Molecular Pharmacology Fast Forward. Published on February 1, 2005 as doi:10.1124/mol.104.009001

The repertoire of G-protein coupled receptors in fully sequenced genomes

Robert Fredriksson and Helgi B. Schiöth

Department of Neuroscience, Uppsala University, BMC, Box 593, 751 24, Uppsala, Sweden.

Molecular Pharmacology Fast Forward. Published on February 1, 2005 as DOI: 10.1124/mol.104.009001
This article has not been copyedited and formatted. The final version may differ from this version.

Running Title Page

Running title: Repertoire of GPCRs

Author for correspondence: Helgi B. Schiöth, Department of Neuroscience, Biomedical Center, Box 593, 75 124 Uppsala, Sweden, Fax: + 46 18 51 15 40, email; helgis@bmc.uu.se

Document statistics

Number of text pages: 29

Number of tables: 4

Number of figures: 3

Number of references: 39

Number of words in Abstract: 225

Number of words in Introduction: 748

Number of words in Discussion: 1498

Molecular Pharmacology Fast Forward. Published on February 1, 2005 as DOI: 10.1124/mol.104.009001
This article has not been copyedited and formatted. The final version may differ from this version.

Abstract

The super family of G-protein coupled receptors (GPCRs) is one of the largest and most studied families of proteins. We created Hidden Markov Models based on sorted groups of GPCRs from our previous detailed phylogenetic classification of human GPCRs and added several other models based on receptors not found in mammals. We used these models to search entire Genscan datasets from 13 species whose genomes are nearly completely sequenced. We found over 5000 unique GPCRs that were divided into 15 main groups, while the largest one, the *Rhodopsin* family, was subdivided into 13 subclasses. The results show that the main families in the human genome; *Glutamate*, *Rhodopsin*, *Adhesion*, *Frizzled* and *Secretin* arose prior the split of nematodes from the chordate lineage. Moreover, several of the subgroups of the *Rhodopsin* family arose prior to the split of the linage leading to vertebrates. We also searched EST databases and identified over 20 000 sequences that match GPCRs. While the GPCRs represent typically 1-2% of the Genscan predictions, the ESTs that match GPCRs are typically only 0.01-0.001%, indicating that GPCRs in most of the groups are expressed at low levels. We also provide searchable datasets that may be used for annotation and further detailed analysis of the GPCR family. This study provides an extensive overview of the expansion of the gene repertoire for families and subgroups of GPCRs.

Introduction

Tremendous amount of primary sequence information has been made available from the recent sequencing projects, providing a near to full coverage of the entire genomes from a diversity of animal species. The fully sequenced genomes include the mammals mouse and human, two species from the bony fish line (pufferfish; *Takifugu rubripes* and zebrafish; *Danio rerio*), two protochordates from the tunicate linage (*Ciona intestinalis* and *Ciona savignyi*) together with two nematodes (*Caenorhabditis elegans* and *Caenorhabditis briggsae*) and the insects, fruitfly (*Drosophila melanogaster*) and mosquito (*Anopheles gambiae*). Moreover, the plants thale cress (*Arabidopsis thaliana*) and rice (*Oryza sativa*) are sequenced as well as several unicellular species of yeast from the fungi linage such as the bakers yeast (*Schizosaccharomyces pombe*) and budding yeast (*Saccharomyces cerevisiae*). The quality of these genomes is constantly being improved with the most recent assembly of the human genome having only 341 gaps (International Human Genome Sequencing Consortium, 2004) and the results of all analyses based on the genomic data, including gene predictions, are dependent on the current gene assembly. Other genomes, such as those of fugu and zebrafish, will however need considerable amount of work to reach the same quality. While only a small fraction of the genes from these genomes are actually annotated all the genomes mentioned above have gene prediction data sets constructed using Genscan (Burge and Karlin 1997).

The super family of G protein-coupled receptors (GPCRs) is one of the largest and most diverse families of proteins in mammals (Bockaert and Pin 1999). GPCRs are the group of proteins that draws most attention in the pharmaceutical industry and it is estimated that 40-50% of all current drug targets are GPCRs. The large number of proteins in this gene family and the complex

Molecular Pharmacology Fast Forward. Published on February 1, 2005 as DOI: 10.1124/mol.104.009001
This article has not been copyedited and formatted. The final version may differ from this version.

structure of GPCRs have, until recently, made it difficult to systematically study their overall evolution. Structurally, the common feature of all GPCRs is a seven α-helical transmembrane region (7TM) that anchor the receptor to the plasma membrane of the cell, with the N-termini exposed to the extracellular space. In addition to the 7TM region, some families of GPCRs have long N-termini containing different kinds of functional-, or ligand binding-domains.

GPCRs can be found in almost any eukaryotic organism, including insects (Hill et al., 2002) and plants (Josefsson 1999) indicating that these proteins are of ancient origin. There is also a light sensing 7TM protein found in bacteria, the bacterial rhodopsin, but it is presently unclear whether this protein has a common origin with GPCRs in eukaryotic organisms, since it does not signal through G-proteins and lacks significant sequence homology to GPCRs (Okada and Palczewski 2001). The human repertoire of GPCRs has recently been described in details (Joost and Methner 2002, Vassilatis et al., 2003, Fredriksson et al., 2003) while there are still additional new human GPCR being annotated (Fredriksson et al., 2003b, Fredriksson et al., 2003c). Also the entire repertoire of GPCRs in mouse (Vassilatis et al., 2003) and malaria mosquito (Hill et al 2002) have been described, while other genomes lack such a whole genome description of the GPCR superfamily. These studies have provided a good overview over the mammalian genomes but there exists considerable confusion about the relationship of GPCR subgroups among different eukaryotes. Several classification systems for GPCRs have been proposed, for examples the A-F system (Kolakowski, 1994) and the 1-5 system (Bockaert and Pin, 1999). These systems attempt to cover the entire GPCR repertoire in several developmental lineages but do not include some of the more recently discovered families. We have previously described a classification system for the GPCR superfamily in the human genome (Fredriksson et al., 2003) where we divided the

Molecular Pharmacology Fast Forward. Published on February 1, 2005 as DOI: 10.1124/mol.104.009001
This article has not been copyedited and formatted. The final version may differ from this version.

receptors into five families using a phylogenetic approach. Here, we use the terminology *Rhodopsin* (also known as A or 1), *Secretin* (B or 2), *Adhesion* (previously included in B or 2), *Frizzled* (F or 5) and *Glutamate* (C or 3), which forms the GRAFS classification system.

In this paper we investigated the origin of the human GPCRs by searching the Genscan datasets for GPCRs from 13 species where a complete genomic sequence is available. We used Hidden Markov Models (HMMs) based on our recent GRAFS classification system for GPCRs (Fredriksson et al., 2003) as well as groups of GPCRs that are not found in mammals, previously identified in other classification systems. The aims was to identify "all" GPCRs in these genomes, group these in families, and determine the relationship of GPCRs in distantly related species and thus reveal the origin and expansion of each group.

Materials and Methods

1. Description of the original datasets

1.1 Human: We used the NCBI build 33 of the Genscan dataset. This genome is largely contiguous, i.e. free of gaps, and includes more than 99% of the genetic material (http://genome.ucsc.edu/). The predicted gene set from this assembly contains around 55 000 genes while the manually reviewed RefSeq dataset contains nearly 20 000 protein sequences (http://www.ncbi.nlm.nih.gov/).

1.2 Mouse: The current assembly of the mouse genome (NCBI build 30) consists of 38 000 contigs with a predicted gene set of 110 000 proteins (http://www.ncbi.nlm.nih.gov/). Approximately 90-95% of the genetic material is present in the assembly (http://genome.ucsc.edu/). The mouse RefSeq dataset currently has around 16 200 protein sequences.

1.3 Fugu: The genome of fugu (*Takifugu rubripes)* used in this study was Ensembl release 17.2.1 which consists of 8 597 contigs, covering almost 320 Mbp i.e. around 95% of the non-repetitive DNA (Aparicio et al., 2003) and has 29 600 Genscan predicted genes (http://www.ensembl.org/Fugu_rubripes/). Very few genes have been manually annotated from this species and the non-redundant protein database at NCBI has less then 50 GPCRs from fugu.

1.4 Zebrafish: The genome of the zebrafish (*Danio rerio*) used here was the Ensemble 17.2.1 release with 1.56 Gbp in 85 700 contigs. The Genscan predicted protein dataset has around 60 000 gene predictions (http://www.ensembl.org/Danio_rerio/). Also from this species, very few genes have been manually annotated, the RefSeq dataset has 1 170 proteins and contains in total 72 GPCRs from zebrafish (http://www.ncbi.nlm.nih.gov/).

1.5 Ciona: The draft genome sequence of the Urochordate *Ciona intestinalis* was published late 2002 and is an 8.2x whole genome shotgun assembly. The sequence contains 2 500 contigs and covers 116 Mbp of non-repetitive sequence, around 90% of the total non-repetitive material (Dehal et al., 2002) and an annotation project is ongoing at (http://genome.jgi-psf.org). Currently the genes are mainly automatically annotated. The current Genscan dataset contains 15 800 genes (http://www.ncbi.nlm.nih.gov/).

1.6 Drosophila: The *Drosophila melanogaster* genome was sequenced in 2000 and is thought to contain around 98% of the 120Mbp of the Drosophila genome (Adams et al., 2000). For this study we used the NCBI Genscan dataset consisting of 14 300 predicted genes. There is a large annotation project at Flybase (http://www.flybase.org/), which lists 13 500 genes and a total of 193 GPCRs.

