The **Secretin GPCRs Descended from the Family of Adhesion GPCRs**

Karl J. V. Nordström,* Malin C. Lagerström,*† Linn M. J. Wallér,* Robert Fredriksson,* and Helgi B. Schioth*

*Department of Neuroscience, Functional Pharmacology, Uppsala University, Uppsala, Sweden; and †Department of Neuroscience, Developmental Genetics, Uppsala University, Uppsala, Sweden

The Adhesion G-protein–coupled receptors (GPCRs) are the most complex gene family among GPCRs with large genomic size, multiple introns, and a fascinating flora of functional domains, though the evolutionary origin of this family has been obscure. Here we studied the evolution of all class B (7 tm2)–related genes, including the Adhesion, Secretin, and Methuselah families of GPCRs with a focus on nine genomes. We found that the cnidarian genome of *Nematostella vectensis* has a remarkably rich set of Adhesion GPCRs with a broad repertoire of N-terminal domains although this genome did not have any Secretin GPCRs. Moreover, the single-celled and colony-forming eukaryotes *Monosiga brevicollis* and *Dictyostelium discoideum* contain Adhesion-like GPCRs although these genomes do not have any Secretin GPCRs suggesting that the Adhesion types of GPCRs are the most ancient among class B GPCRs. Phylogenetic analysis found Adhesion group V (that contains GPR133 and GPR144) to be the closest relative to the Secretin family in the Adhesion family. Moreover, Adhesion group V sequences in *N. vectensis* share the same splice site setup as the Secretin GPCRs. Additionally, one of the most conserved motifs in the entire Secretin family is only found in group V of the Adhesion family. We suggest therefore that the Secretin family of GPCRs could have descended from group V Adhesion GPCRs. We found a set of unique Adhesion-like GPCRs in *N. vectensis* that have long N-termini containing one Somatomedin B domain each, which is a domain configuration similar to that of a set of Adhesion-like GPCRs found in *Branchiostoma floridae*. These sequences show slight similarities to Methuselah sequences found in insects. The extended class B GPCRs have a very complex evolutionary history with several species-specific expansions, and we identified at least 31 unique N-terminal domains originating from other protein classes. The overall N-terminal domain structure, however, concurs with the phylogenetic analysis of the transmembrane domains, thus enabling us to track the origin of most of the subgroups.

**Introduction**

G-protein–coupled receptors (GPCRs) are one of the largest protein families in mammalian genomes with about 800 members in the human genome (Lagerstrom and Schioth 2008). GPCRs are instrumental for hormonal and neurotransmitter signaling and are important in all major physiological systems of the body. GPCRs are among the most studied proteins in mammals given that at least 39 are considered to be major drug targets (Lagerstrom and Schioth 2008). Sequencing and assembly of a number of genomes have provided a comprehensive overview of the gene repertoire in mammals (Fredriksson and Schioth 2005). GPCRs have been grouped by the A–F or 1–5 systems based on structural similarities in receptor size as well as ligand interaction points and phylogeny (Kolakowski 1994; Bockaert and Pin 1999). Overall phylogenetic analysis of the human repertoire provided the GRAFS classification. This system is constituted of five families; Glutamate (G), Rhodopsin (R), Adhesion (A), Frizzled/Taste2 (F), and Secretin (S) (Fredriksson et al. 2003), each with distinct features. The Rhodopsin family of GPCRs is the largest family with about 672 members in the human genome including about 388 olfactory receptors. Most of these receptors have short N-termini and bind peptide-, amine-, and lipid-like compounds in a ligand-binding pocket within the transmembrane (TM) regions of the protein. The Glutamate GPCRs are characterized by the so-called “Venus Flytrap” mechanism, which is found in the N-termini and is crucial for ligand binding. The Frizzled receptors have long cysteine-rich N-termini that interact with the curly twisted Wnt protein and have a role in cell polarity, whereas the Taste 2 receptors lack long N- and C-termini and sense bitter-tasting substances. The Secretin GPCRs all have a hormone-binding domain (HBD) in their N-termini that interact with peptide hormones (Schioth et al. 2007). The Adhesion GPCR family, with 33 members, is the second largest GPCR family in humans. They are characterized by very long serine and threonine-rich N-termini that have multiple domains often found in other types of proteins such as tyrosine kinases. It has been speculated that these long N-termini have a role in cell-to-cell communication allowing them to participate in different types of cell guidance (Bjarnadottir et al. 2007). GPCRs similar to those in mammals are not found in bacteria, but mammalian-like GPCRs can be found in almost any eukaryotic organism. This includes plants (Devoto et al. 1999; Josefsson 1999), insects (Hill et al. 2002), fungi (Versele et al. 2001), and the amoeba *Dictyostelium discoideum* (Prabhui and Eichinger 2006). The light sensing 7TM protein found in bacteria, the bacterial rhodopsin, does not signal through G proteins and has very low sequence identity to GPCRs (Okada et al. 2001). It is thus presently unclear whether this protein has a common origin with GPCRs in eukaryotic organisms. The five main families (see above) are present in considerable numbers in most bilateral species including *Caenorhabditis elegans*, *Drosophila melanogaster*, *Anopheles gambiae*, *Strongylocentrotus purpuratus*, *Ciona intestinalis*, and the vertebrate species, although the hierarchical relationship between these five main families has not been determined (Fredriksson and Schioth 2005; Whittaker et al. 2006; Schioth et al. 2007). We recently observed that the five main families are found in large numbers in *Branchiostoma floridae* (Nordstrom et al. 2008). There are several lineage-specific groups of GPCRs, for example, the nematode chemoreceptors, the gustatory receptors from insects, the

Key words: evolution, GPCR, G-protein, 7TM, EGF, GRAFS.

