Expansion of Secretin-Like G Protein-Coupled Receptors and Their Peptide Ligands via Local Duplications Before and After Two Rounds of Whole-Genome Duplication

Jong-Ik Hwang, Mi Jin Moon, Sumi Park, Dong-Kyu Kim, Eun Bee Cho, Nui Ha, Gi Hoon Son, Kyungjin Kim, Hubert Vaudry, and Jae Young Seong

Graduate School of Medicine, Korea University, Seoul, Republic of Korea
School of Biological Sciences, Seoul National University, Seoul, Republic of Korea
Laboratory of Neuronal and Neuroendocrine Differentiation and Communication, INSERM U982, University of Rouen, Mont-Saint-Aignan, France

These authors contributed equally to this work.

Associate editor: Yuseob Kim

Abstract

In humans, the secretin-like G protein-coupled receptor (GPCR) family comprises 15 members with 18 corresponding peptide ligand genes. Although members have been identified in a large variety of vertebrate and nonvertebrate species, the origin and relationship of these proteins remain unresolved. To address this issue, we employed large-scale genome comparisons to identify genome fragments with conserved synteny and matched these fragments to linkage groups in reconstructed early gnathostome ancestral chromosomes (GAC). This genome comparison revealed that most receptor and peptide genes were clustered in three GAC linkage groups and suggested that the ancestral forms of five peptide subfamilies (corticotropin-releasing hormone-like, calcitonin-like, parathyroid hormone-like, glucagon-like, and growth hormone-releasing hormone-like) and their cognate receptor families emerged through tandem local gene duplications before two rounds (2R) of whole-genome duplication. These subfamily genes have, then, been amplified by 2R whole-genome duplication, followed by additional local duplications and gene loss prior to the divergence of land vertebrates and teleosts. This study delineates a possible evolutionary scenario for whole secretin-like peptide and receptor family members and may shed light on evolutionary mechanisms for expansion of a gene family with a large number of paralogs.

Key words: secretin, G protein-coupled receptor, genome, comparative genomics, evolution, genome duplication.

Introduction

The secretin (SCT)-like (class B or 2) family of G protein-coupled receptors (GPCRs) represents one of the best-studied receptor families, particularly in the gut–brain interaction and neuroendocrine systems. This family of receptors comprises a relatively small population of members and exhibits unique structural features as compared with the rhodopsin-like (class A or 1) GPCRs (Fredriksson et al. 2003). In humans, the SCT-like GPCR family is composed of 15 members that are categorized into five subfamilies based on their phylogenetic relationship: A group contains the corticotropin-releasing hormone receptor 1 (CRHR1) and CRHR2; B group contains the calcitonin receptor (CALCR) and CALCRL; C group comprises the parathyroid hormone receptor 1 (PTH1R) and PTH2R; D group possesses the glucagon receptor (GCR), glucagon-like peptide 1 receptor (GLP1R), GLP2R, and glucose-dependent insulinotropic polypeptide receptor (GIPR); and E group includes the growth hormone-releasing hormone receptor (GHRHR), SCT receptor (SCTR), vasoactive intestinal peptide receptor 1 (VIPR1), VIPR2, and pituitary adenylate cyclase-activating polypeptide receptor (ADCYAP1R1). The SCT-like GPCRs differ in their structures of transmembrane helices (TMH) from the rhodopsin-like (class A) GPCRs. SCT-like receptors share a high degree of amino acid identity in the TMH (Harmar 2001; Harmar et al. 2012) but do not contain residues that are commonly found in the rhodopsin-like GPCRs (Fredriksson et al. 2003; Oh et al. 2006; Krishnan et al. 2012). Particularly, the SCT-like receptors possess a relatively long-structured N-terminal extracellular domain (ECD) of approximately 120 residues with conserved Cys residues that form a network of disulfide bridges, which are important for ligand binding to these receptors (Couvaineau et al. 2004; Moon et al. 2012). In addition, a sushi domain fold consisting of two antiparallel β sheets stabilized by disulfide bridges and a salt bridge is conserved in all class B GPCRs (Grace et al. 2004).

To facilitate receptor binding, the SCT-like peptide family has typical three-dimensional structures, such as a random coiled N-terminus followed by an α-helix in the middle of the
peptide (Jin et al. 2000; Marx et al. 2000; Neidigh et al. 2001; Parthier et al. 2007; Underwood et al. 2010). In particular, the central α-helical portion of the peptide binds to the receptor N-terminal ECD, and this binding is followed by binding of the N-terminal moiety of the peptide to the core domain, including the TMH and extracellular loops of the receptor, allowing receptor activation and G protein coupling (Runge et al. 2003; Castro et al. 2005; Wittelberger et al. 2006). Like their receptors, human SCT-like peptides are categorized into five subfamilies: A group contains CRH, urocortin (UCN), UCN2, and UCN3; B group includes CALC, calcitonin gene-related peptide-I (CGRP-I), CGRP-II, adrenomedullin (ADM), ADM2, and islet amyloid polypeptide (IAPP); C group has PTH, PTH2, and PTH-like hormone (PTH-LH); D group comprises glucagon (Gluc), GLP1, GLP2, and GIP; and E group consists of GHRH, SCT, PACAP, and VIP. Some peptides are encoded by a common precursor gene, which was likely generated by exon duplications during early vertebrate evolution (Sherwood et al. 2000; Vaudry et al. 2009). For instance, the glucagon (CGC) gene encodes a common Gluc-GLP1-GLP2 precursor and produces one or two mature forms of Gluc, GLP1, or GLP2 by a tissue-specific alternative posttranslational process (Mojsov et al. 1986; Kieffer and Habener 1999).