1.7 A. gambiae: The genomic sequence of the mosquito *Anopheles gambiae* was published as about 10x whole genome shotgun in 2002 (Holt et al., 2002) and current assembly contains 278 Mbp in 18 634 contigs (http://www.ensembl.org/Anopheles_gambiae/). The Genscan dataset contains 16 000 predicted genes and there are almost no manually annotated genes for this specie.

1.8 C. elegans: The genomic sequence of the nematode *Caenorhabditis elegans* was published in 1998 (The C. elegans Sequencing Consortium). The most current assembly contains 103 Mbp of genomic sequence in 3 266 contigs. The current Genscan dataset contains 20 200 predicted sequences (http://www.ncbi.nlm.nih.gov/). An annotation project is underway (http://www.wormbase.org/) and the project lists 4 609 confirmed genes but it provides no classification of these.

1.9 O. sativa: The genomic sequence of rice (*Oryza sativa*) is currently in progress (http://rgp.dna.affrc.go.jp/) and the current Genscan dataset contains 2 400 predicted proteins (http://www.ncbi.nlm.nih.gov/). Very few genes are annotated in this species. The annotations are currently almost exclusively automated or semi-automated.

1.10 A. thaliana: The only completed plant genome is the genome of *Arabidopsis thaliana* (The Arabidopsis Initiative 2000) with a Genscan dataset of 6 600 predicted proteins. (http://www.ncbi.nlm.nih.gov/). The annotations for this specie are currently almost exclusively automated or semi-automated.

1.11 P. falciparum: The genome of *Plasmodium falciparum* was sequenced in 2002 and the current assembly contains 23.1 Mb. The current Genscan dataset contains 5 200 predicted proteins. An annotation project is ongoing at http://plasmodb.org/publications.shtml and the current database contains domain and GO annotations of the predicted protein set, but no verified proteins.

1.12 S. pombe: The fission yeast *Schizosaccharomyces pombe* genome was sequenced in 1998 and the current assembly from NCBI consists of 12.6 Mb. The Genscan dataset contains 5 000 predicted proteins. There are several large scale genomics and proteomics projects ongoing aiming at annotating the entire genome of fungi, for example the YPD at https://www.incyte.com/proteome/ and the *Schizosaccharomyces pombe* geneDB at http://www.genedb.org/genedb/pombe/index.jsp. These databases contain annotation information of various qualities from biochemically well characterized proteins to annotations based only on similarity and computer predictions.

1.13: S. cerevisiae: The budding yeast (*Saccharomyces cerevisiae*) genome project is at a similar state as the genome project of *S. pombe* with a current assembly from NCBI consisting of 10.3 Mb with a Genscan dataset of 6 300 predicted proteins. Annotations are available at for example https://www.incyte.com/proteome/ and at http://www.yeastgenome.org/ and vary in quality as the *S. pombe* data.

2. Construction of HMM models

The overall HMM models were constructed from receptors according to Table 1. The accession numbers of each receptor and information regarding which HMM-model each receptor belong to can be found in supplementary data Table 1. We removed the long N- and C-terminals from some of the receptor sequences, as identified by RPS-BLAST searches at http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi. The receptor sequences were subsequently aligned using ClustalW 1.81 (Thompson et al., 1994) using default settings. From the alignments, Hidden Markov Models (HMMs) was constructed using the HMMER 2.2 package (Eddy, 1998). The models were constructed using HMMbuild with default settings and calibrated using HMMcalibrate.

3. Identification of GPCR sequences in the Genscan data-sets

Fasta files containing the protein versions of the Genscan predicted gene-sets were downloaded from ftp.ncbi.nlm.nih.gov/genomes/ (*H. sapiens*, *C. elegans*, *D. melanogaster*, *S. pombe*, *S. cerevisiae*, *P. falciparum*) ftp.ensembl.org/pub/ (*M. musculus*, *T. rubripes*, *D. rerio*, *C. intestinalis*, *A. gambiae*). These predicted protein sets were searched against the HMM-models using HMMsearch from HMMER 2.2 with a cut of at E=10. All hits with an E-value below 0.01 were considered correct and all hits with E-values between 0.01 and 10 were manually inspected to verify that they are true GPCRs using BLASTP searches against the NCBI GenPept dataset. Protein sequences from the list of Fasta tags that were to be manually checked, were extracted using the fastacmd program from the NCBI BLAST-suite. For each predicted protein, the top five hits were manually inspected and a minimum of four of these had to be a GPCR for inclusion in the particular dataset, all other predicted proteins were excluded from further analysis. The criteria used for the rhodopsin family were presence of conserved amino acids, identified from pair wise alignments with the closest known human GPCR, such as for example the NPxxY motif at the end of TM7, the DRY motif at the end of TM3 the conserved N early in TM1. For the other families, which have less established recognisable motifs, we used multiple alignments of the protein(s) of interest together with a selection of the human family members that allowed for identification of conserved residues and motifs

4. Collection of datasets for each species

Lists containing the ID numbers from all significant hits from each species were extracted from the HMMER output files and imported into Microsoft Excel. Here, all predicted proteins that hit more than one model were manually inspected and the hit with the lowest E-value was kept in the dataset. Further, lists containing the "true" GPCR dataset were collected in Microsoft Excel and statistics were calculated. All calculations and graph plotting were performed in Microsoft Excel. The sequence names and the classification of the receptors can be found in supplementary data Table 2. The sequences in Fasta format are available as supplementary file 1.

5. Subdivision of the Rhodopsin GPCRs

A dataset consisting of 614 human *Rhodopsin* GPCRs were divided into 13 groups according to (Fredriksson et al., 2003) and subsequently used to subdivide the *Rhodopsin* GPCRs found in the HMM-searches. The human GPCRs were divided as follows with the group name in bold and the number of receptors in parenthesis: Olfactory receptors, **OLF** (347); Serotonin/Dopamine/Adrenergic/Trace amines, **AMIN** (42); Neuropeptide/Peptide, **PEP** (35); Chemokine like, **CHEM** (42); Purine like, **PUR** (42); Somatostatin/Opioid/Galanin **SOG** (15); Opsin like, **OPN** (9); Glycoprotein binding, **LGR** (8); Prostaglandin, **PTGER** (15);

Melanocortin/EDG/Adenosin/Cannabinoid, **MECA** (22); MRG receptors, **MRG** (8); melatonin, **MTN** (3); melanocyte concentrating hormone receptor, **MCHR** (2). A table listing the accession numbers, names and exact grouping is provided in supplementary data, Table 1. A BLAST database was constructed from the 614 human GPCRs and the *Rhodopsin* GPCRs from the other species were searched against this database. A cut-off value of E=1e-9 was used and the five top hits were manually inspected for each receptor, to elucidate which human GPCR from the database it was most similar to. The requirement for being placed in a given group was to have at least four of the five best hits from that specific group. The receptors that did not match these criteria were grouped as unclassified (**UC**). The sequence names and the classification of the receptors into subgroups can be found in supplementary data Table 3. The sequences in FASTA format are available as supplementary file 1.

6. Expression levels by EST-matches.

The entire gbest was downloaded from ftp://ftp.ncbi.nlm.nih.gov/genebank and entered into an SQL-database using custom made software. From this database FASTA-files containing all high-quality ESTs from each species were extracted, one file per species. The NCBI-BLAST package was used to construct a searchable database for each specie, and all Genscan GPCRs found in a given specie was searched against its EST-database using TBLASTX with a cut-off at E=-40. The blast results were automatically extracted and converted into tables. Since it is likely that a given EST will be hit by several GPCRs, all hits, but the one with the highest E-value, were removed from the tables using custom made software to obtain a non-redundant list. The results were extracted and converted into tables readable by Microsoft EXCEL using custom made software. Data were analyzed and graphs were plotted using Microsoft EXCEL. The software used can be made available from the authors upon request as C++ source code and bash scripts.

Results

Our strategy was to create Hidden Markov Models (HMMs) based on well characterized groups of GPCRs from our phylogenetic classification of the entire set of GPCRs in the human genome (Fredriksson et al., 2003). These main families are *Adhesion* (ADH), *Secretin* (SEC), *Frizzled* (FZD), *Glutamate* (GLR) and *Rhodopsin* (RHOD). Moreover, we also created HMMs of groups of GPCRs that we did not find in the human genome, such as the cAMP binding receptors from slime molds (Dictyostelium), nematode chemoreceptors (Robertson, 1998), the gustatory receptors from insects (Hill et al., 2002) the odorant receptors from drosophila (Hill et al., 2002) MLO receptors in plants (Devoto et al., 2003) and STE2 (Marsh and Herskowitz, 1998) and STE3 (Hagen et al., 1986) from yeast. Some of these families lack significant sequence homology with the mammalian GPCRs. There are also some atypical GPCRs with uncertain relation to other GPCRs, like the ocular albinism gene (OA1) (Shen et al., 2001) and the Vomeronasal receptors found in vertebrates (Lane et al., 2002). A summary of the receptors used to construct the HMMs can be seen in Table 1. Tables with the accession numbers and which HMM model the different GPCRs belong to, are available as supplementary data Table 1.