E-mail: helgi.schioth@neuro.uu.se.

doi:10.1093/molbev/msn228
Advance Access publication October 8, 2008
© The Author 2008. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved.
For permissions, please e-mail: journals.permissions@oxfordjournals.org
odorant receptors from *D. melanogaster*, Mildew-resistance locus O (MLO) receptors in plants, fungal pheromone (STE2 and STE3) from yeast, and the *Methuselah* in *D. melanogaster*, but none of these families are found in vertebrates. Some similarities have been identified between the *Secretin*, *Adhesion*, and *Methuselah* GPCRs (Harmar 2001), and many database (7tm 2 in Pfam [Finn et al. 2006], GPCR _secretin_ in Interpro [Zdobnov and Apweiler 2001], and 7tm 2 in the National Center for Biotechnology Information [NCBI] Conserved domain database [CDD] [Marchler-Bauer and Bryant 2004]) use sequences from all three groups to form common fingerprints for search tools despite the fact that the functional characteristics of these families are highly divergent. Recently, Cardoso et al. (2005) studied evolutionary events that shaped the different branches of the *Secretin* GPCRs and clearly found the relationship between the *Secretin* receptors in *C. elegans* and the vertebrate *Secretin* family subbranches. The evolution of the *Adhesion* family has, however, not been studied in detail. The main reason for this is that the *Adhesion* family is, by far, the most complex group of GPCR sequences. These GPCRs are very large and have a large number of exons. Alternative splicing and complex processing steps, including the putative intracellular cleavage at the GPCR proteolytic site, are also contributing factors to their complexity (Bjarnadottir et al. 2007). The evolutionary relationship between the *Secretin*, *Adhesion*, and *Methuselah* GPCR and other GPCRs that may show relationship to 7tm2 domain (Finn et al. 2006) or the extended GPCR class B (also termed class 2) has not been resolved.

Here we investigated the evolution of the entire set of class B–related genes, including the *Adhesion*, *Secretin*, and *Methuselah* families of GPCRs in nine genomes. Special emphasis has been placed on the genomes of *Tetraodon nigroviridis*, *D. melanogaster*, and *C. elegans* together with the prevertebrate sea anemone *Nematostella vectensis*, whose genome was recently released (Putnam et al. 2007). We have also revisited the “secretin-like” GPCRs from the social amoeba *D. discoideum* and the choanoflagellate *Monosiga brevicollis* genomes (King et al. 2003; Prabhu and Eichinger 2006). These are both single-celled and colony-forming eukaryotes and are thereby interesting from an evolutionary perspective due to their positions in the evolutionary tree as predecessors to multicellular organisms. We have put in considerable effort to manually curate the sequences, which is a prerequisite for correct domain and splice site identification. Therefore, this study provides the most comprehensive overview of the evolutionary events that shaped the extended family of class B (family 2) GPCRs.

Materials and Methods

Sequence Retrieval and Editing

Full-length *Adhesion* sequences for *Homo sapiens*, *Mus musculus*, and *Gallus gallus* were downloaded according to information from previously published articles (Bjarnadottir et al. 2004; Lagerstrom et al. 2006). GPCRs in the remaining genomes were mined with a procedure combining hidden Markov model (HMM) and Blast searches (Altschul et al. 1990) similar to the one used by Fredriksson and Schioth (2005). The genome and the proteome for *T. nigroviridis* (version 7.46), *D. melanogaster* (version BDGP4.3.46), and *C. elegans* (version WB170.46) were downloaded from Ensembl (http://www.ensembl.org). The genome and the proteome for *N. vectensis* (version 1) and *M. brevicollis* (version 1) were downloaded from the Joint Genome Institute (http://www.jgi.doe.gov). The genome and the proteome for *D. discoideum* were downloaded from Dictybase (http://www.dictybase.org) on 26 April 2007. Each proteome was searched against a database consisting of the human RefSeq data set (Pruitt et al. 2007) with all GPCRs marked and representatives from all GPCR families reported on GPCRdb (http://www.gpcr.org/7tm) that are not present in human (nematode chemosensory receptors, the gustatory receptors from insects, the odorant receptors from *D. melanogaster*, MLO receptors from plants, fungal pheromone receptors [STE2 and STE3] from yeast, and vomeronasal receptors [V1R and V3R] and cAMP receptors from *D. discoideum*). If three or more of the top five hits were GPCRs, the protein was sorted out as a putative GPCR. The selected proteins were then searched with the HMMER package (Eddy 1998) against a database consisting of hmm-models adopted from Fredriksson and Schioth (2005) and with Blast against a database consisting of human GPCRs extended with the same nonhuman receptors as above. Proteins that had their best HMM-hit coinciding with three or more of the top five hits of the second Blast search were selected for further analysis. By aligning the sequences against their respective genome and only keeping one sequence from each locus, multiple variants of the same gene were avoided. The potential sequences were manually assembled in Editseq from the DNA-star package version 5.07 (DNASTAR, Madison, WI) according to the canonical splice site GT...AG. Finally, the sequences were aligned with Kalign (Lassmann 2005) and inspected and truncated in Jalview (Clamp et al. 2004) based on human sequences (either *Adhesion* or *Secretin* GPCRs), which had been truncated to only include the 7TM region according to an online rps-Blast search against the CDD. The 7TM region of each new family B receptor was used as bait in Blast searches against its genome, and all hits not coinciding with previously assembled transcripts were reviewed and assembled.

Domain Search

The N-termini of the sequences were assembled in the same manner as mentioned above and then searched for their domain extent in rps-Blast with a cutoff e value of 0.1. The *Methuselah* sequences did not show any domain-specific areas in rps-Blast. These sequences were therefore searched with InterProScan at http://www.ebi.ac.uk/InterProScan (Zdobnov and Apweiler 2001).

Confirmation and Classification

The 7TM regions were aligned to NCBI’s nonredundant (nr) database with Blast, and each of the first five hits had to belong to the extended family of class B (*Secretin*, *Adhesion*, or *Methuselah*). All sequence hits in nr were
extracted and aligned to the Pfam data set. They were considered as class B transcripts if they contained a 7tm_2 domain not overlapped by any higher scoring hit. The criteria for transcripts from *D. discoideum*, *M. brevicollis*, and the *Methuselah* family in *D. melanogaster* were lowered due to their divergence from other species. They were kept on the same basis as those used for categorization of the nr transcripts above. These sequences were then classified into the *Adhesion*, *Secretin*, or *Methuselah* family on the basis of a Blast search against our internal database consisting of all human GPCRs together with the nematode chemosensory receptors, the gustatory receptors from insects, the odorant receptors from *D. melanogaster*, MLO receptors from plants, fungal pheromone receptors (STE2 and STE3) from yeast, vomeronasal receptors (V1R and V3R) and cAMP receptors from *D. discoideum*. The sequences were classified to the respective family with three or more of the top five hits.