In nonmammalian vertebrate species, the ADCYAP1 gene encodes PACAP and a PACAP-related peptide (PRP, also called GHRH-like peptide [GHRH-LP]), and the VIP gene encodes VIP and a VIP-related peptide (VRP) also called either peptide histidine methionine or peptide histidine isoleucine (Sherwood et al. 2000; Tam et al. 2007; Roch et al. 2009).

The SCT-like family peptide and receptor genes are believed to have emerged early in metazoan evolution via multiple gene duplications of a common ancestral gene (Cardoso et al. 2006, 2010). Two rounds of large-scale genome duplication (the 2R hypothesis) during early vertebrate evolution further expanded their subfamily members (Lundin et al. 2003; Lee et al. 2007; Nakatani et al. 2007; Larsson et al. 2008). In addition, some members were likely generated by local duplication following 2R duplication and before the divergence of fish and land vertebrates (Kim et al. 2011). Furthermore, the teleost fish-specific 3R duplication (3R; Jailon et al. 2004) also produced new paralogous members. The pattern of gene duplication and loss, however, differs across vertebrate species such that the number of peptide genes differs by species.

Genome duplication events produce paralogous chromosomal regions, also called paralogons, which share paralogous genes (Lundin 1993; Coulier et al. 2000; Larhammar, et al. 2002; Fredriksson et al. 2003). Currently, the large accumulation of genome sequence information for many invertebrate and vertebrate species combined with recent advances in bioinformatic tools that allow large-scale comparisons of genomes has greatly facilitated the exploration of the origin and relationship of gene families (Lee et al. 2009; Kim et al. 2011, 2012). Presently, synteny-based genomic analyses provide more useful information regarding the evolutionary history of gene families than sequence-based phylogenetic approaches (Abi-Rached et al. 2002; Larhammar et al. 2002; Sundström et al. 2010; Daza et al. 2011; Widmark et al. 2011). Furthermore, comparing entire genomes of evolutionarily distant taxa have allowed reconstruction of the hypothetical chromosomes of the vertebrate and chordate ancestors, which provides another means to trace gene origins (Nakatani et al. 2007; Putnam et al. 2008). Recently, Yegorov and Good (2012) identified linkage groups of the relaxin family hormones and their receptors in vertebrate taxa and then used reconstruction of ancestral vertebrate genomes at distinct evolutionary time periods to trace the origin and duplication history of ligand and receptor genes.

Despite phylogenetic and/or genome comparison analyses of some SCT-like family peptides and receptors (Cardoso et al. 2006, 2010; Roch et al. 2009), the evolutionary origin and the relationship between family members remains unclear. In this study, we performed extensive genome synteny comparisons of the relaxin family chromosomes and matched the chromosomal fragments with conserved synteny to linkage groups of reconstructed vertebrate ancestor chromosomes according to the method of Yegorov and Good (2012) to delineate the presumptive evolutionary mechanism and history of the vertebrate SCT-like peptide and receptor families.

**Results**

**Phylogenetic Analyses of SCT-Like Receptor Subfamilies**

The amino acid sequences of SCT-like receptors from representatives vertebrate taxa as well as from protostomes (Caenorhabditis elegans and Drosophila melanogaster), early deuterostomes (Ciona intestinalis and amphioxus, Brachioptoma floridana), and a basal vertebrate of the Agnatha clade (sea lamprey, Petromyzon marinus) were obtained using the ENSEMBL and NCBI genome browsers (supplementary tables S1 and S2, Supplementary Material online). Phylogenetic analysis of these receptors revealed five main branches that represent five subfamilies (groups) of receptors: A group for CRHR1 and CRHR2; B group for CALC and CALCRL; C group for PTH1R, PTH2R, and PTH3R; D group for GCGR, GIPR, GLP1R, GLP2R, and a novel glucagon receptor-like receptor (GRLR); and E group for SCTR, VIPR1, VIPR2, GHRHR, GHRHR2, GHRHR3, and ADCYAP1R1 (fig. 1). The phylogenetic positions of the relatively newly identified receptors are as follows: PTH3R, a third PTHR form, (Bhattacharya et al. 2011) in C group; GRLR (Irwin and Prentice 2011) in D group, and two GHRHR-like receptors, GHRHR2 and GHRHR3 (Roch et al. 2009; Wang et al. 2010), in E group. As the exact ligand forms of these GHRHR-like receptors are still under investigation (Roch et al. 2009; Wang et al. 2010), we designated these receptors as GHRHR2 and GHRHR3. These new members are not present in mammals but may exist in many teleost and nonmammalian tetrapod species. Therefore, the receptors are likely to have emerged by gene/genome duplications prior to the divergence of teleosts and tetrapods. Furthermore, the teleost-specific 3R duplication likely
contributed to generation of two copies of CALCR, CALCRL, ADCYAP1R1, VIPR1, VIPR2, GCGR, and GHRHR2 (fig. 1). This possibility will be discussed further in the next section with the analysis of teleost-specific 3R paralogons. Some genes have been lost during vertebrate evolution. For instance, GHRHR2, GHRHR3, PTH3R, and GRLR are absent in the human and mouse genomes, whereas GLP1R is missing in teleosts (supplementary table S1, Supplementary Material online). The lack of some genes, such as GHRHR and GHRHR2 in Xenopus tropicalis; GHRHR3 in anoles, medaka, tetraodon, and stickleback; and GRLR in zebrafish, may be due to either actual loss of these genes or incomplete genome information for these species.

