Available online at www.sciencedirect.com
ScienceDirect
Genomics 88 (2006) 263–273
GENOMICS
www.elsevier.com/locate/ygeno

Comprehensive repertoire and phylogenetic analysis of the G protein-coupled receptors in human and mouse

Thóra K. Bjarnadóttir, David E. Gloriam, Sofia H. Hellstrand, Helena Kristiansson, Robert Fredriksson, Helgi B. Schiöth *

Department of Neuroscience, Biomedical Center, Uppsala University, Box 593, 751 24 Uppsala, Sweden

Received 10 January 2006; accepted 1 April 2006
Available online 6 June 2006

Abstract

Understanding differences in the repertoire of orthologous gene pairs is vital for interpretation of pharmacological and physiological experiments if conclusions are conveyed between species. Here we present a comprehensive dataset for G protein-coupled receptors (GPCRs) in both human and mouse with a phylogenetic road map. We performed systematic searches applying several search tools such as BLAST, BLAT, and Hidden Markov models and searches in literature data. We aimed to gather a full-length version of each human or mouse GPCR in only one copy referring to a single chromosomal position. Moreover, we performed detailed phylogenetic analysis of the transmembrane regions of the receptors to establish accurate orthologous pairs. The results show the identity of 495 mouse and 400 human functional nonolfactory GPCRs. Overall, 329 of the receptors are found in one-to-one orthologous pairs, while 119 mouse and 31 human receptors originate from species-specific expansions or deletions. The average percentage similarity of the orthologue pairs is 85%, while it varies between the main GRAFS families from an average of 59 to 94%. The orthologous pairs for the lipid-binding GPCRs had the lowest levels of conservation, while the biogenic amines had highest levels of conservation. Moreover, we searched for expressed sequence tags (ESTs) and identified more than 17,000 ESTs matching GPCRs in mouse and human, providing information about their expression patterns. On the whole, this is the most comprehensive study of the gene repertoire that codes for human and mouse GPCRs. The datasets are available for downloading.

Keywords: GPCR; Glutamate; Rhodopsin; Adhesion; Frizzled; Taste2; Secretin

G protein-coupled receptors (GPCRs) form one of the largest and most studied gene families of mammalian genomes. All GPCRs share a common functional unit in the form of seven α-helical transmembrane regions but many GPCRs also contain various functional domains, in particular within their highly diverse N-terminals. The main role of GPCRs is to recognize a diversity of extracellular ligands such as hormones, proteins, lipids, and pheromones and to transduce their signals into the cell [1]. GPCRs are expressed in virtually all types of tissues in the body and they are involved in most types of physiological and pathological processes. However, they are often expressed at low levels [2] and in specific cell types, which contributes to the fact that they are the most important family of proteins serving as targets in drug discovery. From all marketed drugs, around 45% are targeted at cell membrane receptors, mainly GPCRs [3]. However, these drugs affect only around 30 receptors, leaving a large amount of GPCRs as potential targets for new drugs [4]. The mouse is one of the main experimental models for testing drugs and it is therefore important to establish a complete nonredundant set of the orthologous pairs of mouse and human GPCRs based on validated phylogenetic methods. The differences between the mouse and the human GPCRs are important for design and interpretation of virtually all kinds of physiological and pharmacological studies when conclusions are extended from one species to the other.

A range of methods has previously been used to accumulate repertoires of GPCRs in the human genome.

Abbreviations: BLAST, basic local alignment search tool; BLAT, BLAST-like alignment tool; EST, expressed sequence tag; GPCR, G protein-coupled receptor; HMM, Hidden Markov model; MP, maximum parsimony; RPS-BLAST, reverse positions-specific BLAST; TM, transmembrane region; 7TM, seven transmembrane regions.

* Corresponding author. Fax: +46 18 51 15 40.
E-mail address: helgis@bmc.uu.se (H.B. Schiöth).

0888-7543/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2006.04.001

Early estimations by the Human Genome Sequencing Project were based on automatic genome-wide approaches. One approach was to use protein families (using Gene Ontology and Celera's Panther Classification) as well as protein domains (using Pfam and SMART) to group human genes into families. This led to an estimate of 616 GPCRs in the human genome of the Rhodopsin, Secretin, and metabotropic Glutamate types [5]. Another method was based on analysis with InterPro, a tool for combining sequence-pattern information from four databases (PRINTS, Prosite, Pfam, and Prosite Profile), which resulted in an estimate of 569 Rhodopsin-like GPCRs in total [6]. Takeda and colleagues extracted a high number of open reading frames from the human genome that had 200–1500 amino acid residues similar to those of GPCRs [7]. In another study that included both the human and the mouse GPCRs, a high number of nonsensory GPCRs were collected through TBLASTN and HMM model searches [8]. This study included a number of pseudogenes and partial sequences. Our group also provided an initial list of 342 functional nonolfactory human GPCRs [9] using a combination of manual and semiautomatic search methods.

Various types of progress have been made since these studies, enabling a more improved assessment of the identity and number of GPCRs. First, the human genome assembly has become much more accurate and it now has very few gaps [10]. The mouse genome has also become much more complete. Not only has the genome sequence assembly undergone large improvements, the ENSEMBL and Refseq databases have also expanded rapidly, although much work remains to be done to obtain a conclusive number of all protein coding genes [11]. Second, a number of new human GPCRs have been reported since these previous works [12–15] and several other known receptors were simply missing in previous studies. Overall, the different studies have used different methods, subsets, and criteria, which gives a reason to aim for an all-inclusive list of the human and mouse GPCRs. Third, the continually growing EST (expressed sequence tags) databases and their match to the genome assembly have been very useful in determining which gene predictions code for either a pseudogene or a functional gene. Because of the speed in which ESTs can be generated and the relatively low cost associated with their generation, EST databases have been growing rapidly. In the past 10 years, the National Center for Biotechnology Information (NCBI)-hosted dbEST database has grown from approximately 430,000 to over 28 million ESTs for all species (whereof over 6 million belong to human and 4 million to mouse) [16]. The ESTs can thus today also provide valuable information on in which tissues the genes are expressed.

Here we present a comprehensive dataset of GPCRs in both human and mouse excluding only the olfactory and pheromone receptors of type 1, which have been studied in detail elsewhere [17–19]. We present our own mouse and human dataset acquired by systematic searches using several search tools such as BLAST, BLAT, Hidden Markov models (HMMs), and literature data. Here the emphasis has been on providing a full-length version of each human and mouse GPCR in only one copy referring to a single chromosomal position. We have determined the mouse and human orthologues pairing and made a clear distinction, as far as that can be done, between functional and nonfunctional genes. We also focused on creating accurate phylogenetic trees that are not only the criteria for orthologous pairing but also show resolution between families and groups. The subdivision of this dataset follows the phylogenetic classification of the GRAFS system [9]. This classification system does not consider how the receptors bind ligands or physiological and structural features that have been previously considered for classification [20,21]. The classification

Table 1
The total number of GPCRs in the repertoires of mouse and human

Group	Number in mouse	Number in human	Peptide ligand	Biogenic amine ligand	Lipid ligand	Purin ligand	Other ligand	Orphan
Glutamate	79	22	0	10	0	0	4	65 (8)
Rhodopsin (α)	105	101	8	49 (41)	20	4	6 (8)	18 (20)
Rhodopsin (β)	46	43	36 (37)	0	0	0	0	10 (6)
Rhodopsin (γ)	67	64	49 (51)	0	4	0	2	12 (7)
Rhodopsin (δ)	82	63	43 (22)	0	11	10 (12)	2 (1)	16 (17)
Adhesion	31	33	0	0	0	0	1 (2)	30 (31)
Frizzled	11	11	0	0	0	0	11	0
Taste type 2	34	25	0	0	0	0	1 (4)	33 (23)
Secretin	15	15	15	0	0	0	0	0
V1R	165	3	—	—	—	—	—	—
Olfactory	1037	388	—	—	—	—	—	—
Others	25[a]	23[a]	—	—	—	—	—	—
Total	1697	791	151 (133)	59 (51)	35	14 (16)	27 (32)	184 (112)

The GPCRs are divided into families according to the GRAFS classification system (*Glutamate, Rhodopsin* (α), *Rhodopsin* (β), *Rhodopsin* (γ), *Rhodopsin* (δ), *Adhesion, Frizzled, Taste2, Secretin*). Additionally the most recent published numbers for pheromone receptors type 1 (*V1R*) [29,30] and *Olfactory* receptors [18,19] are given. A ligand preference for the different families is showed in columns 3–8. *Others* (all receptors are found in both species unless otherwise noted): DARC (Duffy) GPR137 (C11ORF4), GPR23, GPR88, GPR120, GPR135, GPR139, GPR141 (and Gpr141b only in mouse), GPR142, GPR146, GPR152, GPR160, GPR151 (GPCR-2037), HGPCR19, GPR149 (IEDA), GPR143 (OA1), GPR172A (PERVAR1) (found only in human), GPR172B (PERVAR2) (found only in human), TRHR (and Trhr2 only in mouse), GPR137B (TM7SF1), GPR137C (TM7SF1L2), TM7SF3, GPR175 (TPRA40).