These HMMs were subsequently used to extract and classify GPCRs from the 13 species, including the human Genscan dataset in order to evaluate the quality of the Genscan datasets regarding GPCRs. During our pilot studies, we found that the HMMs for the *Adhesion* and the *Secretin* families did not easily distinguish between the receptors that hit these models in some species, in particular in fish. We therefore decided to merge them and subsequently separated these based upon the presence of a GPS domain in the N-terminus close to TM1. The GPS domain is found in almost all *Adhesion* GPCRs while it is not found in any *Secretin* GPCR in the

Molecular Pharmacology Fast Forward. Published on February 1, 2005 as DOI: 10.1124/mol.104.009001
This article has not been copyedited and formatted. The final version may differ from this version.

human and mouse genomes (Harmar 2001, Bjarnadottir et al., 2003). The overall results of the HMM searches are displayed in Table 2 and a detailed description of the results is available as supplementary table 2a-l. All hits with an E-value better (lower) than 0.01 were in principal considered to be correct although, the 10 “worst” of these hits were manually controlled using BLASTP searches against the protein database at NCBI and only few errors were found. The hits between 0.01 and 10 were however all manually controlled using BLASTP, and the results of these can be seen in supplementary data Table 4. From this Table, it is evident that the number of positives found for this E-value range was highly dependent on the HMM model. The specificity of the HMMs at high E-values can be seen in supplementary Table 4.

The *Rhodopsin* GPCRs were further subdivided into 13 classes as defined in the Materials and Methods section and as summarized in Table 3. This comparison used BLASTP against a database of all human GPCRs. The results of this subdivision of the *Rhodopsin* family can be seen in Table 4. A detailed description of the results is available as supplementary Table 3a-h.

Our results for the human genome are in good agreement with what we, and others, have published earlier. These earlier published numbers for the human genome are shown in brackets in Table 2 and Table 4. The numbers of GPCRs in the different genomes are also displayed graphically in Figure 1. All previously published numbers are from (Fredriksson et al., 2003) with the exception of olfactory receptors which are from Zozulya et al (Zozulya et al., 2001), the Vomeronasal which are from Kouros-Mehr et al (Kouros-Mehr et al., 2001) and the taste receptors type 2 (TAS2) which are from our own unpublished studies. The main discrepancies are in the olfactory and the vomeronasal groups. This is likely to be related to the number of pseudogenes in the Genscan predictions. It is fairly well established that there are a number of olfactory (Zozulya et al., 2001) and vomeronasal pseudogenes in the Genscan set. The vomeronasal receptors in humans have been shown to be non-functional pseudogenes (Kouros-Mehr et al., 2001). It is also not clear how many functional olfactory receptors there are in the human genome. We also noticed that a number of the human olfactory receptors were longer than 400 bases and when these were investigated further, they were found to correspond to two or three olfactory receptors. These are likely to be erroneously predicted by Genscan due to the small intergenic distance for these genes. If these proteins are counted separately, the number of olfactory receptors increases from 494 to 545. Also in the group of TAS2, we find fewer proteins than have been reported in Genbank. BLAST searches with all known TAS2 receptors against the entire Genscan dataset showed that many of these proteins are not present in the Genscan dataset. This is likely to reflect the inability of Genscan to predict these proteins, since the sequence of these genes is present in the human genome assembly (data not shown). Considering these exceptions, the overall results indicate that the numbers of GPCRs that we identified are in good agreement with previous data.

The results for the mouse genome are similar to that of the human genome with only few exceptions. It is well known that there are much fewer olfactory receptors in the human genome as compared with the mouse genome. This has been studied in a recent paper (Young et al., 2002) but we emphasize that it is still large uncertainties regarding the exact numbers in each genome. It is noteworthy that these receptors lie in large blocks on the chromosomes which cause problems with the assembly of the genomes as such, since many of these receptors are very similar and it is in many cases difficult to distinguish between what is polymorphism and what is truly a unique gene during the assembly of the shotgun data. We also know that we underestimate the number of TAS2 receptors in mouse, most likely for the same reasons as discussed for the human genome. The GLR HMM picks up a considerably higher number of predicted proteins in the mouse genome than the human. This is not related to a large expansion of classical glutamate receptors (mGluRs) as these are found in similar numbers in mouse and human. This is rather related to pheromone receptors that are not found in the human genome. They show similarity to TAS1 receptors (three copies in humans) and have expanded in mouse resulting in at least 80 receptors that are fairly similar but found on at least 11 chromosomal segments (unpublished data). Also for the mouse, the number of olfactory receptors is underestimated due to multiple proteins joined together by Genscan and the number of olfactory receptors is increased from 789 to 827 when these are considered. In a recent paper, the number of GPCRs for endogenous ligands in mouse was determined to be 392 (Vassilatis et al., 2003), which is in agreement with the dataset presented here, which is 391 *Rhodopsin* GPCRs excluding the olfactory receptors.

The fish species have not previously been analysed with regards to the GPCR repertoire except that it has been estimated from the genome sequencing projects that there are in total about 457 *Rhodopsin* GPCRs in fugu (Aparicio et al., 2003). This is significantly higher (twice the number) of what we find in the current Genscan dataset. The GPCRs described by Aparicio et al (Aparicio et al., 2003) was found in a set of 27 779 predicted proteins while the Genscan dataset used in our study has 29 625 proteins. The reason for the large difference in number of GPCRs could be due to differences in the "gene-building pipeline" used by the fugu genome sequencing group and the Genscan program. We know however that the number of GPCRs in the Genscan dataset is similar to the published (manually verified) numbers for mouse and human and we find it likely that the Genscan sets we used are providing good estimate of the GPCR numbers also in fugu. It is notable that zebrafish has two to three fold as many GPCRs in the main families as fugu. When

Molecular Pharmacology Fast Forward. Published on February 1, 2005 as DOI: 10.1124/mol.104.009001
This article has not been copyedited and formatted. The final version may differ from this version.

compared to mammals, the zebrafish has up to twice the number of receptors while fugu has about half the number. The *Frizzled* family is an exception, as it has approximately the same numbers in all mammals and fish. Among the subgroups of *Rhodopsin* GPCRs the picture is similar with zebrafish having about two to three fold more GPCRs than fugu, while the mammalian numbers are in between these. An exception to this is the olfactory receptors, which have a very small number in both fish species, and the MRG receptors that seem to be missing in fish. We have performed detailed separate searches in the fish genomes with the human MRG receptors as baits using TBLSTX and PSI-BLAST without finding any MRG-like sequence in fish (data not shown).

The ciona genome has likewise not been investigated with regards to GPCRs. The main groups, which are all present in ciona, contain three to four fold fewer receptors than the mammalian counterparts, again with the exception of the *Frizzled* family, which has similar numbers in ciona and the vertebrate genomes. Within the *Rhodopsin* family there seems to be several subgroups missing. It is notable that the ciona does not seem to have any receptors that match the olfactory receptors in vertebrates, but also melanocyte concentrating hormone (MCH) and purine (PUR) receptors seem to be absent. It is notable that ciona has fewer members in the GLR group and this is partly explained by the fact that there seems to be no TAS1 genes in ciona (unpublished data).

The two insect species have the same five main families of GPCRs as mammals, fish and ciona. The main difference is that these species also have a large number of gustatory receptors (GUST) and odorant (DMOD) receptors, as previously reported (Hill et al., 2002) and several of the *Rhodopsin* subgroups are missing. Like for ciona, the insects seem to be lacking OLF, MCH and PUR receptors but also chemokine (CHEM) and prostaglandin (PTGER) binding receptors. In a previous analysis, it was described that *A. gambiae* has 276 GPCRs and *D. melanogaster* has 270 (Hill et al., 2002) which fits with the numbers presented here, 260 and 210, respectively. The difference in the number of receptors found in *D. melanogaster* is related to that we find 40 fewer gustatory GPCRs as compared with what was reported by Hill and collegues (Hill et al., 2002). In another paper, it was found that *D. melanogaster* has in total 211 GPCRs (Adams et al., 2000).

The pattern of GPCRs in *C. elegans* shows that the five main families of GPCRs, ADH, SEC, FZD, GLR and RHOD are present, as they are in mammals, ciona, insects and fish. They do have one additional group, namely the nematode chemosensory (NCHM) receptors. These do not show any significant similarities to the olfactory receptors found in vertebrates but also not to the gustatory or the DMOD found in the arthropods. It is notably that these groups are absent in *C. elegans*. It has previously been reported that *C. elegans* has over 800 GPCRs of the chemosensory type, of which 550 seems to be functional (Robertson, 1998). In our searches, we find about 1000 GPCRs of this type. It should be noted that some of the receptors belonging to chemosensory GPCRs can also be detected by the RHOD HMM as well. These were removed from the RHOD dataset manually using BLASTP against a database of known nematode chemosensory receptors and the entire human *Rhodopsin* dataset. Most of the subgroups of the *Rhodopsin* family are absent or only represented by one member. Only the AMIN, PEP and SOG groups have several members in *C. elegans*.