In protostomes, C. elegans has two forms of the SCT-like receptor (C18B12.2 and C18B9.4), and Drosophila contains five forms (DH44-R1, DH44-R2, Pdfr-RA, Dh31-RA, and CG4395-RA) (Hewes and Taghert 2001; Cardoso et al. 2006). The receptors in these protostomes are phylogenetically close to the vertebrate receptors that belong to the A or B group (fig. 1). The early deuterostome amphioxus contains nine receptors that belong to either the A or B group and six receptors that belong to the C group. Ciona has more diversified SCT-like receptor members with receptors that belong to the C group (F6V0N3) and D group (F7AR10 and F7AQZ5) as well as members that belong to the A group (F6X8W4) and B group (D2KVR6). The basal lamprey possesses receptors from all five groups of the SCT-like receptor family: lamCRHR-like1 (lamCRHRL1), lamCRHRL2, and lamCRHRL3 in A group; lamCALCR in B group; lamPTH1R and lamPTH2R in C group; a GIPR-like receptor and lamGCGR in D group; and an SCTR-like receptor in E group (fig. 1 and supplementary table S2, Supplementary Material online).

According to this phylogenetic analysis, receptors belonging to the A and B groups are likely the most ancient forms of the SCT-like receptor family, as CRHR/CALCR-like receptors...
there are present in both *C. elegans* and *Drosophila*. The appearance of members of C group and D group in *Ciona* suggests that these gene subfamilies have emerged by gene duplication of an ancient CRHR/CALCR-like receptor during the emergence of early chordates. Of these receptor subfamilies, the receptors in the E group likely appeared most recently because this group contains no members in species more basal than the lamprey.

**Phylogenetic Analyses of SCT Peptide Families**

The amino acid sequences of mature SCT-like family peptides in vertebrates, lamprey, and *Drosophila* were obtained using the ENSEMBL and NCBI genome browsers; however, we were unable to retrieve the SCT-like family peptide sequences from *C. elegans*, amphioxus, and *Ciona* (Cardoso et al. 2010). The phylogeny relationship of the peptides is slightly different than that of the receptors. For instance, while the C group receptors are more closely related to the D and E group receptors than the A and B group receptors, the C group peptide branch (PTH, PTHLH, and PTH2) is more closely related to the A group (CRH, UCN, UCN2, and UCN3) and B group (CALC, CGRP, IAPP, ADM, and ADM2) branches than to the D group (Gluc/GLP1/GLP2, GlP, and a novel glucagon-related peptide, GCRP) and E group (SCT, VIP/ VRP, PACAP/PRP, and GHRH) branches (fig. 2). The genes encoding these peptides are likely to have emerged before the divergence of teleosts and tetrapods, as the peptides are present in most tetrapods and teleosts. As CALCβ, which encodes CGRP-II, is found only in mammals and is located very close to CALCA (supplementary table S1, Supplementary Material online), CALCβ may have been generated by local gene duplication during the emergence of mammals; however, the second form of CGRP that is found in teleosts appears to have been generated by teleost-specific genome duplication, probably 3R. Teleost-specific genome duplication may also have generated two forms of GCG, ADCYAP1, CRH, ADM2, PTH, and PTHLH in fish.

During our search for all peptide forms, we identified a novel form of GCRP, and this peptide was first reported to be similar to exendin, which was originally discovered in Gila monsters (*Heloderma suspectum*) (Göke et al. 1993; Irwin and Prentice 2011). Our phylogenetic analysis, however, revealed that GCRP can be categorized as a separate member distinct from exendin (data not shown). Although the full-length cDNAs for GCRP are found in a few species, such as *Xenopus* and chicken (Irwin and Prentice 2011), sequences of the mature peptide are found in genomes of both teleosts (medaka, fugu, stickleback, and tetraodon) and a tetrapod (anole) (fig. 2 and supplementary table S1, Supplementary Material online).

To date, the SCT genes have been identified only in mammals and avian species (Cardoso et al. 2010; Wang et al. 2012). Our blast search results, however, indicate the presence of the SCT-like gene in *Xenopus* (scaffold GL173223) as well. Genome syntenic analysis of GL173223 and SCT-harboring human Chromosome 11 clearly shows that these genome fragments share several neighboring genes, including RIC8A, ASCL2, IGF2/3, and PHLDA2 (data not shown), indicating that this gene is likely the *Xenopus* ortholog of the human SCT. The absence of SCT in anoles is likely due to either incomplete genome assembly or the loss of this gene in this species. Interestingly, two avian species, chicken and zebra finch, contain the SCT gene with two mature SCT peptides in different exons (data not shown), suggesting exon duplication during the emergence of birds.