Sequence Identity

Global alignments between all truncated sequences were made using the program package (Rice et al. 2000). The result was compiled using an in-house parser written in Python and viewed and manipulated using the spreadsheet program in the OpenOffice.org suite.

Phylogenetic Analysis

The protein sequences of GPCRs chosen for phylogenetic analysis were aligned using Kalign (Lassmann and Sonnhammer 2005) with default parameter settings. The alignments were retranslated to nucleotides using the genomic sequence procured when assembling the GPCR and correcting its splice sites. We used Markov Chain Monte Carlo (MCMC) analysis with MrBayes (Ronquist and Huelsenbeck 2003), but in order to get at the liberal posterior probability, we used the majority rule. Maximum likelihood branch age with the majority rule was constructed using CONSENSE from the PHYLIP 3.67 package (Felsenstein 1993), and each alignment was analyzed with the general time reversible model with a proportion of invariable sites and a gamma-shaped distribution of rates across sites in MrBayes (nst = 6 and rates = invgamma). Each analysis ran 500,000 generations. Every hundredth tree from the last 100,000 generations was sampled, and a consensus tree was constructed using CONSENSE from the PHYLIP 3.67 package with the majority rule. Maximum likelihood branch lengths were then calculated with DNAML, also from the PHYLIP 3.67 package, and the tree was plotted in the Win32 version of TreeView 1.6.6 (Page 1996). The resulting tree was edited in InkScape.

Splice Sites

The positions of splice sites in the 7TM region were extracted by aligning the edited and truncated sequences to their corresponding genome using BLAST. All sequences were aligned with Kalign. Conserved splice sites were identified by viewing the positions of the splice sites and the alignment together in Jalview. Data were manipulated with OpenOffice.

Results

Sequence Retrieval and Classification

We collected 250 sequences from the extended class B GPCRs in *H. sapiens*, *M. musculus*, *G. gallus*, *T. nigroviridis*, *D. melanogaster*, *C. elegans*, *N. vectensis*, *M. brevicollis*, and *D. discoideum* (see fig. 1 and supplementary fig. 1, Supplementary Material online). GPCRs from *H. sapiens* (33 *Adhesion* and 15 *Secretin*), *M. musculus* (29 *Adhesion* and 15 *Secretin*), and *G. gallus* (18 *Adhesion* and 11 *Secretin*) were retrieved from previously published work (Fredriksson et al. 2003; Bjarnadottir et al. 2004, 2007; Lagerstrom et al. 2006). There are 22 *Adhesion* and 14 *Secretin* GPCRs in *G. gallus*, but only those with a full 7TM region were used here. Although the proteomes of *T. nigroviridis*, *D. melanogaster*, *C. elegans*, *M. brevicollis*, and *D. discoideum* have been, to some extent, mined for GPCRs (Harmar 2001; King et al. 2003; Cardoso et al. 2005; Metpally and Sowdhamini 2005; Prabhu and Eichinger 2006), we revisited them in order to extend these analyses and use the same method for all species. The genomic assemblies and the sequence count of these species have also improved the quality of protein predictions, especially for the *Adhesion* GPCRs. We found 23 *Adhesion* and 22 *Secretin* GPCRs in *T. nigroviridis*, with one additional *Adhesion* GPCR and one novel *Secretin* GPCR in comparison to a search performed by Metpally and Sowdhamini (2005). In *D. melanogaster*, we identified all class B GPCRs found by Harmar (2001) and Cardoso et al. (2005) and moreover one additional *Adhesion* GPCR and six additional *Methuselah* GPCRs, providing a total number of 5 *Adhesion*, 5 *Secretin*, and 15 *Methuselah* GPCRs in this species. Three of the *Methuselah* GPCRs in *D. melanogaster*, CG30018-PC, CG32476-PA, and dm_mth4 did not have any Blast hits in the NCBI nr database outside the insect class among their first 100 hits. A fourth transcript, dm_mth5, hit transcripts without a 7tm_2 domain. These transcripts were kept on the basis that they had a 7tm_2 domain, according to the Pfam search, and hit the other *Methuselah* GPCRs. In *C. elegans*, we found the three transcripts annotated as *Secretin* GPCRs by Harmar (2001) and Cardoso et al. (2005). We found one additional *Adhesion* sequence compared with the three identified by Harmar in *C. elegans*. The *N. vectensis* genome had not previously been mined for class B GPCRs, and we conclude here that it contains 38 class B transcripts of which 37 were classified as *Adhesion* and one, Nv112360, as a *Methuselah* GPCR. Four of the *N. vectensis* *Adhesion* GPCRs, Nv187681, Nv211490, Nv212781, and Nv215376, showed some similarity to *Methuselah* GPCRs. We found six class B transcripts in the *M. brevicollis* genome, which all were classified as *Adhesion* GPCRs of which Mb_21626 was previously identified by King et al. (2003). All of them have their first hit against an *Adhesion* GPCR, but both Mb_7962 and Mb_10341 also hit GPCRs from other families among their first five Blast hits. Three of the top five transcripts of
Mb_7962 did not contain a 7tm_2 domain. However, Mb_7962 contains a 7tm_2 domain according to the Pfam search. Because two of the odd hits were annotated as orthologs to the confirmed class B transcript that were the second best hit of Mb_7962, we suggest that this M. brevicollis transcript belongs to this family. We also revisited the D. discoideum genome and found one class B transcript. This was the same transcript which Prabhu and Eichinger (2006) classified as a class B GPCR. Here, we classified it as an ancestral GPCR because the first 10 hits against our classifying database were all GPCRs. All edited sequences are available in FASTA format in supplementary file 1 (Supplementary Material online), and the original sequences are described in supplementary file 2 (Supplementary Material online).