[a] See Materials and methods.

and the orthologue assignment presented here are thus based solely on strict phylogenetic criteria. The trees are based on careful alignments made from the common region to all GPCRs, the 7TMs. Moreover, to provide a better overview of expression patterns, we searched for ESTs for the receptors and we display matches of more than 17,000 ESTs to the mouse and human GPCRs.

Results

We conducted a comprehensive search for GPCR sequences in the mouse and human genomes. Both automatic and manual methods (see Materials and methods) were used for the purpose of compiling an accurate and up-to-date GPCR dataset. The receptors were classified according to the GRAFS classification system using names of common nature: *Glutamate, Rhodopsin, Adhesion, Frizzled/Taste2,* and *Secretin.* These names (in italic) are used throughout the paper. The *Rhodopsins* were further divided into α, β, γ, and δ groups. Phylogenetic trees were then calculated for each family and group. The number of receptors in each of the families and the ligand of each receptor can be seen in Table 1. In addition, we list numbers for olfactory (*OLF*) [18,19] and pheromone type 1 (*V1R*) receptors [29,30]. When we include the olfactory GPCRs, the overall repertoire covers a total of 1697 mouse GPCR sequences and 791 human sequences, excluding pseudogenes. During the final steps of this work the IUPHAR published a GPCR list containing 359 human nonsensory receptors, including some pseudogenes [31]. All these receptors were found in our dataset (see Supplementary Material 11 and 12) and it is also more extensive that the IUPHAR list.

Overall the largest differences between the human and the mouse repertoires are seen within the *OLF* (1037 mouse and 388 human), the *Glutamate* (79 mouse and 22 human), and the *V1Rs* (165 mouse and 3 human). The other families have much less variation between the species, like the *Adhesions* (31 mouse and 33 human) [13]. Two families have exactly the same number of receptors: *Frizzled* (11 in both mouse and human) and *Secretin* (15 in both mouse and human). An overview of phylogenetic relationships for all GRAFS families can be seen in Fig. 1. The clear majority (329) of these GPCRs make up one-to-one orthologous pairs between mouse and human.

Two of the GRAFS families, the *Adhesion* and the *Glutamate* families, are studied in more detail elsewhere [13,32]. There have not been any new annotations within these families since, except that we have now also listed two human pseudogenes (AB065916, AL512625) for the *Adhesion* family. BLAST searches show that both are most closely related to GPR116 and according to RPS-BLAST they contain *Adhesion*-like domains (data not shown). No mouse orthologue could be found for either of these two sequences. The human and mouse *Frizzled* and *Secretin* receptors are all found in one-to-one orthologous pairs. The repertoire and evolution of the *Secretin* family have recently been addressed by Cardoso and colleagues [33].

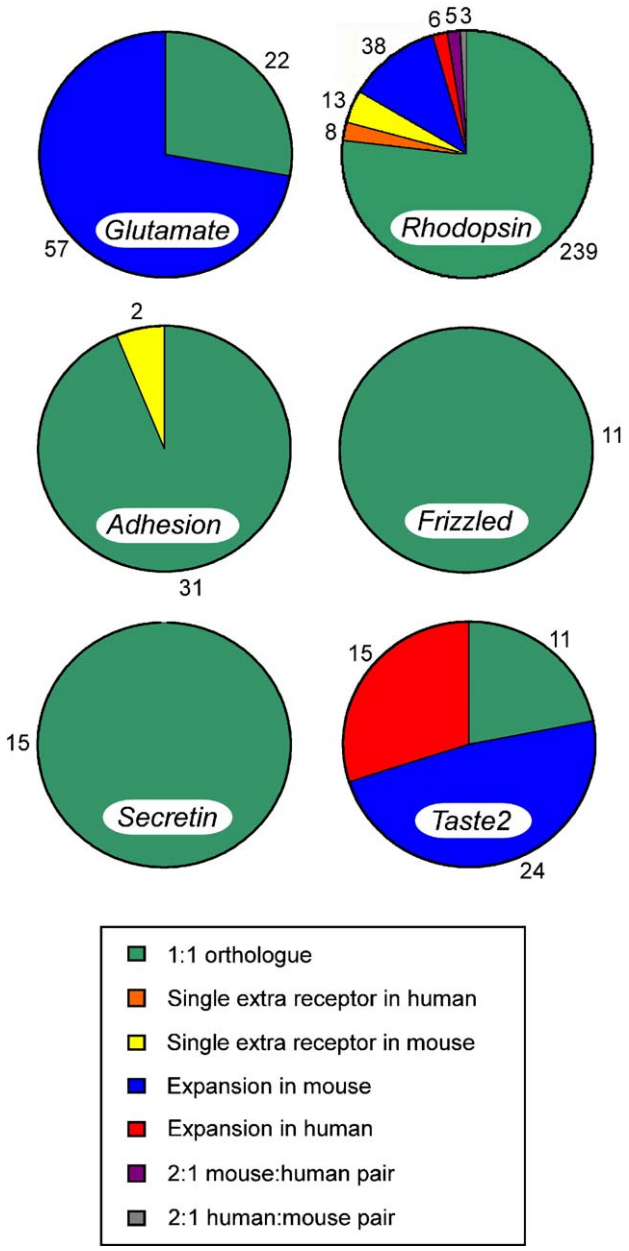


Fig. 1. Summary of the orthologous relationship of mouse and human GPCRs in the different GRAFS families. Green area represents the number of sequences that constitute a one-to-one orthologous relationship, purple a single extra receptor in mouse, yellow a single extra receptor in human, blue an expansion in mouse, and red an expansion in human. The numbers show how many receptors each area represents. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Figs. 2–5 show the phylogenetic trees for each subgroup, α, β, γ, and δ, of the *Rhodopsin* family. GPCRs lacking a functional orthologue are distinguished with a small black dot or box at the end of the branch. The dots represent receptors that could not be found within the genome of the other species and the boxes indicate that the receptors were found as pseudogenes within the genome of the other species. There are 105 mouse and 101 human members in the α group (see Fig. 2), whereof a total of 93 receptors arranged as one-to-one