For lamprey, six genes were retrieved using the ENSEMBL and NCBI genome browsers. These genes include lamGCG1 (for Gluc + GLP1 + GLP2), lamGCG2 (Gluc + GLP2), lamVIP (VIP + VRP), lamADPCYAP1-like (for PACAP-like + GHRH-like), lamGHRH-like (for GHRH-like), and lamADM (supplementary table S2, Supplementary Material online). Interestingly, lamprey GCG, VIP, and ADPCYAP1-like encode two or three mature peptides similar to those of vertebrates, albeit with some differences. For instance, lamGCG2 encodes Gluc and GLP2, while fish GCG2 encodes Gluc and GLP1. LamADPCYAP1-like encodes PACAP-like and GHRH-like peptide sequences. In lamprey, no CRH-like and PTH-like peptide sequences have been identified yet, although lamprey has putative CRHR-like and PTHR-like genes. For *Drosophila*, the six peptides analyzed in this study likely belong to A or B group peptides.

**Clustered Location of SCT-Like Peptide and Receptor Genes in Reconstructed Gnathostome Ancestor Chromosomes**

To determine the evolutionary origin of the genes and the relationship among them, we analyzed the syntenic genomes that contain SCT-like peptide and receptor genes in a variety of vertebrate species. Because of the difficulty in comparing nonvertebrate and vertebrate genomes due with either their marked divergence or the short contiguous genome sequences of *Ciona* and lamprey, we focused only on vertebrate genomes. In particular, we analyzed four tetrapod representatives (human, anole, chicken, and *Xenopus*) as well as four teleost representatives (zebrafish, medaka, tetraodon, and stickleback). In addition, according to the method proposed by Yegorov and Good (2012), we matched human, chicken, and medaka chromosomal segments onto the linkage groups reconstructed by Nakatani et al. (2007) (herein referred to as the N-model) and traced their origins from the gnathostome ancestral chromosome (GAC), which contains the linkage groups of the hypothetical post-2R ancestor of jawed vertebrates (Nakatani et al. 2007). In the N-model, 40 reconstructed GACs (A0–J1) are derived from 10 to 13 vertebrate ancestral chromosomes (VAC), denoted the A–J linkage groups, present in the hypothetical pre-2R vertebrate genome (Nakatani et al. 2007). We also employed the ancestral genome reconstruction of the chordate ancestor, as described by Putnam et al. (2008) as needed (denoted the P model) to study the ancestor of amphioxus.

Our synteny analysis indicates that most SCT-like receptor and peptide genes are contained in genome fragments bearing the HOX, TEA domain (TEAD) family, and opioid receptor (OPR) gene families. Tracing these genome fragments onto
the ancestral vertebrate genome following the N-model indicates that these genomic fragments reside on GACs E0–E3, D0–D3, and B0–B3, respectively (fig. 3).

**HOX-Containing Gene Blocks Contain Most SCT-Like Receptor (Except for D Group Receptor) and D Group Peptide Genes**

The genome fragments of these blocks show conserved synteny across vertebrate species, and paralogs of many gene families, including complement component 1q subcomponent-like (C1QL), microtubule-associated protein (MAP), epidermal growth factor receptor, phosphodiesterase 1, and insulin-like growth factor binding protein, were parallely aligned in each block, supporting the 2R hypothesis for these blocks. These gene blocks mapped to GAC E0, E1, E2, and E3 according to the N-model (16d, 16c, 16b, and 16e according to the P model; fig. 3).

**A Group Receptors (CRHR Subfamily)**

CRHR1 and CRHR2 are distributed on GAC E2 and E1, which contain HOXB and HOXA, respectively. The absence of CRHR forms on GAC E0 and E3 suggests that these CRHR forms were likely lost prior to the divergence of teleosts and land vertebrates.

**B Group Receptors (CALCR Subfamily)**

CALCR and CALCRL are on GAC E1 and E0, respectively. Teleost-specific 3R duplication may have generated
Fig. 3. Putative early paralogs containing SCT-like family peptide and receptor genes with their neighboring genes. For humans, chromosome numbers are on the top of the indicated gene, and the gene locations (megabase) on the chromosome are shown beneath the gene. Paralogs (or orthologs) of each gene on different paralogs were aligned on the same column with the same colors. For chicken and medaka, chromosome numbers of orthologs for the indicated human gene are seen in horizontal bars. For medaka, some genes were duplicated due to the teleost-specific 3R. The positions of each gene block in the GAC linkage groups of the N models (and chordate linkage groups of the P model) are indicated beneath the scaffolds. For humans, chromosome location is either on duplicated medaka chromosomes 17/20 or on medaka chromosome 11, whereas the other medaka chromosomes 2 and 21. Likewise, two copies of CALCRL according to the N-model, are on duplicated medaka chromosomes 17/20 or on medaka chromosome 11, whereas the other CALCRL is on chromosome 7. Thus, the presence of CALCRL on medaka chromosome 7 is likely due to translocation of the gene or to a chromosomal rearrangement. The third and fourth forms of CALCRL appear to have disappeared after 2R duplication.

duplicates of CALCR and CALCRL (fig. 3, supplementary table S1 and fig. S1A, Supplementary Material online). For instance, according to the N-model, CALCRL-containing GAC E0 matched well with regions within 3R-specific medaka chromosomes 2 and 21. Likewise, two copies of CALCRL reside on medaka chromosomes 2 and 21. CALCRL-containing GAC E0 is either on duplicated medaka chromosomes 17/20 or chromosomes 11/16; however, one CALCRL is on medaka chromosome 11, whereas the other CALCRL is on chromosome 7. Thus, the presence of CALCRL on medaka chromosome 7 is likely due to translocation of the gene or to a chromosomal rearrangement. The third and fourth forms of CALCRL appear to have disappeared after 2R duplication.