All the GPCR hits from species that do not belong to bilateria, i.e. plants, yeast and *P. falciparum* were individually investigated by additional manual inspection, to verify their identity as GPCRs and also to assure as accurate classification as possible. In this process 7 putative GPCRs from *S. pombe*, originally identified at low scores in the HMM-searches, were removed since BLASTP searches against the nr database at NCBI indicated that these proteins most likely are membrane transporters. That left only two GPCRs, which are the well-known pheromone GPCRs STE2 and STE3. One putative GPCR from *S. cerevisiae,* tentatively placed in the ADH group, were found to correspond to a protein previously shown to be able to bind G-proteins in a yeast two hybrid system (Yun et al., 1997). This protein does seem to have very low, if any, similarity to the regular 7TM receptors. It consists of a 930 amino acid long open reading frame and seems to have between 12 and 14 hydrophobic segments in a Kyle Doolitle hydrophobicity plot (data not shown). We left this protein in the dataset, but will not discuss it further in relation to 7TM GPCRs. Apart from this GPCR, we only identify the two STE2 and STE3 GPCRs from *S. cervisa.* No GPCRs were found in the malaria parasite *P. falciparum.* In plants, we found GPCRs of the MLO-family (Devoto et al., 2003) but also one GPCR that matched the ADH/SEC model in *A. thaliana*. This GPCR was previously reported (Josefsson and Rask 1997) but its relationship to vertebrate GPCRs has not been clearly elucidated in terms of the overall classification of GPCRs.

To more clearly visualize the results, we display the distribution of GPCRs in percentages between the different subfamilies of GPCRs in Fig. 1 (top panel of each diagram). The figure shows that certain families like *Frizzled* receptors have approximately the same percentages of the total number of GPCRs in all species while others, like *Rhodopsin* GPCRs in mammals and *Glutamate* in mouse, has expanded in some lineages. The figure also shows the same kind of analysis of the different groups of the *Rhodopsin* family (lower panel of each diagram). The database of GPCR-predictions can be obtained from the authors upon request.

The *Rhodopsin* family can be divided into four groups with in total 13 main branches, or clusters, termed α, β, γ and δ. The largest of the four main groups (α) contains the large AMIN cluster that includes many receptors that bind monoamines such as adrenalin, serotonin, dopamine and histamine. The amine cluster (Fig. 1) is highly represented ranging from 18 to 57 members in bilateral species with the exception of zebrafish whose trace amine receptors have undergone large expansion (unpublished data). The other branches in the α-group, such as and the MECA branch, that includes peptide and lipid binding receptors, and the opsin branch are also found in all bilateria, albeit not in as high numbers as the receptors in the amine branch. The prostaglandin receptors also belong to the α-group but these were only found in vertebrates, suggesting that these arose later in the evolution than the other branches of the α-group. The β-group contains many peptide receptors and several of these are receptors for neuropeptides such as NPFF, NPY, orexin, neurotensin, and TRH. This group contains only one branch, and it is found in fairly similar numbers in all the bilateria species. This may indicate that the ancestors of the receptors that bind peptides, which regulate many "higher" functions, did not arise later than for example the amine receptors, whose ligands are considerably less complex than the peptide ligands of the β-group. There are three branches within the γ-group. Only one of those is found in all bilateria species. This is the SOG branch that contains receptors that bind several peptides such as opioids, RF-peptides, neuropeptide W (GPR7 and 8) and somatostatins. This provides further evidence for high representation of receptors that possibly bind complex neuropeptides among pre-vertebrates. Indeed there are examples of GPCRs in pre-vertebrates that bind, albeit with low affinity, peptides of mammalian origin, for example the NPY-binding receptor in drosophila (Li et al., 1992). This “NPY-receptor” is found in our PEP group as expected. It is very difficult to track the origin of most mammalian peptides that bind GPCRs in pre-vertebrates due to fact that their conserved motifs are very short and not well preserved in the primary structure. This analysis shows however that it is indeed possible to track the ancestors to the mammalian GPCR that bind

Molecular Pharmacology Fast Forward. Published on February 1, 2005 as DOI: 10.1124/mol.104.009001
This article has not been copyedited and formatted. The final version may differ from this version.

peptides. The other branches in the γ-group appear for the "first time" in *C. elegans*, while the MCH receptors are only found in vertebrates. This may suggest that the SOG branch includes the ancestors to the entire receptor repertoire in this group. The δ-group has four main branches but only one of them, the LRG, are found in all bilateria species investigated. This group in mammals includes the LH and FSH receptors that have, unlike most other *Rhodopsin* GPCRs, long N-termini in addition to the 7TM regions. Indeed many of the GPCRs that we predict to be the ancestral genes in the invertebrates species, do have long N-terminal stretches that show similarities to the mammalian counterparts, even though they do not show recognizable "domains" in for example RPS-BLAST (data not shown). The other branches in this group were only found in vertebrates and it is particularly interesting that the purine and the olfactory receptor branches appears for the "first time" in fish, while the MRG group are only found in mammals. In order to further verify the nature of the sequences classified as purine and olfactory GPCRs from fish, we calculated phylogenetic trees using the neighbour-joining method. All the sequences classified as purine receptors from fugu were aligned together with the entire delta-group of human rhodopsin GPCRs for this calculation. This tree shows that all the purine sequences from fugu place inside the purine cluster and it is also clear that most of the fugu sequences have a clear ortholog in the human genome. We took a similar approach for the olfactory receptors and combined all the fugu and zebrafish sequences classified as olfactory with 20 randomly selected human olfactory sequences and the entire human delta-group and calculated a phylogenetic tree. This tree shows that all the olfactory receptors from fugu place on the same main branch as the human olfactory receptors, although the fugu receptors seem to form clusters of their own, distinct from the human receptors. This phylogenetic analysis strongly supports the conclusion that both olfactory and purine GPCRs are present in teleost fish. These trees are available from the authors upon request. Taken together it seems evident that the ancestors of many of the peptide and amine receptors that are found in mammals have a long evolutionary history involving multiple members. It does not appear that the gene repertoire for these has taken any drastic changes during the evolution of bilateria, rather it seems that the numbers have undergone gradually increase in "higher" species.

EST data

In order to investigate the level of expression for the different classes of GPCRs, we used the Genscan GPCRs from each species to search against an internal BLAST-database containing all ESTs from that particular species. These EST data are collected from hundreds of cDNA projects and the tissue sampling is highly different between species. We do thus not discuss the tissue origin of the ESTs in any detail but rather focus on the number of ESTs for certain families, as displayed in Fig. 2 and 3. The large numerical range of the results makes it difficult to display these data. In Fig. 2, we show the percentage of GPCRs in the Genscan dataset. This percentage is multiplied by a constant (10000) and the logarithm of that number is plotted as a striped bar. The non-striped bar represents the percentage of GPCRs in the EST dataset, again multiplied by 10000 and converted to logarithms. This means that when the striped bar is higher than the non-striped bar, the number of GPCRs found in the EST dataset is lower than expected as compared with the number of predicted GPCR genes found in the Genscan set. As this is the case for most of the GPCRs in all species investigated, this indicates that GPCRs are generally expressed in low numbers, which is a known fact for many well-studied proteins of this family. Here, we can see that this is a general feature for most groups of GPCRs in most lineages. There are few exceptions to this, as *Rhodopsin* GPCRs in insects seems to be relatively highly expressed. In order to see the relationship of the expression level of GPCRs between the different families, we plotted the number of ESTs containing GPCRs from each family as a fraction of the total number of GPCR-ESTs, as can be seen in Fig. 3. This analysis shows that the majority of the GPCR-ESTs are from the *Rhodopsin* family with two exceptions, the chemosensory receptors in *C. elegans,* which constitutes almost 75% of all GPCR-ESTs in this species and the vomeronasal (VR) receptors in humans. It was a surprise to us to see that there is such a high expression of VR in humans since these are considered pseudogenes as their mRNA does not contain a full-length open reading frame. It is in particular notable that the expression is much higher in humans as compared with mouse, which has VR receptors that are clearly functional. Further investigation showed that these ESTs originates from 186 different libraries and are found in numerous tissues (data not shown).

Discussion

This analysis highlights the tremendous success of GPCRs through evolution of "higher" species. It is evident that all major GPCR families in the human genome, according to our previously published GRAFS system, the *Glutamate*, *Rhodopsin*, *Adhesion*, *Frizzled* and *Secretin*, arose prior the split of nematodes from the chordate lineage. Moreover, the majority of the GPCRs in each of the vertebrate species belong to these five families. The overall similarity of GPCR-repertoire between the lineage leading to arthropods and chordates are also remarkable, not only for the main families but also for several of the subgroups within the large *Rhodopsin* family. There are only few of the groups that are clearly lineage specific in bilateral species. These are the chemosensory receptors in the nematodes that are not found in any other species and represent about 87% percent of the GPCRs in *C. elegans*. The gustatory receptors are only found in the two species of insects where they represent about 20% and 28% of the repertoire in the fruitfly and mosquito, respectively. There are only two families, the VR and the TAS2 genes that have arisen after the split of ciona from the lineage leading to vertebrates. Both these groups are found in fish. No new *families* of GPCRs seem to have arisen during the last 450 million years in the vertebrate linage.

