The gonadal primordium first appears as a thickening of the coelomic epithelium at the lateral sides of the mesentery. Histological observations suggested that these epithelial cells migrate mediodorsally to form the primitive urogenital ridge (1–3), which subsequently separates into the future gonad and mesonephros. Sexually dimorphic events occur thereafter in the genital ridge in response to transient expression of sex-determining region on the Y chromosome (*Sry*). A number of studies has attempted to elucidate the mechanisms underlying gonad development and sex differentiation. Disruption of some genes in mice and humans leads to hypoplastic gonads or complete dysgenesis, whereas disruption of other genes leads to sex reversal (4–6). Nevertheless, the early phase of gonad development is still poorly understood.

The paired-like homeobox gene *Emx2* is a mouse homologue of the *Drosophila empty spiracles* gene (7). Mouse knockout (KO) studies demonstrated that *Emx2* is implicated in the development of multiple tissues, including the kidney, gonad, reproductive tracts, and central nervous system. A detailed study of the kidney defect showed that ureteric bud extension is disrupted in the metanephric mesenchyme of *Emx2* KO mice, leading to the disappearance of the kidney (8). In addition, proliferation of neuronal progenitors and area specification of the neocortex are disrupted in *Emx2* KO animals (9–13). Some of the observed abnormalities in these animals appeared to be closely related to defects in the spatial expression of growth factors (14–17). Although these studies revealed important insights into how *Emx2* functions during kidney and central nervous system development, its role in gonad development is still unclear.

Epithelial cells are polarized by establishing functionally specialized apical, lateral, and basal surfaces and adhere tightly one another through tight and adherens junctions at their lateral interfaces. These junctions are comprised of transmembrane proteins, such as junctional adhesion molecules, cadherins, occludin, and claudin, and intracellular membrane-associated proteins, such as β-catenin and zonula occludens (ZO) (18, 19). Mammalian homologues of the *Caenorhabditis elegans* partitioning-defective proteins (PARs) PAR3 and PAR6 position tight junctions by forming a complex with atypical protein kinase C (aPKC). Multiple extracellular signals regulate the formation and localization of the PAR3/PAR6/aPKC complex via phosphorylation of the components (20, 21).

ErbB (erythroblastic leukemia viral oncogene homolog) receptors comprise a family of four structurally related tyrosine kinase receptors (22) that are activated by a variety of ligand molecules, including epidermal growth factor (EGF). Upon ligand binding, the receptor is dimerized and the kinase activity triggers numerous downstream signaling pathways (23). One family member, EGF receptor (EGFR), is expressed in the epithelial cells of a variety of tissues, where it plays fundamental roles in tissue development through regulating cell proliferation, cellular polarity formation, and epithelial cell migration (24). A number of studies revealed the presence of phospholipase C-γ, PKC-mediated cascades, mitogen-activated protein cascades, and small GTPase downstream of EGFR (23). Importantly, a recent study demonstrated that EGFR is implicated in the regulation of tight junction assembly via tyrosine phosphorylation of sarcoma viral oncogene homolog (c-Src)/Yamaguchi sarcoma viral oncogene homolog 1 (c-Yes), which subsequently phosphorylates PAR3 to regulate PAR3/PAR6/aPKC complex formation (25).

Here, we examine early stages of gonad development in *Emx2* KO embryos and show that tight junction assembly and migration of the epithelial cells of the gonad are significantly affected. Interestingly, microarray analysis of the epithelial cells of the embryonic gonad indicates that *Egfr* is dramatically up-regulated in *Emx2* KO mice. This ectopic *Egfr* expression is accompanied by aberrant c-Src tyrosine phosphorylation. Our data strongly suggested that *Emx2* is required for tight junction assembly and migration of epithelial cells at the early stage of gonadal development possibly through suppression of *Egfr* expression.

Materials and Methods

Experimental animals

Emx2 KO mice (accession no. CDB0018K; http://www.cdb.riken.jp/arg/mutant%20mice%20list.html) (8) were crossed to B6/J Jcl mice (Clea, Tokyo, Japan) for five generations. Genotypes were determined by PCR using the primers, empty spiracles homeobox 2 (Emx2)-sense (S) (5′-CCACCTTAGAGACCATTGCT-3′), Emx2-antisense (AS) (5′-TTCTCAAAAGCGTGCTCTAG-3′), and phosphoglycerate kinase-AS (5′-GCTACCGGTGGATGTGGAATG-3′). A wild-type allele is amplified with Emx2-S and Emx2-AS and a KO allele with Emx2-S and phosphoglycerate kinase-AS. The sex of the mice was determined by PCR with primers for *Sry*, Sry-M5 (5′-GTGGTGAGAGGCACAAGTTGGC-3′) and Sry-M3 (5′-CTGTGTAGGATCTTCAATCTCT-3′). Embryos were dissected between embryonic d (E)10.0 and E12.5. To stage the embryos accurately, tail somites (tss) were counted; ts stages are indicated in the figure legends. All protocols for animal experiments were approved by the Institutional Animal Care and Use Committee of the National Institute for Basic Biology.

Preparation of antibody for EMX2

Full-length mouse *Emx2* cDNA was amplified by PCR with primers, 5′-ACACACCTCGAGATGTTTCAGCCGGCGCCCAAG-3′ and 5′-ACACACACGCGGCCGCGCCTTAATCGTCTGAGGTCACATC-3′, and then cloned into pET-28a (Strat-

agene, La Jolla, CA) to produce His-tagged EMX2 recombinant protein. The recombinant EMX2 was purified with Ni-agarose (Invitrogen, Carlsbad, CA). Rabbits were immunized with the purified His-tagged EMX2 as described previously (26).

Scanning electron microscopy (SEM), histology, immunohistochemistry, and *in situ* hybridization

SEM of E11.5 embryos was performed using a Hitachi S-800 (Hitachi, Tokyo, Japan), as previously described (27). Histochemical and immunohistochemical analyses were performed as previously described (28). Rabbit antibodies to EMX2, Ad4BP/SF-1 [*Adrenal-4 Binding Protein* (29), *Steroidogenic Factor-1* (30), and *NR5A1* (31)] (26), aristaless-related homeobox (32), aPKC (PKCζ) (Santa Cruz Biotechnology, Inc., Santa Cruz, CA), suppressor gene for Wilms' tumor (WT1) (Santa Cruz Biotechnology, Inc.) (33, 34), laminin (Sigma, St Louis, MO), ZO-1 (Zymed, South San Francisco, CA), occludin (Zymed), Src phosphrylated at tyrosine 418 (Biosource, Camarillo, CA), EGFR phosphorylated at tyrosine 845 (Abcam, Cambridge, MA), and EGFR phosphrylated at tyrosine 1068 (Abcam), sheep antibody to EGFR (Upstate, Charlottesville, VA), goat antibody to globin transcription factor (GATA) binding protein-4 (GATA4) (Santa Cruz Biotechnology, Inc.) (35), mouse antibodies to bromodeoxyuridine (BrdU) (Roche, Indianapolis, IN) and hemaglutinin (HA) (Sigma), and guinea pig antibody to CBX2/M33 (Polycomb component) (36) were used. Biotinylated antirabbit, antigoat, antisheep, and antiguinea pig antibodies (Jackson ImmunoResearch, West Grove, PA), Alexa Fluor 488-labeled antirabbit and antigoat (Molecular Probes, Eugene, OR) antibodies, and Cy3-labeled and Cy5-labeled antimouse antibodies (Jackson ImmunoResearch) were used as secondary antibodies. Antigen-antibody complexes were detected using Histofine kit (Nichirei, Tokyo, Japan) or directly by fluorescence. *In situ* hybridization for *LIM homeobox gene 9* (*Lhx9*) (37) was performed as previously described (38).

Cell proliferation and apoptosis assays

Pregnant females received an ip injection of BrdU (Sigma) (50 mg/kg body weight) at E10.0, E10.5, and E11.0 (39) and were killed 2 h after injection. Paraffin sections of the embryos were double immunostained for BrdU and GATA4. In brief, after the sections were boiled in 10 mM citrate (pH 6.0) for 20 min (40), they were incubated with the mouse anti-BrdU antibody, and thereafter with the Cy3-labeled antimouse antibody. Subsequently, the sections were incubated with goat anti-GATA4 antibody and with Alexa Fluor 488-labeled antigoat antibody. Nuclei in the sections were stained with propidium iodide (PI) (Molecular Probes). The number of BrdU-immunoreactive gonadal epithelial cells was counted in more than 10 sections for every gonad. Apoptosis in E11.0 and E12.0 embryonic gonads was assayed using the ApopTag Plus Peroxidase kit (CHEMICON, Temecula, CA). After apoptotic cells were detected with rhodamine-labeled anti-digoxigenin antibody, the sections were stained by goat anti-GATA4 and Alexa Fluor 488-labeled antigoat antibodies.

Cell fate mapping with organ culture

After the abdominal tissues of E10.25 embryos were removed, the coelomic epithelial cells were labeled with 20 mM 5-(and-6)-carboxy-2′,7′-dichlorofuluorescein diacetate, succinimidyl ester (CCFSE) (Molecular Probes) in DMEM (Invitrogen) for 1 h. Subsequently, they were cultured for 5 or 24 h in DMEM containing 10% fetal bovine serum and antibiotics under a humidified atmosphere of 5% CO_2 in air at 37 C. After fixation in 4% paraformaldehyde for 5 min, the embryos were frozen sectioned and stained with Ad4BP/SF-1 or laminin antibody. The number of CCFSE-positive cells that migrated through the laminin layer was counted in more than 10 sections for every gonad.

Preparation of Emx2-expressing M15 cells and knockdown of Emx2 expression by short interfering RNA (siRNA)

Full-length cDNA for mouse *Emx2* was cloned into pOZ-FH retroviral vector (41). The construct, pOZ-FH-Emx2, encodes FLAG-HA-tagged EMX2 [Emx2(HA)] and IL-2 receptor with an internal ribosomal entry site. Recombinant viruses prepared with pOZ-FH-Emx2 were transfected into M15 cells derived from mesonephric epithelial cells (42). To prepare M15 cells expressing EMX2 [M15-Emx2(HA)], the infected cells were sorted by anti-IL-2 receptor monoclonal antibody (Upstate) conjugated with magnetic beads (Dynal Biotech, Oslo, Norway) (41, 43); 10^5 original M15 and M15-Emx2(HA) cells were plated on a six-well dish. After 24 h, cells were transfected with 100 pmol siRNA for *Emx2* (S, 5′-UUCGAAUCCGCUUUGGCUUUCUGGC-3′ and AS, 5′-GCCAGAAAGCCAAAGCGGAUUCGAA-3′) or control siRNA (Invitrogen) using lipofectamine 2000 (Invitrogen). The cells were collected for RT-PCR and Western blot analysis (26) after another 24-h incubation. For immunohistochemistry, the cells were grown on poly-L-lysine-coated glass (IWAKI, Tokyo, Japan), then fixed with 4% paraformaldehyde and incubated with anti-HA and anti-EGFR antibodies.

Microarray and quantitative RT-PCR analyses

Microarray analyses were performed essentially as described (44, 45). E10.5 wild-type and *Emx2* KO embryos were frozen in OCT compound (Sakura Finetechnical, Tokyo, Japan) without fixation. They were sectioned (30 μm), stained with hematoxylin, and air dried. Some were used for GATA4 immunostaining to locate gonadal primordia. The epithelial cells of the gonadal primordia were obtained using a Laser Microdissection System (Leica, Wetzlar, Germany). The specimens prepared from three individuals were combined into one group. Total RNA was prepared from three groups, and 15 ng total RNA was subjected to two-cycle amplification and biotin labeling using MessageAmp II aRNA Amplification and MessageAmp II-Biotin Enhanced kit (Ambion, Austin, TX), respectively. The labeled aRNA was fragmented and hybridized to GeneChip Mouse Genome 430 2.0 array according to the manufacturer's instructions (Affymetrix, Santa Clara, CA). Signals were scanned and scaled using Affymetrix GCOS 1.1 software. The scaled values were then analyzed by GeneSpring software (Silicon Genetics, Redwood City, CA). Pairwise comparison analysis was performed with Affymetrix GCOS 1.2 to identify differentially expressed transcripts. Each sample (n = 3) was compared with each reference samples (n = 3), resulting in nine pairwise comparisons. This approach, which is based on Mann-Whitney pairwise comparison test, allows ranking of differentially expressed genes as well as calculation of significance ($P < 0.05$) of each identified change. The microarray data have been deposited in the Gene Expression Omnibus of the National Center for Biotechnology Information (accession no. GSE10216; http://www.ncbi.nlm.nih.gov/geo/). Quantitative RT-PCR with TaqMan probes for mouse *Egfr* (Mm00433021_m1) and *Gapdh*