C Group Receptors (PTHR Subfamily)

PTHR1, PTHR2, and PTHR3 are located on GAC E1, E0, and E2, respectively, although PTHR3 is absent in humans. In zebrafish, PTHR2 may have been doubled probably by teleost-specific 3R duplication. The fourth form of PTHR has been lost in vertebrates.

E Group Receptors (GHRHR Subfamily)

Many members (VIPR2, VIPR1, GHRHR1, GHRHR2, and ADcyAP1R1) of this subfamily reside on GAC E1, whereas SCTR and GHRHR3 are on GAC E0 and E2, respectively. The presence of multiple members in post-2R GAC E1 suggests that these receptors are likely to have emerged through local tandem duplications after 2R duplication. Two copies of VIPR2, VIPR1, GHRHR2, and ADcyAP1R1 in fish have been generated by teleost-specific 3R duplication as revealed by synteny between two duplicated teleost genome fragments. The fourth form of the GHRHR subfamily is absent in GAC E3. Therefore, GAC E3 lacks all members of SCT-like receptor genes in tetrapods and teleosts.
D Group Peptides (GCG Subfamily)

GCG, GIP, and GCRP are localized on GAC E0, E2, and E3, respectively. In teleosts, GCG has been doubled through 3R duplication. The fourth form of the GCG subfamily in GAC E1 was lost before the divergence of teleosts and land vertebrates (supplementary fig. S1A, Supplementary Material online).

TEAD-Containing Gene Blocks Have A Group (UNC2 and UCN3), B Group, C Group, and D Group (SCT) Peptide Genes

In addition to SCT-like peptide genes, these gene blocks share family members of various genes, including tubby-like protein, interferon regulatory factor, and ras oncogene family genes. In the N-model, the chromosome segments harboring these genes are in GAC D0, D1, D2, and D3; however, in the P model, the positions of these chromosome segments are not obvious, and they may be in either linkage group 14 or 11 (fig. 3). This discrepancy is likely due to the different model organisms used for their analyses (Nakatani et al. 2007; Putnam et al. 2008).

A Group Peptides (CRH Subfamily)

In this family of genes, only UCN2, and UCN3 are assigned to GAC D1 and D0, respectively, while CRH and UCN are on GAC B0 and B2, respectively (fig. 3 and supplementary figs. S1B and C, Supplementary Material online). Interestingly, although these genes are on different GAC Ds and Bs, all members of this subfamily have a GATA-binding protein (GATA) family gene as a neighbor, showing some relationship of pre-2R duplications from a common ancestral gene. However, this issue is needed to be further investigated. For example, PTH local duplications are complex, as indicated by the presence of two copies of PTH genes within the same linkage group (15d). Indeed, gene location studies reveal that GCGR, GLP2R, and GRLR are likely aligned on the same chromosomal fragments of many vertebrates (fig. 3 and supplementary fig. S1C, Supplementary Material online). Thus, these genes are likely to have emerged through local duplication after 2R. Interestingly, the P model assigns the GIPR-containing fragment to a linkage group 15a. Indeed, synteny between 15a and 15d was observed (fig. 3).

In addition, GLP1R-containing chromosome fragment shares common neighbors, such as PCNK, DHX, and/or FAM83 family members with GIPR-containing fragment. Thus, paralogs (or ohnologs) for GIPR neighbors split into I1, UN, and B2 linkage groups (fig. 3), raising a possibility that GCGR, GLP2R, GCGR, and possibly GLP1R have been generated through local duplications from a common ancestral gene. However, this issue is needed to be further investigated.

E Group Peptides (GHRH Subfamily)

VIP, GHRH, and ADCYAP1 are distributed on GAC B2, B1, and B0, respectively, indicating that these genes are ohnologs that are duplicated by 2R. Unexpectedly, SCT is in GAC D1. Thus, SCT may have been translocated to GAC D1 after 2R duplication. Alternatively, the SCT ancestor that is distinct from that of VIP, GHRH, and ADCYAP1 may have already been present in pre-2R VAC D. In this case, the ohnologs of SCT were lost after 2R duplication. SCT has been lost in teleosts following the divergence of fish and land vertebrates. The presence of two copies of ADCYAP1 is due to teleost-specific 3R duplication (supplementary fig. S1C, Supplementary Material online).

D Group Receptors (GCGR Subfamily)

This gene subfamily comprises GCGR, GIPR, GLP2R, GRLR, and GLP1R. Unexpectedly, members in this subfamily are on different GAC blocks. For instance, GLP1R is on B2, GIPR is on G2, GCGR is on I1, and GRLR and GLP2R are on UN. Thus, speculating the mechanism of emergence of these family members is not simple; however, according to the P model, GAC G2 and I1 are in the same linkage group (15d). Indeed, gene location studies reveal that GCGR, GLP2R, and GRLR are likely aligned on the same chromosomal fragments of many vertebrates (fig. 3 and supplementary fig. S1C, Supplementary Material online). Thus, these genes are likely to have emerged through local duplication after 2R. Interestingly, the P model assigns the GIPR-containing fragment to a linkage group 15a. Indeed, synteny between 15a and 15d was observed (fig. 3).