The *Rhodopsin* family has had the largest evolutionary success, representing about 60% of the entire GPCR repertoire in the bilateria species. The *Rhodopsin* family can be divided in four main groups (α-, β- γ- and δ) with 13 main branches (Fredriksson et al., 2003). Members within each of the four main groups are clearly found in all eight bilateria species, while the representation of each of the main branches is more variable. The four other main families are also found in all the

Molecular Pharmacology Fast Forward. Published on February 1, 2005 as DOI: 10.1124/mol.104.009001
This article has not been copyedited and formatted. The final version may differ from this version.

bilateria species. The *Frizzled* receptors are found in fairly constant numbers ranging from 5-14 members. The *Frizzled* receptors control cell fate, proliferation, and polarity that are basic functions within metazoan development (Gho and Schweisguth, 1998) and that could contribute to the evolutionary pressure that keeps their numbers relatively constant. The *Adhesion* GPCRs were discovered rather recently and comparatively little is known about their functional role. The long N-termini of these receptors are likely to interact with other membrane bound proteins, perhaps enabling cell-to-cell communications without soluble ligands (Kwakkenbos et al., 2004). The results show that these receptors arose early, having multiple members in nematodes and they have at least 5 members in each of the bilateral species. It seems thus likely that GPCRs evolved the ability of this type of N-terminal based cell-to-cell interactions long before the presence of vertebrates. The *Secretin* GPCRs have hormone-binding domains in their N-termini that interact with rather large peptides. These peptides act in most cases in a paracrine manner while the *Glutamate* GPCRs, which also have a ligand-binding domain in the long N-termini, interacts with small molecules such as the neurotransmitter glutamate, GABA, Ca^{2+} ions and taste molecules. Our preliminary analyses of the GPCRs in pre-vertebrate species, indicates that the characteristic "ligand" domain within the N-termini of these four main groups of GPCRs are indeed present in all of these families, at least in some of these proteins (unpublished data). This suggests that not only the 7TM domains but also the specific functional domains within all these groups of GPCRs appeared in pre-vertebrates and that their principle functions have been maintained.

There are several groups of GPCRs that have undergone seemingly rapid expansions that are species specific. These include the olfactory receptor group (in the *Rhodopsin* family) in human and mouse, the chemosensory receptors in C. elegans, the gustatory receptors in insects and the pheromone receptor group in mouse (in the *Glutamate* family). These groups share little inter-relationships, considering their amino acid identity or functional motifs. They do however share some general structural features such as absence of any functional domains beyond the 7TM regions (according to for example RPS-BlAST searches), they have short N-termini, and they do not show clear motifs within their groups in the TM regions that can easily earmark them (data not shown). Another common feature for these groups of GPCRs that seem to have very "dynamic" gene repertoire is that they bind small ligands such as odor, taste and chemosensory molecules. Moreover, it is interesting to note that these ligands chemically belong to groups with many structurally similar members within the organism. It is possible that the number of interaction points and the structural constraints of the ligand binding pocket of the receptors for these small molecules are fewer than for receptors that bind large ligands. Therefore, constraints on the 3D structure of these receptors that binds a variety of small ligands is likely to be relatively low. A duplicated/mutated copy of such relatively "promiscuous" receptors in an environment of multiple ligands that play an important physiological role may affect the ability of these genes to survive after duplication events. It is possible that this could be the reason for the unusual evolution of these groups of GPCRs.

There is only minor sequence homology between the GPCRs in plants and fungi and those in the bilateria. There is obviously a very large evolutionary difference between those species as it is estimated that they diverged more than 1000 million year ago. It is however very interesting that there is a sequence in *A. thaliana* that shows resemblance to the ADH/SEC model. This sequence in *A. thaliana* does not have a long N-terminal and is therefore missing all the domains that are specific for the *Adhesion* receptors in mammals (Bjarnadottir et al., 2004). It is tempting to speculate that this is the only sequence that links the GPCR repertoire in bilateria with more

Molecular Pharmacology Fast Forward. Published on February 1, 2005 as DOI: 10.1124/mol.104.009001
This article has not been copyedited and formatted. The final version may differ from this version.

evolutionary distant species, thus providing evidence for common ancestor of all eukaryotic GPCRs.

Our analysis of the expression pattern using the growing EST databases shows that the relative number or percentage of EST sequences is lower than the percentage of gene predictions, with the exception of the *Rhodopsin* GPCRs in insects. This indicates that GPCRs are in general expressed at low levels, at least when considering the mRNA. Remarkably, despite the high diversity within the GPCR gene family this phenomenon of relative low expression level is found for all families in all the species with only a few exceptions. The representation of GPCRs in the EST database correlates fairly well with the relative number of predicted GPCRs, i.e. the larger the number of predicted proteins, the higher is the number of ESTs in the database. Another important observation from these data is that the expression level of most of the main families of GPCRs seems to be conserved between the species, even though they are separated by more than 500 million years. One could speculate that the subdivision of the functional roles between the main families of GPCRs have been similar through the evolution, even if the number of GPCRs in the different families have changed several fold. This notion correlates well to the conservation of functional domains within the five main families that we mentioned earlier.

The EST pattern for human and mouse was fairly similar except for the VR genes that had over 1800 hits in the human EST database. The high copy number of human VR ESTs is quite remarkable considering the fact that the human VR genes are thought to be pseudogenes (Kouros-Mehr et al., 2001). These VR ESTs are found in 186 different libraries and are found in large variety of tissues. It seems that these genes have been pseudogenized rather recently and that the promotor has not yet mutated to become non-functional. The relative high expression of the VR pseudogenes in humans as compared with the VR expression in mouse can possibly be related to lack of negative feedback from a functional VR protein that subsequently down regulates the expression. There are also some examples that the expression of certain genes or entire groups seems to be very low or non-existing. It was for example highly surprising that ESTs for the gustatory and odorant receptors from the mosquito seem to be completely missing in the databases. Since the total number of ESTs is approximately the same in the two species, this could indicate a much lower expression of gustatory and odorant receptors in the mosquito. It was also surprising that the TAS2 receptors in both mammals did only match very few ESTs. This could be due to the fact that these receptors have a very restricted expression pattern and that these "specific" tissues or cells are not represented in the EST databases.

In summary, this analysis provides extensive overview of the expansion of the repertoire of GPCRs in many important genomes. The analysis covers a larger number of genes than has previously been simultaneously analysed in evolutionary perspective for GPCRs and perhaps any protein family. The databases we generated will provide tremendously valuable source for further detailed analysis, assembly and annotation of individual GPCR genes.

Molecular Pharmacology Fast Forward. Published on February 1, 2005 as DOI: 10.1124/mol.104.009001
This article has not been copyedited and formatted. The final version may differ from this version.

Acknowledgments

We thank Christian A. Murray, Uppsala University for assistance with handling the EST data. We thank Thóra K. Bjarnadóttir and David E. Gloriam, both Uppsala University, for providing insight into their preliminary results regarding detailed analysis, assembly and annotation of several subgroups of GPCRs.

REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al (2000) The genome sequence of Drosophila melanogaster. *Science*. **287(5461)**:2185-2195.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Gelpke, M.D., et al. (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science*. **23;297(5585):**1301-1310

Bjarnadottir, T.K., Fredrisson, R., Hoglund, P.J., Gloriam, D., Lagerström, M.C., Schiöth, H.B. (2004) Identification of G-protein coupled receptors in mouse belonging to adhesion family; expression pattern, phylogeny and domains. *(In press Genomics)*

Bockaert, J. and Pin, J..P., (1999) Molecular tinkering of G protein-coupled receptors: an evolutionary success. *EMBO J*. **18**:1723-1729.

Burge, C., and Karlin, S. (1997) Prediction of complete gene structures in human genomic *DNA. J Mol Biol*. **268**:78-94.

Dehal, P., Satou, Y., Campbell, R.K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., Di Gregorio, A., Gelpke, M., Goodstein, D.M., et al. (2002) The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. *Science*. **13;298(5601)**:2157-2167.

Devoto, A., Hartmann, H.A., Piffanelli, P., Elliott, C., Simmons, C., Taramino, G., Goh, C.S., Cohen, F.E., Emerson, B.C., Schulze-Lefert, P., Panstruga, R. (2003) Molecular phylogeny and evolution of the plant-specific seven-transmembrane MLO family. *J Mol Evol.* **56(1)**:77-88

Eddy, S.R.. (1998) Profile hidden Markov models. *Bioinformatics*.;**14(9)**:755-763.

Fredriksson, R., Lagerström M.C. Höglund P.J. and Schiöth, H.B. (2002) Novel human G protein-coupled receptors with long N-terminals containing GPS domains and Ser/Thr-rich regions *FEBS Letters*, **531**:407-414.

Fredriksson, R., Lagerström M.C., Lundin L.G., Schiöth, H.B. (2003a) The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol. Pharmacol*., **63**:1256-1272

Fredriksson, R., Hoglund, P.J., Gloriam, D.E., Lagerstrom, M.C., Schioth, H.B. (2003b) Seven evolutionarily conserved human rhodopsin G protein-coupled receptors lacking close relatives. *FEBS Lett*. **20;554(3)**:381-388.

Fredriksson, R., Gloriam, D.E. Höglund, P.J. Lagerström, M.C., Schiöth, H.B. (2003c) Novel human G protein-coupled receptors with long N-terminals containing GPS domains and Ser/Thr-rich regions. *Biochem. Biophys. Res. Commun*. **301**:725-734.

Gho, M., Schweisguth, F. (1998) Frizzled signalling controls orientation of asymmetric sense organ precursor cell divisions in Drosophila. *Nature.***14;393(6681)**:178-181.

Hedges, B.S. and Kumar, S. (2003) Genomic clocks and evolutionary timescales. *TRENDS in genetics* **19:4**:200-206

Hagen, D.C., McCaffrey, G., Sprague, G.F. Jr., (1986) Evidence the yeast STE3 gene encodes a receptor for the peptide pheromone a factor: gene sequence and implications for the structure of the presumed receptor. *Proc Natl Acad Sci U S A.* **83(5)**:1418-1422.