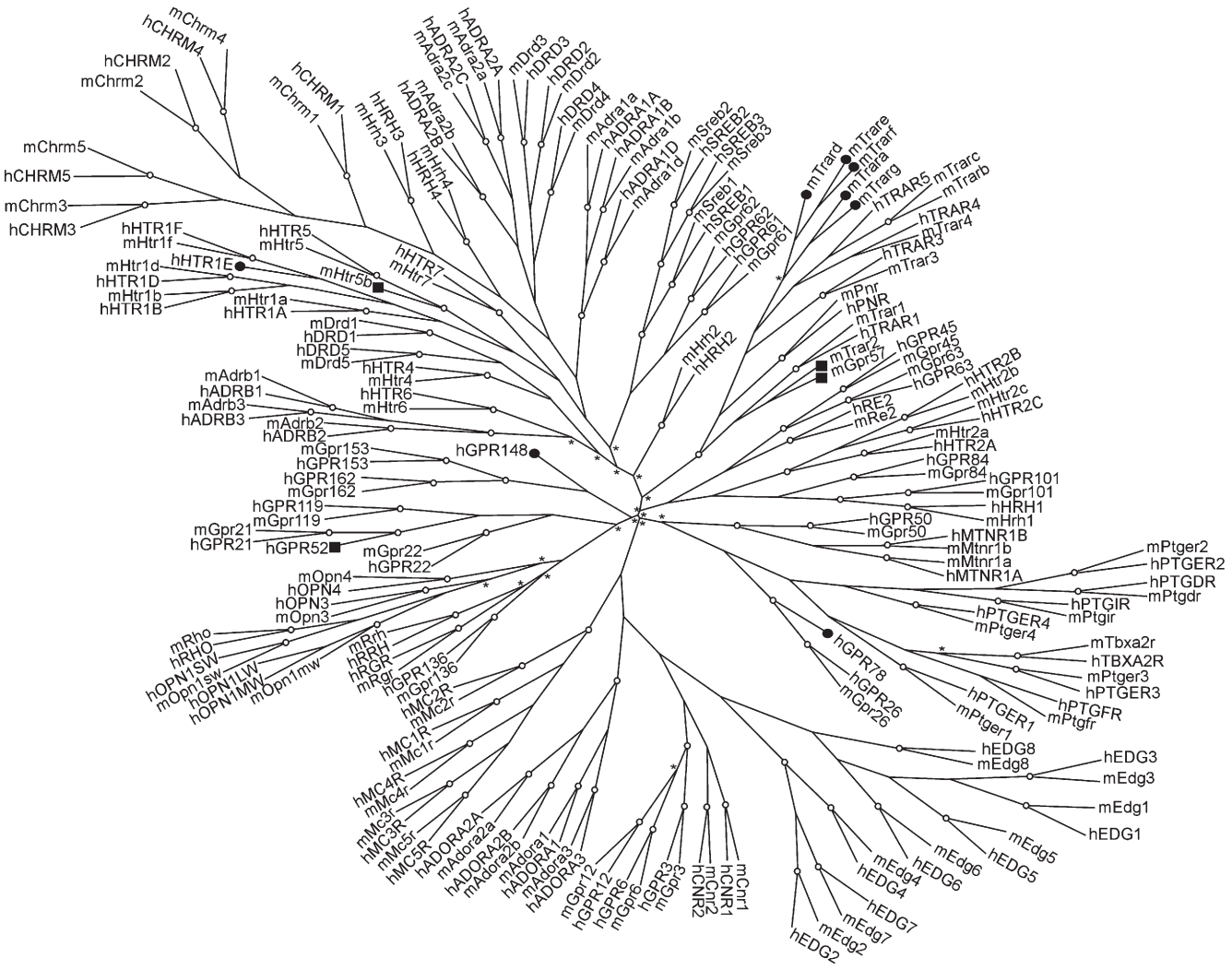


Fig. 2. A phylogenetic analysis of 105 mouse and 101 human *Rhodopsin* α-group GPCRs. The seven transmembrane regions from TMI to the end of TMVII were used to obtain a phylogenetic tree using 100 bootstrap replicas of the dataset in a Maximum Parsimony analysis. Nodes with bootstraps of 100 are indicated with a white circle, whereas nodes with bootstraps under 50 are denoted with a star. A black dot at the end of a branch represents a receptor not found in the genome of the other species and a black box at the end of a branch means that the sequence was found as a pseudogene in the genome of the other species.

orthologous pairs. They all pair up with a bootstrap value of 100. A subgroup expansion in mouse has resulted in several trace amine receptors that are lacking a human orthologue (mTrara, mTrarb, mTrarc, mTrard, mTrare, mTrarf, and mTrarg). Two trace amine receptors could be found only as pseudogenes in humans (orthologues of mTrar2 and mGpr57) [34,35]. One receptor in the same group could be found only in the human genome (hTRAR5). Apart from the trace amine expansion, 3 receptors could not be found within the mouse genome (hGPR78, hGPR148, and hHTR1E) and 1 was found as a pseudogene in mouse (orthologue to hGPR52). There exist two "2:1 orthologous pairs," with an extra copy of the receptor in either mouse or human (mHtr5b, hOPN1LW). The β group has 46 mouse and 43 human members. Of these, 39 make up one-to-one orthologous pairs, all with bootstrap values of 100. One of the mouse receptors has a pseudogene orthologue in human (hNPY6R). The same is true for 3 mouse receptors (mGpr165, mPgr15l, and mGprN1). Two receptors are found in only one of the two species, hMLNR and mGprN1. 2:1 orthologous pairs can be seen in mGpr103a, mGpr103b, and GPR103 as well as in GNRHR, Gnrhr, and GNRHRII. The γ group consists of 67 mouse and 64 human family members. A majority of the 55 receptors make up one-to-one orthologous pairs. Five receptors could be found only within the human genome (hFPRL1, hFPRL2, hGPR8, hGPR145 (hMCHR2), and hGPR32) and 6 only in the mouse genome (mGpr33, mFprl1, mFprrs2, mFprrs3, mFprrs4, and mFprrs6). Orthologous pairs of 2:1 appear for hAGTR1 and hCCR1 (both have two copies in mouse). The last *Rhodopsin* group, the δ group, has 82 mouse and 63 human members. One-to-one orthologous pairs exist for 52 members. Orthologous pairs of 2:1 are seen for P2Y10 (two in mouse) and HM74 (two in human). mGpr90 is found only in mouse and the following 4 are found only in human: hMRG, hP2Y11, hP2Y8, and hTG1019. An expansion is seen for the MAS-related GPCRs that accounts for 26 mouse receptors and 4 human receptors.