In addition, GLP1R-containing chromosome fragment shares common neighbors, such as PCNK, DHX, and/or FAM83 family members with GIPR-containing fragment. Thus, paralogs (or ohnologs) for GIPR neighbors split into I1, UN, and B2 linkage groups (fig. 3), raising a possibility that GCGR, GLP2R, GCGR, and possibly GLP1R have been generated through local duplications from a common ancestral gene. However, this issue is needed to be further investigated.

Proposed Evolutionary History of the SCT-Like Receptor and Peptide Families

Our genome comparison analyses revealed that most SCT-like receptor and peptide genes were aligned within post-2R GAC Bs, Ds, and Es. The ancestral forms of these receptor and peptide families are likely CRHR/CALCR and their peptide...
ligands, as these family members are found in both protostome lineages (C. elegans and Drosophila) and deuterostome lineages (figs. 1 and 2). Thus, these genes were present before the divergence of protostomes and deuterostomes (fig. 4). The presence of the majority of the receptors (A, B, C, and E group receptors) in GAC Bs suggests that local duplications of ancestral receptor genes occurred within pre-2R VAC E. Although the D group receptors occur in GAC B2 or other GACs (G2, I1, and UN), the chromosome fragments bearing D group receptors share some common neighbors that are also found in GAC Bs, suggesting a possible fusion of these other GACs (G2, I1, and UN) with GAC Bs. Thus, the D group receptor ancestor appeared to have moved from VAC E to VAC B through chromosome rearrangement or translocation (fig. 4). Because the majority of the peptide genes (A, B, and C group peptides) are found in GAC Ds, the local duplication of the ancestral peptide genes is likely to have occurred within VAC D. In addition, local duplications of A and B group peptide genes further produced second forms of ancestral genes (A2 and B2 group peptides). Then, some genes moved to VAC B (A2 and E group peptides) or VAC E (D group peptides) through chromosome rearrangements or gene translocation. These rearranged VACs have undergone two rounds of whole-genome duplication to generate four copies of corresponding GACs. After 2R duplication and before the divergence of teleosts and tetrapods, local gene duplications or gene loss occurred. In a reconstructed osteichthyan ancestor chromosome in the N-model, GAC E1–B0 and GAC E0–B2 are linked on the same chromosomes (Nakatani et al. 2007). Thus, in the proposed model, these possible links are indicated by broken lines between GAC E1 and B0 and between GAC E0 and B2 (fig. 4). Of the E group peptides, only SCT is found in GAC D1; and thus, the evolutionary history of SCT is unclear. SCT was likely translocated from one of the GAC Bs to GAC D1. Another possibility is that the SCT ohnologs produced by 2R duplication disappeared before the divergence of teleosts and tetrapods (fig. 4). Likewise, GIPR on GAC G2/15a and an ancestor of GRLR, GLP2R, and GCGR on GAC I1/UN/

---

**Fig. 4.** Proposed evolutionary history of SCT-like family peptide and receptor genes. The ancestral genes of the SCT-like family peptides and receptors were tandem duplicated before the divergence of protostomes and deuterostomes. These genes were further doubled locally within the chromosomes to generate the ancestral forms of receptor genes that belong to the A to E groups of receptor and peptide genes during early (or pre-) vertebrate evolution. Then, chromosome rearrangements or translocations of the ancestral genes occurred to generate pre-2R VACs. Two rounds of whole-genome duplication produced four copies of post-2R GACs. Before divergence of fish and land vertebrates, additional tandem local duplications and gene losses occurred. R–A to R–E, and P–A to P–E indicates ancestral forms of receptor and peptide groups, respectively. Local duplications are indicated by “x.”
of peptide genes in five E group receptors in 15d seem to have translocated from GAC Bs. The presence of five E group receptors in GAC E1 suggests local duplications of these members. The three D group receptors in GAC (II/UN/15d) are likely to have occurred through local duplications. The absence of any receptor genes in GAC E3 and the absence of peptide genes in GAC D2 indicate loss of these genes after 2R duplication.

**Discussion**

The SCT-like GPCR and their peptide ligand families exhibit a variety of functions in the brain, gut, and endocrine tissues. The SCT-like GPCRs emerged prior to the divergence of protostomes and deuterostomes (Cardoso et al. 2006). In particular, CRHR-like (A group) and CALCR-like (B group) receptors are likely the earliest members of this family, and these receptors are present in both protostomes and early deuterostomes (Cardoso et al. 2006). Both a recent study and our phylogenetic study revealed the presence of A, B, and C group receptor members and the absence of D and E group subfamilies in the chordate representative amphioxus (Nordstrom et al. 2008). GCGR-like (D group) subfamily members emerged in the early deuterostome Ciona. E group members appeared in the early vertebrate lamprey. Thus, E group members are likely the most recent member of the SCT-like receptor family.