Harmar, A.J. (2001) Family-B G-protein-coupled receptors. *Genome Biol*.,**2(12)**

Hill, C.A., Fox, A.N., Pitts, R.J., Kent, L.B., Tan, P.L., Chrystal, M.A., Cravchik, A., Collins, F.H., Robertson, H.M., Zwiebel, L.J. (2002) G protein-coupled receptors in Anopheles gambiae. *Science*. **4;298(5591):**176-178.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M., Wides, R., Weissenbach, J. et al., (2002) The genome sequence of the malaria mosquito Anopheles gambiae. *Science.* **4;298(5591):**129-149.

International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature,* **21;431(7011)**: 931-45.

Joost, P., Methner, A. (2002) Phylogenetic analysis of 277 human G-protein-coupled receptors as a tool for the prediction of orphan receptor ligands. *Genome Biol*. **17;3** (11).

Josefsson, L.G., Rask, L. (1997) Cloning of a putative G-protein-coupled receptor from Arabidopsis thaliana. *Eur J Biochem.* **15;249(2)**:415-420.

Josefsson, L.G., Evidence for kinship between diverse G-protein coupled receptors. (1999) *Gene.* **1;239(2)**:333-340.

Kolakowski LF Jr. (1994) GCRDb: a G-protein-coupled receptor database. *Receptors Chanels,* **2(1)**:1-7

Kouros-Mehr, H., Pintchovski, S., Melnyk, J., Chen, Y.J., Friedman, C., Trask, B., Shizuya, H. (2001) Identification of non-functional human VNO receptor genes provides evidence for vestigiality of the human VNO. *Chem Senses*. **26(9)**:1167-1174.

Kwakkenbos, M.J., Kop, E.N., Stacey, M., Matmati, M., Gordon ,S., Lin, H.H., Hamann, J. (2004) The EGF-TM7 family: a postgenomic view. *Immunogenetics*., **55(10)**:655-666.

Lane, R.P., Cutforth, T., Axel, R., Hood, L., Trask, B.J. (2002) Sequence analysis of mouse vomeronasal receptor gene clusters reveals common promoter motifs and a history of recent expansion. *Proc Natl Acad Sci U S A.* **8;99(1)**:291-6.

Li, X.J., Wu, Y.N., North, R.A., Forte, M. (1992) Cloning, functional expression, and developmental regulation of a neuropeptide Y receptor from Drosophila melanogaster. *J Biol Chem.* Jan **5;267(1):**9-12.

Marsh, L., Herskowitz, I,. (1988) STE2 protein of Saccharomyces kluyveri is a member of the rhodopsin/beta-adrenergic receptor family and is responsible for recognition of the peptide ligand alpha factor. *Proc Natl Acad Sci U S A*. **85(11)**:3855-3859.

Okada, T., Palczewski, K. (2001) Crystal structure of rhodopsin: implications for vision and beyond. *Curr Opin Struct Biol*. **11(4)**:420-426.

Robertson HM. (1998) 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*.**May;8(5):**449-63.

Shen, B., Samaraweera, P., Rosenberg, B., Orlow, S.J. (2001) Ocular albinism type 1: more than meets the eye. *Pigment Cell Res*. **14(4)**:243-248.

Springer, M.S., Murphy, W.J., Eizirik, E., O'Brien, S.J.. (2003) Placental mammal diversification and the Cretaceous-Tertiary boundary. *Proc Natl Acad Sci U S A*. **4;100(3)**:1056-1061

The Arabidopsis Initiative, Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. (2000) *Nature*. **14;408(6814)**:796-815.

The C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. (1998) *Science*. **11;282(5396)**:2012-2018.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (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*. **11;22(22)**:4673-4680.

Vassilatis, D.K., Hohmann, J.G., Zeng, H., Li. F., Ranchalis, J.E., Mortrud, M.T., Brown, A., Rodriguez, S.S., Weller, J.R., Wright, A.C., Bergmann, J.E., Gaitanaris, G.A. (2003) The G protein-coupled receptor repertoires of human and mouse. *Proc Natl Acad Sci U S A*.**15;100(8)**:4903-8.

Young, J.M., Friedman, C., Williams, E.M., Ross, J.A., Tonnes-Priddy, L., Trask, B.J. (2002) Different evolutionary processes shaped the mouse and human olfactory receptor gene families. *Hum Mol Genet*. **1;11(5)**:535-546

Yun, C.W., Tamaki, H., Nakayama, R., Yamamoto, K., Kumagai, H. (1997) G-protein coupled receptor from yeast Saccharomyces cerevisiae. *Biochem Biophys Res Commun.* **17;240(2)**:287-292.

Zozulya, S., Echeverri, F., Nguyen, T. (2001) The human olfactory receptor repertoire. *Genome Biol.*;**2(6)**

Footnotes

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

Molecular Pharmacology Fast Forward. Published on February 1, 2005 as DOI: 10.1124/mol.104.009001
This article has not been copyedited and formatted. The final version may differ from this version.

Legends for figures

Fig 1.

Evolutionary tree with the number of GPCRs in different species indicated in graphs. The graph displays the logarithm of the number of GPCRs at the Y-axis, with the main class above the X-axis and the subdivision of *Rhodopsin* GPCRs below the X-axis (striped). The numbers at the nodes indicates the time in million years since the split at that node occurred, based on fossil data according to Hedges and Kumar 2003, *Dehal et al., 2002, **Springer et al., 2003.

Fig 2.

Bar graphs representing the relative number of Genscan proteins containing GPCRs (striped) and the relative number of ESTs containing GPCRs (non-striped) in the seven genomes where significant amount of EST-data is available. The number of GPCRs was calculated as the percentage of GPCRs in the respective dataset, EST or Genscan dataset, multiplied by 10000 (to obtain a number greater than 1) and then we used the natural logarithm of that number to display the relative number of GPCRs. In all cases but two, the striped graph is higher than the non-striped graph, i.e. the relative number of GPCRs in the Genscan dataset is higher than the relative number of GPCRs in the EST dataset. This would indicate that GPCRs are frequenlty expressed in low numbers.

Fig 3.

Circle graphs representing the absolute number of ESTs containing GPCRs in the seven genomes where significant amount of EST-data is available. One circle represents the total number of GPCRs in that genome (100%) and the different classes are fractions of that total.

Molecular Pharmacology Fast Forward. Published on February 1, 2005 as DOI: 10.1124/mol.104.009001
This article has not been copyedited and formatted. The final version may differ from this version.

Table 1. Description of the dataset used to build the Hidden Markov Models (HMMs) used for mining and classification into the 15 main families of GPCRs in the genomes investigated. The source of the dataset is indicated in parenthesis.

Model Name	Number of GPCRs	Description
ADHSEC	46	13 human secretin like GPCRs (Fredriksson et al., 2003) 33 Adhesion like GPCRs (Fredriksson et al., 2002; Fredriksson et al., 2003b; Fredriksson et al., 2003c)
CAMP	4	cAMP binding GPCRs from *Dictyostelium* (source: http://www.gpcr.org/7tm/)
DMODOR	39	Drosophila odorant receptors (source: http://www.gpcr.org/7tm/)
FZD	11	Human Frizzled 1-10 and Smoothened (Fredriksson et al., 2003)
GUST	72	Gustatory receptors from *Anopheles Gambiae* (Hill et al., 2002)
GLR	18	8 mGluRs, 3 Taste receptors type 1, 2 GABA receptors, CASR, 4 Orphan GPCRs (Fredriksson et al., 2003)
MLO	16	Plant GPCRs of the MLO type (Devoto et al., 2003)
NCHM	34	Nematode chemokine receptors (source: http://www.gpcr.org/7tm/)
OA1	2	Ocular albinism genes from mouse and human (Fredriksson et al., 2003)
RHOD	260	Human *Rhodopsin* receptors(Fredriksson et al., 2003)
STE2	4	Yeast pheromone receptors of the STE2 type (source: http://www.gpcr.org/7tm/)
STE3	4	Yeast pheromone receptors of the STE2 type (source: http://www.gpcr.org/7tm/)
TAS2	13	Human taste receptors, type 2 (Fredriksson et al., 2003)
VR	42	Vomeronasal receptors from mouse and rat (source: http://www.gpcr.org/7tm/)

Table 2. The number of GPCRs in the different main classes in the genomes investigated. --- indicates that no GPCRs of this class was found. The last row indicates the percent of total predicted genes.

	H.sapiens	*M.musculus*	*D.reiro*	*T.rubripes*	*C.intestinalis*	*D.melanogaster*	*A.gambiae*	*C.elegans*	*A.thaliana*	*O.sativa*	*S.cerevisiae*	*S.pombe*
ADH	27(31)	13	22	6	48	5	13	5	---	---	1	---
CAMP	---	---	---	---	---	---	---	---	---	---	---	1
DMODOR	---	---	---	---	---	58	85	---	---	---	---	---
FZD	10 (11)	11	14	10	7	7	7	5	---	---	---	1
GLR	24 (18)	112	52	26	8	9	8	6	---	---	---	2
MLO	---	---	---	---	---	---	---	---	5	1	---	2
NCHM	---	---	---	---	---	---	---	1006	---	---	---	---
OA1	1 (1)	1	---	1	---	---	1	---	---	---	---	---
RHOD	752 (614)	1106	591	224	139	76	77	124	---	---	---	1
SEC	20(15)	28	55	18	6	13	1	5	1	---	---	---
STE2	---	---	---	---	---	---	---	---	---	---	1	1
STE3	---	---	---	---	---	---	---	---	---	---	1	1
TAS2	13 (25)	3	2	---	---	---	---	---	---	---	---	---
VR	18 (25)	44	1	1	---	---	---	---	---	---	---	---
GUST	---	---	---	---	---	42	76	---	---	---	---	---
Total	865	1318	737	286	208	210	268	1149	6	1	3	9
% of total	1.60	1.19	1.23	0.97	1.31	1.47	1.66	5.69	0.09	0.04	0.05	0.18

Molecular Pharmacology Fast Forward. Published on February 1, 2005 as DOI: 10.1124/mol.104.009001
This article has not been copyedited and formatted. The final version may differ from this version.