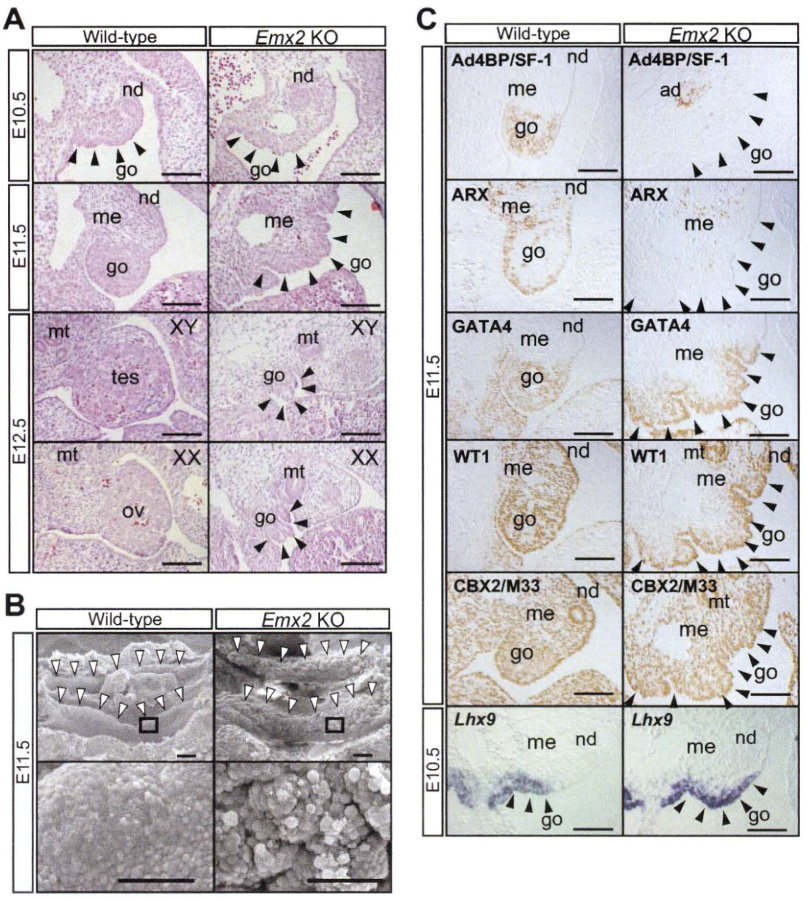

FIG. 1. Structural defects and altered marker gene expression in *Emx2 KO* gonads. A, Developing gonads of wild-type and *Emx2* KO embryos. Sections of wild-type and *Emx2* KO embryonic gonads at E10.5 (ts 8), E11.5 (ts 18–19), and E12.5 (ts 27–32) [male (XY) and female (XX)] were stained with hematoxylin and eosin. Developing and degenerating gonads are indicated by *closed arrowheads*. *Scale bars*, 100 μm. B, SEM of E11.5 (ts 19) wild-type and *Emx2* KO embryonic gonads (*open arrowheads*). The *regions enclosed by squares* are enlarged in the *lower panels*. *Scale bars*, 100 μm (*upper*) and 50 μm (*lower*). C, Gonadal marker gene expression in wild-type and *Emx2* KO gonads. The expression of gonadal markers was examined in wild-type and *Emx2* KO embryos by immunohistochemistry for Ad4BP/SF-1, ARA, GATA4, WT1, and CBX2/M33 at E11.5 (ts 18–21) and by *in situ* hybridization for *Lhx9* at E10.5 (ts 7). Gonadal regions are indicated by *closed arrowheads*. *Scale bars*, 100 μm. go, Gonad; nd, nephric duct (feature Wolffian duct); me, mesonephrose; mt, mesonephric tubule; ad, adrenal primordium; tes, testis; ov, ovary.

(TaqMan Rodent glyceraldehyde-3-phosphate dehydrogenase Control Regents; Applied Biosystems, Foster City, CA) was performed using QuantitectProbe RT-PCR kit (QIAGEN, Valencia, CA) and PM9000 cermal cycler (Applied Biosystems). Total RNAs prepared from M15 and M15-Emx2 cells were used for quantitative RT-PCR using Power SYBR Green PCR Master mix (Applied Biosystems). PCR primers for mouse *Egfr* are mEgfr-S (5′-CACAGCGCTACCTTGTTATCCA-3′) and mEgfr-AS (5′-CCTCCATGTCCTCTTCATCCA-3′). Primers for *β-actin* were as previously described (46).

Results

Abnormal surface of Emx2 KO embryonic gonad

Although the gonads of *Emx2* KO embryos disappear by E12.5 (8), the mechanism by which this defect arises was not analyzed. As shown in Fig. 1A, the urogenital primordium is observed as a thickening of eplithelia at E10.5 in both wild-type and *Emx2* KO embryos. By E11.5, the developing urogenital primordium separates into a sexually indifferent gonad and mesonephros and thereafter develops into either the testis or ovary. The gonads of *Emx2* KO embryos are underdeveloped at E11.5 and largely disappear by E12.5 in both sexes. We noted that during the degeneration process, the gonadal surface of *Emx2* KO embryos appears to be abnormal at E11.5. Thus, the gonadal surface was examined by SEM (Fig. 1B). The gonads of wild-type embryos exhibit smooth surfaces, whereas those of *Emx2* KO embryos show irregularly protruding cellular clusters.

To characterize the *Emx2* KO gonad, we examined the expression of various gonadal marker genes. The expressions of Ad4BP/SF-1 required for the gonadal and adrenal development (47, 48) and aristaless-related homeobox required for the development of testicular Leydig cells (32) disappeared from the *Emx2* KO gonads by E11.5, whereas the expression of Ad4BP/SF-1 in the adrenal primordium was unaffected (Fig. 1C). This loss of Ad4BP/SF-1 expression in the gonads but not the adrenal glands of *Emx2* KO mice is consistent with the observation that although *Ad4BP/SF-1* KO mice failed to develop both the gonads and adrenal glands (49), *Emx2* KO mice failed to develop the adrenal gland. The expression of GATA4, WT1, and CBX2/M33, all of which are expressed in the developing gonad of wild-type and involved in the gonadal development, was unaffected. Interestingly, the epithelial expression of *Lhx9* required for the gonadal development (37) was expanded laterally in *Emx2* KO.

Abnormal cell proliferation and apoptotic cell death induced in Emx2 KO gonads

The structural abnormalities in the gonadal surface of *Emx2* KO strongly suggested that *Emx2* is critical for the development of the epithelial cells of the gonad. Therefore, the expression of EMX2 in the developing gonad was examined. Immunohistochemical studies demonstrated that EMX2 is expressed in the gonadal epithelial and adjacent

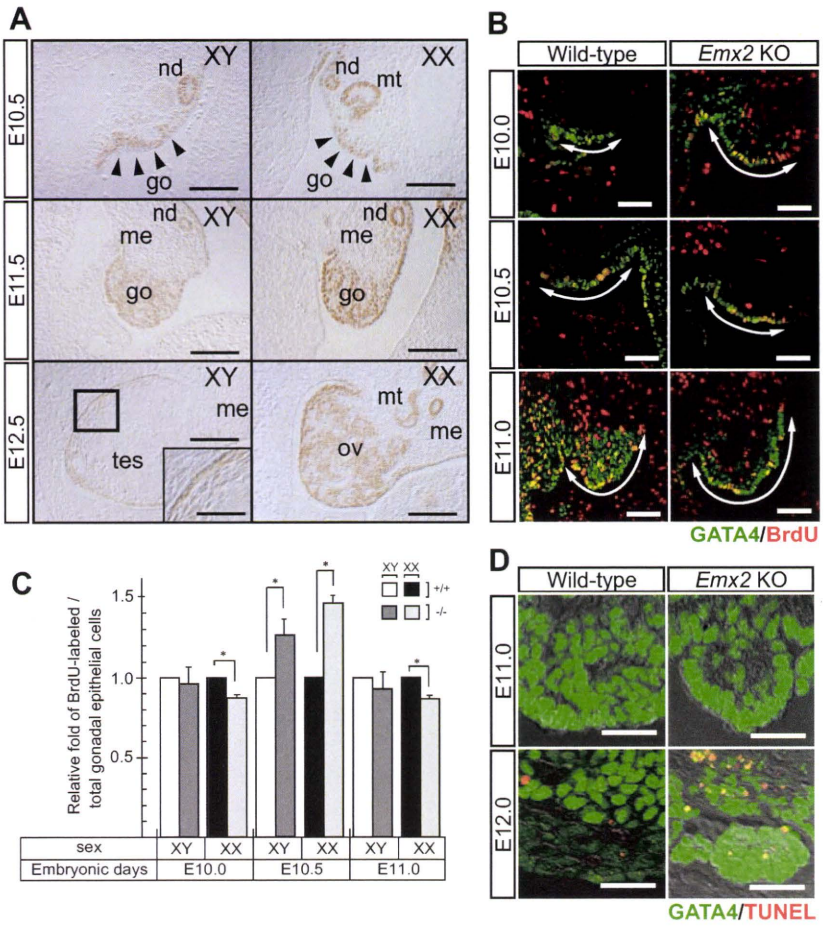

FIG. 2. Up-regulated epithelial cell proliferation and apoptosis in *Emx2 KO* embryonic gonads. A, Expression of EMX2 during gonadal development. The expression of EMX2 was examined immunohistochemically in male (XY) and female (XX) at E10.5 (ts 7–8), E11.5 (ts 18–19), and E12.5 (ts 32). *Arrowheads* in E10.5 indicate the early developing gonads. The *region enclosed by a square* in the E12.5 male gonad is enlarged as an *inset*. *Scale bars*, 100 and 50 μm (*inset*). go, Gonad; nd, nephric duct (feature Wolffian duct); me, mesonephrose; mt, mesonephric tubule; tes, testis; ov, ovary. B, BrdU labeling in wild-type and *Emx2* KO embryonic gonads. Cell proliferation was assessed by BrdU labeling as described in *Materials and Methods*. The BrdU-labeled gonads of wild-type and *Emx2* KO embryos were sectioned at E10.0 (ts 4–6), E10.5 (ts 9–11), and E11.0 (ts 13–17) and stained with anti-BrdU (*red*) and GATA4 (*green*) antibodies. As indicated by *arrows*, gonadal regions were determined by GATA4 staining and morphology. *Scale bars*, 50 μm. C, Transient up-regulation of BrdU incorporation into epithelial cells of *Emx2* KO embryonic gonads. The total number of BrdU-labeled gonadal epithelial cells in the *areas indicated by arrows* in A was counted at E10.0 (ts 4–6), E10.5 (ts 9–11), and E11.0 (ts 13–17). Data were obtained only when wild-type and *Emx2* KO embryos of the same sex were in the same litter. The relative fold changes of the number of BrdU-positive cells in the gonadal epithelial cells are plotted, with the number in wild-type embryos of each sex set at 1 for each stage. The number of gonads used in this study is as follows: two wild-type and 6 KO gonads for E10.0 males, two wild-type and 4 KO gonads for E10.0 females, five wild-type and 5 KO gonads for E10.5 males, five wild-type and five KO gonads for E10.5 females, four wild-type and 4 KO gonads for E11.0 males, and three wild-type and three KO gonads for E11.0 females. Values are the means ± SD; *, $P <$ 0.001. D, Ectopically increased apoptosis in *Emx2* KO gonads. A TUNEL assay (*red*) was used to detect apoptotic cells in the wild-type and *Emx2* KO gonads at E11.0 (ts 15) and E12.0 (ts 25). Gata4 immunostaining (*green*) was used to detect the developing gonads. Gonadal regions are indicated by *arrows*. *Scale bars*, 25 μm.

mescenchymal cells, nephric duct, and mesonephric tubule at E10.5 and expressed in all somatic cells in the gonad but not in the mesonephrose except tubular structure at E11.5 (Fig. 2A). Although the expression was similar between the two sexes before gonadal sex differentiation, the expression became different between the testis and ovary at E12.5. The expression was down-regulated in the testis except testicular tunica albuginea and underneath mesenchymal cells, whereas the expression was still evident in the whole ovary at E12.5. Because the testis initiates to synthesize testosterone as early as E12.5, the sexually dimorphic expression of *Emx2* established by E12.5 seems to be independent of gonadal sex steroid.

Next, we examined whether epithelial cell proliferation is affected in the *Emx2* KO gonad. After BrdU was injected into pregnant females, embryos were collected at E10.0, E10.5, and E11.0 and their gonads stained with an antibody against BrdU (Fig. 2B, *red*). Because the gonads at the stage are structurally primitive, it is difficult to discriminate between the future gonadal and mesonephric areas only by the morphology. For the gonad at the stages, Ad4BP/SF-1 and GATA4 are known to be potential gonadal markers. However, the expression of Ad4BP/SF-1 was affected significantly in the *Emx2* KO gonad, and thus GATA4 immunostaining was performed to evaluate the gonadal area (Fig. 2B, *green*) (35). In addition, considering that GATA4 is expressed in the mesentery, the gonadal area was eventually determined as GATA4 immunoreactive but not the mesentery cells. The number of BrdU-positive proliferating epithelial cells was similar between wild type and *Emx2* KO at E10.0 in both sexes (Fig. 2, B and C). By E10.5, the number of BrdU-positive cells in *Emx2* KO gonads was increased by approximately 1.3-fold in males and 1.5-fold in females compared with wild type. This increase in epithelial cell proliferation was not observed at E11.5.

Apoptosis was examined using terminal deoxynucleotidyltransferase-mediated 2′-deoxyuridine 5′-triphosphate

nick end labeling (TUNEL) labeling. Although the KO gonad showed obvious structural defects at E11.5, the number of TUNEL-positive apoptotic cells was not increased at E11.0. By E12.0, however, the number of TUNEL-positive cells was significantly increased (Fig. 2D).

Migration of gonadal epithelial cells affected in Emx2 KO

It has long been surmised that the coelomic epithelia at the both sides of the mesentery proliferate and migrate mediodorsally to give rise to the gonadal primordia. As described above, the BrdU incorporation study indicated that epithelial cell proliferation was transiently up-regulated in *Emx2* KO gonads. However, if migration of the epithelial cells is affected, it is assumed that the two BrdU-positive daughter cells remain in the epithelial compartment, thus causing an apparent increase in the number of the BrdU-positive epithelial cells. We therefore examined whether migration of epithelial cells to the mesenchymal compartment was affected in the KO.