Unlike the receptor genes, the SCT-like peptide family genes were not easy to retrieve from nonvertebrate species, probably due to the considerable variation in amino acid sequences of the nonvertebrate peptides. Indeed, the Drosophila peptides, DH44, DH31, PDF, and amnesiac, exhibit a high degree of amino acid sequence variation compared with the vertebrate SCT-like peptide family, although these peptides interact with Drosophila SCT-like GPCRs DH44R, DH31R, and PDFR (Johnson et al. 2005; Lear et al. 2005; Mertens et al. 2005). No sequence homologs for any of these peptides nor the members of vertebrate SCT-like peptide family were found in C. elegans or Ciona (Cardoso et al. 2010).

The SCT-like peptide family is likely to have evolved to preserve their three-dimensional structures rather than to maintain their amino acid sequences. Indeed, vertebrate peptide members share a common structural feature, an N-terminal random coiled structure followed by α-helical regions in the middle and C-terminal portions that correspond to the typical structure of the SCT-like receptors (Jin et al. 2000; Marx et al. 2000; Neidigh et al. 2001; Parthier et al. 2007; Underwood et al. 2010). The α-helical region binds the long N-terminal ECD of the SCT-like receptors, whereas the N-terminal region interacts with the helical core domain of the receptor, and this binding is crucial for receptor activation, supporting the two domain hypothesis for SCT-like receptor ligand binding and activation (Pioszak et al. 2008; Pioszak and Xu 2008; Stroop et al. 1995; Moon et al. 2010, 2012). Recently, the helix N-capping motif in the N-terminus of these peptides was proposed to be critical for ligand-induced receptor activation (Neumann et al. 2008). This motif is common to all SCT-like peptide family members. For instance, the N-capping motif is formed by hydrophobic amino acids at positions 6 and 10 and a short-chain polar amino acid at position 7 in the SCT, Gluc, and CRH subfamilies, whereas there are additionally hydrophobic amino acids at position 2, 3, 6, and 7 in the PTH subfamily and two conserved Cys residues in the CALC subfamily (Neumann et al. 2008). Thus, the functional structure of the peptides has been maintained, whereas the amino acid sequences have varied during evolution of the SCT-like peptide subfamily. This variation within a short sequence is likely the reason for our inability to retrieve the sequences of SCT-like peptide family members from nonvertebrate genomes. Indeed, despite the absence of their sequence similarity, the arthropods Drosophila and sand fly contain the genes amnesiac and maxadilan, respectively, which are thought or known to be homologs of vertebrate PACAP (Hashimoto et al. 2002). This is strengthened by evidence that the maxadilan peptide activates the mammalian ADCYAP1R1 receptor in vitro (Tatsuno et al. 2001), indicating that the ADCYAP1R1 likely recognizes the conserved structure of the maxadilan peptide rather than the amino acid sequence.

Sequence-based phylogenetic analysis of related genes often provides incorrect information despite high bootstrap values (Abi-Rached et al. 2002; Larhammar et al. 2002). Thus, synteny-based analysis often offers more accurate information about the origin and relationship of related genes (Kim et al. 2011). For instance, fugu and medaka GIPRs that were defined by phylogenetic analysis (Roch et al. 2009) were found to be GRLR orthologs in our synteny-based study. Indeed, no GIPR orthologs were identified in these fish species. The zebrafish GHRHR that was described by Roch et al. (2009) is likely a GHRHR3 ortholog, according to synteny of the genome fragment that harbors these genes. As no genetic GLP1R ortholog has been identified in fish, the fish GLP1Rs that mentioned by Roch et al. (2009) are likely a copy of duplicated GCGR that was generated by teleost-specific 3R duplication. Because some GCGRs in fish can respond to the GLP1 peptide, these receptors are likely to be described as GLP1R (Roch et al. 2009). Thus, the confusing nomenclature is likely due to a high degree of sequence identity among GCGR and GHRHR subfamily members. Our chromosome analyses revealed that GRLR, GLP2R, and GCGR are likely to have been on one chromosome. Likewise, VIPR1, VIPR2, ADCYAP1R1, GHRHR, and GHRHR2 are also likely to have been on the same chromosome. These observations indicate that the above-mentioned genes have emerged relatively recently via local tandem duplication just prior to the divergence of teleosts and tetrapods. Thus, these genes share a very high degree of sequence identity.

Delineating the evolutionary mechanism for a gene family with a large number of paralogs is particularly difficult due to scattered distribution of the genes on many different chromosomes. Small scale synteny analysis is useful for understanding the paralogous relationships of gene, but using large scale synteny comparisons that reconstruct ancestral genomes, as shown by Yegorov and Good (2012), provides additional information about the origin and timing of gene duplication events. One of the highlights of our study is the identification of local duplications of SCT-like peptide and
receptor ancestor genes in pre-2R VACs. Five receptor groups (with 17 paralogs) and five peptide groups (with 18 paralogs) are clustered on only three VACs (E, D, and B). In addition, the colocalization of GCG peptide subfamily genes with receptor genes in GAC Es, the split of A and D group peptide genes into GAC Ds and Bs, and the presence of a D group receptor in GAC B2 suggest possible connections among VAC E, D, and B before a split or rearrangement of the VAC. Indeed, the reconstructed osteichthyian chromosome (post-2R stage) reveals connections between GAC E1 and B0 and between GAC E0 and B2 (Nakatani et al. 2007). Thus, we postulate that five sets of the ancestral forms of each receptor and peptide group may have arisen by multiple local gene duplications within three related pre-2R VACs during early vertebrate evolution. Local duplications after 2R duplication are also evident for D and E group receptors. Thus, our study clearly shows contribution of local duplications before and after 2R duplication to expand the SCT-like peptide and receptor gene families during early vertebrate evolution.