Table 3. Description of the human dataset used for sub classification of *Rhodopsin* GPCRs. The classification is based on (Fredriksson et al., 2003) except for OLF, which are from (Zozulya et al., 2001).

Model Name	Number of GPCRs	Description
AMIN (α-group)	42	Bioamine GPCRs binding 5-HT, Dopamine, Histamine, Trace amines, Adrenalin and Acetylcholine
MEC (α-group)	22	Receptors for phospholipids (EDG), melanocortin, cannabinoids and somatostatin receptors together with 3 orphan GPCRs
MTN (α-group)	3	Melatonin and orphan receptor GPR50
OPN (α-group)	9	Opsin/putative opsin receptors and orphan receptors GPR21 and GPR52
PTGR (α-group)	15	Prostaglandin receptors and orphan receptors SREB 1-3, GPR26, GPR61, GPR62 and GPR78
PEP (β-group)	35	Receptors for NPY, Tachykinins, Neurotensin, orexin, neuromedin, NPFF, PrRP, GnRH, CCK etc.
CHEM (γ-group)	42	Bradykinin receptors and receptors/putative receptors for chemokines
MCHR (γ-group)	2	Receptors for melanocyte concentrating hormone
SOG (γ-group)	15	Somatostatin, Opsin and Galanin receptors
LGR (δ-group)	8	Orphan LGR receptors and receptors for Relaxin, FSH, TSH and LH
MRG (δ-group)	8	MRG and MAS receptors
OLF (δ-group)	347	Olfactory receptors according to (Zozulya et al., 2001).
PUR (δ-group)	42	Purin/putative purin receptors, formyl-peptide receptors, retinoic acid receptors and orphan gPCRs

Table 4. Subdivision of *Rhodopsin* GPCRs. The classification is according to Fredriksson et al. 2003. The numbers in parenthesis in the human column shows the number of GPCRs from each group as published in Fredriksson et al. 2003.

	H.sapiens	*M.musculus*	*D.reiro*	*T.rubripes*	*C.intestinalis*	*C.elegans*	*D.melanogaster*	*A.gambiae*
AMIN (α-group)	44(42)	57	122	28	23	20	21	18
MEC (α-group)	18(22)	23	35	11	17	1	1	2
MTN (α-group)	3(3)	2	6	3	3	-	2	2
OPN (α-group)	11(9)	6	31	21	3	1	8	12
PTGER (α-group)	13(15)	11	22	13	2	-	-	-
PEP (β-group)	43(35)	43	67	32	16	31	21	29
CHEM (γ-group)	43(42)	51	77	23	7	-	-	-
MCHR (γ-group)	1(2)	1	4	-	-	-	-	-
SOG (γ-group)	10(15)	15	34	9	18	10	5	3
LGR (δ-group)	7(8)	8	8	3	36	1	4	3
MRG (δ-group)	7(8)	22	-	-	-	-	-	-
OLF (δ-group)	494(347)	789	31	20	-	-	-	-
PUR (δ-group)	35(42)	49	80	26	-	-	-	-
Unclassified (UC)	20(17)	32	74	35	26	60	17	8
Total	749(607)	1109	591	224	151	124	79	77

Molecular Pharmacology Fast Forward. Published on February 1, 2005 as DOI: 10.1124/mol.104.009001
This article has not been copyedited and formatted. The final version may differ from this version.

Supplementary Table 1. Accession numbers of the sequences belonging to each HMM. (Hill et al., 2002) indicates that these sequences are available from Hill et al 2002 (Hill et al., 2002), supporting table S1. NA indicates that this information is not available.

Supplementary Table 2a-l. Results from the HMM-searches in the different genomes. Each species is placed in a separate sub-table, with receptors from each family placed in separate columns. The contents of the tables are as follows: Supplementary Table 2a, *H. sapiens*; Supplementary Table 2b, *M. musculus*; Supplementary Table 2c, *D. rerio*; Supplementary Table 2d, *T. rubripes*; Supplementary Table 2e, *C. intestinalis*; Supplementary Table 2f, *D. melanogaster*; Supplementary Table 2g, *A. gambiae*; Supplementary Table 2h, *C. elegans*; Supplementary Table 2i, *S. cerevisiae*; Supplementary Table 2j, *S. pombe*; Supplementary Table 2k, *A. thaliana*; Supplementary Table 2l, *O. sativa*. The sequence names match the names in the FASTA file in supplementary file 1.

Supplementary Table 3a-h. Results from the subdivision of the *Rhodopsin* family. Each species is placed in a separate sub-table, with receptors from each sub-group placed in separate columns. The contents of the tables are as follows: Supplementary Table 3a, *H. sapiens*; Supplementary Table 3b, *M. musculus*; Supplementary Table 3c, *D. rerio*; Supplementary Table 3d, *T. rubripes*; Supplementary Table 3e, *C. intestinalis*; Supplementary Table 3f, *D. melanogaster*; Supplementary Table 3g, *A. gambiae*; Supplementary Table 3h, *C. elegans*; The sequence names match the names in the FASTA file in supplementary file 1.

Figure 1

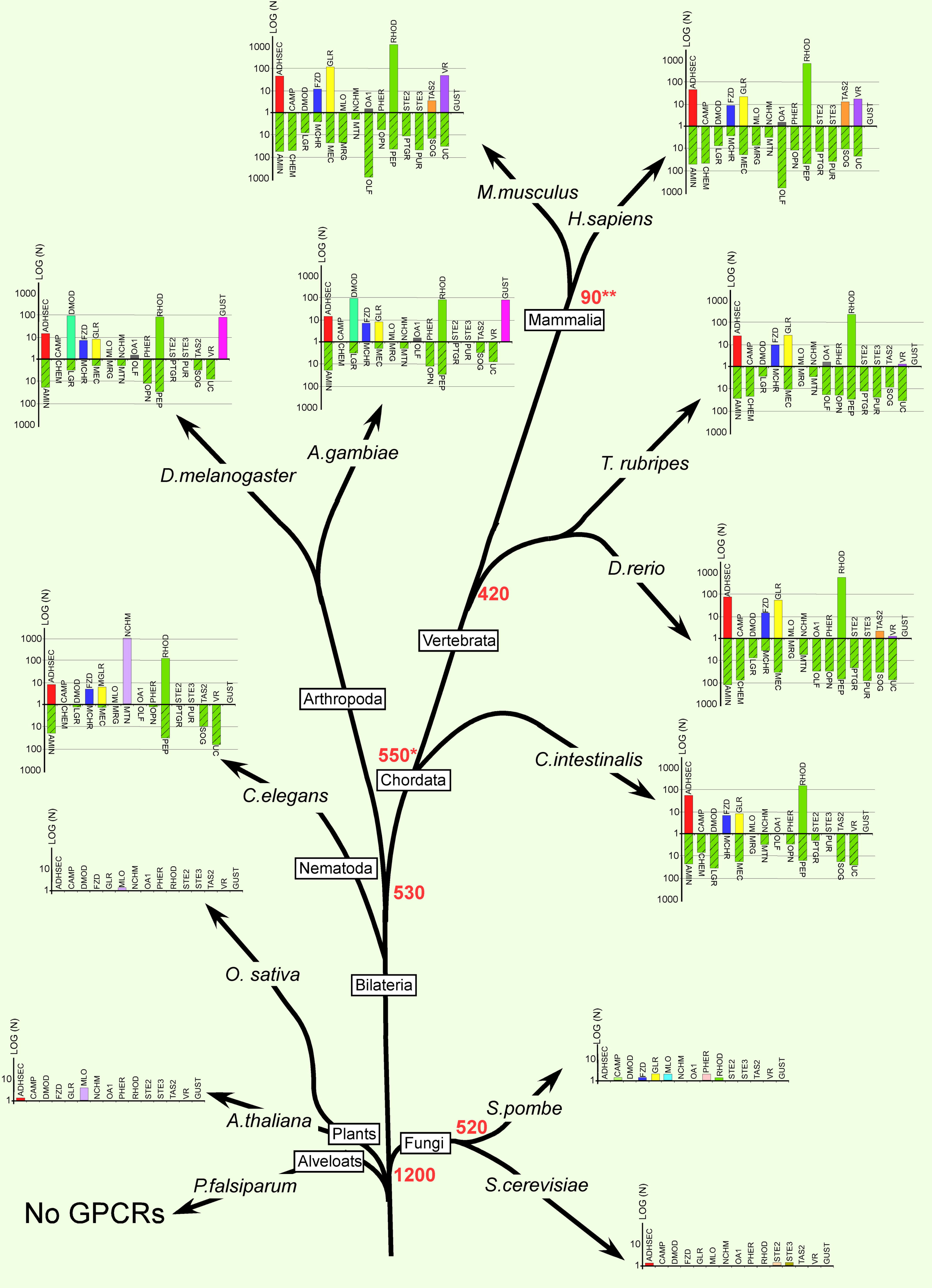

M.musculus
H.sapiens
Mammalia
90**
T. rubripes
D.rerio
420
Vertebrata
A.gambiae
D.melanogaster
Arthropoda
C.intestinalis
550*
Chordata
C.elegans
Nematoda
530
Bilateria
O. sativa
A.thaliana
Plants
Alveloats
P.falsiparum
No GPCRs
1200
Fungi
520
S.pombe
S.cerevisiae
LOG (N)
1000
100
10
1
ADHSEC
CAMP
DMOD
FZD
GLR
MGLR
MLO
NCHM
OA1
PHER
RHOD
STE2
STE3
TAS2
VR
GUST
AMIN
CHEM
LGR
MCHR
MEC
MRG
MTN
OLF
OPN
PEP
PTGR
PUR
SOG
UC