Phylogenetic Analysis

We constructed a phylogenetic tree of the Adhesion GPCRs (fig. 2 and supplementary fig. 2, Supplementary Material online) using a bootstrapped nucleotide method with MrBayes and estimated the branch lengths using DNAML from the PHYLIP 3.67 package. According to Cardoso et al. (2005), the two most ancient groups of Secretin family GPCRs are the corticotropin-releasing hormone receptors (CRHRs) (group A) and the calcitonin receptors (CALCRs) (group E). We constructed a preliminary tree of the Secretin GPCRs with an outgroup of Adhesion GPCRs and identified the Secretin group A as the group most closely related to the Adhesion GPCRs. Based on this, we included all the prevertebrate and group A Secretin GPCRs in the tree of the Adhesion GPCRs. This demonstrates that the Adhesion groups I–VIII classification, we presented based on the human and the mouse sequences, is well conserved among the vertebrates (Bjarnadottir et al. 2004). The Adhesion tree has two major nodes for which bifurcal topology could not be reached. The node closest to the outgroup consists of group VI, the very long G-protein–coupled receptors (VLGR1s); a branch with groups III, VIII, and GPR128; and a branch with 12 N. vectensis GPCRs (see fig. 2 and supplementary fig. 2, Supplementary Material online). The latter branch, marked NvX, contains both the gene classified as Methuselah and the three genes with similarities to the Methuselah from N. vectensis. The other major node holds one branch with GPR144, part of group V, and the Secretin GPCRs, one branch with groups I and II, and two branches with groups IV and VII. The GPR133 orthologs from group V are also placed in this node.

Group I consists of three lectomedin receptors (LECs) and one EGF-TMVII-latrophilin-related receptor (ETL) in H. sapiens. Gallus gallus is missing LEC1 and T. nigroviridis LEC2 although there is one T. nigroviridis GPCR, TnLec3-3, basal to the other LECs. There is also one C. elegans sequence, A_ce4k, which clusters with LEC3. Group II holds EGF-like module–containing receptors (EMRs) and a CD97 antigen receptor (CD97). Homo sapiens has an expansion of four EMRs compared with T. nigroviridis, which has one ortholog in TnCD97-1e. These two species also have one ortholog of CD97 each. Gallus gallus and the prevertebrate species are not present in this group. Groups I and II cluster together in the tree and basal to both groups, we find three prevertebrate sequences. Closest to groups I and II is one N. vectensis sequence, Nv24490, and basal to this and the two groups is a clade consisting of one D. melanogaster GPCR, sDm2, and one C. elegans GPCR, A_ce3. Group III is made up of GPR123, GPR124, and GPR125 in H. sapiens. All the vertebrates we searched have the same three sequences found in H. sapiens, whereas T. nigroviridis has a duplication of one of the three, which is placed basal to the other vertebrate sequences. Basal to group III, there are five sequences; sdm_6 from D. melanogaster, Mb_37365 from M. brevicollis, and Nv238736, Nv199785, and Nv200012 from N. vectensis. Group IV is represented by three EGF LAG seven-pass G-type receptor (CELSRs) in H. sapiens, whereas G. gallus has two orthologs, one to CELSR1 and one to CELSR3. There are also two sequences in T. nigroviridis, one basal to CELSR1 and one that clusters with the human CELSR2. Basal to group IV, there are sdm1 from D. melanogaster and Nv84228 from N. vectensis. Group V consists of GPR133 and GPR144 in H. sapiens, and G. gallus has the same configuration. Tetraodon nigroviridis only

---

**FIG. 1.—**A schematic tree showing the total numbers of class B GPCRs in each species in evolutionary perspective. Each family is represented by a color; blue for the Adhesion family, red for the Secretin family, and green for the Methuselah family.
contains a GPR144 ortholog. In group VI, H. sapiens has a repertoire of five GPCRs, GPR110, GPR111, GPR113, GPR115, and GPR116, whereas G. gallus has two; an ortholog to GPR116 and ggGPR115, which falls basal to GPR110 and GPR115 in mammals. Tetraodon nigroviridis also has two group VI sequences, TnGPR116-1 basal to GPR110, GPR111, and GPR115 and TnGPR116-2 basal to the whole group and thereby closest to GPR113. We could not find evidence for group VI outside of the vertebrate species. Group VII consists of the brain-specific angiogenesis-inhibitory receptors (BAIs) which are present in three copies in vertebrates. The G. gallus ortholog of BAI2 is not included in the tree as its 7tm-domain was incomplete. According to our tree, BAI1 falls basal to BAI2 and BAI3. The human members of group VIII are GPR56, GPR64 (or H6e), GPR97, GPR112, GPR114, and GPR126. Only GPR112 and GPR126 are found in G. gallus. GPR56 and GPR114 form one subbranch together with GPR97. Basal to this clade are TnGPR97-1 and TnGPR56-1 from T. nigroviridis. Tetraodon nigroviridis also has one GPR126 ortholog in TnGPR126-1. Between GPR64 and GPR112 are TnGPR112-3 and TnGPR112-2 from T. nigroviridis, and basal to the entire group is TnGPR126-3 from T. nigroviridis.

Among the ungrouped Adhesion GPCRs, VLGR1 and GPR128 are both present in H. sapiens and G. gallus and form two stable clades. Mb_21626 from M. breviceps is located basal to the VLGR1 node. The M. breviceps GPCR Mb_22592 is placed in the collapsed node holding the N. vectensis expansion, groups III, VI, and VIII. The D. discoideum GPCR, dd_1, shares a node with sdm8 and sdm9 from D. melanogaster. There are six sequences, A_ce5k from C. elegans and Nv81203, Nv125574, Nv133486, Nv145494, and Nv17885 from N. vectensis, basal to the clade consisting of group V and the Secretin GPCRs but with bootstrap values below 50%. The N. vectensis GPCR Nv78835 is placed basal to group VII with a bootstrap value of 25%.