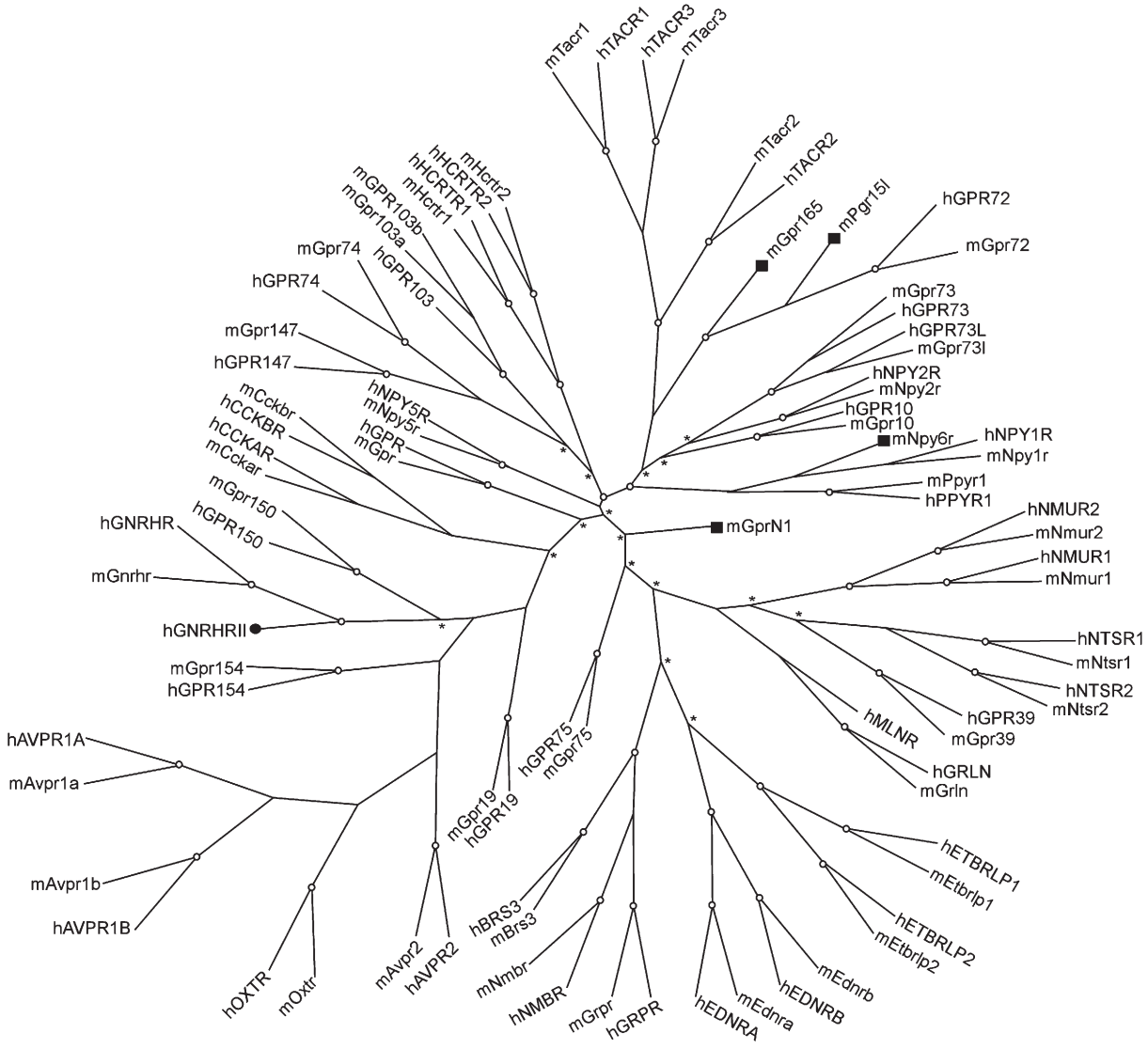


Fig. 3. A phylogenetic analysis of 46 mouse and 43 human *Rhodopsin* β-group GPCRs. The seven transmembrane regions from TMI to the end of TMVII were used to obtain a phylogenetic tree using 100 bootstrap replicas of the dataset in a Maximum Parsimony analysis. Nodes with bootstraps of 100 are indicated with a white circle, whereas nodes with bootstraps under 50 are denoted with a star. A black dot at the end of a branch represents a receptor not found in the genome of the other species and a black box at the end of a branch means that the sequence was found as a pseudogene in the genome of the other species.

We performed a thorough EST search for the whole dataset. It resulted in a total of 7467 mouse and 10,244 human ESTs from various tissues, each matching a unique GPCR. The number of ESTs differs greatly between individual receptors, ranging from none to 250 ESTs. We show this in two overview figures. In Fig. 6A we highlight the receptors that have a highly tissue-specific tissue pattern in mouse. A total of 62 receptors were found to match a high number of ESTs in only 1 of the 13 "major" organs or tissues in the body. A majority of these receptors are expressed in the brain (34), eye (7), and female reproductive tissue (6). The highest number of ESTs, a total of 250, was found for the Rho receptor in eye. Other ESTs found in high numbers include Mrgb1 (92) and Mrgb10 (36) in skin, Rgr (76) and Opn1sw (26) in eye, and Cnr1 (36), Cckbr (31), and Ntsr2 (36) in brain. Fig. 6B illustrates a total of 72 receptors with wide tissue distribution in the mouse or with expression in at least 5 different tissues. Two receptors, P2y5, and Gpr137b (Tm7sf1) are found in all 14 of the tissues and Gpr56 is found in 13 of them.

Discussion

We present a detailed and up-to-date dataset for the mouse and human GPCR receptor repertoires with carefully constructed high-resolution phylogenetic trees for each of the family. This repertoire contains receptors not included in previous studies of the human [7,9] and mouse [8] GPCRs. Thus it holds the most comprehensive set of full-length functional mouse and human GPCRs, including 495 mouse and 400 human nonolfactory GPCRs. Most of the nonolfactory receptors, as we have shown earlier, could be classified into the five GRAFS families according to phylogeny [9]. Only 25 mouse receptors and 23 human receptors could not be classified.

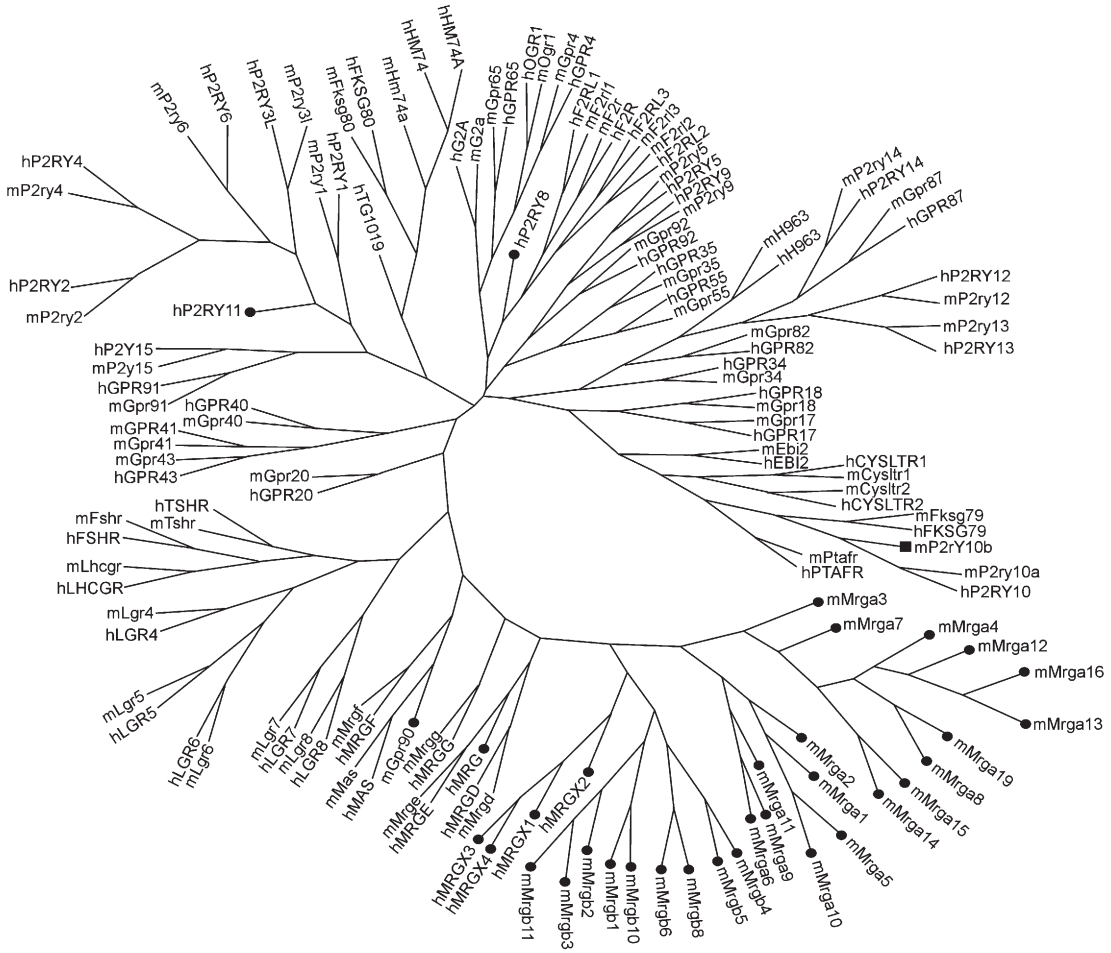


Fig. 4. A phylogenetic analysis of 67 mouse and 64 human *Rhodopsin* γ-group GPCRs. The seven transmembrane regions from TMI to the end of TMVII were used to obtain a phylogenetic tree using 100 bootstrap replicas of the dataset in a Maximum Parsimony analysis. Nodes with bootstraps of 100 are indicated with a white circle, whereas nodes with bootstraps under 50 are denoted with a star. A black dot at the end of a branch represents a receptor not found in the genome of the other species and a black box at the end of a branch means that the sequence was found as a pseudogene in the genome of the other species.