Gonadal epithelial cells are thought to pass through the basement membrane during their migration into the mesenchymal compartment. The gonads of wild-type and *Emx2* KO embryos were sectioned and stained with laminin antibody to visualize the basement membrane (Fig. 3A, *green*), whereas nuclei were stained with PI (Fig. 3A, *red*). As expected, cells were frequently seen in the basement membrane in wild-type embryos (Fig. 3A, *arrowheads*) but approximately 6-fold less frequently in the KO gonads (Fig. 3B). Epithelial cells can undergo polarized cell division, with one daughter cell retaining epithelial cell features while the other cell losing them and migrating into the mesenchymal compartment through the basement membrane (21, 50, 51). The decreased number of cells localized to the basement membrane in the *Emx2* KO gonad strongly suggests that polarized cell division and cell migration are affected in these mutants. Moreover, we noticed that the basement membrane of the KO gonad is not tightly lining the epithelial cells when compared with wild type. This unusual basement membrane might affect the epithelial cell migration.

Therefore, we examined this migration defect using organ culture. Embryonic trunk tissue containing the developing gonads was prepared at E10.25, and whole coelomic epithelial cells, including the gonadal epithelia, were labeled with a fluorescent dye, CCFSE, to chase the epithelial cells (Fig. 3C, *green*). After the labeled trunks were cultured for 5 h, the gonads were sectioned and immunostained with antilaminin antibody (Fig. 3C, *red*). Expectedly, the gonadal epithelia of wild-type embryos migrated through the laminin layer, whereas those of *Emx2* KO

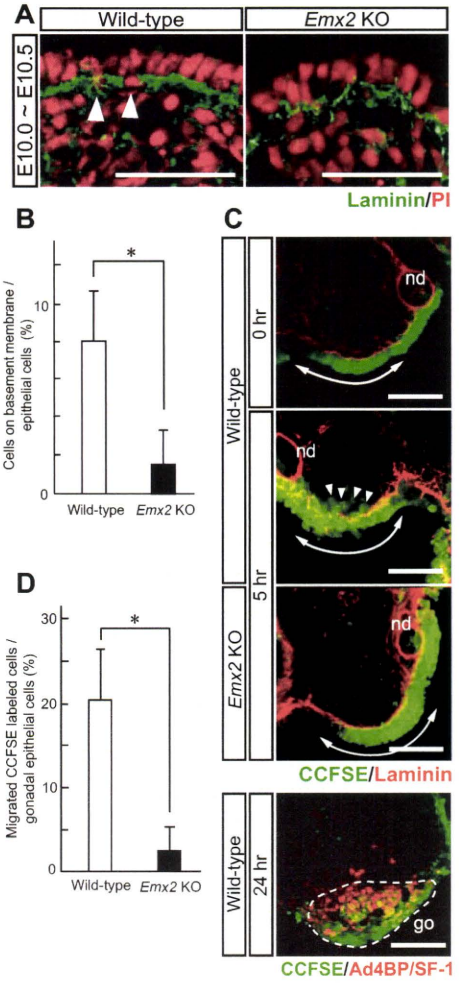

FIG. 3. Aberrant migration of gonadal epithelial cells to the mesenchymal compartment in *Emx2 KO*. A and B, A decreased number of cells was localized to the basement membrane in *Emx2* KO embryonic gonads. Gonadal sections at E-10.0–E10.5 (ts 4–7) were stained with laminin (*green*) and PI (*red*) to mark the basement membrane and nuclei, respectively. The cells localized near the basement membrane are indicated by *open arrowheads*. *Scale bars*, 50 μm. The number of cells near the basement membrane is shown relative to the total number of epithelial cells. Ten sections from six wild-type and 10 *Emx2* KO gonads were examined. Values are expressed as the means ± SD; *, $P < 0.01$. C and D, Aberrant migration of epithelial cells in the *Emx2* KO embryonic gonad. The abdominal epithelial cells of wild-type and *Emx2* KO embryos at E10.25 (ts 5–6) were labeled with CCFSE (*green*) (0 h), and thereafter cultured for 5 or 24 h (5 or 24 h). Specimens at 0 and 5 h were then sectioned and stained with an antibody against laminin (*red*), whereas those at 24 h were stained with an antibody against Ad4BP/SF-1 (*red*). The developing gonadal regions are indicated by *arrows*, which are determined through GATA4 staining (data not shown) and morphology. The migrated cells are indicated by *arrowheads* (5 h). As shown in the *lowermost panel* (24 h), the migrated cells became immunoreactive for Ad4BP/SF-1. The *area enclosed by a dotted line* is the developing gonad. *Scale bars*, 100 μm. go, Gonad; nd, nephric duct. The gonadal epithelial cells and cells that had migrated through the basement membrane after 5 h of incubation were counted. The number of migrated cells is shown relative to the total epithelial cell numbers. Six wild-type and six *Emx2* KO embryonic gonads were used in this study. Values are the means ± SD; *, $P < 0.0001$.

scarcely migrated. The number of migrating cells seen in wild-type gonads was approximately 9-fold higher than in *Emx2* KO (Fig. 3D). This migration defect possibly results in a substantial increase of epithelial cells and, at the same time, decrease of mesenchymal cells. Taken together, the observations above suggest that the loss of *Emx2* blocks gonadal development by restricting epithelial cell migration.

Based on the observation above, a question arose as to whether the migrating cells differentiate into gonadal somatic cells or not. Thus, the gonad sections were stained with anti-Ad4BP/SF-1 antibody after further incubation up to 24 h. Ad4BP/SF-1 was expressed in many CCFSE-positive migrating cells (Fig. 3C). These results demonstrate directly for the first time that the gonadal epithelial cells at around E10.25 have the potential to migrate and to form gonadal mesenchyme after migration.

Abnormal tight junction assembly of the gonadal epithelia

Because epithelial cells are characterized by specialized cellular junctions, we examined whether tight junction assembly is affected in the gonadal epithelia of *Emx2* KO. The intracellular component ZO-1 normally interacts with the homomeric tight junction protein occludin, and together, they localize to the apicolateral region in epithelial cells. aPKC forms a complex with PAR3 and PAR6 that localizes to tight junctions and regulates tight junction assembly (21, 22). These marker proteins were normally localized to the apicolateral region of the gonadal epithelia in wild-type embryos at E10.5 (Fig. 4) but were disrupted in *Emx2* KO. Occludin distribution is expanded deeply to the lateral domain, whereas ZO-1 and aPKC are localized irregularly to whole aspects of the cellular surface. These data suggest that the *Emx2* KO gonadal epithelia have lost their cell polarity.

Affected gene expression in Emx2 KO embryonic gonads

To assess the effect of *Emx2* gene disruption, gonadal epithelia were microdissected from wild-type and *Emx2* KO embryos at E10.5 to prepare total RNA. After biotinylation, these samples were used as probes for microarray analysis. Genes that satisfied a pairwise comparison test and displayed a more than 4-fold change in expression are summarized in Table 1. Because *Emx2* KO gonad showed affected cell polarity, genes encoding the components of tight and adherens junctions, as well as *Snail* and *Slug*, which regulate transcription of the junction component genes (21, 52), were expected to be affected in the KO gonad. However, none of these genes showed differential expression greater than 4-fold.

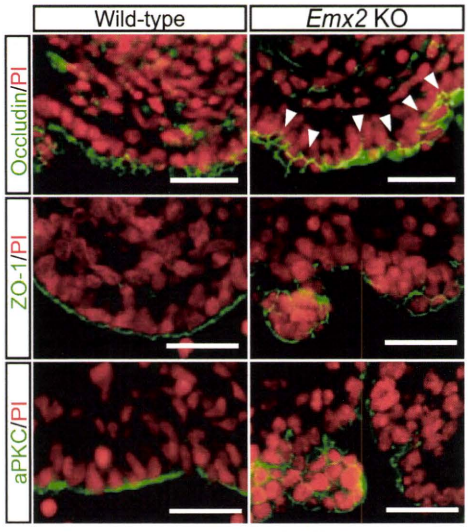

FIG. 4. Abnormal distribution of tight junction marker proteins. Wild-type (*left*) and *Emx2* KO (*right*) embryonic gonads at E10.5 (ts 8–9) were sectioned and stained with antibodies against occludin, ZO-1, and aPKC (*green*). The nuclei were stained with PI (*red*). The apicolateral distribution of marker proteins observed in wild-type embryos was disrupted in *Emx2* KO. *Arrowheads* indicate that occludin distribution expanded deeply to the lateral domain of cell surface. *Scale bars*, 25 μm.

EGFR ectopically induced in Emx2 KO embryonic gonads

Our microarray data listed *Egfr* as the top-scored gene induced in the *Emx2* KO gonad. Quantitative RT-PCR revealed an approximately 64-fold increase in *Egfr* expression in *Emx2* KO tissue compared with wild type (Fig. 5A). Furthermore, immunohistochemistry demonstrated that EGFR is expressed at a low level in the gonadal epithelia of wild-type embryos, whereas the expression is high in both the epithelial and mesenchymal cells of the *Emx2* KO gonad (Fig. 5B). In contrast, EGFR was not induced in the nephric duct and mesonephric tubules of KO.

The microarray study strongly suggested that *Egfr* gene is suppressed by EMX2. Therefore, we used M15 cells derived from mesonephric epithelial cells to test this hypothesis. Western blot analyses using an anti-EMX2 antibody showed that EMX2 is not expressed in M15 cells (Fig. 5D), whereas immunohistochemistry, Western blotting, and RT-PCR revealed that EGFR is expressed in the cells (Fig. 5, C–E). As expected, when Emx2(HA) was overexpressed, EGFR expression was reduced, as assayed by immunohistochemistry, Western blotting, and RT-PCR. Moreover, when the Emx2(HA)-overexpressing cells were treated with siRNA for *Emx2*, the expression of EGFR was significantly up-regulated. Such up-regulation was never observed with control siRNA.

TABLE 1. Down-regulated and up-regulated genes in *Emx2* KO gonadal epithelia

Gene symbol	Log2 fold change	P	Description	GenBank accesion no.
Down-regulated				
Cbln1	−4.42	0.00115262	Cerebellin 1 precursor protein	NM_019626
Inhbb	−3.47	0.00095948	Inhibin β-B	NM_008381
Dct	−3.34	0.00599337	Dopachrome tautomerase	NM_010024
Pnlip	−2.76	0.00564109	Pancreatic lipase	NM_026925
Enpep	−2.53	0.03860764	Glutamyl aminopeptidase	NM_007934
Myh6	−2.37	0.00233953	Myosin, heavy polypeptide 6, cardiac muscle, α	NM_010856
Sept4	−2.22	0.00394216	Septin 4	NM_011129
Hpgd	−2.12	0.00095948	Hydroxyprostaglandin dehydrogenase 15 (NAD)	NM_008278
Up-regulated				
Egfr	4.61	0.00003547	EGFR	NM_007912 NM_207655
Tbx18	3.73	0.00095948	T-box18	NM_023814
Crh	3.45	0.00643128	Corticotropin releasing hormone	NM_205769
Gpm6a	2.55	0.00156104	Glycoprotein m6a	NM_153581
Fut9	2.45	0.00115262	Fucosyltransferase 9	NM_010243
Slitrk6	2.20	0.00494608	SLIT and NTRK-like family, member 6	NM_175499
GA17	2.10	0.00886440	Dendritic cell protein GA17	NM_145380
Epha3	2.00	0.00864173	Eph receptor A3	NM_010140

Gene expression was compared between wild-type and *Emx2* KO gonadal epithelia at E10.5 (ts 7) by microarray analysis. Genes showing more than 4-fold change are listed ($P < 0.05$). All expression data have been deposited in the Gene Expression Omnibus of NCBI (accession no. GSE10216; http://www.ncbi.nlm.nih.gov/geo/). SLIT, *Drosophila* slit gene homolog; NTRK, neurotrophic tyrosine receptor kinase.

Ectopic activation of c-Src in Emx2 KO gonad

Recently, activated EGFR was shown to phosphorylate a tyrosine residue of c-Src/c-Yes, and phosphorylated c-Src/c-Yes in turn phosphorylates a tyrosine residue of PAR3. Through this successive tyrosine phosphorylation, EGFR signaling is thought to fine-tune tight junction assembly (25, 53). Therefore, we examined phosphorylation of EGFR and c-Src in the *Emx2* KO gonads. Interestingly, tyrosine phosphorylation of c-Src is clearly activated in the *Emx2* KO gonad in a pattern that overlaps with that of ectopically induced EGFR. A low level of c-Src phosphorylation is detected in the epithelial cells of wild-type gonads (Fig. 5F). It has been well established that EGFR is phosphorylated at the tyrosine 845 by activated c-Src and autophosphorylated at tyrosine 1068 upon ligand binding followed by dimerization (54). As indicated in Fig. 5G, phosphorylation of EGFR at the tyrosine 845 was clearly elevated in the *Emx2* KO gonad, whereas that at the tyrosine 1068 was unlikely elevated.