We constructed a phylogenetic tree of the Secretin GPCRs with the same method (see fig. 3 and supplementary fig. 3, Supplementary Material online). According to the Adhesion tree (see fig. 2 and supplementary fig. 2, Supplementary Material online), group V was the Adhesion group closest to the Secretin family, and therefore, we included it in the tree together with six N. vectensis GPCRs showing similarities to group V. The human Frizzled GPCRs were used as an outgroup. The tree places Ce_C18B12 and Ce_13B9.4 from C. elegans sequences together with one D. melanogaster sequence (CG13758-PA) at the root of the Secretin clade. The pair of CG8422-PA and CG12370-PA from D. melanogaster are also placed at the root. The most basal vertebrate GPCRs are present in group A (CRHRs) in the same node, followed by group E (CALCRs), D (parathyroid hormone receptors [PTHRs]), and finally B (growth hormone–releasing hormone receptor [GHRHR], secretin receptor [SCTR], vasoactive intestinal peptide receptors [VIPRs], and pituitary adenyl cyclase–activating protein [PACAP]) and C (glucagon-like peptide receptors [GLPRs], glucagon receptor [GCGR], and gastric inhibitory polypeptide receptor [GIPR]). Group A consists of CRHR1 and CRHR2, which are present in both mammals and G. gallus. Tetraodon nigroviridis has one ortholog to CRHR1 in Tn_22293. In group B, the mammals and G. gallus have the complete set of GHRHR, SCTR, VIPR1, VIPR2, and PACAP. The mammalian GHRHRs and the G. gallus homolog do not cluster in the same node. There is one T. nigroviridis sequence, Tn_12645, basal to the mammalian GHRHRs, whereas there are two T. nigroviridis sequences, Tn_11830 and Tn_34494, that group basally with the G. gallus GHRHR. Beside this, T. nigroviridis has a complete set with duplicate versions of PACAP, VIPR1, and VIPR2. In group C, all vertebrates have one ortholog of GLP1R, but G. gallus lacks GLP1R. Tn_19752 and Tn_23238 from T. nigroviridis are placed basal to a clade of GCGR and GIPR and the clade of GLP1R. The mammals have one ortholog each of GCGR and GIPR, whereas G. gallus lacks GCGR. For group D, the mammals and G. gallus both have one ortholog of PTHR1 and PTHR2. The T. nigroviridis GPCR Tn_14484 falls basal to the mammalian PTHR2s. There is also one node holding ggPTHR2, Tn_16129, a clade of ggPTHR1 and Tn_35038, and a clade of the mammalian PTHR1s. Group E is the only Secretin group that has prevertebrate GPCRs associated with it. It has two D. melanogaster GPCRs, CG4395-PA and CG32843-PA, basal to the other GPCRs. Both G. gallus and the mammals have one ortholog each of CALCRL and CALCRL. There are also four T. nigroviridis GPCRs; Tn_32973 basal to CALCRL and Tn_10414 and Tn_16864 and Tn_20073 basal to CALCRL.

We also calculated a tree containing the Methuselah GPCRs, the new expansion of N. vectensis GPCRs and the Adhesion GPCRs in groups III, VI, and VIII (see fig. 3 and supplementary fig. 3, Supplementary Material online). Because the domain structure (see below) of the N. vectensis expansion is similar to an expansion in B. floridae (Nordstrom et al. 2008), these sequences were also included. We excluded GPCRs from G. gallus in groups III and VI. The resulting tree (see fig. 3 and supplementary fig. 3, Supplementary Material online) inserts the Methuselah family closest to the outgroup. The expansions in N. vectensis and B. floridae are placed in the same node as groups III, VI, and VIII. Bf133112 from B. floridae is placed basal to the N. vectensis expansion with a bootstrap value below 50%.

Domain Search

We searched all Adhesion GPCRs for conserved domains and the domains of the N-terminus are presented together with abbreviations in figure 4. The domain structure of H. sapiens and M. musculus has been thoroughly described in our previous paper (Bjarnadottir et al. 2007). Gallus gallus—in group I, ggETL has a GPS domain; ggLEC2 has a GPS, a HBD, an OLFL, and a GBL domain; and ggLEC3 has the same setup as ggLEC2 excluding the HBD domain. No G. gallus GPCRs could be found in group II. In group III, ggGPR123 has no domains in the N-terminus, ggGPR124 has an Ig and an LRR domain, whereas ggGPR125 has a GPS domain. There are two group IV sequences: ggCELSR1 with one GPS, one HBD, one EG-F_Lam, one LamG, and one EGF domain and ggCELSR3 with one GPS followed by one HBD, one
EGF_Lam, one EGF, one LamG, another EGF, one LamG, two more EGF, and seven CA domains. There are one or tholog to GPR133 and GPR144 each, but none of them con- tains any domains in their N-termini. In group VI, ggGPR115 contains one GPS domain and ggGPR116 has a GPS and an Ig domain. Group VII contains three GPCRs in G. gallus. ggBAI1 has one HBD domain and four TSP1 domains; ggBAI2 does not have any domains in its N-terminus; and ggBAI3 has one GPS, one HBD, and four TSP1 domains. In group VIII, ggGPR112, ggGPR116, and ggHE6 have one GPS domain each. The ortholog to GPR128 has no domains in its N-terminus, and the ortholog to VLGR1 has one GPS and three Calx_beta domains.

**Tetraodon nigroviridis**

The group structure of the Adhesion GPCRs is well conserved in T. nigroviridis with sequences from all groups although orthologs to GPR128, GPR133, and VLGR1 are missing. In group I, the ortholog to ETL has gained an HprK domain, and we could not find OLF or GBL domains in the end of the N-termini of TnLec3-1 although these are present in all the human LECs. The two sequences from T. nigroviridis in group II have three conserved EGF domains each. In group III, TnGpr124-1 does not have a GPS do- main. Both TnGpr124-1 and TnGpr125-1 have lost the HBD domain compared with their human orthologs. TnCelsr1-1 only has six CA domains in the end of the N-terminus compared with seven in the human ortholog but has an insertion of a TNFR domain after the EGF_Lam domain. The domain structure of TnCelsr2-1 is the same as its human ortholog. The ortholog to GPR144 has a PTX and an Urease_beta domain but no GPS domain. In group VI, TnGpr116-1 and TnGpr116-2 have only GPS domains. In group VII, TnBai2-1 has one GPS, one OPF, one HBD, and three TSP_1 domains. TnBai3-1 has the same N-termi- nus setup except for the OPF domain. TnBai3-3 only has four TSP_1 domains. TnGpr56-1, TnGpr97-1, and TnGpr126-3 of group VIII only have the GPS domain. In the same group, TnGpr112-2 and TnGpr112-3 have one GPS and one PTX domain, whereas TnGpr126-1, in addition to this setup, sports a CUB domain.

**Drosophila melanogaster**

sDm2, which falls basal to groups I and II, has both a GPS and a GBL domain. In group III, the sDm6 has one HBD, one Ig domain, and two LRR domains. The sDm1 of group IV has one HBD, two EGF_Lam, and one Metallotoho_PEC domain, followed by two pairs of alter- nating LamG and EGF domains. The N-terminus ends with seven CA domains. sDm8 and sDm9 could not be as- signed to a group based on our phylogenetic examinations. sDm8 has a GPS domain.