Figure 2

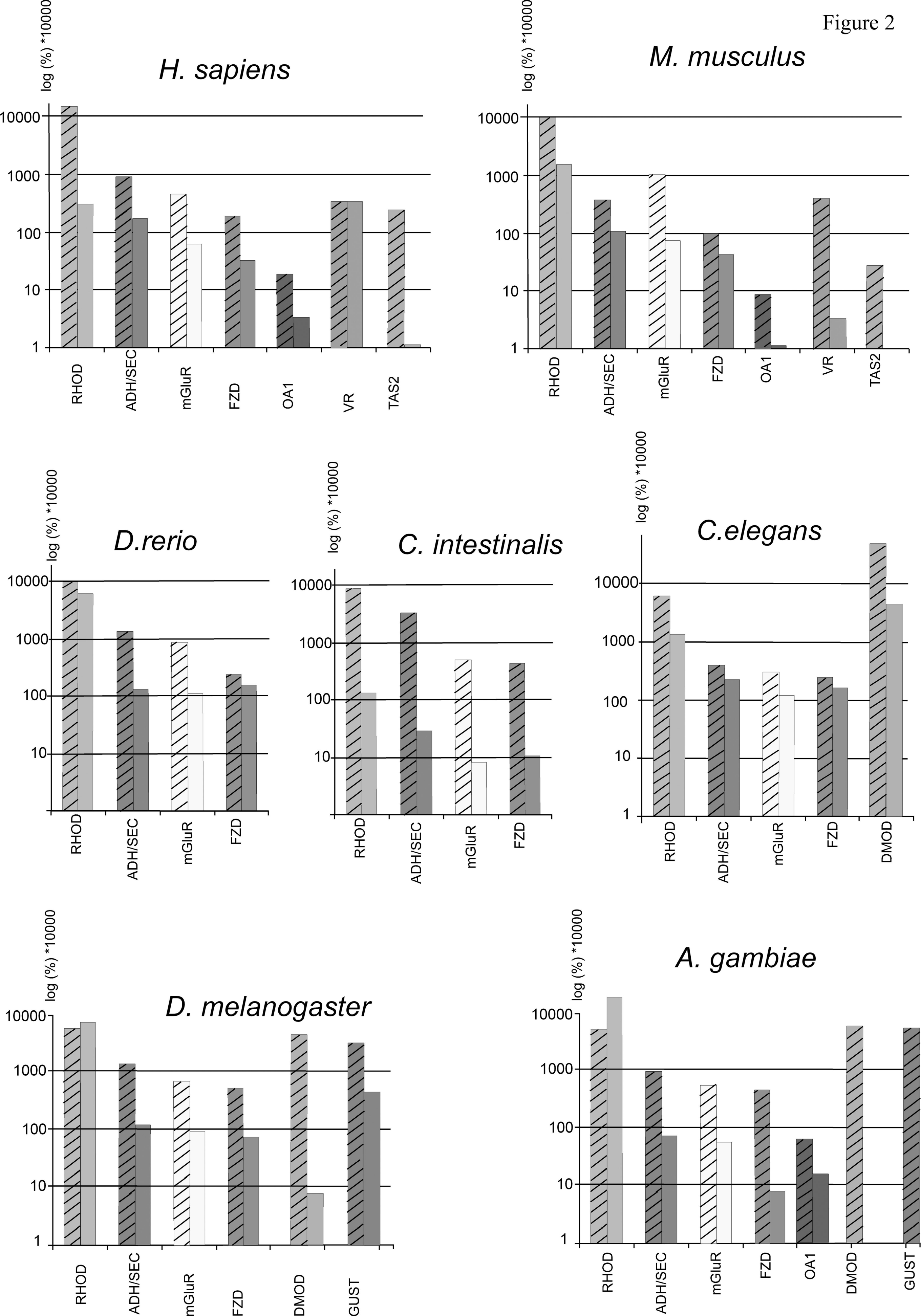
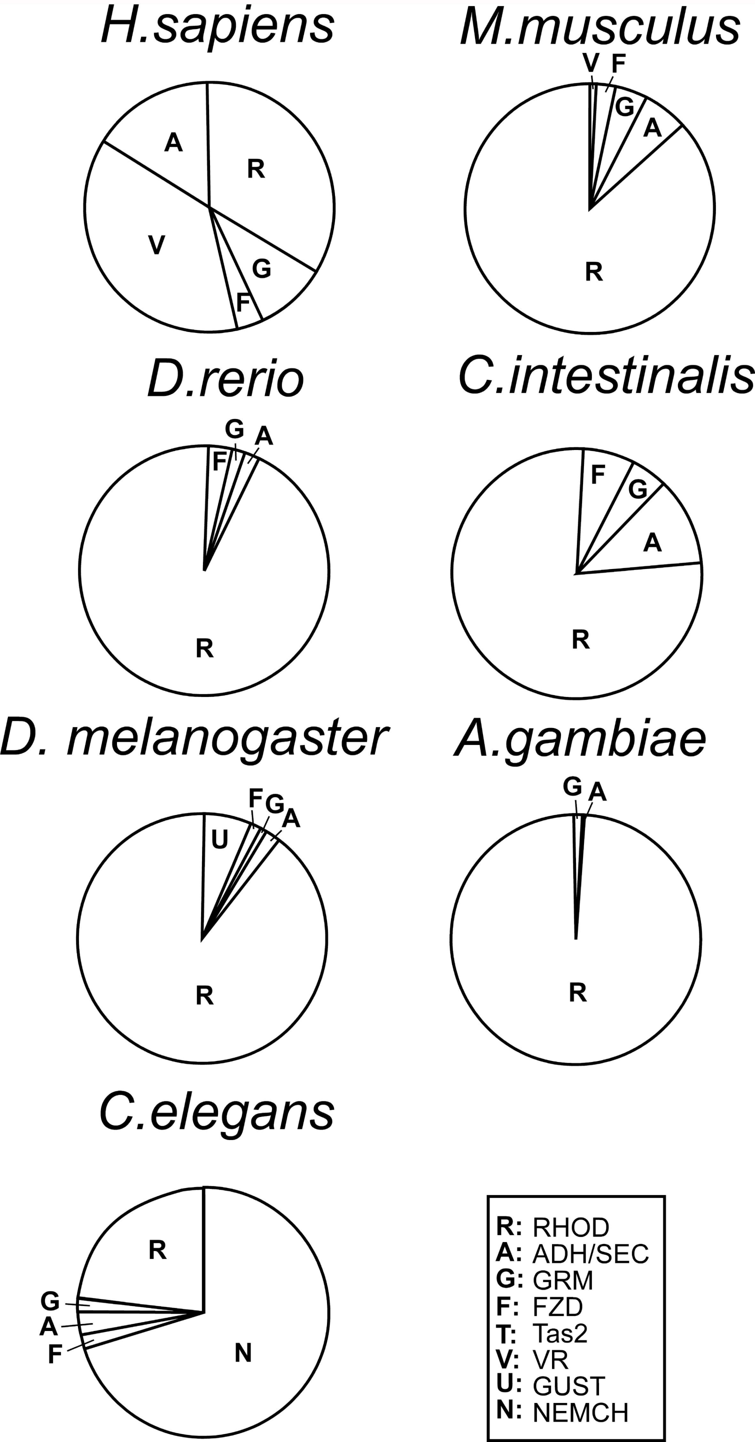

H. sapiens
log (%) *10000
10000
1000
100
10
1
RHOD
ADH/SEC
mGluR
FZD
OA1
VR
TAS2
M. musculus
log (%) *10000
10000
1000
100
10
1
RHOD
ADH/SEC
mGluR
FZD
OA1
VR
TAS2
D.rerio
log (%) *10000
10000
1000
100
10
RHOD
ADH/SEC
mGluR
FZD
C. intestinalis
log (%) *10000
10000
1000
100
10
RHOD
ADH/SEC
mGluR
FZD
C.elegans
log (%) *10000
10000
1000
100
10
1
RHOD
ADH/SEC
mGluR
FZD
DMOD
D. melanogaster
log (%) *10000
10000
1000
100
10
1
RHOD
ADH/SEC
mGluR
FZD
DMOD
GUST
A. gambiae
log (%) *10000
10000
1000
100
10
1
RHOD
ADH/SEC
mGluR
FZD
OA1
DMOD
GUST

Figure 3


H.sapiens
A
R
V
G
F
M.musculus
V
F
G
A
R
D.rerio
G
A
F
R
C.intestinalis
F
G
A
R
D. melanogaster
F
G
A
U
R
A.gambiae
G
A
R
C.elegans
R
G
A
F
N
R: RHOD
A: ADH/SEC
G: GRM
F: FZD
T: Tas2
V: VR
U: GUST
N: NEMCH