Discussion

Migration of coelomic epithelial cells to the mesenchymal compartment during development of the gonadal primordium

Based on histological observations, it has long been surmised that regions of the coelomic epithelia at the both sides of the mesentery proliferate and migrate mediodorsally to give rise to the gonadal primordium (1–3). However, this has not been addressed directly by cell fate mapping. Here, we chemically labeled coelomic epithelial cells at E10.25 and found that the labeled epithelial cells migrated through the basement membrane. After migration, they began to express *Ad4BP/SF-1*, a marker gene for Sertoli and Leydig cell lineages (47), strongly suggesting that the migrated epithelial cells differentiated into these gonadal somatic cells.

Transition of the gonadal epithelial cells to mesenchyme was previously demonstrated with mouse embryonic gonad at around E11.5 (55). Interestingly, the migrated cells at E11.2– E11.4 differentiated into Sertoli and interstitial cells, whereas those at E11.5– E11.7 no longer developed into Sertoli cells. Unfortunately, our study with E10.25 embryos failed to culture the gonads until Sertoli and Leydig cells differentiate, and thus it remains unsolved whether the epithelial cells at the earlier stage develop into Sertoli and Leydig cells. A new culture system, which enables to culture the early gonadal primordium for a longer period, is required to resolve the issue.

Emx2 implicated in the maintenance of epithelial polarity and the epithelial-to-mesenchyal transition

Many tissues are known to undergo epithelial-to-mesenchymal transition and/or mesenchymal-to-epithelial transition during the development. These transitions are closely correlated with the assembly and disassembly of tight junctions. In the present study, we have demonstrated that ectopic tight junctions are formed in *Emx2* KO gonadal epithelia, and thus it is assumed that the aberrant persistence of tight junctions inhibits the epithelial-

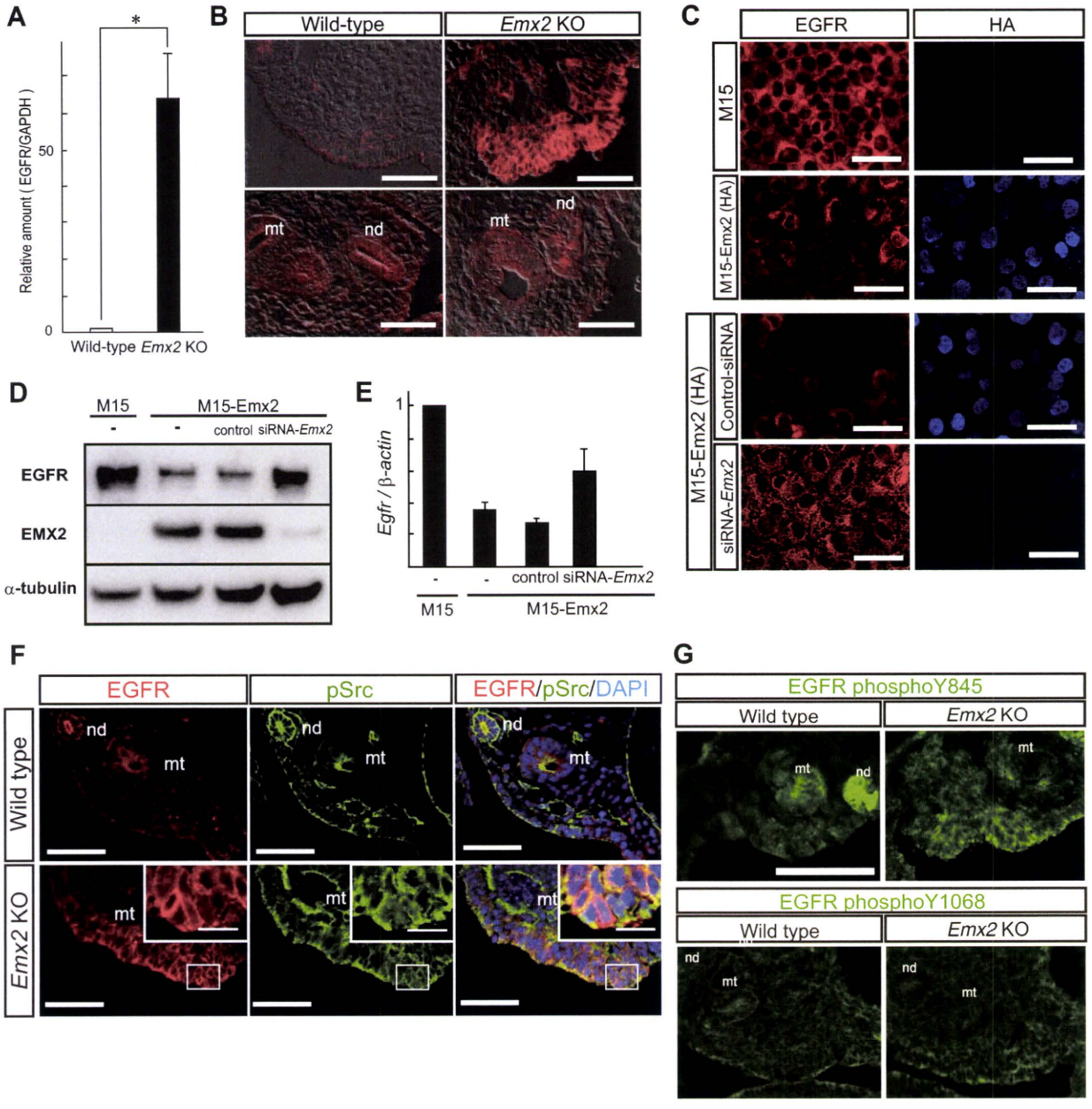

FIG. 5. EGFR gene expression suppressed by *Emx2*. A, Increased *Egfr* mRNA revealed by quantitative RT-PCR. Total RNA used in microarray analysis was subjected to RT-PCR for *Egfr*. *Gapdh* was used as a control. The amount of *Egfr* relative to that of *Gapdh* is shown, with wild-type levels set at 1. Values are the means ± SD; *, $P < 0.001$. B, Increased EGFR protein revealed by immunohistochemical staining. Sections prepared from wild-type and *Emx2* KO embryos at E10.5 (ts 9) were stained with EGFR antibody (*red*). Images for the gonadal (*upper*) and mesonephric regions (*lower*) are shown. *Scale bars*, 50 μm. nd, Nephric duct; mt, mesonephric tubule. C, Immunohistochemical examination of EMX2 and EGFR. After M15 cells overexpressing Emx2(HA) [M15-Emx2(HA)] were cultured for 48 h, they were immunostained with antibodies to HA-tag (*blue*) and EGFR (*red*). After M15-Emx2(HA) cells were cultured for 24 h, they were treated with siRNA for *Emx2* (siRNA-*Emx2*) or control siRNA (control siRNA) for 24 h. These cells were immunostained as above. *Scale bars*, 50 μm. D, Immunoblotting of M15 and M15-Emx2(HA) cells. Total cell extracts (15 μg) prepared from M15, M15-Emx2(HA), and M15-Emx2(HA) treated with siRNA-*Emx2* or control siRNA were used. EGFR, EMX2, and α-tubulin were detected with specific antibodies. E, Quantitative RT-PCR analysis of EGFR in M15 and M15-Emx2(HA) cells. Total RNA prepared from M15, M15-Emx2(HA), and M15-Emx2(HA) cells treated with siRNA-*Emx2* or control siRNA was used. The amount of *Egfr* mRNA relative to *β-actin* is plotted with levels in control M15 cells set at 1. Values are the means ± SD. F, Abnormal phosphorylation of c-Src in *Emx2* KO gonads. Sections prepared from wild-type and *Emx2* KO embryos at E10.5 (ts 7) were stained with EGFR antibody (*red*) and tyrosine-phosphorylated c-Src (pSrc) (*green*). Nuclei were stained with 4′,6-diamidino-2-phenylindole (DAPI) (*blue*). *Superimposed images* are shown at the *right*. The *regions enclosed by squares* are enlarged. G, Tyrosine phosphorylation of EGFR. Sections prepared from wild-type and *Emx2* KO embryos at E10.5 (ts 9–11) were stained with an antibody to EGFR phosphorylated at tyrosine 845 (Y845) or tyrosine 1068 (Y1068) (*green*). *Scale bars*, 50 μm. nd, Nephric duct; mt, mesonephric tubule.

to-mesenchymal transition during early gonadal development. As described below, this defect may correlate with ectopic EGFR expression in the developing *Emx2* KO gonad.

In addition to the gonad, *Emx2* is expressed in the epithelial cells of the ureteric bud, Wolffian duct, Müllerian duct, and mesonephric tubule (56). Interestingly, extension and branching of the ureteric bud are affected in *Emx2* KO. Similarly, Wolffian duct and mesonephric tubules degenerate after the structures initially develop, whereas Müllerian duct fails to develop (8). Given that these structures are formed via a mesenchymal-to-epithelial transition, *Emx2* is thought to be involved in both directions of transitions. However, because the expression of *Egfr* was unaffected in the tubular structures, *Emx2* may regulate two transition processes through differential target gene expression.

Potential function of EGFR in developing gonadal epithelial cells

Microarray studies clearly demonstrated that the *Egfr* expression was up-regulated in the *Emx2* KO gonads, and consequently, the question of why *Egfr* should be negatively regulated by *Emx2* arises from our studies. Given that EGFR is not required throughout gonad development, one would expect the *Egfr* locus to be silenced, possibly through inactivation by a suppressive chromatin state. However, considering that *Egfr* expression is activated by *Emx2* gene disruption, *Egfr* gene is not structurally silenced in the early gonad; instead, it is kept in a state that is ready to be activated.

Recently, Wang *et al.* (25) demonstrated that activated EGFR phosphorylates a tyrosine residue of c-Src/c-Yes, and subsequently, the phosphorylated c-Src/c-Yes phosphorylates a tyrosine residue of PAR3. Because tight junction assembly is delayed but not blocked with a phosphorylation-defective mutant of PAR3, EGFR signaling is thought to fine-tune tight junction assembly through successive tyrosine phosphorylation. Consistent with these observations, our data show that tyrosine-phosphorylated c-Src accumulates in the *Emx2* KO embryonic gonad, in which *Egfr* is ectopically up-regulated.

EGFR is phosphorylated at multiple tyrosine residues. Upon ligand binding and dimerization, autophosphorylation of EGFR occurs at several tyrosine residues, including Y1068. Likewise, activated c-Src phosphorylates several tyrosine residues, including Y845 (54). The present study showed that phosphorylation at Y1086 was not elevated in the *Emx2* KO gonad, whereas that at Y845 was elevated. This phosphorylation status strongly suggested that EGFR is phosphorylated by the activated c-Src but not EGFR itself. In fact, c-Src was phosphorylated at tyrosine residue and thus activated in the KO gonad. Taken together, the overexpressed EGFR was activated possibly by c-Src in the KO gonad, although it remains to be clarified how *Emx2* gene disruption causes c-Src activation.

Importantly, our immunohistochemical studies reproducibly detected EGFR and tyrosine-phosphorylated c-Src at low levels in the gonadal epithelial cells of wild type. Considering that the epithelial cells migrate to the mesenchymal compartment at an early stage of gonadal development, tight junctions of the epithelial cells should be disassembled sporadically before migration. EGFR signaling may act as the cue for this disassembly. Taken together, our data demonstrate that Emx2 guarantees proper gonadal development by regulating *Egfr* expression and thus modulating tight junction assembly.

Microarray analyses identified a number of genes whose expression was up- or down-regulated in the *Emx2* KO gonadal epithelial cells. Among the genes that showed up- and down-regulated expressions in the *Emx2* KO gonads, we highlighted the up-regulated expression of *Egfr* in the KO. However, this does not necessarily exclude the possible role of other genes described above in gonad development. Through its regulation of EGFR and possibly other genes, *Emx2* appears to play a crucial role in regulating epithelial cellular junctions of the early developing gonad.

Acknowledgments

Address all correspondence and requests for reprints to: Ken-ichirou Morohashi, Ph.D., Department of Molecular Biology, Graduate School of Medical Sciences, Kyushu University, Fukuoka 812-8582, Japan. E-mail: moro@cell.med.kyushu-u.ac.jp.

Present address for M.K.: Department of Bioscience and Bioinformatics, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, Iizuka 820-8502, Japan.

Present address for Y.K.-F.: Department of Aging Intervention, National Institute for Longevity Science, National center for Geriatrics and Gerontology, Obu 474-8511, Japan.

Present address for H.O.: Kobe Advanced ICT Research Center, National Institute of Information and Communications Technology, Kobe 651-2492, Japan.

Present address for N.S.: Department of Anatomy and Developmental Biology, Graduate School of Medical Science, Kyoto Prefecture University of Medicine, Kyoto 602-8566, Japan.

Present address for Y.Su.: Department of Pharmaceutical Biochemistry, Faculty of Life Sciences, Kumamoto University, Kumamoto 862-0973, Japan.

This work was supported by Grants-in-Aid for Scientific Research on Priority Areas and Grant-in-Aid for Scientific Research

from the Ministry of Education, Culture, Sports Science, and Technology of Japan.

Disclosure Summary: The authors have nothing to disclose.