**Caenorhabditis elegans**

Ce3 and Ce4 are associated to group I, and both have one GPS, one HBD, and one GBL domain. In addition to this, Ce5 also has a CLECT domain. Ce2, which only has a GPS domain, did not place in any group, whereas the do- main setup for Ce5 is similar to group IV. This sequence has one GPS, followed by one HBD, three EGF, two LamG, another EGF, and six CA domains.

**Nematostella vectensis**

The phylogenetic trees (see figs. 2 and 3 and supple- mentary fig. 2 and 3, Supplementary Material online) show an expansion with 12 N. vectensis sequences of which three have a Somatomedin_B domain in the N-termini. There is a fourth N. vectensis sequence, Nv212781, that contains this domain, and this sequence was grouped with the other 12 in the expansion. The remaining nine in the expansion lack the Somatomedin_B domain, but they also lack a stop codon in the end of the transcript. When we extracted the genome sequence downstream of these transcripts, trans- lated in the three forward reading frames, and searched for conserved domains, we found reminiscent traces of So- matomedin_B domains. The group II sequence Nv24490 contains a GPS domain in its N-termini. Nv_242046 has a HBD and an Ig domain, which is similar to group III. Based on the phylogenetic analysis, Nv_200012, without domains in the N-terminus, Nv_238726, with a GPS, an Ig, a WSC, and a DUF885 domain, and Nv_199785, with a GPS, an fn3 domain, and three F5_F8_type_C domains, belong to group III. In group IV, Nv_84226 has a series of a GPS, a HDB, an EGF_Lam, two EGF, one LamG, one EGF, one LamG, one EGF, and eight CA domains. The group V sequence Nv20791 has only a GPS domain. Among the sequences associated with group V, Nv_204814 and Nv_201898 have a GPS and a CLECT domain. The CLECT is present in a group I sequence in C. elegans and in the human GPR144, but the phylo- genetic analysis puts Nv_204814 closer to group V. Nv_242264 has one GPS domain followed by 16 Calx_beta, one LamG, and another two Calx_beta do- mains and is likely an ortholog to VLGR1 because of the Calx_beta domains. Fifteen N. vectensis sequences did not place into any of the Adhesion groups. Of these, seven contain a GPS domain.

**Monosiga brevicollis**

Mb_7962 has one GPS and seven Calx_beta domains, whereas Mb_10341 has a GPS and 12 Calx_beta domains suggesting that they are orthologs to VLGR1. Mb_12491 has one GPS, four EGF, and one ADH_Zinc_N domain and is a putative ortholog to group IV. We could not place
Mb_21626, Mb_22592, or Mb_37365 in any group, but all these sequences hold one GPS domain each in the N-termini.

**Dictyostelium discoideum**

The dd1 GPCR in *D. discoideum* does not have any other domains except the 7tm_2 domain.

**Splice Sites**

We analyzed the number and position of splice sites in the 7TM region of the class B GPCRs. We can confirm the results of Cardoso et al. (2005) considering the positions of the splice sites in the Secretin family GPCRs. There are seven well-conserved splice sites in the vertebrate groups A–D Secretin GPCRs and six in group E Secretin GPCRs (see fig. 5 and supplementary fig. 4, Supplementary
Material online). The first conserved splice site (css1) can be defined by one amino acid upstream to a conserved arginine located directly after the end of transmembrane helix 1 (TM1). The second conserved splice site (css2) is located in the first extracellular loop, three amino acids upstream to a conserved cysteine followed by either an arginine or a lysine. The third conserved splice site (css3) is located in the middle of TM4 in a tryptophan between two conserved glycines. The position of the fourth conserved splice site (css4) is one amino acid upstream to a conserved cysteine in the second extracellular loop. The fifth conserved splice site (css5) is located in the middle of TM5, two amino acids upstream to a conserved asparagine. The sixth conserved splice site (css6) is located at the end of the third intracellular loop and can be defined as four amino acids upstream to a position with either an arginine or a lysine in the beginning of TM6, and it is this splice site that is missing in group E. The seventh conserved splice site (css7) is located in the middle of TM7, one amino acid upstream to a conserved glutamine. The invertebrate Secretin GPCRs in C. elegans have fewer splice sites but generally share them with the vertebrates, whereas the splice sites in D. melanogaster are more dispersed.

These conserved splice sites are also present in the Adhesion family, although there are distinct differences between the groups. Groups I and II both have a similar setup having the four splice sites, css2, css4, css5, and css6, conserved. There are four prevertebrate GPCRs in group I. The N. vectensis gene Nv24490 has five splice sites, css3–css7, although the sites at css5 and css6 are shifted 6 amino acids and all 11 amino acids downstream, respectively. The D. melanogaster sequence in group I and A. ce4k has a splice site at css1. A. ce3k has a splice site between css1 and css2 and a second splice site at css4. Sdm_2 has one splice site between css2 and css3 and one at css5 and one at css6.