In general the receptors within the GRAFS families form well-conserved orthologous pairs and a clear majority of these pair with the highest possible bootstrap values. The *Rhodopsin* family is the largest of the GRAFS families and it holds many interesting differences between the human and the mouse repertoire. Single receptor gene deletions or duplications are not uncommon in the *Rhodopsin* family and it also contains clusters of receptors originating from large expansions. Most of the receptors in the *Adhesion* and *Frizzled/Taste2* families are technically orphans, while the receptors in the *Rhodopsin* family are known to bind to various types of ligands. Each of the *Rhodopsin* α, β, γ, and δ subgroups contains receptors that bind different kind of ligands (see Table 1) despite the fact that all of these four groups contain receptors that bind peptide ligands. The receptors that bind lipids are the only ligand group that has an exactly equal number of receptors in each subfamily in human and mouse. This might suggest that the lipid-binding GPCRs are more highly conserved between the two species than the receptors that bind other ligands. However, calculation of the percentage of identical and positively matching amino acids in the TM regions shows that this is not the case. The receptor pairs that bind lipids have averages of 86% identical amino acids and 91% positive matching amino acids, which are lower than for each of the other ligand groups (see Supplementary Material 10). It seems thus that the presence of lipid-binding GPCRs is well conserved between the mouse and the human genomes even though their amino acid identities are somewhat lower than for the other *Rhodopsins.* This may reflect that the lipid ligands in general do not provide the same evolutionary constraints on the primary structure of their receptors, while it does not suggest a less important functional role for these receptors, which could be implicated for the receptors that are missing an orthologue in one of the species. We calculated branch lengths for the *Rhodopsin* trees (see supplementary material). A few receptors in the α and β groups had proportionally longer branches than the others. Three of these, GPR75, GPR148, and GPR153, have orthologues in fish (data not shown), suggesting that they are ancient. Orthologues for GPR162 could be found in both chicken and mammals. GPR62 and mGprN1 were found neither in fish nor in chicken, but in mammals only, suggesting that they evolved after the split of Aves but

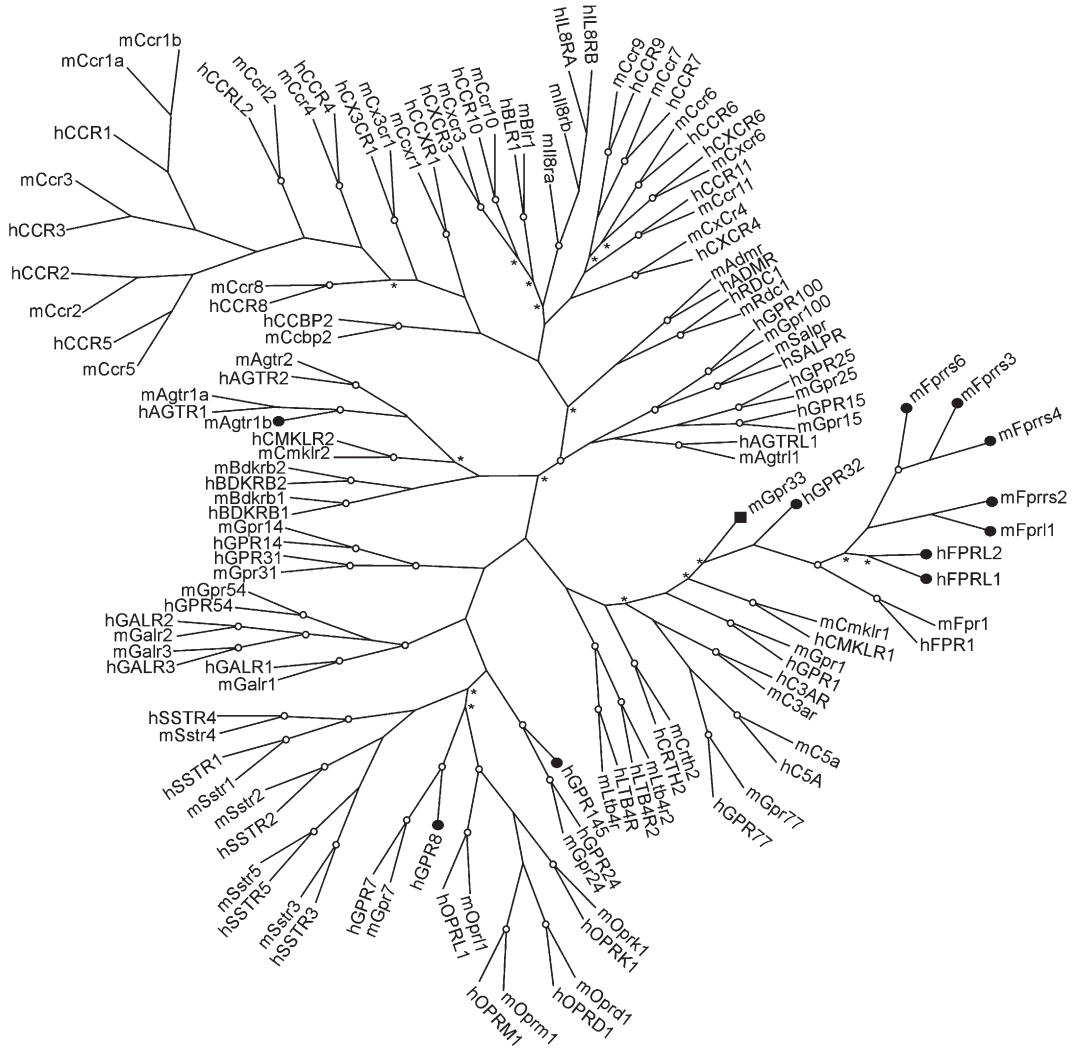
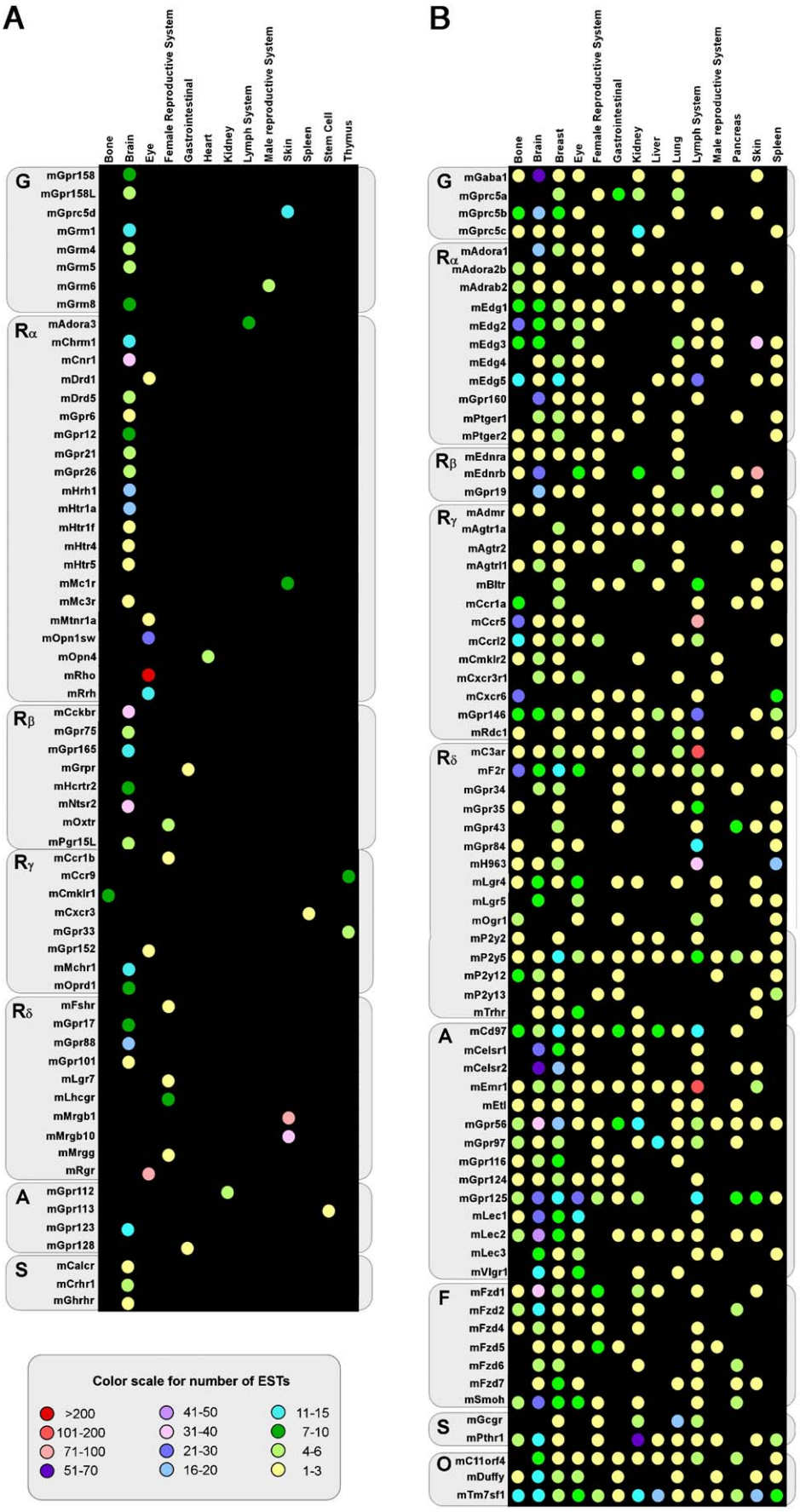