References

1. **Allen BM** 1904 The embryonic development of the ovary and testis of the mammals. Am J Anat 3:89–154
2. **Pelliniemi LJ** 1975 Ultrastructure of gonadal ridge in male and female pig embryos. Anat Embryol 147:20–34
3. **Wartenberg H** 1982 Development of the early human ovary and role of the mesonephros in the differentiation of the cortex. Anat Embryol 165:253–280
4. **Swain A, Lovell-Badge R** 1999 Mammalian sex determination: a molecular drama. Genes Dev 13:755–767
5. **Wilhelm D, Palmer S, Koopman P** 2007 Sex deter mination and gonadal development in mammals. Physiol Rev 87:1–28
6. **Ross AJ, Capel B** 2005 Signaling at the crossroads of gonad development. Trends Endocrinol Metab 16:19–25
7. **Simeone A, Gulisano M, Acampora D, Stornaiuolo A, Rambaldi M, Boncinelli E** 1992 Two vertebrate homeobox genes related to the *Drosophila* empty spiracles gene are expressed in the embryonic cerebral cortex. EMBO J 11:2541–2550
8. **Miyamoto N, Yoshida M, Kuratani S, Matsuo I, Aizawa S** 1997 Defects of urogenital development in mice lacking Emx2. Development 124:1653–1664
9. **Gangemi RM, Daga A, Marubbi D, Rosatto N, Capra MC, Corte G** 2001 Emx2 in adult neural precursor cells. Mech Dev 109:323–329
10. **Heins N, Cremisi F, Malatesta P, Gangemi RM, Corte G, Price J, Goudreau G, Gruss P, Götz M** 2001 Emx2 promotes symmetric cell divisions and a multipotential fate in precursors from the cerebral cortex. Mol Cell Neurosci 18:485–502
11. **Galli R, Fiocco R, De Filippis L, Muzio L, Gritti A, Mercurio S, Broccoli V, Pellegrini M, Mallamaci A, Vescovi AL** 2002 Emx2 regulates the proliferation of stem cells of the adult mammalian central nervous system. Development 129:1633–1644
12. **Shinozaki K, Miyagi T, Yoshida M, Miyata T, Ogawa M, Aizawa S, Suda Y** 2002 Absence of Cajal-Retzius cells and subplate neurons associated with defects of tangential cell migration from ganglionic eminence in Emx1/2 double mutant cerebral cortex. Development 129:3479–3492
13. **O'Leary DD, Chou SJ, Sahara S** 2007 Area patterning of the mammalian cortex. Neuron 56:252–269
14. **Fukuchi-Shimogori T, Grove EA** 2003 Emx2 patterns the neocortex by regulating FGF positional signaling. Nat Neurosci 6:825–831
15. **Ligon KL, Echelard Y, Assimacopoulos S, Danielian PS, Kaing S, Grove EA, McMahon AP, Rowitch DH** 2003 Loss of Emx2 function leads to ectopic expression of Wnt1 in the developing telencephalon and cortical dysplasia. Development 130:2275–2287
16. **Kimura J, Suda Y, Kurokawa D, Hossain ZM, Nakamura M, Takahashi M, Hara A, Aizawa S** 2005 Emx2 and Pax6 function in cooperation with Otx2 and Otx1 to develop caudal forebrain primordium that includes future archipallium. J Neurosci 25:5097–5108
17. **Shimogori T, Banuchi V, Ng HY, Strauss JB, Grove EA** 2004 Embryonic signaling centers expressing BMP, WNT and FGF proteins interact to pattern the cerebral cortex. Development 131:5639–5647
18. **Shin K, Fogg VC, Margolis B** 2006 Tight junctions and cell polarity. Annu Rev Cell Dev Biol 22:207–235
19. **Niessen CM** 2007 Tight junctions/adherens junctions: basic structure and function. J Invest Dermatol 127:2525–2532
20. **Etienne-Manneville S, Hall A** 2003 Cell polarity: Par6, aPKC and cytoskeletal crosstalk. Curr Opin Cell Biol 15:67–72
21. **Suzuki A, Ohno S** 2006 The PAR-aPKC system: lessons in polarity. J Cell Sci 119:979–987
22. **Harris RC, Chung E, Coffey RJ** 2003 EGF receptor ligands. Exp Cell Res 284:2–13
23. **Singh AB, Harris RC** 2005 Autocrine, paracrine and juxtacrine signaling by EGFR ligands. Cell Signal 17:1183–1193
24. **Sibilia M, Kroismayr R, Lichtenberger BM, Natarajan A, Hecking M, Holcmann M** 2007 The epidermal growth factor receptor: from development to tumorigenesis. Differentiation 75:770–787
25. **Wang Y, Du D, Fang L, Yang G, Zhang C, Zeng R, Ullrich A, Lottspeich F, Chen Z** 2006 Tyrosine phosphorylated Par3 regulates epithelial tight junction assembly promoted by EGFR signaling. EMBO J 25:5058–5070
26. **Morohashi K, Zanger UM, Honda S, Hara M, Waterman MR, Omura T** 1993 Activation of CYP11A and CYP11B gene promoters by the steroidogenic cell-specific transcription factor, Ad4BP. Mol Endocrinol 7:1196–1204
27. **Owaribe K, Kodama R, Eguchi G** 1981 Demonstration of contractility of circumferential actin bundles and its morphogenetic significance in pigmented epithelium in vitro and in vivo. J Cell Biol 90: 507–514
28. **Zubair M, Ishihara S, Oka S, Okumura K, Morohashi K** 2006 Two-step regulation of Ad4BP/SF-1 gene transcription during fetal adrenal development: initiation by a Hox-Pbx1-Prep1 complex and maintenance via autoregulation by Ad4BP/SF-1. Mol Cell Biol 26: 4111–4121
29. **Honda S, Morohashi K, Nomura M, Takeya H, Kitajima M, Omura T** 1993 Ad4BP regulating steroidogenic P-450 gene is a member of steroid hormone receptor superfamily. J Biol Chem 268:7494–7502
30. **Ikeda Y, Shen WH, Ingraham HA, Parker KL** 1994 Developmental expression of mouse steroidogenic factor-1, an essential regulator of the steroid hydroxylases. Mol Endocrinol 8:654–662
31. **Committee NRN** 1999 A unified nomeclature system for the nuclear receptor superfamily. Cell 97:161–163
32. **Kitamura K, Yanazawa M, Sugiyama N, Miura H, Iizuka-Kogo A, Kusaka M, Omichi K, Suzuki R, Kato-Fukui Y, Kamiirisa K, Matsuo M, Kamijo S, Kasahara M, Yoshioka H, Ogata T, Fukuda T, Kondo I, Kato M, Dobyns WB, Yokoyama M, Morohashi K** 2002 Mutation of ARX causes abnormal development of forebrain and testes in mice and X-linked lissencephaly with abnormal genitalia in humans. Nat Genet 32:359–369
33. **Pelletier J, Schalling M, Buckler AJ, Rogers A, Haber DA, Housman D** 1991 Expression of the Wilms' tumor gene WT1 in the murine urogenital system. Genes Dev 5:1345–1356
34. **Rackley RR, Flenniken AM, Kuriyan NP, Kessler PM, Stoler MH, Williams BR** 1993 Expression of the Wilms' tumor suppressor gene WT1 during mouse embryogenesis. Cell Growth Differ 4:1023–1031
35. **Viger RS, Mertineit C, Trasler JM, Nemer M** 1998 Transcription factor GATA-4 is expressed in a sexually dimorphic pattern during mouse gonadal development and is a potent activator of the Mullerian inhibiting substance promoter. Development 125:2665–2675
36. **Katoh-Fukui Y, Tsuchiya R, Shiroishi T, Nakahara Y, Hashimoto N, Noguchi K, Higashinakagawa T** 1998 Male-to-female sex reversal in M33 mutant mice. Nature 393:688–692
37. **Rétaux S, Rogard M, Bach I, Failli V, Besson MJ** 1999 Lhx9: a novel LIM-homeodomain gene expressed in the developing forebrain. J Neurosci 19:783–793
38. **Sato Y, Baba T, Zubair M, Miyabayashi K, Toyama Y, Maekawa M, Owaki A, Mizusaki H, Sawamura T, Toshimori K, Morohashi K, Katoh-Fukui Y** 2008 Importance of forkhead transcription factor Fkhl18 for development of testicular vasculature. Mol Reprod Dev 75:1361–1371
39. **Schmahl J, Eicher EM, Washburn LL, Capel B** 2000 Sry induces cell proliferation in the mouse gonad. Development 127:65–73
40. **Nomura M, Kawabe K, Matsushita S, Oka S, Hatano O, Harada N, Nawata H, Morohashi K** 1998 Adrenocortical and gonadal expression of the mammalian Ftz-F1 gene encoding Ad4BP/SF-1 is independent of pituitary control. J Biochem 124:217–224
41. **Ogawa H, Ishiguro K, Gaubatz S, Livingston DM, Nakatani Y** 2002

A complex with chromatin modifiers that occupies E2F- and Myc-responsive genes in G0 cells. Science 296:1132–1136

42. **Larsson SH, Charlieu JP, Miyagawa K, Engelkamp D, Rassoulzadegan M, Ross A, Cuzin F, van Heyningen V, Hastie ND** 1995 Subnuclear localization of WT1 in splicing or transcription factor domains is regulated by alternative splicing. Cell 81:391–401
43. **Nakatani Y, Ogryzko V** 2003 Immunoaffinity purification of mammalian protein complexes. Methods Enzymol 370:430–444
44. **Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, Finnell R, Zettel M, Coleman P** 1992 Analysis of gene expression in single live neurons. Proc Natl Acad Sci USA 89:3010–3014
45. **Sugimoto Y, Tsuboi H, Okuno Y, Tamba S, Tsuchiya S, Tsujimoto G, Ichikawa A** 2004 Microarray evaluation of EP4 receptor-mediated prostaglandin E2 suppression of 3T3-L1 adipocyte differentiation. Biochem Biophys Res Commun 322:911–917
46. **Manuylov NL, Fujiwara Y, Adameyko II, Poulat F, Tevosian SG** 2007 The regulation of Sox9 gene expression by the GATA4/FOG2 transcriptional complex in dominant XX sex reversal mouse models. Dev Biol 307:356–367
47. **Hatano O, Takayama K, Imai T, Waterman MR, Takakusu A, Omura T, Morohashi K** 1994 Sex-dependent expression of a transcription factor, Ad4BP, regulating steroidogenic P-450 genes in the gonads during prenatal and postnatal rat development. Development 120:2787–2797
48. **Hatano O, Takakusu A, Nomura M, Morohashi K** 1996 Identical origin of adrenal cortex and gonad revealed by expression profiles of Ad4BP/SF-1. Genes Cells 1:663–671
49. **Luo X, Ikeda Y, Parker KL** 1994 A cell-specific nuclear receptor is essential for adrenal and gonadal development and sexual differentiation. Cell 77:481–490
50. **Chalmers AD, Strauss B, Papalopulu N** 2003 Oriented cell divisions asymmetrically segregate aPKC and generate cell fate diversity in the early *Xenopus* embryo. Development 130:2657–2668
51. **Blanpain C, Horsley V, Fuchs E** 2007 Epithelial stem cells: turning over new leaves. Cell 128:445–458
52. **Huber MA, Kraut N, Beug H** 2005 Molecular requirements for epithelial-mesenchymal transition during tumor progression. Curr Opin Cell Biol 17:548–558
53. **Kang ES, Oh MA, Lee SA, Kim TY, Kim SH, Gotoh N, Kim YN, Lee JW** 2007 EGFR phosphorylation-dependent formation of cell-cell contacts by Ras/Erks cascade inhibition. Biochim Biophys Acta 1773:833–843
54. **Morandell S, Stasyk T, Skvortsov S, Ascher S, Huber LA** 2008 Quantitative proteomics and phosphoproteomics reveal novel insights into complexity and dynamics of the EGFR signaling network. Proteomics 8:4383–4401
55. **Karl J, Capel B** 1998 Sertoli cells of the mouse testis originate from the coelomic epithelium. Dev Biol 203:323–333
56. **Pellegrini M, Pantano S, Lucchini F, Fumi M, Forabosco A** 1997 Emx2 developmental expression in the primordia of the reproductive and excretory systems. Anat Embryol 196:427–433

van der Horst *et al. BMC Bioinformatics* 2010, **11**:316
http://www.biomedcentral.com/1471-2105/11/316

BMC Bioinformatics

RESEARCH ARTICLE Open Access

A novel chemogenomics analysis of G protein-coupled receptors (GPCRs) and their ligands: a potential strategy for receptor de-orphanization

Eelke van der Horst[†1], Julio E Peironcely[†1], Adriaan P IJzerman[1], Margot W Beukers[1], Jonathan R Lane[1], Herman WT van Vlijmen[1], Michael TM Emmerich[2], Yasushi Okuno[3] and Andreas Bender*[1,4]

Abstract

Background: G protein-coupled receptors (GPCRs) represent a family of well-characterized drug targets with significant therapeutic value. Phylogenetic classifications may help to understand the characteristics of individual GPCRs and their subtypes. Previous phylogenetic classifications were all based on the sequences of receptors, adding only minor information about the ligand binding properties of the receptors. In this work, we compare a sequence-based classification of receptors to a ligand-based classification of the same group of receptors, and evaluate the potential to use sequence relatedness as a predictor for ligand interactions thus aiding the quest for ligands of orphan receptors.

Results: We present a classification of GPCRs that is purely based on their ligands, complementing sequence-based phylogenetic classifications of these receptors. Targets were hierarchically classified into phylogenetic trees, for both sequence space and ligand (substructure) space. The overall organization of the sequence-based tree and substructure-based tree was similar; in particular, the adenosine receptors cluster together as well as most peptide receptor subtypes (*e.g.* opioid, somatostatin) and adrenoceptor subtypes. In ligand space, the prostanoid and cannabinoid receptors are more distant from the other targets, whereas the tachykinin receptors, the oxytocin receptor, and serotonin receptors are closer to the other targets, which is indicative for ligand promiscuity. In 93% of the receptors studied, de-orphanization of a simulated orphan receptor using the ligands of related receptors performed better than random (AUC > 0.5) and for 35% of receptors de-orphanization performance was good (AUC > 0.7).