The sequences belonging to group III have four splice sites. Of these, one is located at css1 and one at css4. The remaining two do not coincide with the conserved splice sites found among the Secretin GPCRs but are associated with the second intracellular loop, which in group III is extended, and more proline rich compared with other class B GPCRs. The first of these splice sites is located in the beginning of TM3, two amino acids downstream of the cysteine defining css2. The second splice site follows the second extracellular loop, one amino acid upstream to a conserved arginine. The two conserved splice sites css1 and css4 are also present in the four N. vectensis GPCRs in this group, although Nv200012 has css1 shifted upstream. The first intracellular loop, where css1 is located, is extended in Nv200012. The D. melanogaster GPCR (sdm_6) lacks splice sites in the 7TM region but is the only prevertebrate gene with the second intracellular loop that is extended in the vertebrates. The N. vectensis GPCRs also lack the group III–specific splice sites associated with this loop. However, three of them, Nv199785, Nv200012, and Nv242046, have a splice site at css3. Nv200012 also has splice site at the beginning of the second intracellular loop. Group IV has five splice sites. Four of these coincide with css2, css5, css6, and css7. The fifth splice site is located close to css3 but has been shifted seven amino acids downstream and is located one amino acid upstream of another conserved glycine. The N. vectensis GPCR Nv84228 has six splice sites at css1–css5 and css7. The D. melanogaster GPCR sdm1, phylogenetically located basal to this group, has only one splice site in the 7TM region, and it is located at css4. The C. elegans GPCR A. ce5k has two splice sites, one close to css3 and one in TM6 close to css6. The M. brevicollis GPCR Mb_12491 putatively associated to this group lacks splice sites in the 7TM region. Group V is split between two neighboring nodes in the phylogenetic tree, but both have similar patterns of splice sites. All three vertebrate orthologs to GPR144 have six of the conserved Secretin splice sites, with only css4 is missing. The N. vectensis GPCR Nv201898 and Nv204814 both have the complete setup of conserved Secretin splice sites. All the vertebrate orthologs to GPR133 have all the seven conserved splice sites found in the Secretin family. Nv20971 has splice sites at css4–css7. The three N. vectensis GPCRs, Nv78835, Nv125574, and Nv217885, placed in the node holding groups I, II, IV, V, and VII, also have all seven splice sites css1–css7. The sequences in group VI only have one splice site per GPCR. This is located at css7. There are six to seven splice sites in group VII of which the first five splice sites are located at css1–css5. Like group III, group VII has an extended third extracellular loop, and the sixth splice site is located at the beginning of this loop between css5 and css6. Four vertebrate GPCRs in group VII are missing the third extracellular loop, and these are the ones with only six splice sites. The remaining seven vertebrate GPCRs have a seventh splice site at the end of the loop, which also coincides with css6. TnBAI2-1 also has an eighth splice site, which is located between css6 and css7. The sequences in group VIII, in general, have four splice sites at css1, css3, css4, and css7.

Discussion

We found that the N. vectensis genome has a remarkably rich set of Adhesion GPCRs with many of the domains found in mammalian GPCRs. Both the phylogeny (see fig. 2 and supplementary fig. 2, Supplementary Material online) and the domain composition (see fig. 4) show that N. vectensis has members, which clearly belong to group III and groups IV and V of the mammalian Adhesion GPCRs. Intriguingly, the N. vectensis genome does not have any Secretin GPCRs. Neither did we find any Secretin GPCRs in M. brevicollis or D. discoideum although both these genomes have Adhesion-like GPCRs. We can thus draw the conclusion that the Adhesion GPCRs are a more ancient GPCR family than the Secretin GPCRs (see fig. 1 and supplementary fig. 1, Supplementary Material online). This is the first time that an evolutionary hierarchy has been clearly delineated among the five main families of vertebrate GPCRs. This is also interesting because the biology of the Secretin receptors has been much more widely studied than that of Adhesion GPCRs.

We searched for evidence that the Secretin GPCRs, which are found in both D. melanogaster and C. elegans, might have originated from one of the ancient branches of Adhesion GPCRs. Our overall phylogenetic tree (see fig. 2 and supplementary fig. 2, Supplementary Material online).
suggests that group V (that contains GPR133 and GPR144 in humans) is the closest relative to the Secretin family in the Adhesion family. Interestingly, group V sequences in *N. vectensis* (Nv_201898 and Nv_204814) both share the same splice site setup as the Secretin GPCRs, and this splice site setup is not shared by any of the other ancient groups. One of the most conserved motifs in the whole Secretin family is PL(L/F)G found in TM6. This motif is highly conserved in Secretin groups A–C and E and also present in group D but with lower degree of conservation (see supplementary fig. 5, Supplementary Material online). If we consider all other Class B groups present in *N. vectensis*, this specific Secretin family motif is only found in group V of the Adhesion family. Taken together, this provides strong evidence that the Secretin family of GPCRs could have originated from group V Adhesion GPCRs.

We also find it likely that Adhesion groups I, IV, and V share a common ancestor because they group together in the phylogenetic analysis (see fig. 2 and supplementary fig. 2, Supplementary Material online). This hypothesis is also strengthened by the fact that three of their common conserved splice sites (css2, css5, and css6) are missing in the other groups present in *N. vectensis* (see fig. 5 and supplementary fig. 4, Supplementary Material online). This scenario also suggests that group VII, which we only find in vertebrates, arose from this branch of the evolutionary tree, most likely from group I or IV according to the phylogeny but have domain similarity that matched the particular group. The number above the members of the Secretin family denotes the number of transcripts with the given domain structure. Each domain is marked with a symbol, and the explanation is found in the legend in the lower left corner. The following domains were found: 7TM (7 transmembrane receptor [Secretin family]), GPS (GPCR proteolytic site), HBD (hormone-binding domain), OLF (olfactomedin domain), GBL (galactose-binding lectin domain), EGF_CA (calcium-binding epidermal growth factor-like domain), LRR (leucine-rich repeat), CA (cadin repeaters), TSP1 (thrombospondin repeats, type I), PTX (Pentraxin domain), LamG (laminin G domain), EGF_Lam (laminin type epidermal growth factor domain), Calx-b eta domain, OPF (Oligoendopetidase F), CUB (ClrC1s urinary epidermal growth factor and bone morphogenetic domain), Ig (immunoglobulin domain), SEA (sea urchin sperm protein domain), TNFR (tumor necrosis factor receptor domain), HprK (serine kinase of the HPr protein), EPTP (epitakin protein domain), CLECT (C-type lectin-like domain), Urease_beta (Urease beta subunit), Metallothio_PEC, Herpes_gp2 (equine herpesvirus glycoprotein gp2 domain), FA58C (coagulation factor 5/8 C-terminal domain), fn3 (fibronectin type 3 domain), SIN3 (histone deacetylase complex), Somatomedin B domain, Methuselah_N domain, DUF885, WSC domain, and ADH_Zinc_N (Zinc-binding dehydrogenase).
N-terminus, but its first 10 hits in a Blast search against our database of Class B GPCRs are all Adhesion GPCRs, suggesting that this gene is more similar to Adhesion GPCRs than any other branch of GPCRs.