Fig. 5. A phylogenetic analysis of 82 mouse and 63 human *Rhodopsin* δ-group GPCRs. The seven transmembrane regions from TMI to the end of TMVII were used to obtain a phylogenetic tree using 100 bootstrap replicas of the dataset in a Maximum Parsimony analysis. Nodes with bootstraps of 100 are indicated with a white circle, whereas nodes with bootstraps under 50 are denoted with a star. A black dot at the end of a branch represents a receptor not found in the genome of the other species and a black box at the end of a branch means that the sequence was found as a pseudogene in the genome of the other species.

before the split of rodents and humans. Thus these receptors arose relatively recently and according to the branch lengths of the trees they seem to have undergone relatively more evolutionary changes than the other *Rhodopsins,* suggesting that they are evolving at a higher rate.

There are also some important differences in the level of conservation between the different families. The group that has the highest conservation in the TM regions between the orthologue pairs (see Supplementary Table 1) is the *Frizzled* GPCRs. On average, the *Frizzled* human–mouse pairs have 94% identical amino acids. This can be contrasted with the *Taste2* receptors, which have the lowest conservation between the orthologous pairs of 59%. This reflects the evolutionary history of these groups, the *Frizzled* receptors being highly conserved in vertebrates and even in invertebrates, while the *Taste2* group seems to be rapidly evolving and expanding in number [36]. The *Taste2* family appears to have undergone a complex combination of local duplications, interchromosomal duplications, and deletions [37]. The expansion in the *Taste2* family is similar to those seen for distinct clusters of the *Rhodopsin* family: olfactory, trace amine, and chemosensory receptors, as well as pheromone receptors [34,38–40]. These rapid species-specific expansions coincide with the fact that they all bind relatively small molecular ligands. It can be speculated that there are only relatively few receptor residues that are involved in binding, which could contribute to the 3D structure being not as fixed as in receptors binding larger complex molecules, such as in the case of Wnt protein bound by the *Frizzled* receptors. It is thus possible that this type of ligand-binding mode contributes to the "dynamic" gene repertoire.

There are several receptors that could not be included in the GRAFS families. These have divergent compositions in the TM region with low amino acid identity to other GPCRs. Thus search tools like RPS-BLAST often fail to detect their TM regions. In our attempts to classify the atypical receptors comprising the group *Others* (see Table 1) we used BLAST searches and alignments. We found that many of the receptors showed similarities to TM regions of


A
Bone
Brain
Eye
Female Reproductive System
Gastrointestinal
Heart
Kidney
Lymph System
Male reproductive System
Skin
Spleen
Stem Cell
Thymus
G
mGpr158
mGpr158L
mGprc5d
mGrm1
mGrm4
mGrm5
mGrm6
mGrm8
Rα
mAdora3
mChrm1
mCnr1
mDrd1
mDrd5
mGpr6
mGpr12
mGpr21
mGpr26
mHrh1
mHtr1a
mHtr1f
mHtr4
mHtr5
mMc1r
mMc3r
mMtnr1a
mOpn1sw
mOpn4
mRho
mRrh
Rβ
mCckbr
mGpr75
mGpr165
mGrpr
mHcrtr2
mNtsr2
mOxtr
mPgr15L
Rγ
mCcr1b
mCcr9
mCmklr1
mCxcr3
mGpr33
mGpr152
mMchr1
mOprd1
Rδ
mFshr
mGpr17
mGpr88
mGpr101
mLgr7
mLhcgr
mMrgb1
mMrgb10
mMrgg
mRgr
A
mGpr112
mGpr113
mGpr123
mGpr128
S
mCalcr
mCrhr1
mGhrhr
Color scale for number of ESTs
>200
101-200
71-100
51-70
41-50
31-40
21-30
16-20
11-15
7-10
4-6
1-3
B
Bone
Brain
Breast
Eye
Female Reproductive System
Gastrointestinal
Kidney
Liver
Lung
Lymph System
Male reproductive System
Pancreas
Skin
Spleen
G
mGaba1
mGprc5a
mGprc5b
mGprc5c
Rα
mAdora1
mAdora2b
mAdrab2
mEdg1
mEdg2
mEdg3
mEdg4
mEdg5
mGpr160
mPtger1
mPtger2
Rβ
mEdnra
mEdnrb
mGpr19
Rγ
mAdmr
mAgtr1a
mAgtr2
mAgtrl1
mBltr
mCcr1a
mCcr5
mCcrl2
mCmklr2
mCxcr3r1
mCxcr6
mGpr146
mRdc1
Rδ
mC3ar
mF2r
mGpr34
mGpr35
mGpr43
mGpr84
mH963
mLgr4
mLgr5
mOgr1
mP2y2
mP2y5
mP2y12
mP2y13
mTrhr
A
mCd97
mCelsr1
mCelsr2
mEmr1
mEtl
mGpr56
mGpr97
mGpr116
mGpr124
mGpr125
mLec1
mLec2
mLec3
mVlgr1
F
mFzd1
mFzd2
mFzd4
mFzd5
mFzd6
mFzd7
mSmoh
S
mGcgr
mPthr1
O
mC11orf4
mDuffy
mTm7sf1

more than one *Rhodopsin* subgroup, making the classification ambiguous. The four orphan receptors GPR141, GPR146, GPR152, and GPR151 can, for example, be aligned with the members of both the γ and the δ groups. The receptors grouped in *Others* are thus GPCRs that have TM regions that differ from the amino acid sequence of those we have grouped.A majority of the members in *Others* exist in both species with a one-to-one orthologous relationship. Only two of the receptors have an extra copy in mouse and thus exist in a 2:1 orthologous pair. One is GPR141 in human, which exists in two copies in mouse. mGpr141a and mGpr141b are located on chromosome 13qA2 in close proximity to one another, which suggests a local duplication in mouse. The other is TRHR, which has two mouse orthologues on different chromosomes. Interestingly, several of the receptors seem to be highly conserved in evolutionary terms and we find orthologues to GPR23, GPR135, GPR139, GPR146, GPR149 (IEDA), GPR143 (OA1), GPR172A (PERVAR1), and GPR137B (TM7SF1) in teleost fish (data not shown), suggesting that they arose early in vertebrate evolution at least 450 million years ago. It seems thus clear that the group of *Others* does not primarily hold receptors that diverged rapidly within the mammalian lineage; several of these may represent unique subgroups or families that arose early but did not expand into multiple members like the GRAFS families that we show in the phylogenetic trees.

The use of EST sequences has proven to be very useful to indicate gene expression, in particular for those genes that were not previously studied with detailed methods in many tissues. Considering the number of EST libraries and the number of different tissues that are now found in the EST databases, it was no surprise that the GPCRs in our lists matched numerous ESTs from various tissues. The numbers of ESTs found for each receptor vary widely from none up to 250. Fig. 6A shows receptors that we found in high numbers in 1 of the 13 "major" tissues in mouse. The rhodopsin and opsin receptors, opsin 1 short-wave sensitive (Opn1sw), rhodopsin (Rho), peropsin (Rrh), and retinal GPCR (Rgr), are, for example, found in abundant numbers in the eye, as expected. Interestingly, more than 50% of these abundant GPCRs are expressed exclusively in brain tissues, emphasizing the important roles of GPCRs in CNS functions. The result also highlight receptors that have high number of ESTs in a single tissue, such as the Mas-related GPCRs Mrgb1 and Mrgb10 that are found in skin. Neither of these receptors has, however, a human orthologue. It is also notable that the orphans Gpr88, Gpr123, and Gpr165 are all expressed highly and exclusively in brain tissue. Several of the receptors seem to be expressed abundantly in many tissues and we display those with widest tissue distribution in Fig. 6B (i.e., those that are expressed in at least 5 tissues). Two receptors, P2y5 and Gpr137b (Tm7sf1), are found in all 14 of the tissues and Gpr56 is found in 13 of them. There are several receptors with widespread distribution but also high peaks in specific tissues, including Gaba1, which has a high number of ESTs in brain (65), Ednrb in skin (80), C3r in the lymph system (107), Celsr2 in brain (56), Emr1 in the lymph system (144), and Pthr1 in kidney (65). Even though these receptors are widely distributed none of them shows a housekeeping gene pattern because of the variation between the tissues.