Conclusions: We constructed a phylogenetic classification of GPCRs that is solely based on the ligands of these receptors. The similarities and differences with traditional sequence-based classifications were investigated: our ligand-based classification uncovers relationships among GPCRs that are not apparent from the sequence-based classification. This will shed light on potential cross-reactivity of GPCR ligands and will aid the design of new ligands with the desired activity profiles. In addition, we linked the ligand-based classification with a ligand-focused sequence-based classification described in literature and proved the potential of this method for de-orphanization of GPCRs.

Background

G protein-coupled receptors (GPCRs) comprise a large family, more than 800 in human [1], of cell surface receptors that consist of seven transmembrane (TM) helices. These receptors are activated by a variety of external stimuli, including light, ions, small molecules, lipids, and proteins; moreover, the majority of therapeutic drugs act on GPCRs [2]. Because of the limited number of target crystal structures [3-6], GPCR drug design relies largely on ligand-based approaches [7] such as property-based

* Correspondence: bendera@lacdr.leidenuniv.nl
[1] Division of Medicinal Chemistry, Leiden/Amsterdam Center for Drug Research, Leiden University, Einsteinweg 55, 2333CC, The Netherlands
[†] Contributed equally
Full list of author information is available at the end of the article

methods [8], pharmacophore models [9], and substructure methods [10]. These methods do not require any knowledge about the target protein; however, combining them with target information often increases their potential. The resulting so-called 'chemogenomics' approaches thus involve both ligand-based and target-based aspects [11]. They do not focus on a single group of ligands and one individual target, but rather on groups of ligands against groups of targets. The central idea is that similar targets have similar ligands [12,13]. Therefore, relationships between targets from the sequence side can be exploited to search for novel receptor ligands on the chemical structure side.

Traditionally, the GPCR superfamily has been classified based on sequence homology of the receptors. Kolakowski grouped all seven transmembrane (7-TM) proteins into classes A to F for receptors proven to bind G-proteins and class O for the other 7-TM proteins [14]. Class A receptors resemble rhodopsin and form the largest cluster. Later, Fredriksson *et al.* proposed a more elaborate classification for known and predicted human GPCRs [1]. Surgand *et al.* presented a sequence-based phylogenetic classification of GPCRs viewed from a ligand perspective [15]. By selecting residues pointing inwards into the generic binding pocket of GPCRs, the authors assembled a set of 30 residues most likely to be accessible for ligand binding. Based on these residues, phylogenetic clustering was performed. Although only a subset of residues was used, the classification was similar to classifications based on the full sequence. Applications of a grouping such as proposed by Surgand *et al.* constitute ligand design for related receptors, as well as de-orphanization of GPCRs [15]. However, the study by Surgant *et al.* is somewhat limited by the scarcity of structural protein data where the identification of binding site residues was solely based on the structure of bovine rhodopsin. It could not yet take into account recent advances that yielded three pharmacologically relevant X-ray crystal structures, namely those of the human β_2 and turkey β_1 adrenoceptors, as well as of the human adenosine A_{2A} receptor [3,5,6,16]. Building further on Surgand's work, Gloriam *et al.* proposed an extended set of ligand-accessible residues, derived from visual inspection of the newly available X-ray GPCR crystal structures, from supporting mutagenesis data and from the evaluation of previously established residue sets [17]. The resulting set of 44 residues was then applied to cluster class A GPCRs into a phylogenetic tree, which reflected similarities in binding site of the receptors.

Complementary to these sequence-based classifications are the ligand-based classifications of GPCRs. Approaches that use ligand similarity measures for target classification have been previously described [18,19]. Keiser *et al.* related targets by pair-wise comparison of their ligands [20]. From a set of 65 k ligands, a network was constructed connecting almost all 246 targets through sequential linkage. From this, previously unknown antagonism of methadone on the muscarinic M_3 receptor and of emetine on the α_2-adrenoceptor was identified.

While sequence-based similarity relies on comparison of the residues at certain positions in the sequence, there is no unambiguously defined method to measure ligand-based similarity. One way of defining ligand similarity is to consider the overlap of substructures in the molecules. Frequent substructure mining is a method for finding the most common substructures in a set of molecules [21-23]. It evaluates all possible substructures, not only discrete fragments that are present in the molecules; it is therefore an exhaustive approach, resulting in a more complete view on the structural features in the set.

In this study, we employ frequent substructure mining to determine the similarity between groups of ligands in a thorough and unbiased manner. This substructural similarity is then used for classification of GPCRs according to relatedness of substructure profiles of their ligands. The substructure-based classification of GPCRs visualizes relatedness of receptors in the form of a phylogenetic tree, which is then compared to the sequence-based phylogenetic classifications of GPCRs. The differences in tree organization are examined with methods that visualize changes in target position. Taken together, we present a (GPCR) classification from the small molecule (ligand) perspective, which facilitates analysis of target similarities and differences in ligand-binding behavior. In addition, we explore the potential of our ligand-based classification in receptor de-orphanization, *i.e.* the prediction of new ligands for orphan receptors.

Results and Discussion

Sequence-based classification

Three types of sequence-based phylogenetic trees were built, namely: one tree that was based on the full 7-TM sequence, one tree employing 30 residues described by Surgand *et al.* [15], and one tree which was based on the set of 44 residues described by Gloriam *et al.* [17]. Note that the three sequence-based trees presented here are different from those published in the referenced original work [1,15,17], since in the current study orphan receptors, receptors with a low number of ligands, and singleton receptors were left out. Singleton receptors are receptors that are the only (available) member in their respective subfamily. Due to the chemogenomic nature of this study, we focus on the phylogenetic tree based on the set of Gloriam *et al.* since it represents the ligand perspective best; this set is referenced as the GSK set [17]. The two other trees are provided for reference purposes in Additional file 1 - Phylogenetic trees based on 7TM

domain and selected residues. The tree that was built based on the multiple sequence alignment of the GSK set is shown in Figure 1. The GPCR subtypes in this tree are grouped as branches in the tree according to subfamily and target since it resembles the sequence-based phylogenetic tree on which GPCR classification is based [1]. For instance, the opioid receptor subtypes δ, κ, μ, and NOP cluster together, as well as the α- and β-adrenoceptor subtypes. The fact that clustering follows the receptor classification is expected since the classification of GPCRs was based on sequence similarity [24,25]. Four clusters are clearly defined in the tree: the aminergic receptors, the adenosine receptors, the prostanoid receptors, and the peptide-binding receptors.

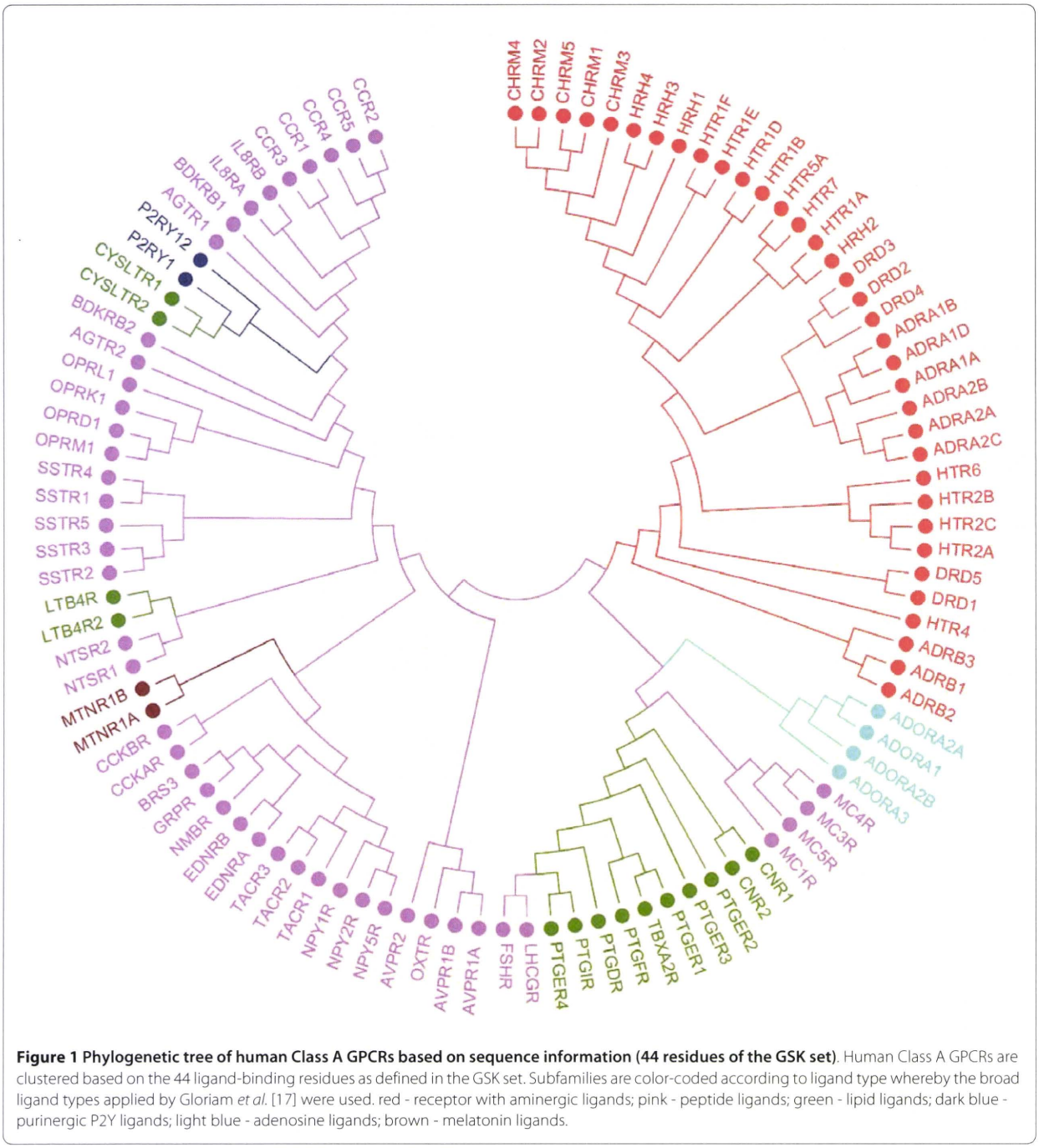

Figure 1 Phylogenetic tree of human Class A GPCRs based on sequence information (44 residues of the GSK set). Human Class A GPCRs are clustered based on the 44 ligand-binding residues as defined in the GSK set. Subfamilies are color-coded according to ligand type whereby the broad ligand types applied by Gloriam *et al.* [17] were used. red - receptor with aminergic ligands; pink - peptide ligands; green - lipid ligands; dark blue - purinergic P2Y ligands; light blue - adenosine ligands; brown - melatonin ligands.

Ligand-based classification

The ligand-based receptor classification, which we will compare to the sequence-based classification, is provided in Figure 2. Subfamilies in this tree are more scattered; however, most subfamilies cluster together. For instance, except for the two purinergic receptors ($P2Y_1$ and $P2Y_{12}$) and the two glycoprotein hormone receptors (FSH and LH), all other receptors represented by only two subtypes, such as the melatonin or the leukotriene B_4 receptors, are clustered together. The adenosine receptors A_1 (ADORA1), A_{2A} (ADORA2A), A_{2B} (ADORA2B), and A_3 (ADORA3) group together, indicating overlap in ligand profiles. This may imply that ligands for these receptor subtypes are non-selective, such as the adenosine receptor antagonists caffeine and theophylline. Additionally, receptor selectivity may vary with relatively small