Interestingly, we found a set of unique Adhesion-like GPCRs in *N. vectensis*, and these are shown in green in figure 4. These 13 genes do not have any GPS domain but have a TM domain that can readily be aligned with Adhesion GPCRs (amino acid identities range from 21% to 30%). These sequences have long N-termini containing one Somatomedin_B domain each. This Somatomedin_B domain is not found in any mammalian Adhesion GPCR but is, interestingly, found in a set of Adhesion-like GPCRs found in *B. floridana* (Nordström et al. 2008). The TM regions of these two groups do not group with each other or any of the other branches of class B sequences (see fig. 3 and supplementary fig. 3, Supplementary Material online). However, this unique N-terminal composition with no GPS domains and their relatively high amino acid identities (21–38%) suggests that these two groups could be related. One of the mysteries of the class B GPCRs is the origin of the Methuselah GPCRs. These genes are only found in insects (i.e., *D. melanogaster* and *A. gambiae*; Finn et al. 2006). The Methuselah GPCRs have long N-termini but have only one type of domain in their N-termini, the Methuselah_N domain. The Methuselah genes do not cluster with other branches of class B in phylogenetic analyses. We found, however, that one of the Somatomedin_B domain containing sequences (Nv112360) in *N. vectensis* had three of its five best hits as Methuselah in a Blast search among class B sequences. Moreover, an additional four of this type of *N. vectensis* sequences had Methuselah sequences among their top five best hits. The sequence similarity between the Methuselah and these Somatomedin_B containing *N. vectensis* sequences is about 15–32%. There are no conserved motifs in the 7TM region between these groups, and there is no obvious link between the expansion in *N. vectensis* and *B. floridana* and Methuselah with any of the groups of the classical Adhesion GPCRs, but the phylogeny shows closer relationship to the Adhesion branch shown on the right in figure 2 (groups III, VI, and VIII) as compared with the groups on the left.

The mammalian Adhesion GPCRs are known to have very complex genomic structure consisting of multiple introns and have large genomic size, whereas most GPCRs and, in particular, the Rhodopsin GPCRs are much simpler, often coded by a single exon. The splice sites in Adhesion GPCRs may play a role in forming alternative splice variants with different set of N-terminal domains (Bjarnadottir et al. 2004), and this could be important for interaction with other proteins. Interestingly, the overall complexity of the Adhesion GPCRs seems to be well conserved through evolution and thus likely to be important for their overall functions. Several splice sites are highly conserved in the extended class B type of sequences within the 7TM region (see fig. 5 and supplementary fig. 4, Supplementary Material online), and we did not find any gene with good genomic sequence coverage that did not have at least several introns (Nordström KJV, unpublished data).

There are two major contrasting theories about why mammalian Rhodopsin GPCRs have relatively small number of introns compared with invertebrates (Brosius 1999; Gentles and Karlin 1999). One report has suggested that there was a major loss of introns within the GPCR family, whereas we have argued that formation of new genes through RNA-based mechanisms explains the lower intron density in mammalian GPCRs. It is notable that no major expansion of the class B GPCRs that seem to originate from RNA-based mechanisms, as is observed for Rhodopsin GPCRs. Based on the high degree of conservation of the splice sites, it is likely that the expansion of Class B took place mainly through DNA-based mechanisms like duplications of the whole or parts of the genome. There are, however, examples of parts of the genes that are intron free in which introns seem to be lost, such as for the 7TM region of group VI, and there is no general pattern showing a higher number of introns in vertebrates compared with invertebrates (Nordström KJV, unpublished data). Moreover, there are examples even in the TM regions where acquisition of a loop in the second intracellular region in group III GPCRs is associated with the gain of one splice site and the movement of another, comparing the *N. vectensis* to the vertebrate sequences (see fig. 5 and supplementary fig. 4, Supplementary Material online). In group VII, some transcripts have an extra exon extending the third intracellular loop in a similar manner that an extra exon has been gained in some N-termini mammalian Adhesion GPCRs (Bjarnadottir et al. 2007). In general, the Class B sequences seem not to follow the 2R pattern (tetraploidization events) although several of the human Adhesion genes are found in the specific paralogues groups (Fredriksson et al. 2003). However, we find four clear examples where each gene is available in two copies in *T. nigroviridis*, which is thought to have gone through a third whole-genome duplication (3R) together with the other teleosts (Jailon et al. 2004). These genes are GPR123, PACAP, VIPR1, and VIPR2 and in Secretin group C, there is a node with two *T. nigroviridis* GPCRs basal to the clade containing GIPR and GCGR, which also are the result of putative 3R duplications. In that case, this is a duplication of a GPCR ancestral to both GIPR and GCGR. The fact that we do not see more duplications is consistent with theories of heavy gene loss reported to take place after a whole-genome duplication (Brunet et al. 2006). Finally, it is interesting to note that the repertoire of Adhesion GPCRs in *N. vectensis* is more closely related to that of the vertebrates compared with that of the Ecdysozoan *C. elegans* and *D. melanogaster*, which is in agreement with the conclusion that a large number of gene families have been lost in the Ecdysozoa lineage (Putnam et al. 2007).

In summary, we have performed detailed mining of the most multifaceted family of GPCRs in nine genomes. The repertoire of Adhesion GPCRs is uniquely complex with at least 31 unique N-terminal domains, lengths up to 7,042 amino acids, and intron numbers up to 100. At least 84% of the extended class B sequences have identifiable domains in their N-termini. The overall N-terminal domain structure fits remarkably well to the phylogenetic analysis of the TM domains enabling us to track the origin of most of the subgroups. We provide compelling evidence for the ancient origin of the Adhesion GPCR family. Moreover, it is likely that the biologically well-studied Secretin family of receptors, which mediate many key hormonal functions through binding hormones in long N-termini, originated from ancestors to the Adhesion GPCR family.
Supplementary Material

Supplementary files 1 and 2 and figures 1–4 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Chris Pickering for help with the language of the article and also Anders Hellström for initial data mining. The sequence data for *N. vectensis* and *M. brevicollis* were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/. The studies were supported by the Swedish Research Council, The Novo Nordisk Foundation, Swedish Royal Academy of Sciences, and Magnus Bergvall Foundation. M.C.L. was supported by the Swedish Brain Research Foundation. R.F. was supported by the Göran Gustafsson's Foundation.

Literature Cited

Brosius J. 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene. 238:115–134.


Billie Swalla, Associate Editor

Accepted September 27, 2008