In summary, analysis of vertebrate genome fragments with conserved synteny followed by matching these fragments with reconstructed early vertebrate chromosomes proved to be a useful tool to explore the evolutionary mechanism for expansion of a gene family with a large number of paralogs. This study supports a provisional model for the evolution of SCT-like family peptides and receptors during metazoan evolution. However, some difficulties were encountered in resolving the origin of some genes using ancestral genome reconstructions in cases when 1) the gene has an independent origin from its expected ohnologs or 2) the gene underwent a single gene translocation that caused it to move from its authentic chromosomal fragments after 2R duplication (Yegorov and Good 2012). In this case, SCT is likely either of 1 or 2. Further, some ancestral linkage groups can be inaccurately reconstructed. For instance, although the D group receptor-containing chromosome fragments revealed synteny to some extent, these were assigned to different linkage groups. In the future, complete assembly of genomes of nonvertebrate and early vertebrates, the identification of nonvertebrate SCT-like peptides, and higher accuracy of reconstructed ancestral chromosomes will provide a better opportunity to further explore the evolutionary scenario of these ligand–receptor pairs.

Materials and Methods

Data Retrieval

The amino acid sequences of SCT-like family peptides and receptors were downloaded from either the Ensembl Genome Browser (http://www.ensembl.org) or the GenBank database using the Entrez data retrieval tool (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Specifically, search tools for orthologous or paralogous genes provided by Ensembl Genome Browser were mainly used for identification of genes from nonmammalian vertebrate and invertebrate species. This data set was considered the start set. To identify the members of SCT-like family peptides and receptors in vertebrates and invertebrates, genes that were not already part of the data set were manually searched against the genome database of human, mouse, chicken, anole lizard, Xenopus tropicalis, zebrafish, medaka, fugu, stickleback, tetraodon, lamprey, amphioxus, Ciona, C. elegans, and Drosophila with the TBLASTN algorithm. To determine the full-length open-reading frame (ORF) sequence, the exons and splice junctions were defined using both the Ensembl protein report service and an HMMgene (v.1.1) program (http://www.cbs.dtu.dk/services/HMMgene/) that was provided by the CBS Prediction Service. Putative signal peptides were predicted using SignalP v3.0 (http://www.cbs.dtu.dk/services/SignalP/). Duplicates were removed from this data set using a crude phylogenetic analysis.

Phylogenetic Analysis of SCT Family Peptides and Receptors

The full set of SCT-like receptor sequences were aligned using MUSCLE (Edgar 2004) as implemented in MEGA v5.05. The default alignment parameters were applied. Alignments were bootstrapped 1,000 times, and a maximum likelihood phylogenetic tree was constructed using the Jones–Taylor–Thornton model. Trees were unrooted and plotted using MEGA v5.05. The full set of SCT-like peptide sequences were aligned using the Windows version of ClustalX-2.1 (Gonnet matrix, gap opening penalty 10, and gap extension 0.2). Alignments were bootstrapped 1,000 times, and neighbor-joining trees were obtained using MEGA v5.05. As the sequences of the peptides are relatively short in length and variable across the subfamilies, we were unable to generate a good tree using the maximum likelihood method, although the overall pattern of the tree generated by the maximum likelihood method was similar to that generated using the neighbor-joining tree method. Because of the complexity of figures 1 and 2, the bootstrap values were removed from the figures.

Genome Synteny Analysis and Tracing of the Duplication History of SCT-Like Receptor and Peptide Families

The genome synteny analysis was performed by comparing the Contig Views of genome regions containing the SCT-like family peptide and receptor loci. The information for chromosome localization of orthologs/paralogs of neighboring genes was obtained from the Ensembl Genome Browser. According to the method of Yegorov and Good (2012), chromosome fragments with reliable synteny were matched with the reconstructed protochromosome models by Nakatani et al. (2007) and Putnam et al. (2008). The N-model provides conserved vertebrate linkage blocks that are displayed along individual human chromosomes, which also indicate the locations of human ohnologs and orthologs of medaka, chicken, and mouse (Nakatani et al. 2007). Thus, we were able to compare the chromosome fragments with linkage groups shown for each chromosome of the three taxa (human, chicken, and medaka) to resolve the positions of the gene blocks at consecutive stages of vertebrate genome evolution (e.g., pre-2R VAC and post-2R GAC). The P-model
also provides reconstructed ancestral chordate linkage groups on human chromosomes (Putnam et al. 2008). As a result, we were able to define the positions of gene blocks with conserved synteny on reconstructed protochromosomes.

**Supplementary Material**

Supplementary tables S1 and S2 and figure S1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

This work was supported by the Brain Research Program of the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science, and Technology (2011-0019205). The authors appreciate Dr Larhammar at Uppsala University, Sweden, for his critical review of the manuscript. They also appreciate Drs Yegorov and Good at the University of Winnipeg, Canada, for their kind comments and suggestions during the revision of the manuscript.

**References**