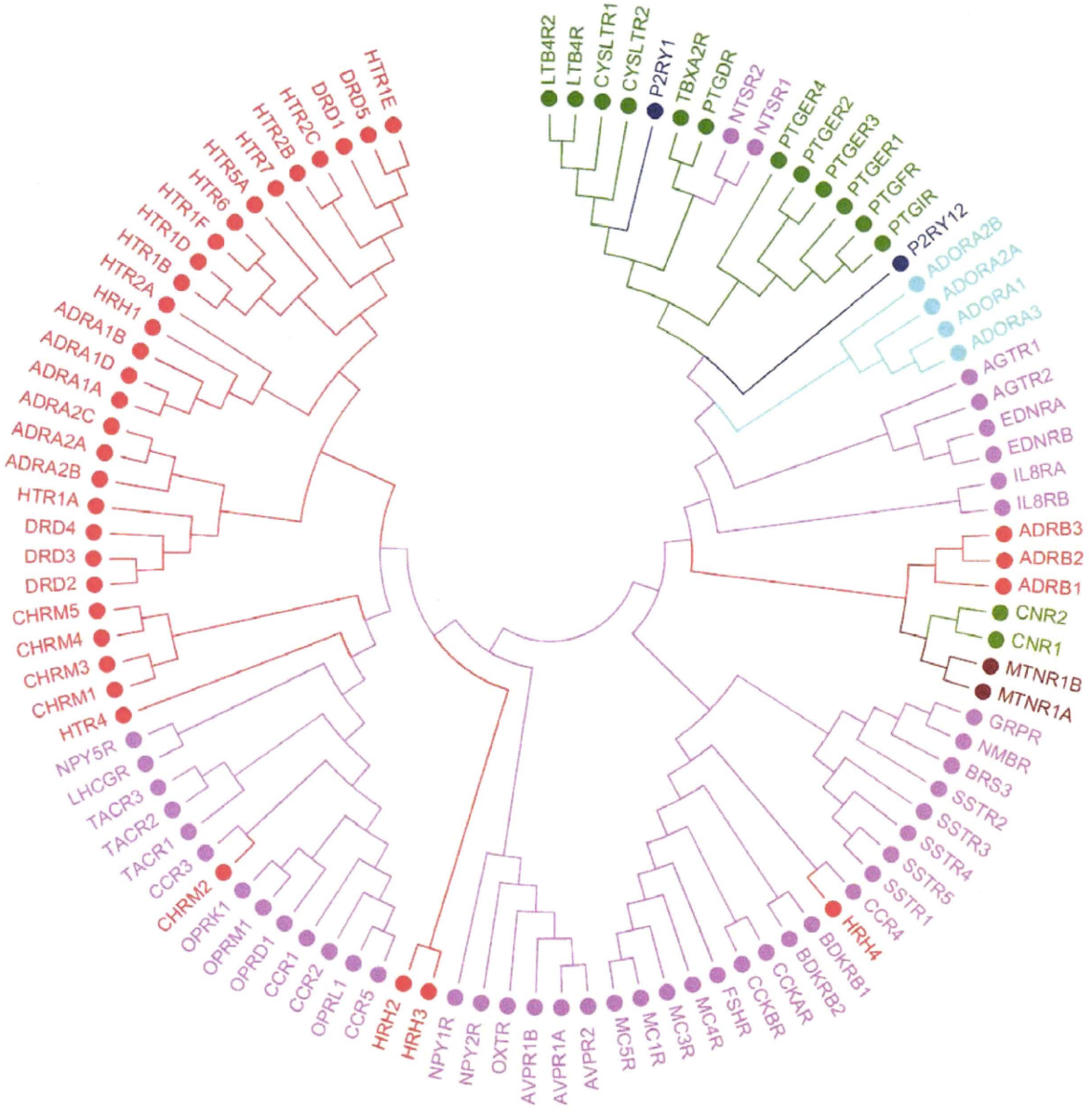

Figure 2 Phylogenetic tree of human Class A GPCRs based on ligand information (frequent substructure mining). Human Class A GPCRs are clustered based on the frequent substructure analysis. Subfamilies are color-coded according to ligand type whereby the broad ligand types applied by Gloriam *et al.* [17] were used. red - receptor with aminergic ligands; pink - peptide ligands; green - lipid ligands; dark blue - purinergic P2Y ligands; light blue - adenosine ligands; brown - melatonin ligands.

changes in ligand structure: an 8-cycloalkyl substituent on theophylline confers A_1 receptor selectivity, whereas a phenylstyryl substituent on the same position in caffeine renders these compounds selective for the A_{2A} receptor. The purinergic receptor $P2Y_{12}$ is found near the adenosine receptors owing to the purine core typical for ligands of both these subfamilies. In agreement with the ligand selectivity reported for the α_1-, α_2-, and β-adrenoceptor subfamilies, these receptors form three distinct clusters [26]; furthermore, the α_{1B} and α_{1D} receptors are the closest in the distance matrix. The muscarinic acetylcholine receptors M_1, M_3, M_4, and M_5 (CHRM1/3/4/5, in Figure 2) cluster together as one group, supporting the low subtype selectivity of muscarinic antagonists [27]. However, the acetylcholine receptor M_2 is found more distant from this cluster. This indicates the presence of distinct chemical classes in the ligand set of the M_2 receptor, which may be the result of inclusion of allosteric ligands. For instance, gallamine is an allosteric modulator of the muscarinic M_2 receptor [28] that is also present in the GLIDA database [29], classified as an M_2 antagonist. In general, the remaining aminergic receptors (serotonergic, dopaminergic, histaminergic and cholinergic) are more scattered throughout the substructure tree. This means that targets share ligands or ligand substructures among subfamilies/subtypes, which is in line with the high level of polypharmacology observed for these aminergic GPCRs [30]. For instance, the serotonin receptor $5\text{-}HT_{1A}$ clusters together with the D_2 dopamine receptor, which fits with reports on antipsychotic compounds combining dopamine D_2 receptor antagonism and serotonin $5\text{-}HT_{1A}$ receptor agonism [31,32]. Structurally similar ligands may act on diverse targets, for instance, when ligands have a GPCR-privileged structure at their core [33,34]. The grouping of the eight prostanoid receptors (Figure 2) indicates similarity in substructure profiles of the ligands. This is based on the fact that most prostanoid receptor ligands are direct derivatives of the endogenous ligands [35,36], the so-called eicosanoids. These ligands are highly similar, all consisting of large aliphatic, lipophilic alkyl chains. The presence of the leukotriene and cannabinoid receptors in this lipid cluster may seem strange at first. Leukotrienes are however also eicosanoids, which clarifies the position of the leukotriene B_4 and cysteinyl-leukotriene receptors in this cluster [37,38]. In addition, arachidonic acid is the common precursor for eicosanoids and two derivatives of arachidonic acid, anandamide and 2-arachidonylglycerol, both of which are endogenous ligands ('endocannabinoids') of the cannabinoid receptors.

The relationship between target clustering in the substructure tree (Figure 2) and ligand promiscuity suggests that the substructure tree may be used to identify possible side effects on receptors that are close neighbors in this tree. For instance, off-target activity of ligands can be identified. If inspection reveals a ligand to bind to receptor(s) that are phylogenetically related to the target of interest, a more detailed experimental follow-up with respect to receptor selectivity would be worthwhile.

Tree comparison

Visual comparison of the sequence tree (Figure 1) with the substructure tree (Figure 2) reveals that the overall phylogenetic organization is similar. For instance, with the exclusion of the glycoprotein, P2Y, angiotensin, and bradykinin receptors, all other receptors represented by two subtypes occur in pairs in both the ligand tree and the sequence tree. This is also true for receptors with three subtypes present in the dataset, *e.g.* the three members of the α_1, the α_2, and the β_1 adrenoceptors, as well as the bombesin receptors. Exceptions to this rule are the neuropeptide Y and vasopressin receptors. In addition, the prostanoid receptors largely group together in both trees, as do most of the aminergic receptors.

The clear distinction between the two dopamine receptor types, i.e. D_1 and D_5 (D_1-like) versus D_2, D_3, and D_4 (D_2-like), exists both in the sequence-based classification and ligand-based classification. This is in agreement with a previous study [39] and also known from drugs on the market such as the benzazepines that favor D_1-like over D_2-like dopamine receptors. Similarly, antipsychotics such as chlorpromazine have a higher affinity for the D_2-like subtypes than D_1-like receptors [40].

The fact that many clusters arise in both trees indicates that the receptors in these clusters have similar sequences and similar ligands, that is, ligands with substantially overlapping substructure sets. However, there are also receptor targets for which this is clearly not the case. The (qualitative) similarities and differences among sequence and substructure trees are discussed in the following. A delta-delta plot was constructed to compare how pairs of receptors change. This plot, provided in Figure 3 (and described in detail in the Materials and Methods section), visualizes how receptor distances deviate between the sequence-based tree and the ligand-based classification of receptors. In sequence space, receptor distances indicate the (dis)similarly between protein sequences, while in ligand space, receptor distances reflect the overlap in structural features found in ligands for these receptors. For each receptor, the mean distance to all other receptors is plotted. From the delta-delta plot, it becomes apparent that the prostanoid receptors and $P2Y_1$ receptor are on average the most distant receptors from the rest of the classes. The distances of the purine $P2Y_1$ receptor, the prostanoid FP receptor, and leukotriene receptor $CysLT_2$

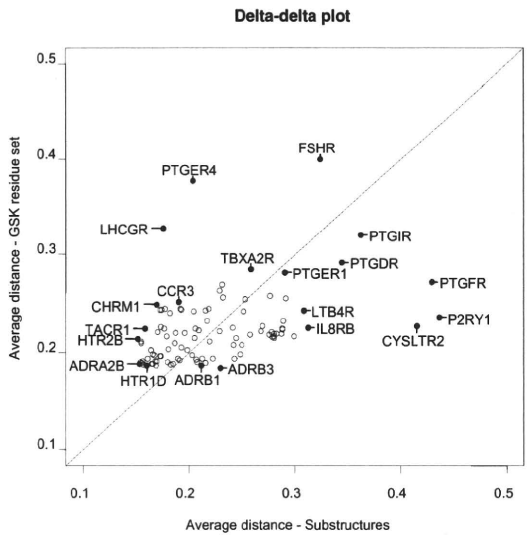

Figure 3 Delta-delta plot visualization of receptor distances in sequence and substructure space. The delta-delta plot visualizes how target distances differ between sequence-based classification (GSK set, y-axis) and substructure-based classification (x-axis). The average distance towards the other targets is plotted for sequence and substructure space. A few targets are highlighted in the plot to serve as examples. These are marked by a black dot and a label that denotes the gene symbol. Targets that are, on average, more distant from the rest are plotted further away from the origin; targets plotted above the diagonal are more distant in sequence space, while targets plotted below the diagonal are more distant in substructure space. For example, the FSH receptor (FSHR) is positioned relatively far from the origin and above the diagonal. This indicates that this receptor is, in general, more distant from the other receptors, most prominent in sequence space.

towards the other classes are all larger in substructure space than in sequence space, implicating that overall their ligands show little resemblance with ligands of the other GPCRs. In contrast, for most aminergic receptors, *e.g.* for the α_{2B}-adrenoceptors and the 5-HT_{2B} serotonin receptor in Figure 3, distances are smaller in substructure space compared to sequence space. This, again, corresponds with the high polypharmacology found for aminergic ligands, such as for most atypical antipsychotics [41], with clozapine as a prominent example [42]. With the exception of a few targets (FSH, LH), the distribution of targets in the delta-delta plot is more scattered along the x-axis (substructure space) than the y-axis (sequence space). This may be a reflection of the evolutionary relationship between sequences, which results in coverage of a small region of the overall sequence space. The ligands for these targets do not have such a direct relationship and thus cover a broader range in overall substructure space.

The difference between ligand-based and target-based classifications may be due to convergent evolution [43]. Functional convergence denotes how proteins that differ in sequence may fulfill the same protein function. The protein sequence of GPCR subtypes will be similar in parts that are involved in the endogenous ligand recognition but may be different in other parts, for instance those parts that play a role in recognition of other, exogenous, ligands (*e.g.* synthetic drugs). These may therefore have a different selectivity profile compared to the endogenous ligand.

Validation

To validate how well our method performed as a chemogenomics method, *i.e.* how well it connects sequence space with small molecule space and how applicable the relationship is in practice, we conducted a 'virtual de-orphanization exercise'. For each receptor in the dataset, we pretended not to know any of its ligands by excluding them from the datasets (we 'orphanized' the receptor in this particular run of the protocol). We next predicted its ligands by considering a model derived from the closest neighbors of the receptor in sequence space (we attempted to 'de-orphanize' the receptor whose ligands we omitted from the study in the previous step). For this calculation, the distance matrix for the GSK residue set was used. The cumulative number of correctly identified ligands of every receptor is plotted against the number of closest neighbors (sequences) included to find these ligands. The (relative) area under the curve (AUC) and shape of the curve are measures of the performance of our method. In 93% of the studied receptors, de-orphanization of the pretended orphan receptor using the ligands of related receptors performed better than random (AUC > 0.5) and for 35% of receptors de-orphanization performance was good (AUC > 0.7). All AUC plots could be divided into four categories according to curve shape and AUC (the complete set of plotted scores is available as additional material in Additional file 2 - Plotted scores for the *leave-one-out* validation). Typical examples of the four categories are given in Figure 4. The first category is most abundant and consists of curves with a convex shape and an AUC above 0.5, marking good performance. An example of this category is the muscarinic acetylcholine receptor M_1 (CHRM1 in Figure 4) with an AUC of 0.7990. Curves of the second category display a gradual rise that is approximately equal to the diagonal of the plot. These plots have an AUC near 0.5, indicating performance that is equal to random prediction. An example is the plot of the angiotensin receptor AT_1 (AGTR1 in Figure 4) with an AUC of 0.5120. Curves of the third category perform worse than random and are characterized by a concave shape and an AUC below 0.5. Clearly the worst example is the $P2Y_1$ purinoceptor with an AUC value of 0.0857 (P2RY1 in Figure 4). In contrast to the first three categories, curves of the fourth category do not