In summary we present an up-to-date dataset for nonolfactory mouse and human GPCRs. Phylogenetic trees with high resolution were calculated for each of the GRAFS families and here we highlight the differences between the two species. The comprehensive EST list indicates the individual expression patterns for the receptors, which are in particular valuable for those receptors for which no or only very simple tissue distribution studies have been performed. This comprehensive dataset of GPCRs is a solid platform for functional analysis and drug discovery of GPCRs for which comparison between results from mouse and human is necessary.

Materials and methods

Identification of sequences

Initial dataset

Known human protein sequences were used as bait for search of "all" sequences belonging to the family of GPCRs, in the genome databases of human (*Homo sapiens*) and mouse (*Mus musculus*). The initial dataset was obtained from our previous publication [9]. The olfactory receptors were excluded from the analysis since they have been extensively analyzed elsewhere [22].

Identifying GPCR sequences with BLAST, BLAT, and Hidden Markov models

Three different approaches were used to identify GPCR-related sequences: BLAST [23] searches, BLAT searches, and HMMs. The BLAST searches were carried out in NCBI's database and the Celera Genomics database (http://www.celera.com). HMMs tend to be more sensitive search tools than BLAST in the identification of more distant homologues. An initial dataset of truncated GPCRs (only the TM regions) was aligned using ClustalW 1.81 [24]. HMMs were constructed from the alignments, using the HMMER 2.2 package [25]. These HMMs were subsequently used to search Genscan datasets of the public human (NCBI build 32) and mouse (NCBI build 30) genome sequences, downloaded from the NCBI ftp site (at ftp.ncbi.nlm.nih.gov/genbank/). The HMMsearch cutoff was $E = 1 \times 10^{-4}$. The HMM was constructed using default settings and calibrated using HMMcalibrate. An RPS-BLAST was carried out to verify that all the sequences contained a seven-helical-transmembrane region.

Fig. 6. (A) Receptors with many ESTs in only one of the following 13 tissues: bone, brain, eye, female reproductive system, gastrointestinal, heart, kidney, lymph system, male reproductive system, skin, spleen, stem cell, and thymus. The receptors are divided into the following families/groups: *Glutamate* (G), α-group *Rhodopsin* (R_α), β-group *Rhodopsin* (R_β), γ-group *Rhodopsin* (R_γ), δ-group *Rhodopsin* (Rδ), *Adhesion* (A), and *Secretin* (S). The color scale indicates how many ESTs support each gene. See Materials and methods for further details. (B) Receptors with the widest tissue distribution in the following 14 tissues: bone, brain, breast, eye, female reproductive system, gastrointestinal, kidney, liver, lung, lymph system, male reproductive system, pancreas, skin, and spleen. The receptors are divided into the following families/groups: *Glutamate* (G), α-group *Rhodopsin* (R_α), β-group *Rhodopsin* (R_β), γ-group *Rhodopsin* (R_γ), δ-group *Rhodopsin* (Rδ), *Adhesion* (A), *Frizzled* (F), *Secretin* (S), and *Others* (O). The color scale indicates how many ESTs support each gene. See Materials and methods for further details.

Manual inspection of sequences

All the sequences found in the above searches were checked thoroughly to ensure that they had both an intact transmembrane region and an N-terminus of appropriate length (i.e., pheromone receptors should have ANF domains according to RPS-BLAST). Those with incomplete regions were inspected further. BLAT searches of incomplete genes were carried out in UCSC's Genome Bioinformatic database (http://www.genome.ucsc.edu/) to find the entire coding regions. The May 2004 human and mouse genome assemblies were used. In many cases the nucleotide sequences contained the complete coding regions and then the nucleotide sequences were collected by assembling all exons manually (in some cases up 25 exons), verifying intron–exon boundaries, and splicing exons together.

Chromosomal localization of human and mouse GPCR protein sequences and identification of EST clones

DNA sequences downloaded

The DNA sequences as well as chromosomal position for all sequences in the dataset were downloaded from UCSC's Genome Bioinformatic home page (http://www.genome.ucsc.edu/) using BLAT. Intron sequences were thereafter removed.

Collecting ESTs

ESTs were collected to receive an overview of the receptor's tissue distribution. There are currently about 19 million sequenced ESTs available for all species, whereof 6 million are derived from human and 4 million are derived from mouse (http://www.ncbi.nlm.nih.gov/dbEST/index.html). The full coding domain DNA sequences of all the GPCRs were used to search against the human and mouse EST database at www.ncbi.nlm.nih.gov/BLAST/using BLASTN with a cutoff $E = 1 \times 10^{-12}$. We made two figures to display the mouse ESTs. Fig. 6A shows receptors that were found in abundance in only one tissue. Moreover, if a receptor had three times more ESTs in a specific tissue compared to the sum of all the other tissues, it was also included in the figure (for example: mNtsr2 has 36 ESTs in brain, 1 in spinal cord medulla, and 3 of unknown origin, thus, $36 > 12$ ($3 \times (1 + 3)$)). Fig. 6B shows receptors that are widely distributed. The criteria for this category were that the GPCR was found in five or more tissues and had more than 10 ESTs total.

Phylogenetic analysis

The total dataset for mouse and human includes over 850 sequences, which makes it difficult to achieve high resolution in a single phylogenetic tree. Thus, according to the GRAFS classification system, the sequences were divided into families of *Glutamate, Rhodopsin, Adhesion, Frizzled/Taste2,* and *Secretin.* Furthermore the *Rhodopsins* were divided into α, β, δ, and γ. Trees were calculated for each of the groups. A small set of receptors, named *Others,* are too divergent to be grouped in a specific phylogenetic group and could not be included in the phylogenetic analysis. The set of *Others* contains the following GPCRs (existing in both species unless otherwise is noted): GPR137 (C11ORF), DARC (Duffy), GPR23, GPR88, GPR120, GPR135, GPR139, GPR141 (Gpr141b, second copy found only in mouse), GPR142, GPR146, GPR152, GPR160, GPR151 (GPCR-2037), HGPCR19, GPR149 (IEDNA), GPR143 (OA1), GPR172A (PERVAR1, found only in human), GPR172B (PERVAR2, found only in human), TRHR (mTrhr2, second copy found only in mouse), GPR137B (TM7SF1), GPR137 (TM7SF1L1), GPR137C (TM7SF1L2), and GPR175 (TPRA40).

The seven transmembrane regions are the part common in all GPCRs. Therefore these regions were used for the phylogenetic analyses and both N- and C-termini were removed from all sequences in the dataset. The *Glutamate* sequences were cut according to a previous suggestion of the TM boundaries [26]. The *Rhodopsin* sequences were cut using the known 7TM regions of bovine rhodopsin [27]; *Adhesion, Frizzled/Taste2,* and *Secretin* were cut according to RPS-BLAST searches (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb).

The sequences that were unclassified in the original human *Rhodopsin* dataset were categorized by a local BLAST search against a database containing all the classified *Rhodopsin* sequences. The proteins that could not be classified were not included in the phylogenetic analysis. Two GPCRs not included in the phylogenetic trees grouped with BLAST, hGPR142 in the γ group and hGPR160 in the α group, but these were too divergent to include in alignments and phylogenetic trees. Additionally, four of the GPCRs from *Others* (GPR141, GPR146, GPR152, GPR151 (GPCR-2037)) are similar to members of both the γ and the δ subgroups and thus display relationships to two groups. Phylogenetic trees were calculated using neighbor joining, with 1000 bootstraps of the dataset for all families, while we also used one additional phylogenetic method (maximum parsimony (MP)) for the *Rhodopsin* subgroups in attempt to obtain a better resolution.

