Influence of base composition on quantitative estimates of gene evolution. Methods Enzymol 183:570–583

Salvignol I, Calvas P, Socha WW, Colin Y, Le Van Kim C, Bailly P, Ruffie J, Cartron JP, Blancher A (1995) Structural analysis of the RH-like blood group gene products in nonhuman primates. Immunogenetics 41:271–281

Seack J, Pancer Z, Muller IM, Muller WE (1997) Molecular cloning and primary structure of a Rhesus (Rh)-like protein from the marine sponge Geodia cydonium. Immunogenetics 46:493–498

Socha WW, Ruffié J (1983) The Rhesus system. In: Blood groups of primates. Theory, practice, and evolutionary meaning. Alan R. Liss, New York, pp. 75–90

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

Wagner FF, Gassner C, Muller TH, Schonitzer D, Schunter F, Flegel WA (1998) Three molecular structures cause rhesus D category VI phenotypes with distinct immunohematologic features. Blood 91: 2157–2168

Westhoff CM, Wylie DE (1994) Investigation of the human Rh blood group system in nonhuman primates and other species with serologic and Southern blot analysis. J Mol Evol 39:87–92

Xie SS, Huang CH, Reid ME, Blancher A, Blumenfeld OO (1997) The glycophorin A gene family in gorillas: Structure, expression, and comparison with the human and chimpanzee homologues. Biochem Genet 35:59–76

Yang Z (1994) Estimating the pattern of nucleotide substitution. J Mol Evol 39:105–111