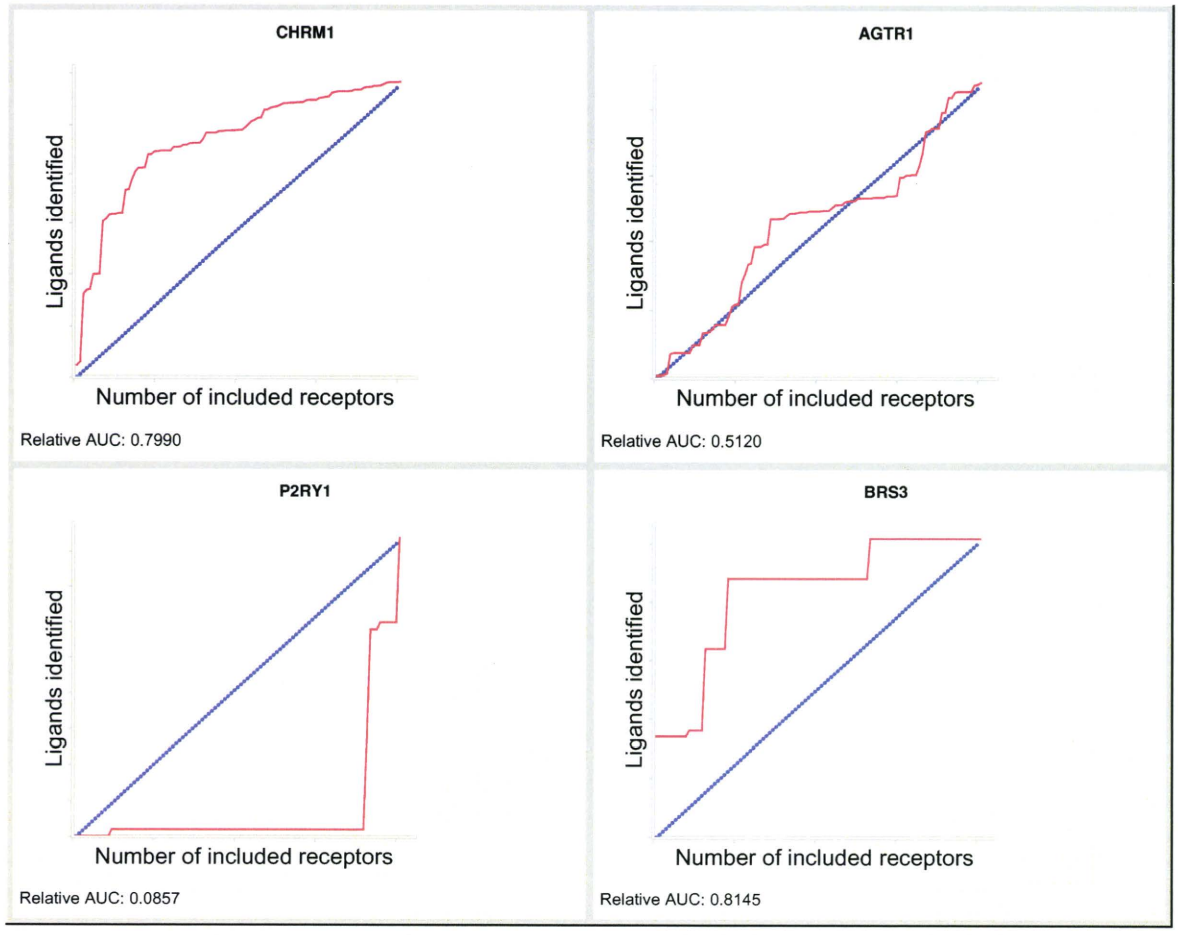

Figure 4 Examples of plotted scores for the *leave-one-out* validation. Example plots expressing the performance of the simulated receptor deorphanization. Performance plots for the following receptors are provided (from left to right and from top to bottom): CHRM1 - muscarinic acetylcholine receptor M_1 (first category); AGTR1 - angiotensin receptor AT_1 (second category); P2RY1 - $P2Y_1$ purinoceptor (third category); BRS3 - bombesin receptor BB_3 (fourth category). These examples are discussed in the text. The full set of plotted scores is provided in Additional file 2 - Plotted scores for the *leave-one-out* validation. For each plot, receptors are ordered along the x-axis (labeled "Number of included receptors") in order of increasing distance in sequence space to the receptor under study. On the y-axis (labeled "Ligands identified"), the cumulative number of retrieved ligands is depicted, normalized linearly to the interval [0;1]. The red curve indicates the number of active ligands that are retrieved when including all (closest) receptors that are listed along the x-axis up to that point. For example, the plot of the muscarinic acetylcholine receptor M_1 (CHRM1) displays a steeply rising curve near the origin, indicating that many of its ligands are retrieved using a small number of closest receptors. The blue diagonal illustrates recovery of ligands when performance is equal to random prediction. The relative area under the curve (AUC) of the red curve is stated at the bottom of each plot. An AUC above 0.5 indicates good performance, while poor performance is indicated by an AUC of 0.5 or below.

have a clear AUC range. This category consists of curves that are divided into several discrete parts of alternating rises and plateaus, as shown in the plot of bombesin receptor BB_3 (BRS3 in Figure 4), with an AUC of 0.8145. Performance varies from good (BRS3) to worse than random, depending on the value of the AUC. An example of such a plot with an AUC value below 0.5 is the FSH receptor (not shown, see: Additional file 2 - Plotted scores for the *leave-one-out* validation) with an AUC of 0.4428. The steep rises are caused by a few receptors identifying the majority of ligands. Some of these curves are steeply rising at the start, which suggests that part of its ligand set could be readily identified even though this is not reflected in the AUC. The poor performance concerning the $P2Y_1$ receptor is probably due to the nature of its ligands: this set consists of a small number of highly similar ligands that all possess a phosphate group, a feature not found in other ligands in the database. The number of features (substructures) shared with ligands of this receptor and other receptors is therefore small. Interestingly,

the adenosine A_1 and A_3 receptors, which are also purinergic, identify most (28 out of 42) of the $P2Y_1$ ligands. However, in sequence space these receptors are at great distance (at positions 91 and 92, respectively).

Overall, our method proves useful for receptor de-orphanization, since for 93% of receptors studied de-ophanization performed better than random selection (AUC > 0.5) and for 35% of receptors de-orphanization performed well (AUC > 0.7).

Limitations of the work

In the present study, some targets were excluded due to insufficient availability of ligand data in the source databases. The absence of a receptor may influence the order of other receptors in the trees. Scarcity of ligand data is reflected in the substructure profiles, thereby influencing the correlations among receptors. The issue of data (in) completeness and its effect on interaction networks was recently discussed by Mestres *et al.* [44]. Using three datasets of increasing complexity (more connections) that linked ligands to targets based on full chemical identity, the authors showed that an increase in the number of connections rapidly leads to shifts in connection patterns. However, our study linked targets based on overlap in substructures; as a consequence sharing of substructures rather than of ligands is sufficient for targets to be identified as related. Bender *et al.* and Keiser *et al.* already showed that overlapping ligands are not necessary to predict whether targets are close in ligand space [19,20]. In addition, our method employs an exhaustive approach to analyze the structural features of ligands. Frequent substructure mining considers all possible substructures that occur in the ligands and is therefore unbiased, *i.e.* all possible substructures were evaluated, not only those intuitive to chemists, such as functional groups, ring systems (e.g. a phenyl ring), and linkers [45]. However, in the present study less 'obvious' substructures such as ethyl or isobutyl are also considered [21]. For a complete discussion on substructure generation and evaluation, see ref. [46]. Our method is not limited to GPCRs alone; it is easily extended to other protein families for analysis of the differences between subfamily phylogenies, given that sufficient ligand information is available. For instance, it can be applied to the realm of enzymes to complement other chemogenomics analyses [47].

Conclusions

In this work, we presented a ligand-based phylogenetic classification that complements the well-established sequence-based classification of proteins, and applied our method to classification of GPCRs. This alternate view may contribute to our understanding of GPCR classification since it reveals relationships that are unnoticed with conventional phylogeny. Targets were analyzed based on the substructure profiles of their ligands using an unbiased approach. The overall organization of the sequence tree and the substructure tree was similar; however, substantial differences were also discovered. In the substructure tree, several clusters of subtypes were identified. For instance, it was found that the adenosine receptors group together, and that certain GPCR subfamilies that do not share sequence homology cluster because of ligand similarity. Thus, receptor similarities that signal for potential off-target effects, such as for the serotonergic receptors, are readily identified. In addition, combined with sequence-based classification, the ligand-based classification presented has proven potential (93% of receptors with AUC > 0.5 and 35% with AUC > 0.7) for de-orphanization of receptors.

Methods

Datasets

Ligands

Ligands for human GPCRs were collected from three publicly available data sources: the StARLITe database, as made available by ChEBI (EMBL-EBI) as part of the ChEMBL database [48], GLIDA [29], and KiDB [49]. ChEMBL consists of a collection of more than 500,000 small molecules annotated with activity. Here, only activity values measured directly from binding studies were included. Compounds with K_i, IC_{50}, or EC values below 10 μM were considered active. GLIDA provides biological information on GPCRs (sequences) and chemical information about ligand structures. It has links to several external databases, GPCRDB [25], UniProt [50], PubChem [51], and DrugBank [52]. A reported affinity in one of these source databases classifies a compound as active, independent of the reported binding affinity. Ligands are annotated with an activity type, namely: full agonist, partial agonist, agonist, antagonist or inverse agonist. In the present study, we focused only on binding affinity and not on the activity type. This allowed us to merge the set with the rest of the data. KiDB provides information on drugs and molecular compounds that interact with GPCRs, ion channels, transporters, and enzymes. The entries in KiDB are annotated with ligand, K_i value, radiolabeled ligand, receptor name, source & tissue, species, and PubMed link to the publication(s). Our dataset consisted of ligands from all three sources, by selecting human GPCR ligands with a molecular weight between 50 and 700 Da. Only targets that had 20 or more ligands listed were used. In this study, we focused on class A (rhodopsin-like) GPCRs since the majority of targets are from class A and only a minor part from class C; combining both classes would have negatively affected homogeneity of the phylogenetic trees, thereby hampering comparison. For the same reason, we removed two singleton targets (targets that are the only member in a subfamily), the gonadotrophin-

releasing hormone receptor and the ghrelin receptor. The final set consisted of 102 targets (provided in Table 1 of Additional file 3 - List of GPCRs used in this study) with 37350 unique ligands in total.

Sequences

The multiple sequence alignment of (specific residues of) the 7-TM domain was obtained from GPCRDB [25,53]. Only human receptors that were non-olfactory and not orphan were used.

Tree generation

Frequent Substructure Mining

For the ligands of each receptor, the most frequently occurring substructures were determined. This was accomplished by using the frequent subgraph-mining algorithm [54], which finds all frequent substructures in a set of molecular graphs [23]. For a description and a quantitative comparison of recent substructure mining algorithms, see [55]. Briefly, starting from the smallest substructure, namely the single atoms, the algorithm finds the number of molecules in which the substructure occurs. If this occurrence is above a user-defined minimum, the minimum support value, the substructure is stored. Stored substructures are stepwise extended, and tested in a systematic manner, with the aim of testing all possible substructures that have at least one of the stored substructures as their basis. The algorithm seeks ways to test only those substructures that actually occur in the set, and that have a frequency above the set minimum. An important concept of frequent substructure mining is the *a priori* principle, originating from frequent item set mining [56]. Algorithms based on the *a priori* principle exploit that the frequency of a substructure will be equal or lower than the frequency of the substructures it contains. Therefore, whenever the occurrence of a substructure is below the minimum support, all extensions of that substructure are discarded.

Structures were represented as labeled graphs with a special type for aromatic bonds. In this study, the minimum support value was set to 30% of the number of ligands in each activity set. At this value, the algorithm provided a large group of substructures while still being computationally feasible to work with. In addition, molecular structures were sorted in ascending order according to the number of bonds. This allowed the algorithm to prune scarce, complicated substructures that consisted of a large number of bonds, thereby reducing memory requirements. If the set of generated substructures is disproportionately large (more than 1000 times larger) compared to the majority of the other classes, the generated substructures are discarded except for those that also occur in other classes. This step was performed in order to prevent single targets from dominating the analysis. Since in practice most classes generated sets of less than 1000 substructures, a cut-off of 1 M substructures was used. Substructures with molecular weight below 50 Dalton were discarded. The frequent substructures of all classes were merged into one set, removing any duplicates. For all substructures in this set, the frequency in each subfamily was determined. To calculate the correlation between two targets, we used the substructure frequencies as features for that target. A correlation matrix was constructed by calculating the Pearson correlation coefficient for each pair of targets. Finally, a distance matrix was constructed by subtracting the values of the correlation matrix from unity and normalizing the results linearly to the interval [0;1].

Phylogenetic Trees

To study receptor organization, receptors were clustered into a phylogenetic tree using the Neighbor-Joining (NJ) method (Neighbor from the PHYLIP package [57]). This method infers phylogenies from the pair-wise distances between receptors. Phylogenetic trees built from distance matrices facilitate tree comparison across domains. In addition, NJ clusters each domain equally well since it does not involve an 'evolutionary clock', a concept rooted in evolutionary biology. Two distance matrices represented the similarities of the receptors: according to the frequent substructures of their ligands and the 7-TM domain sequence alignment, both were visualized as a phylogenetic tree, with receptors as leaves of the tree. The number of branches between two leaves in the tree grows with dissimilarity of these two leaves.

The protein distances between the aligned sequences were calculated with Protdist from the PHYLIP package version 3.6. using the Jones-Taylor-Thornton matrix (default) [57]. Both the sequence-based and ligand-based phylogenetic trees were constructed using the neighbor.exe program from the PHYLIP package. Tree construction might be influenced by the order in which targets are provided to the tree constructor. To minimize the influence on the resulting phylogenetic tree, target input order was randomized 10 times and 10 new trees were generated. From these, a consensus tree was built. MEGA4 [58] was used for editing the layout of the trees and for visualization. Trees were rooted on the midpoints, that is, a root is placed at the mid-point of the longest distance between two taxa of the unrooted tree. Taxa were arranged for balanced shape and trees were visualized as circular trees showing only topology, *i.e.* branch lengths do not reflect evolutionary distance in a quantitative manner.

Tree comparison

For the comparison of trees, several methods and visualizations are available; however, there is not a single definitive measure for tree difference. To visualize how the