Alignment

The datasets were aligned using the UNIX version of ClustalW 1.82 [24]. The default alignment parameters were applied.

Sequence bootstrapping and randomization

The alignments were then bootstrapped 1000 times (100 times for MP trees) using SEQBOOT from the Win32 version of the Phylip 3.6 package [28], providing a total of 1000 different alignments (100 for MP trees). The SEQBOOT program generates a multiple dataset consisting of the resampled versions of the original input file. It is run in conjunction with the programs below.

Neighbor-joining trees

The protein distances were calculated using PROTDIST from the Win32 version of the Phylip 3.6 package. The Jones–Taylor–Thornton matrix was used for the calculation. The trees were calculated from the distance matrix previously generated by PROTDIST, using NEIGHBOR from the Win32 version of the Phylip 3.6 package, and resulted in 1000 trees (100 for MP). These were merged using the GNU-UNIX "cat" command and the resulting file was analyzed using CONSENSE (http://www.cmbi.kun.nl/PHYLIP/consence-1.html) from the Win32 version of the Phylip 3.5 package to get one bootstrapped consensus tree. The tree was plotted using TREEVIEW (http://www.taxonomy.zoology.gla.ac.uk/rod/treeview.html) and then manually edited in CANVAS 8.0. Weighted branch lengths were calculated for the MP trees for the four *Rhodopsin* subgroups, using Tree-Puzzle (version 5.2). MP trees were used for the tree search procedure, parameter estimate was set to slow, the Mueller–Vingron 2000 model of substitution was chosen as well as the mixed (1 invariable + 8 Gamma rates) model of rate heterogeneity.

Acknowledgments

These studies were supported by the Swedish Research Council (VR, Medicine) and the Swedish Society for Medical Research (SSMF), Svenska Läkaresällskapet, Åke Wikberg Foundation, Lars Hiertas Foundation, Thurings Foundation, Novo Nordisk Foundation, and Magnus Bergwall Foundation.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2006.04.001.

References

[1] J. Bockaert, J.P. Pin, Molecular tinkering of G protein-coupled receptors: an evolutionary success, EMBO J. 18 (1999) 1723–1729.
[2] R. Fredriksson, H.B. Schioth, The repertoire of G-protein-coupled receptors in fully sequenced genomes, Mol. Pharmacol. 67 (2005) 1414–1425.
[3] J. Drews, Drug discovery: a historical perspective, Science 287 (2000) 1960–1964.
[4] A. Wise, K. Gearing, S. Rees, Target validation of G-protein coupled receptors, Drug Discovery Today 7 (2002) 235–246.

[5] J.C. Venter, et al., The sequence of the human genome, Science 291 (2001) 1304–1351.
[6] E.S. Lander, et al., Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.
[7] S. Takeda, et al., Identification of G protein-coupled receptor genes from the human genome sequence, FEBS Lett. 520 (2002) 97–101.
[8] D.K. Vassilatis, et al., The G protein-coupled receptor repertoires of human and mouse, Proc. Natl. Acad. Sci. USA 100 (2003) 4903–4908.
[9] R. Fredriksson, et al., The G-protein-coupled receptors in the human genome form five main families: phylogenetic analysis, paralogon groups, and fingerprints, Mol. Pharmacol. 63 (2003) 1256–1272.
[10] International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome, Nature 431 (2004) 931–945.
[11] T.P. Larsson, et al., Comparison of the current RefSeq, Ensembl and EST databases for counting genes and gene discovery, FEBS Lett. 579 (2005) 690–698.
[12] D.E. Gloriam, H.B. Schioth, R. Fredriksson, Nine new human Rhodopsin family G-protein coupled receptors: identification, sequence characterisation and evolutionary relationship, Biochim. Biophys. Acta 1722 (2005) 235–246.
[13] T.K. Bjarnadottir, et al., The human and mouse repertoire of the adhesion family of G-protein-coupled receptors, Genomics 84 (2004) 23–33.
[14] R. Fredriksson, et al., There exist at least 30 human G-protein-coupled receptors with long Ser/Thr-rich N-termini, Biochem. Biophys. Res. Commun. 301 (2003) 725–734.
[15] R. Fredriksson, et al., Seven evolutionarily conserved human rhodopsin G protein-coupled receptors lacking close relatives, FEBS Lett. 554 (2003) 381–388.
[16] C.G. Murray, et al., Evaluation of EST-data using the genome assembly, Biochem. Biophys. Res. Commun. 331 (2005) 1566–1576.
[17] H. Matsunami, L.B. Buck, A multigene family encoding a diverse array of putative pheromone receptors in mammals, Cell 90 (1997) 775–784.
[18] Y. Niimura, M. Nei, Evolutionary changes of the number of olfactory receptor genes in the human and mouse lineages, Gene 346 (2005) 23–28.
[19] Y. Niimura, M. Nei, Evolution of olfactory receptor genes in the human genome, Proc. Natl. Acad. Sci. USA 100 (2003) 12235–12240.
[20] T.K. Attwood, J.B. Findlay, Fingerprinting G-protein-coupled receptors, Protein Eng. 7 (1994) 195–203.
[21] L.F. Kolakowski Jr., GCRDb: a G-protein-coupled receptor database, Recept. Channels 2 (1994) 1–7.
[22] P. Mombaerts, Genes and ligands for odorant, vomeronasal and taste receptors, Nat. Rev. Neurosci. 5 (2004) 263–278.
[23] S.F. Altschul, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[24] J.D. Thompson, D.G. Higgins, T.J. Gibson, ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 11 (1994) 4673–4680.
[25] S.R. Eddy, Profile hidden Markov models, Bioinformatics 14 (1998) 755–763.
[26] J.P. Pin, T. Galvez, L. Prezeau, Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors, Pharmacol. Ther. 98 (2003) 325–354.
[27] K. Palczewski, et al., Crystal structure of rhodopsin: a G protein-coupled receptor, Science 289 (2000) 739–745.
[28] J. Felsenstein, PHYLIP (phylogenetic inference package), version 3.5c, Department of Genetics, University of Washington, Seattle, 1993.
[29] X. Zhang, et al., Odorant and vomeronasal receptor genes in two mouse genome assemblies, Genomics 83 (2004) 802–811.
[30] J.M. Young, et al., Divergent V1R repertoires in five species: amplification in rodents, decimation in primates, and a surprisingly small repertoire in dogs, Genome Res. 15 (2005) 231–240.
[31] S.M. Foord, et al., International Union of Pharmacology. XLVI. G protein-coupled receptor list, Pharmacol. Rev. 57 (2005) 279–288.
[32] T.K. Bjarnadottir, et al., The gene repertoire and the common evolutionary history of glutamate, pheromone (V2R), taste(1) and other related G protein-coupled receptors, Gene 362 (2005) 70–84.
[33] J.C. Cardoso, et al., The secretin G-protein-coupled receptor family: teleost receptors, J. Mol. Endocrinol. 34 (2005) 753–765.
[34] D.E. Gloriam, et al., The repertoire of trace amine G-protein-coupled receptors: large expansion in zebrafish, Mol. Phylogenet. Evol. 35 (2005) 470–482.
[35] L. Lindemann, et al., Trace amine-associated receptors form structurally and functionally distinct subfamilies of novel G protein-coupled receptors, Genomics 85 (2005) 372–385.
[36] H.B. Schioth, R. Fredriksson, The GRAFS classification system of G-protein coupled receptors in comparative perspective, Gen. Comp. Endocrinol. 142 (2005) 94–101.
[37] C. Conte, et al., Evolutionary relationships of the Tas2r receptor gene families in mouse and human, Physiol. Genom. 14 (2003) 73–82.
[38] I. Rodriguez, et al., Multiple new and isolated families within the mouse superfamily of V1r vomeronasal receptors, Nat. Neurosci. 5 (2002) 134–140.
[39] H.M. Robertson, Two large families of chemoreceptor genes in the nematodes Caenorhabditis elegans and Caenorhabditis briggsae reveal extensive gene duplication, diversification, movement, and intron loss, Genome Res. 8 (1998) 449–463.
[40] J.M. Young, et al., Different evolutionary processes shaped the mouse and human olfactory receptor gene families, Hum. Mol. Genet. 11 (2002) 535–546.